

# CACMIL 2023

**Conference Proceedings**

2023 2nd Asia Conference on  
Algorithms, Computing and  
Machine Learning

Shanghai, China

March 17-19, 2023



The Association for Computing Machinery  
1601 Broadway, 10th Floor  
New York, NY 10019-7434

ACM ISBN: 978-1-4503-9944-9

**ACM COPYRIGHT NOTICE.** Copyright © 2023 by the Association for Computing Machinery, Inc. Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Publications Dept., ACM, Inc., fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

For other copying of articles that carry a code at the bottom of the first or last page, copying is permitted provided that the per-copy fee indicated in the code is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, +1-978-750-8400, +1-978-750-4470 (fax).

# Table of Contents

## Article 1

Virtual Human Talking-Head Generation

*Wenchao Song, Qiang He, Guowei Chen*

*Page: 1-5*

## Article 2

Hierarchical Monte Carlo Tree Search for Latent Skill Planning

*Yue Pei*

*Page: 6-12*

## Article 3

Adaptive Model Fusion Algorithm for Decision Trees and Association Rules

*Hui Zhang,H,Zhang, Zhiling Nie,ZL,Nie, Hongwei Xiao,Hw,Xiao*

*Page: 13-18*

## Article 4

Speech Image Data Mining Algorithm Based on Multimodal Decision Fusion

*Cong Lu,C,Lu, Danxing Wang,Dx,Wang, Daquan Zhang,Dq,Zhang*

*Page: 19-24*

## Article 5

Optimization Model Analysis of Blockchain PoW Protocol under Long Delay Attack

*Tao Feng, Yufeng Liu*

*Page: 25-30*

## Article 6

ADCapsNet: An Efficient and Robust Capsule Network Model for Anomaly Detection

*Xiangyu Cai, Ruliang Xiao, Zhixia Zeng, Ping Gong, Shi Zhang*

*Page: 31-39*

## Article 7

Customer Service Hot Event Discovery Based on Dynamic Dialogue Embedding

*Fei Li, Yanyan Wang, Ying Feng, Qiangzhong Feng, Yuan Zhou, Dexuan Wang*

*Page: 40-48*

## Article 8

MergeTree: A Tree Model with Merged Nodes for Threat Induction

*Ping Chen, Jingjing Hu, Zhitao Wu, Ruoting Xiong, Wei Ren*

*Page: 49-53*

## Article 9

Heart Sound Classification Algorithm Based on Sub-band Statistics and Time-frequency Fusion Features

*Xiaoqing, Zhang, Weilian, Wang*

*Page: 54-59*

## Article 10

Garment Metaverse: Parametric Digital Human and Dynamic Scene Try-on

*Hua Wang, Xiaoxiao Liu, Minghua Jiang, Changlong Zhou*

*Page: 60-65*

## Article 11

Multi-dimensional Analysis of Urban Shrinkage Problem in Liaoning Province Based on Multi-index System, Grey Correlation Analysis and BP Neural Network with Particle Swarm Optimization

*Zhenyu Fang, Junpeng Li, Junyu Xiong, Xin Wang*

*Page: 66-72*

## Article 12

Helmet Wear Detection Based on YOLOV5

*Jun Liu, Jiacheng Cao, Changlong Zhou*

*Page: 73-77*

**Article 13**

An Intrusion Detection Model with Attention and BiLSTM-DNN

*Yongcai Tao, Jitao Zhang, Lin Wei, Yufei Gao, Lei Shi*

*Page: 78-83*

**Article 14**

Deep Reinforcement Learning with Copy-oriented Context Awareness and Weighted Rewards for Abstractive Summarization

*Caidong Tan*

*Page: 84-89*

**Article 15**

An Autoencoder-based Fast Online Clustering Algorithm for Evolving Data Stream

*Dazheng Gao*

*Page: 90-95*

**Article 16**

Estimation of Distribution Algorithm with Discrete Hopfield Neural Network for GRAN3SAT Analysis

*Yuan, Y.G., Gao, Chengfeng, C.Z., Zheng, Ju, J.C., Chen, Yueling, Y.G., Guo*

*Page: 96-101*

**Article 17**

Face Anti-spoofing Method Based on Deep Supervision

*Hongxia Wang, Li Liu, Ailing Jia*

*Page: 102-106*

**Article 18**

Genetic Algorithm in Hopfield Neural Network with Probabilistic 2 Satisfiability

*Ju, J.C., Chen, Chengfeng, C.Z., Zheng, Yuan, Y.G., Gao, Yueling, Y.G., Guo*

*Page: 107-111*

**Article 19**

Construction of Scene Library System for Commercial Vehicle Products Based on Multidimensional Terminal

*Zheng Yunshuang, Z., Zheng, Hu Shilan, H., Hu, Xue Nannan, X., Xue*

*Page: 112-116*

**Article 20**

Research on Epidemic Big Data Monitoring and Application of Ship Berthing Based on Knowledge Graph-community Detection

*Shang Dongfang, Li Yuesong, Xu Jianshuai, Bao Kexin, Wang Ruixi, Qin Liu*

*Page: 117-124*

**Article 21**

Early Warning of Corporate Financial Crisis Based on Sentiment Analysis and AutoML

*Wei Cheng, Shiyu Chen, Xi Liu, Jiali Kang, Jiahao Duan, Shixuan Li*

*Page: 125-130*

**Article 22**

Graph Representation Learning and Software Homology Matching Based a Study of JAVA Code Vulnerability Detection Techniques

*Yibin Yang, Xin Bo, Zitong Wang, Xinrui Shao, Xinjie Xie*

*Page: 131-142*

**Article 23**

PhyGNNet: Solving Spatiotemporal PDEs with Physics-informed Graph Neural Network

*Longxiang Jiang, Liyuan Wang, Xinkun Chu, Yonghao Xiao, Hao Zhang*

*Page: 143-147*

**Article 24**

Multi-strategy Improved Multi-objective Harris Hawk Optimization Algorithm with Elite Opposition-based Learning

*Fulin Tian, Jiayang Wang, Fei Chu, Lin Zhou*

*Page: 148-153*

**Article 25**

Elastic Detection Mechanism Aimed at Hybrid DDoS Attack

*Yubo Wang, Jinyu Wang*

*Page: 154-160*

**Article 26**

Detecting Arbitrary-oriented Objects in Remote Sensing Imagery with Segmentation-Aware Mask

*Jiali Wei, Bo Hua, Fei Gao, Huan Zhang, Jiangwei Fan, Shuran Zhang*

*Page: 161-166*

**Article 27**

A Review of Routing Optimization Techniques for Quality of Service Assurance in Software-defined Networks

*Guozhu Yan, Jingchao Wang, Shuangyin Ren, Chao Xue*

*Page: 167-174*

**Article 28**

TIRec: Transformer-based Invoice Text Recognition

*Yanlan Chen*

*Page: 175-180*

**Article 29**

Two-channel Conformance Test Analysis of S-band Dual-polarization Radar

*Yuxin Gong, Qian Zhang, Weijia Sun, Chuancheng Ma, Yucheng Gong, Juxiu Wu, Xiqiang Yuan*

*Page: 181-187*

**Article 30**

Feature Selection Based on Improved Principal Component Analysis

*Zhangyu Li, Yihui Qiu*

*Page: 188-192*

**Article 31**

A Water Quality Parameter Prediction Method Based on Transformer Architecture and Multi-sensor Data Fusion

*Bo Fang, Hao Liu, Wei He, Dexin Li, Chengzhao Liu*

*Page: 193-199*

**Article 32**

Unknown Radar Signals Deinterleaving Based on TCN Network

*Liying Ma, Xueqiong Li, Yuhua Tang*

*Page: 200-204*

**Article 33**

A Component for Query-based Object Detection in Crowded Scenes

*Shuo Mao*

*Page: 205-209*

**Article 34**

Explainable Deep Learning for Medical Image Segmentation with Learnable Class Activation Mapping

*Kaiyue Wang, Sixing Yin, Yining Wang, Shufang Li*

*Page: 210-215*

**Article 35**

An Emotion Recognition Method Based on Feature Fusion and Self-supervised Learning

*Xuanmeng Cao, Ming Sun*

*Page: 216-221*

**Article 36**

A Modified Fuzzy K-nearest Neighbor Using the Improved Sparrow Search Algorithm for Two-classes and Multi-classes Datasets

*Chengfeng Zheng, Yuan Gao, Ju Chen, Mohd.Asyraf Mansor*

*Page: 222-227*

**Article 37**

An Analysis Software for Visual Position and Attitude Measurement Algorithm

*Xu Tao, Zhang Jing, Cai Bin, Wang Yafei*

*Page: 228-235*

**Article 38**

Heuristic Search for DNN Graph Substitutions

*FeiFei Deng, HongKang Liu*

*Page: 236-241*

**Article 39**

Discrimination of Seismic and Non-seismic Signal Using SCOUTER

*Kang Wang, Ji Zhang, Jie Zhang*

*Page: 242-246*

**Article 40**

ENOSE Performance in Transient Time and Steady State Area of Gas Sensor Response for Ammonia Gas: Comparison and Study

*Geng Kuan, Ata Jahangir Moshayedi, Chen Jing, Hu Jiandong, Zhang Hao*

*Page: 247-252*

**Article 41**

MM-UNet: Multi-attention Mechanism and Multi-scale Feature Fusion UNet for Tumor Image Segmentation

*Yaozheng Xing, Jie Yuan, Qixun Liu, Shihao Peng, Yan Yan, Junyi Yao*

*Page: 253-257*

**Article 42**

A Hybrid Aquila Optimizer Sine Cosine Algorithm for Numerical Optimization

*Fei Chu, Jiayang Wang, Fulin Tian*

*Page: 258-263*

**Article 43**

Comparative Research on Embedding Methods for Video Knowledge Graph

*Zhihong Zhou, Qiang Xu, Hui Ding, Shengwei Ji*

*Page: 264-270*

**Article 44**

Performance Evaluation of an Extradosed Cable-stayed Bridge with Corrugated Web Based on Machine Learning Algorithms

*Zeyu Du, Zhenhua Pan, Zhihua Xiong, Lei He, Haipeng Wang, Houda Zhu, Jiangbo Wang*

*Page: 271-276*

**Article 45**

An Adaptive Gradient Privacy-preserving Algorithm for Federated XGBoost

*Hongyi Cai, Jianping Cai, Lan Sun*

*Page: 277-282*

**Article 46**

Generate Earthquake Catalog Using the VAE Method

*Zhangyu Wang, Jie Zhang*

*Page: 283-286*

**Article 47**

Robust Hypergraph-augmented Graph Contrastive Learning for Graph Self-supervised Learning

*Zeming Wang, Xiaoyang Li, Rui Wang, Changwen Zheng*

*Page: 287-293*

**Article 48**

Quantum Kernel Subspace Alignment for Unsupervised Domain Adaptation

*Xi He, Feiyu Du*

*Page: 294-297*

**Article 49**

Cross-modal Audio-text Retrieval via Sequential Feature Augmentation

*Fuhu Song, Jifeng Hu, Che Wang, Jiao Huang, Haowen Zhang, Yi Wang*

*Page: 298-304*

**Article 50**

Health Monitoring System for Elderly People Based on Raspberry Pi

*Qingsong Peng*

*Page: 305-308*

**Article 51**

Multiple Frequency Bands Temporal State Representation for Deep Reinforcement Learning

*Che Wang, Jifeng Hu, Fuhu Song, Jiao Huang, Zixuan Yang, Yusen Wang*

*Page: 309-315*

**Article 52**

Federated Learning-based Intrusion Detection Method for Smart Grid

*Bin Dongmei, Li Xin, Yang Chunyan, Han Songming, Ling Ying*

*Page: 316-322*

**Article 53**

Deep Learning AD Detection Model Based on a Two-layer Ensemble Module with Data Augmentation and Contrastive Learning

*Weicheng Wang*

*Page: 323-328*

**Article 54**

RhySpeech: A Deployable Rhythmic Text-to-speech Based on Feed-forward Transformer for Reading Disabilities

*Yixuan Lin*

*Page: 329-337*

**Article 55**

Research on Natural Scene Vehicle Nameplate Text Detection Based on Improved DBNet

*Du Yucheng, Dong Jinsong*

*Page: 338-345*

**Article 56**

Performance Evaluation of Agricultural Logistics Enterprises Based on GA Algorithm

*Yebin*

*Page: 346-350*

**Article 57**

SSGAR: A Genetic-based Routing Solution for Aeronautical Networks Aided by Software Defined Satellite Network

*Kaixuan Sun, Ke Wu, Wenke Yuan, Guangyuan Wei, Huasen He*

*Page: 351-356*

**Article 58**

Mathematical Models of Colony Population Dynamics and Hive Placement

*Zixuan Zhang, Dongyi He, Hanwen Zhang*

*Page: 357-365*

**Article 59**

Global-local Framework for Medical Image Segmentation with Intra-class Imbalance Problem

*Yifan Zhou, Bing Yang, Xiaolu Lin, Risa Higashita, Jiang Liu*

*Page: 366-370*

**Article 60**

End-to-end Parking Behavior Recognition Based on Self-attention Mechanism

*Penghua Li, Dechen Zhu, Qiyun Mou, Yushan Tu, Jinfeng Wu*

*Page: 371-376*

**Article 61**

Foreign Object Recognition Method of Transmission Line Based on Improved Outlier Rate Method

*Dongmei Liu, Zhongwang Zhu, Bo Chen*

*Page: 377-381*

**Article 62**

Improved YOLOv5 UAV Target Detection Algorithm by Fused Attention Mechanism

*Yan, YH, He, Yanni, YNz, Zhao, Hongfei, Hfn, Nie*

*Page: 382-388*

**Article 63**

Haze Video Image Clarity Processing Based on Optical Flow Threshold

*Chen Ru, Wang Xijuan*

*Page: 389-393*

**Article 64**

Gaussian-guided Character Erasure for Data Augment of Industrial Characters

*Hongchao Gao, Chao Yao, Zhennan Wang*

*Page: 394-401*

**Article 65**

A 3D Discrete Memristive Chaotic Map and Its Application in Image Encryption

*Junwei Shen*

*Page: 402-412*

**Article 66**

CBAM-based Method in YOLOv7 for Detecting Defective Vacuum Glass Tubes

*Zeyu Sheng, Haiguang Chen, Zifeng Qi*

*Page: 413-418*

**Article 67**

Image Generation Model Applying PCA on Latent Space

*Myung Keun Song, Asim Niaz, Kwang Nam Choi*

*Page: 419-423*

**Article 68**

An Interpretable Brain Network Atlas-based Hybrid Model for Mild Cognitive Impairment Progression Prediction

*Xianglong Guan, Li Ma, Suqin Tang, Tinghui Li, Yunyou Huang*

*Page: 424-428*

**Article 69**

Improved Convolutional Neural Networks by Integrating High-frequency Information for Image Classification

*Chengyuan Zhuang, Xiaohui Yuan, Xuan Guo, Zhenchun Wei, Juan Xu, Yuqi Fan*

*Page: 429-434*

**Article 70**

Comparison of Regional Monitoring Methods for Grassland Degradation Based on Remote Sensing Images

*Haoran Wang, Tianyu Xue, Zhaoran Wang, Xiangyu Bai*

*Page: 435-439*

**Article 71**

Multi-modal Fusion Object Tracking Based on Fully Convolutional Siamese Network

*Ke Qi, Liji Chen, Yicong Zhou, Yutao Qi*

*Page: 440-444*

**Article 72**

An Ensemble Model Using Face and Pose Tracking for Engagement Detection in Game-based Rehabilitation

*Xujie Lin, Siqi Cai, Patrick P. K. Chan, Longhan Xie*

*Page: 445-449*

**Article 73**

A U-Net Based Self-supervised Image Generation Model Applying PCA Using Small Datasets

*Sang Hun Han, Asim Niaz, Kwang Nam Choi*

*Page: 450-454*

**Article 74**

An Unmanned Lane Detection Algorithm Using Deep Learning and Ordered Test Sets Strategy

*Zhang Shenwei, Lin Xiaoyan, Zhang Mingwei, Zhang Zhen, Hou Yun, Ning Honglong, Qiu Tian*

*Page: 455-461*

**Article 75**

Intelligent Perception Recognition and Positioning Method of Distribution Network Drainage Line

*Shuzhou Xiao, Qiuyan Zhang, Qiang Fan, Jianrong Wu, Chao Zhao*

*Page: 462-466*

**Article 76**

KRE: A Key-retained Random Erasing Method for Occluded Person Re-identification

*HongXia Wang, Yao Ma, Xiang Chen*

*Page: 467-473*

**Article 77**

Digital Image Denoising by Partial Differential Equation Based on P-M Model and Its Fuzzy Evaluation Method System

*Jingying, L, and Liu, Yang, H, and Hu*

*Page: 474-479*

**Article 78**

Deep Vision Network Based CT Image Detection for Aiding Lumbar Herniated Disc Diagnosis

*Wenzhe Xie, Feiwei Qin, Yanli Shao*

*Page: 480-487*

**Article 79**

MCSC-UTNet: Honeycomb Lung Segmentation Algorithm Based on Separable Vision Transformer and Context Feature Fusion

*Wei Jianjian, Li Gang, He Kan, Li Pengbo, Zhang Ling, Wang Ronghua*

*Page: 488-494*

**Article 80**

Research on Colorization of Qinghai Farmer Painting Image Based on Generative Adversarial Networks

*Chunyan Peng, C.P, and Peng, Xueya Zhao, X.Z, and Zhao, Guangyou Xia, G.X, and Xia*

*Page: 495-503*

**Article 81**

A Histo-puzzle Network for Weakly Supervised Semantic Segmentation of Histological Tissue Type

*Tengyun, Ma, Guotian, He, Lin, Chen, Yuanchang, Lin*

*Page: 504-509*

**Article 82**

Pose Estimation of Space Targets Based on Geometry Structure Features

*Xiwen Liu, Shuling Hao, Kefeng Xu*

*Page: 510-514*

**Article 83**

Detecting Respiratory Events with End-to-end ConvNet

*Yangping Shuai, Zhangbo Li, Xingjun Wang, Hanrong Cheng*

*Page: 515-519*

**Article 84**

Human Activity Recognition Based on Transformer in Smart Home

*Xinmei Huang, Sheng Zhang*

*Page: 520-525*

**Article 85**

Research on Constant Perturbation Strategy for Deep Reinforcement Learning

*Jiamin Shen, Li Xu, Xu Wan, Jixuan Chai, Chunlong Fan*

*Page: 526-533*

**Article 86**

Twitter Stance Detection Using Deep Learning Model with FastText Embedding

*Yongqing Deng, Yongzhong Huang*

*Page: 534-541*

**Article 87**

An Objective Reduction Evolutionary Multiobjective Algorithm Using Adaptive Density-based Clustering for Many-objective Optimization Problem

*Mingjing Wang, Long Chen, Huiling Chen*

*Page: 542-546*

**Article 88**

Infrared Small Target Detection Based on the Combination of Single Image Super-resolution Reconstruction and YOLOX

*Wang Zhiyong, Xiang Xuefu, Zeng Kan, Zhang Zhenyu, Li Yanan, Song Dengpan*

*Page: 547-552*

**Article 89**

An Encryption Scheme Using Multi-scroll Memristive Chaotic System

*Fan Wu, Musha Ji'e, Lidan Wang, Shukai Duan*

*Page: 553-560*

**Article 90**

FlowTexNet: Fast Texture Synthesis for Massive Flow Field Visualization

*Zijian Kang, Wenyao Zhang, Na Wang*

*Page: 561-568*

**Article 91**

CIP-ES: Causal Input Perturbation for Explanation Surrogates

*Sebastian Steindl, Martin Sturner*

*Page: 569-574*

**Article 92**

Research on Identification Method of Gap Nonlinear Vibration

*Jialiang, S, and Sun, Jingying, L, and Liu*

*Page: 575-580*

**Article 93**

Real-time Emulation of MASQUE-based QUIC Proxying in LTE Networks Using ns-3

*Donát Scharnitzky, Zsolt Krämer, Sándor Molnár, Attila Mihály*

*Page: 581-586*

# CACML 2023 Committee

## Conference Chair

Prof. David Zhang, Dapeng, Chinese University of Hong Kong (Shenzhen), China (IEEE Fellow and CS Fellow/ IET Fellow/ AAIA Fellow)

Prof. Shuanghua Yang, University of Reading, UK (IEEE Senior Member / IET Fellow)

## Program Committee Chair

Prof. Witold Pedrycz, University of Alberta, Canada (IEEE Life Fellow)

Prof. Giancarlo Succi, University of Bologna, Italy

## Steering Committee Chair

Prof. David Greenhalgh, University of Strathclyde, UK

Prof. Ljiljana Trajkovic, Simon Fraser University, Canada

## Publication Chair

Prof. Debao Zhou, University of Minnesota Duluth, USA

Prof. Priti Srinivas Sajja, Sardar Patel University, India

Assoc. Prof. Zhiyu Jiang, Northwestern Polytechnical University, China

## Publicity Chair

Assoc. Prof. Gahangir Hossain, University of North Texas, USA

Assoc. Prof. Komalpreet Kaur, Salem State University, USA

## Technical Program Committee

Prof. Farshad Khorrami, New York University, USA

Prof. HAW SU CHENG, Multimedia University, Malaysia

Prof. Ahrar Husain, Jamia Millia Islamia, India

Prof. Ljiljana Trajkovic, Simon Fraser University, Canada

Prof. Hari Mohan Srivastava, University of Victoria, Canada

Prof. Imran Zuolkernan, University of Minnesota and Pennsylvania State University, USA

Prof. B.Sharmila, Sri Ramakrishna Engineering College, India

Prof. Qiu Chen, Kogakuin University, Japan

Prof. Pavlo Maruschak, Ternopil Ivan Puluj National Technical University, Ukraine

Prof. Michael Opoku Agyeman, University of Northampton, UK

Prof. Sunny Joseph Kalayathankal, Jyothi Engineering College, India

Prof. Janusz Szpytko, AGH University of Science and Technology, Poland

Prof. Debao Zhou, University of Minnesota Duluth, USA

Prof. Ishak b. Aris, Universiti Putra Malaysia, Malaysia

Prof. Priti Srinivas Sajja, Sardar Patel University, India  
 Assoc. Prof. REN Hongliang, National University of Singapore, Singapore  
 Assoc. Prof. CHUA FANG FANG, Multimedia University, Malaysia  
 Assoc. Prof. Francesco Colace, University of Salerno, Italy  
 Assoc. Prof. Komalpreet Kaur, Salem State University, USA  
 Assoc. Prof. Voltaire Mistades, De La Salle University, Philippines  
 Assoc. Prof. Libor Pekař, Tomas Bata University in Zlín, Czech Republic  
 Assoc. Prof. Chandra Shekhar, Birla Institute of Technology & Science Pilani, India  
 Assoc. Prof. El-Said Mamdouh Mahmoud Zahran, Ain Shams University, Egypt  
 Assoc. Prof. Mohd Ashraf Ahmad, University Malaysia Pahang, Malaysia  
 Assoc. Prof. KHALDI Amine, Universite Kasdi Merbah Ouargla, Algeria  
 Assoc. Prof. Md Baharul Islam, American University of Malta, Malta  
 Assoc. Prof. Arthur Daniel Limantara, Universitas Kadiri, Indonesia  
 Assist. Prof. Jianhui Yue, Michigan Technological University, USA  
 Assist. Prof. P. K. Paul, Raiganj University, India  
 Assist. Prof. Olarik Surinta, Mahasarakham University, Thailand  
 Assist. Prof. P. Aruna, Coimbatore Institute of Technology, India  
 Assist. Prof. Ferddie Quiroz Canlas, Muscat College, Oman  
 Assist. Prof. Dariusz Jacek Jakobczak, Koszalin University of Technology, Poland  
 Assist. Prof. Yagya Raj Pandeya, Jeonbuk National University, South Korea  
 Assist. Prof. Abidalrahman Mohd, Eastern Illinois University, USA  
 Assist. Prof. Shahzad Ashraf, Hohai University, China  
 Dr. GOH HUI NGO, Multimedia University, Malaysia  
 Dr. Moises Almeida Castelo Branco, University of Waterloo, Canada  
 Dr. Xuechao Li, Auburn University, USA  
 Dr. Farshad Badie, Aalborg University, Denmark  
 Dr. Stefania Tomasiello, University of Salerno, Italy  
 Dr. Woo Chaw Seng, University of Malaya, Malaysia  
 Dr. Xiaoye Liu, University of Southern Queensland, Australia  
 Dr. M. Mujiya Ulkhaq, Diponegoro University, Indonesia  
 Dr. Zhenyu Zhang, University of Southern Queensland, Australia  
 Dr. Roohallah Azarmi, Eindhoven University of Technology, The Netherlands  
 Dr. Mohd Aliff Afira Hj Sani, Universiti Kuala Lumpur, Malaysia  
 Dr. Olarik Surinta, Mahasarakham University, Thailand  
 Dr. Zati Hakim Binti Azizul Hasan, University of Malaya, Malaysia

# Virtual Human Talking-Head Generation

Wenchao Song

State Key Laboratory of Media  
Convergence and Communication,  
Communication University of China,  
Beijing, China  
songwenchao@cuc.edu.cn.

Qiang He

State Key Laboratory of Media  
Convergence and Communication,  
Communication University of China,  
State Key Laboratory of Media  
Convergence Production Technology  
and Systems, Beijing, China  
heqiang@xinhua.org.

Guowei Chen\*

State Key Laboratory of Media  
Convergence and Communication,  
Communication University of China,  
Beijing, China  
cuc\_chenguowei@cuc.edu.cn.

## ABSTRACT

*Abstract:* Virtual humans created by computers using deep learning technology are being used widely in a variety of fields, including personal assistance, intelligent customer service, and online education. Human-computer interaction systems integrate multi-modal technologies like speech recognition, dialogue systems, speech synthesis, and virtual digital human video synthesis as one of the applications of virtual humans. In this paper, we first design the framework for a human-computer interaction system based on a virtual human; next, we classify the talking head video synthesis model according to the generation of a virtual human's depth; finally, we conduct a systematic review of the technical developments in talking head video generation over the last five years, highlighting seminal work.

## CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI); • **Computing methodologies** → Artificial intelligence; Computer vision.

## KEYWORDS

Virtual Human, Talking-head Generation, Multi-modal Human Computer Interaction

### ACM Reference Format:

Wenchao Song, Qiang He, and Guowei Chen\*. 2023. Virtual Human Talking-Head Generation. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590004>

## 1 INTRODUCTION

A virtual human is an artificial intelligence technique that creates a digital representation of human properties in a computer environment, such as geometric and behavioral. [1, 2] With the advancement of deep generation technologies such as computer vision (CV) and natural language processing (NLP), the concept of

text-to-image, text-to-video, virtual human video synthesis, and so on has gained immediate attention in academic and industrial fields. Meanwhile, virtual human video synthesis has been widely used in human-computer interaction, online teaching, and computer games.

Audio-driven lip synthesis is a well-liked research area in talking head synthesis. The lower dimensional speech or text signal is dynamically mapped into a higher dimensional video signal by inputting the corresponding audio and any mesh vertex, facial image, and video to synthesize the lip-synced talking head video. Note that this task naturally extends to text-driven lip synthesis.

With the rapid improvement of computing power, the task of talking head synthesis based on deep learning has attracted widespread attention, which has promoted the vigorous development of this field. This paper mainly makes a systematic review of the talking head video synthesis model based on deep learning in the past five years. Fig.1 shows the development venation of the talking-head generation algorithm in recent years. Along the timeline, the number of methods in different generation technical routes has increased dramatically.

According to the content of the model input, we can divide talking-head generation models into 2D-based methods and those based on 3D approach. However, in the method of synthesizing talking head video, most models take a relatively long time to generate video, and only a small part of models can output results in a short time, such as DCK [28]. More details are in the third part.

The mainly contributions of this paper are summarized as follows.:

- (1) We have put a system framework for multimodal human and computer-interaction, a novel methodology for the talking-head generation.
- (2) We summarized the 2D and 3D based talking-head generation models.

## 2 SYSTEM ARCHITECTURE

The system strives for multi-modal interaction with low-latency, high-fidelity anthropomorphic virtual humans while relying on artificial intelligence technologies like natural language processing, speech processing, and image processing. The system consists primarily of four modules, as shown in Fig. 2 (1) the system converts the user's voice information into text information using the ASR module; (2) the text information produced by the ASR module is used as the dialogue system's input; (3) the dialogue system's responses are converted into realistic speech information using the TTS module. (4) Preprocess the picture, video or Blendshepe as the image of the speaking head to extract its facial features, Then map

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590004>

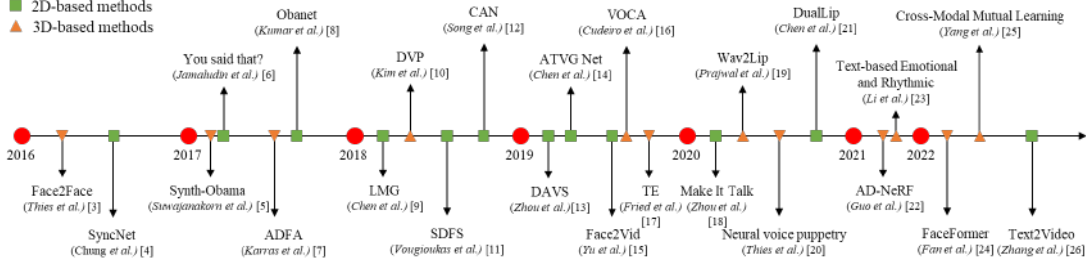


Figure 1: development venation of the talking-head generation algorithm in recent years

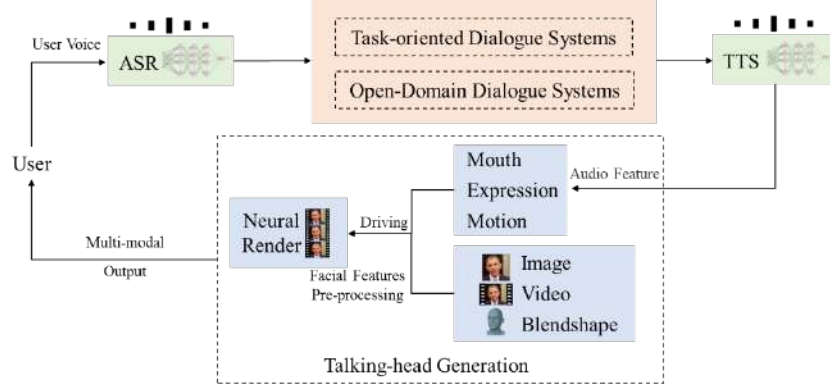


Figure 2: A system architecture of multi-modal human computer interaction

the lower-dimensional voice signal output by the TTS module to the higher-dimensional mouth, expression, motion and etc. video signals, and finally use the rendering system to fuse the features, output multi-modal video, and display it on the user side.

## 2.1 Speech Module

The ASR and TTS of the speech module correspond to the human hearing function and language function respectively. After decades of research, speech recognition and text-to-speech synthesis have been widely used in various commercial products. We can choose API services provided by commercial companies, such as Baidu, Sogou, iFLYTEK, etc.

## 2.2 Dialogue Module

Our dialogue module needs to have the ability to have multiple rounds of dialogue, it not only needs to answer domain-specific questions, but also meet the needs of users to chat. As shown in Fig. 2, after the user's voice passes through the ASR, the question is passed to the dialogue module, and the dialogue module needs to retrieve or generate matching answers from the knowledge base according to the user's question.

## 2.3 Talking-head Generation

The facial appearance data of the talking head generation module mainly comes from real person photos, videos or other character model coefficients. Taking video as an example, we first perform video preprocessing on these facial appearance data, and then map

the audio signal of TTS in Fig. 2 to higher-dimensional signals such as human face lip shape, facial expression and facial action, and finally use neural network. The model performs video rendering and outputs multi-modal video data.

## 3 TALKING HEAD GENERATION

Talking head video generation, i.e., the purpose of lip motion generation is to synthesize the lip motion sequence of talking-head based on driving data (a segment of audio or a segment of text). Fig. 1 shows that from the image dimension of talking head, it can be divided into two categories: 2D-based approaches and 3D-based approaches.

### 3.1 2D-based approaches

In 2D-based methods, talking head synthesis mainly used facial landmarks, semantic graph, or other methods of image-like representations to solve the problem. Bregler et al. [27] use of the audio's morpheme to modify the lip shape's action parameters to synthesize video can be traced back to 1997. In contrast, Chen et al. [14] used landmarks as an intermediate layer for mapping from low-dimensional audio to high-dimensional video, and divided the whole method into pipeline; Chung et al. [6] used two decoders to the voice and speaker identity are decoupled so that the video is synthesized without the influence of the speaker identity, and the lip synthesis is also performed in the manner of image-to-image translation.

Table 1. The main model of talking-head generation in recent years. ID: The model can be divided into three types: identity-dependent (D), identity-independent(I) and hybrid(H). Driving Data: Audio(A), Text(T) and Video(V).

References	Key Idea	Driving Data	ID D/I	Model Dimension
Suwajanakorn [5]	Audio to mouth editing to video	A	D	3D
[6, 9]	Joint embedding of audio and identity features	A	I	2D
Karras [7]	From audio and emotion-state to 3D vertices	A	D	3D
Kumar [8]	Text to audio to mouth key-points to video	T	D	2D
Kim [10]	DVP: parameter replacement and facial reenactment with cGAN to video	V	I	3D
Vougioukas [11]	Aduio-driven GAN	A	I	2D
Zhou[13]	Joint embedding of person-id and word-id features	V or A	I	2D
Chen[14]	From Audio to facial landmarks to video synthesis	A	I	2D
Yu[15]	From text or audio feature to facial landmarks to video synthesis	A and T	I	2D
Cudeiro [16]	VOCA: from audio to FLAME head model with facial motions	A	I	3D
Fried [17]	3D reconstruction and parameter recombination	T	D	3D
Zhou [18]	Audio-driven landmark prediction	A	I	2D
Prajwal [19]	Wav2Lip: aduio-driven, based GAN lip-sync discriminator	A	I	2D
Thies [20]	NVP: from the fusion of audio expression feature extraction and intermediate 3D model to video	A	H	3D
Guo [22]	AD-NeRF: Audio to video generation via two individual neural radiance fields	A	D	3D
Li [23]	TE: text-driven to video generation combine phoneme alignment, viseme search and parameter blending	T	D	3D
Fan [24]	FaceFormer: Audio to 3D Mesh to video	A	I	3D
Yang [25]	A unified framework for visual-to-speech recognition and audio-to-video synthesis	A	I	3D
Zhang [26]	Text2Video: GAN+phoneme-pose dictionary	T	I	3D

### 3.2 3D-based approaches

Traditional approaches use pre-built 3D models of particular people, which are then rendered. This method can have better control over motion than 2D techniques. The impact of changing a new identity cannot be guaranteed due to the high construction cost of such a 3D model. These works [5, 8] synthesize realistic speaking facial videos for Barack Obama’s videos by building 3D facial models in advance and using learned audio-to-video mapping to power the model. There are also numerous generative speech head models based on 3DMM parameters [7, 16, 17, 20], and models like Flame [29] and Blend shape [16] are used with audio as the models’ input for the independent identity.

## 4 DISCUSSION

The two main categories of the technical approach to virtual human synthesis are the depth generation method and the method based on computer graphics (CG). The CG-based virtual human synthesis method, which has the advantages of lifelike characters and quick synthesis time but has a high cost, is currently widely used in industrial production. Deep learning-based techniques are a subset of virtual human synthesis techniques. The virtual human talking head generation model based on deep learning can realistically generate a variety of adult faces using speech. However, it still lacks some details necessary for conversational realism, such as head movement, eye movement, and expression. Additionally, these methods’ evaluation indicators focus only on image quality

and audio-visual coherence, ignoring the importance of the virtual human video synthesis time.

The long synthesis time of the virtual human talking head is a limitation of the deep generative model. To address this shortcoming, Ye et al. [28] performed data preprocessing on the deep generative model’s multi-modal video input. The approach introduces a novel dynamic convolution kernel (DCK) in the mapping process from low-dimensional audio to high-dimensional video. However, because DCK directly overlays the generated lip shape on the background image, the model’s time to synthesize the video is reduced. However, the output video authenticity needs to be improved. Guo et al. [22] proposed that the NeRF-based talking head generation model AD-NeRF can produce more realistic virtual human videos in terms of video rendering effects. The problem of long video rendering time also needs to be addressed. Of course, [31–34] suggested optimization strategies that could increase the speed of video synthesis NeRF-based to get over its comparatively slow video synthesis time. In addition to the pure depth production technique, Chen et al. [30] Speech-to-animation (S2A) system for generating face animation from speech combined with the transformer and rendering engine. The S2A system’s fastest video inference speed and the most improvement is 17 times greater than the most recent method.

Currently, the response, such as the dialogue and text-to-speech systems, can already be finished with little latency. The module that costs the most time in the virtual human-based human-computer communication system is the talking-head generation model for

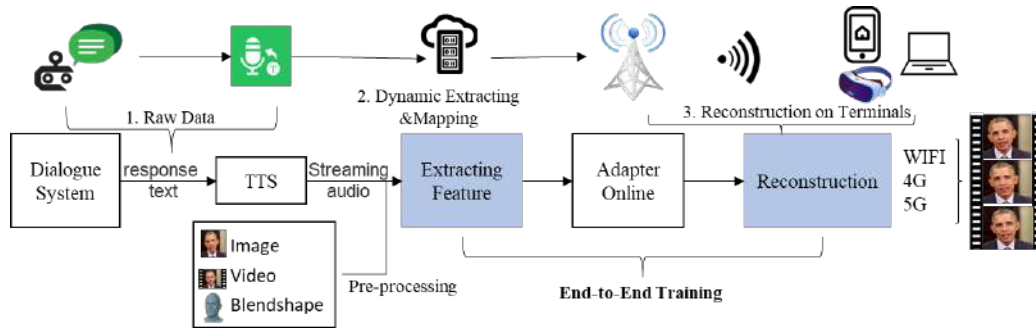


Figure 3: System workflow and components of virtual human video reconstruction

virtual humans. Additionally, real-time transmission to the client uses a lot of bandwidth after high-quality virtual human video generation. Huang et al. [35] propose the approach of AITransfer, a bandwidth-aware and adoptive transfer system, which can solve this issue. Inspired by literature [28, 35], as shown in Fig. 3 in the feature extraction and data mapping stages, we can preprocess the multimodal video and utilize the approach of voice segmentation to map streaming audio into the video key point information. Finally, the real-time video reconstruction work is performed on the client terminal.

## 5 CONCLUSION

In our works, (1) we present a system framework for multimodal human-computer interaction, which provides a new idea for the application of speech head generation models. (2) We summarized the 2D and 3D based speech head synthesis models. In the future, we will take the quality of synthetic video and rendering speed as indicators to study the method of generating digital human speech head video. In addition, we also discuss methods for generating virtual human videos in real time.

## ACKNOWLEDGMENTS

This work was supported by the State Key Laboratory of Media Convergence Production Technology and Systems (SKLM-CPTS2020012), Beijing, China, 100803.

## REFERENCES

- [1] Wang Zhaoqi, "A review of virtual human synthesis", Journal of Chinese Academy of Sciences, vol. 17, no. 2, pp. 89, 2000.
- [2] Chen Qixiang and Wei Kejun, Research on virtual human technology China water transportation, Academic, pp. 5, 2006.
- [3] Thies J, Zollhofer M, Stamminger M, et al. Face2face: Real-time face capture and reenactment of rgb videos[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2387-2395.
- [4] J. S. Chung, A. Zisserman, Out of time: automated lip sync in the wild, in: Asian conference on computer vision (ACCV), 2016, pp. 251–263.
- [5] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM ToG, vol. 36, no. 4, pp. 1–13, 2017.
- [6] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?" in BMVC, 2017.
- [7] Karras T, Aila T, Laine S, et al. Audio-driven facial animation by joint end-to-end learning of pose and emotion[J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-12.
- [8] Kumar R, Sotelo J, Kumar K, et al. Obamanet: Photo-realistic lip-sync from text[J]. arXiv preprint arXiv:1801.01442, 2017.
- [9] Chen L, Li Z, Maddox R K, et al. Lip movements generation at a glance[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 520-535.
- [10] Kim H, Garrido P, Tewari A, et al. Deep video portraits[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 1-14.
- [11] Vougioukas K, Petridis S, Pantic M. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs[C]//CVPR Workshops. 2019: 37-40.
- [12] Song Y, Zhu J, Li D, et al. Talking face generation by conditional recurrent adversarial network[J]. arXiv preprint arXiv:1804.04786, 2018.
- [13] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in AAAI, vol. 33, no. 01, 2019, pp. 9299–9306.
- [14] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in CVPR, 2019, pp. 7832–7841.
- [15] Yu L, Yu J, Ling Q. Mining audio, text and visual information for talking face generation[C]//2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019: 787-795.
- [16] Cudeiro D, Bolkart T, Laidlaw C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10101-10111.
- [17] Fried O, Tewari A, Zollhofer M, et al. Text-based editing of talking-head video[J]. ACM Transactions on Graphics (TOG), 2019, 38(4): 1-14.
- [18] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "Makeltalk: speaker-aware talking-head animation," ACM TOG, vol. 39, no. 6, pp. 1–15, 2020.
- [19] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Nambodiri, and CV Jawahar. 2020. A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In Proceedings of the 28th ACM International Conference on Multimedia. 484–492.
- [20] Thies J, Elgharib M, Tewari A, et al. Neural voice puppetry: Audio-driven facial reenactment[C]//European conference on computer vision. Springer, Cham, 2020: 716-731.
- [21] W. Chen, X. Tan, Y. Xia, T. Qin, Y. Wang, and T.-Y. Liu, "Dualip: A system for joint lip reading and generation," in ACM MM, 2020, pp. 1985–1993.
- [22] Guo Y, Chen K, Liang S, et al. Ad-nerf: Audio driven neural radiance fields for talking head synthesis[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 5784-5794.
- [23] Li L, Wang S, Zhang Z, et al. Write-a-speaker: Text-based emotional and rhythmic talking-head generation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 1911-1920.
- [24] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speechdriven 3d facial animation with transformers," arXiv:2112.05329, 2021.
- [25] C.-C. Yang, W.-C. Fan, C.-F. Yang, and Y.-C. F. Wang, "Crossmodal mutual learning for audio-visual speech recognition and manipulation," in AAAI, 2022.
- [26] S. Zhang, J. Yuan, M. Liao and L. Zhang, "Text2video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme - Pose Dictionary," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 2659-2666.
- [27] Bregler C, Covell M, Slaney M. Video rewrite: Driving visual speech with audio[C]//Proceedings of the 24th annual conference on Computer graphics and interactive techniques. 1997: 353-360.
- [28] Ye Z, Xia M, Yi R, et al. Audio-driven talking face video generation with dynamic convolution kernels[J]. IEEE Transactions on Multimedia, 2022.
- [29] Li T, Bolkart T, Black M J, et al. Learning a model of facial shape and expression from 4D scans[J]. ACM Trans. Graph., 2017, 36(6): 194:1-194:17.
- [30] Chen L, Wu Z, Ling J, et al. Transformer-S2A: Robust and Efficient Speech-to-Animation[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 7247-7251.
- [31] Hong Y, Peng B, Xiao H, et al. Headnerf: A real-time nerf-based parametric head model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 20374-20384.

- [32] Neff T, Stadlbauer P, Parger M, *et al.* DONERF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks[C]//Computer Graphics Forum. 2021, 40(4): 45-59.
- [33] Yu A, Li R, Tancik M, *et al.* Plenotrees for real-time rendering of neural radiance fields[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 5752-5761.
- [34] Martin-Brualla R, Radwan N, Sajjadi M S M, *et al.* Nerf in the wild: Neural radiance fields for unconstrained photo collections[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7210-7219.
- [35] Huang Y, Zhu Y, Qiao X, *et al.* Aitransfer: Progressive ai-powered transmission for real-time point cloud video streaming[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 3989-3997.

# Hierarchical Monte Carlo Tree Search for Latent Skill Planning

Yue Pei

YUP20@pitt.edu

University of Pittsburgh

Pittsburgh, PA, United States

## ABSTRACT

Monte Carlo Tree Search (MCTS) continues to confront the issue of exponential complexity growth in certain tasks when the planning horizon is excessively long, causing the trajectory’s past to grow exponentially. Our study presents Hierarchical MCTS Latent Skill Planner, an algorithm based on skill discovery that automatically identifies skills based on intrinsic rewards and integrates them with MCTS, enabling efficient decision-making at a higher level. In the grid world maze domain, we found that latent skill search outperformed the standard MCTS approach that do not contain skills in terms of efficiency and performance.

## CCS CONCEPTS

• Computing methodologies → Planning and scheduling.

## KEYWORDS

deep reinforcement learning, monte carlo tree search

### ACM Reference Format:

Yue Pei. 2023. Hierarchical Monte Carlo Tree Search for Latent Skill Planning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590005>

## 1 INTRODUCTION

Model-based Reinforcement Learning approaches have attracted a great deal of interest and achieved significant progress in recent years [5, 32, 36]. Both data from real environments and ‘imagined data’ from the model can be utilized to train the policy, enabling agents to efficiently reason about what will occur in order to prevent potentially undesirable outcomes [26].

The planning algorithm is one of the critical parts of model-based methods, as it determines what actions to take. Monte Carlo Tree Search (MCTS) [14] has demonstrated excellent performance in a range of complicated games without requiring domain expertise and has established itself as a standard approach for tackling a variety of planning problems.

However, MCTS continues to suffer the problem of exponential complexity development in some situations. When the planning horizon is excessively lengthy, the state space and history of the planned trajectory get significantly larger, resulting in the curse

of dimension. One possible explanation for this phenomenon is that the number of previously visited nodes in the search tree rises exponentially as a result of action branching after each step.

Several previous studies have demonstrated that by introducing macro-actions with a coarser time resolution and multiple levels of abstraction, planning complexity can be decreased. Hierarchical planning [15] improves the performance of planning algorithms by dividing the domain into sub-goals [7, 31].

Numerous methods, on the other hand, presuppose that all sub-goals are known in advance [22, 29]. Manual effort, such as a comprehensive specification of sub-goals based on domain expertise, restricts these solutions to specific domains, which are typically time-consuming and difficult to generalize. Meanwhile, considerable recent research has been conducted on the subject of unsupervised skill discovery, providing strategies for identifying skills that are not rewarded by external tasks [9, 16, 25]. A skill is a strategy that operates over an extended period and develops behaviors capable of decoding the latent variable [25] (We will refer to the combination of the original action sequences used to accomplish these sub-goals as skill later in this work). Techniques for unsupervised skill discovery alleviate some of the load associated with manually specifying rewards for each behavior.

Some recent studies assume accessibility of sub-goals in advance, which is indispensable for collecting data to approximate [9]. Meanwhile, the design of hierarchical online planners remains an open challenge. Therefore, we present a hierarchical planning method that alleviates the demand for human labor to predefine sub-goals or extra work for generating sub-goals while automatically acquiring macros for planners.

Humans usually set task-relevant goals and challenges as intrinsic motivation when they have no relevant reinforcement signal, abstract their actions into behavior or skill, and organize these skills to solve tasks to decompose complex problems [4]. Inspired by this learning pattern, we propose a Hierarchical MCTS Latent Skill Planner (H-MCTS), a technique based on skill discovery that automatically identifies skills via intrinsic reward function. We also combine MCTS planning to perform reasoning with skills, which enables higher-level efficient decision-making. We find that skill-based MCTS is more efficient than standard MCTS and achieves a higher success rate in the Grid-World Maze domain.

## 2 RELATED WORK

Hierarchical Reinforcement Learning (HRL) and temporal abstraction methods [3] usually abstract temporal sequences of primitive actions into macro-actions by dividing the original goal into sub-goals, thereby decomposing a complex task into sub-tasks. There are many decision-making strategies based on temporal abstraction: some algorithms adhere to the options framework [29], in which each option has an internal policy that is performed until

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590005>

a termination condition is fulfilled, given particular states. Value Function Decomposition (MAX-Q) was proposed by [7], which decompose the target Markov Decision Process (MDP) into smaller MDPs. These frameworks presuppose thorough prior knowledge about sub-tasks, which is difficult to acquire without extra regularization.

With the introduction of deep reinforcement learning, some research about HRL has reemerged in recent years. Specifically, [30] presents FeUdal Networks (FuNs). And Hindsight Experience Replay (HER) is a method introduced by [2] that enables efficient learning from sparse rewards. Hierarchical Actor-Critic (HAC) is introduced by [17] as a solution to the instability difficulties that occur when agents attempt to simultaneously learn multiple levels of policies. In addition, there are RL algorithms that intend to utilize skills as pretraining for episodic learning. The majority of current skill discovery algorithms try to maximize the mutual information between skills and states, which might result in an unsupervised diversity maximization objective [1, 8, 10, 25, 37].

Meanwhile, several methods have investigated learning through planned actions. It is feasible to improve reasoning by combining search and learning [11, 13, 28, 35], among which Alpha Zero is a remarkable example of this approach [26]. In addition, there is a significant amount of research on planning with sub-goals or skills [20, 23, 33], which are hierarchical planning techniques.

The authors of [9] provide a method to generic subgoal-based temporal abstraction in MCTS that utilizes a predetermined sub-goal generator as a high-quality heuristic. Divide-and-Conquer Monte Carlo Tree Search (DC-MCTS) is a planning technique proposed by [21] that leverages learnt intermediate sub-goals to hierarchically split initial tasks into simpler ones that are then addressed independently and recursively. [6] offers the Subgoal Search (kSubS) technique, which is a domain-independent hierarchical planner for complicated domains. The approach was inspired by the human mental process of shifting from one concept to a related one.

For comparison, our technique included skill discovery into hierarchical MCTS planning, in which each skill leads to a sub-goal and is selected locally by a high-level planning policy. We assume a constant number and length of skills. Instead of a pre-designed sub-goal space, the meaning of skills leading to a sub-goal is freely given or found during training; hence, we are tolerant of differences in the sub-goal arrival process. Our strategy does not need additional manual labor or complex sub-goal design, other than the training of a skill-based policy. As experimental results, we found that H-MCTS improves planning efficiency and performance in contrast to standard MCTS.

## 3 BACKGROUND

### 3.1 Terminology

We describe a two-level hierarchical planning setup as Markov decision processes (MDPs). We assume the properties of MDPs  $\mathcal{M}_1$  is defined by the tuple  $(\mathcal{S}, \mathcal{Z}, \mathcal{P}, \gamma, R)$  for the high-level policy, as well as  $\mathcal{M}_2$  is defined by  $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R_L)$  for low-level policy. Where  $\mathcal{S}$  is the state space,  $\mathcal{A}$  is the action space. Let  $\mathcal{Z}$  denote a set of skills (latent variables), which is the macro space for high-level policy.  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$  is the environment transition probability.

$\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  is the reward function, and  $\gamma \in [0, 1]$  is the discount factor.

We assume MDPs to be discrete and to have deterministic state transitions. The goal is to find a low-level policy  $\pi : \mathcal{S} \times \mathcal{Z} \rightarrow \mathcal{A}$  outputs primitive actions to produce useful behavior by optimizing low-level intrinsic reward function  $R_L$ :

$$\max_{\pi} \mathbb{E}_{z \sim \mu} \left[ \sum_{k=1}^K \mathbb{E}_{\tau_k \sim \pi, P} [R_L(z_k, \tau_k)] \right] \quad (1)$$

Where  $\tau_k$  is the  $k$ -th trajectory segment. Also to find a high-level policy,  $\mu : \mathcal{S} \rightarrow \mathcal{Z}$  learns to select skills to optimize extrinsic reward function  $\mathcal{R}$ :

$$\max_{\mu, \pi} \mathbb{E}_{z \sim \mu} \left[ \mathbb{E}_{s_t, a_t \sim \pi, P} \left[ \sum_{t=1}^T \gamma^t R(s_t, a_t) \right] \right] \quad (2)$$

### 3.2 Monte Carlo Tree Search

The Monte Carlo Tree Search algorithm is a well-known method for Monte Carlo Planning that has been used to a variety of challenging domains [27].

MCTS explores various possible future states and actions using a simulator or a model of the environment including the reward function and the dynamics transition function, in order to determine the best action to perform from the current node. MCTS also needs heuristics including a value function  $V(s)$  that provides a long-term evaluation or expected return of the tree's leaf node without further rollouts, as well as a search policy function  $\pi_{\text{tree}}$  that acts as a search prior over actions [35].

MCTS generates a search tree progressively, storing the visit counts and values for each simulated state and action. Each iteration of MCTS consists of three phases: selection, expansion, and backup [11].

During the selection phase, MCTS expands the search tree, starting with the root node as the current state and taking steps according to the search policy until a leaf state node is reached. During the expansion phase, the leaf state node  $s_t$  is expanded by a new node representing the next state  $s_{t+1}$ , which is reached after simulating a random action. Then the new state  $s_{t+1}$  is added to the search tree, and its value  $V(s_{t+1})$  is estimated via the state-value function. Eventually, the backup phase begins, during which the value of  $s_{t+1}$  is used to update the value estimates of its parent states in the simulated path [11].

After expanding the nodes for a defined number of time steps, the MCTS returns the improved policy to the root node, resulting in a potentially better policy than the current network.

### 3.3 Skill Discovery

Several previous methods maximize the mutual information (MI) between latent variables and states as objective, which adopted a probabilistic approach to learning diverse skills. The MI objective  $I(Z; S)$  can be written with the variational lower bound as

follows [8]:

$$\begin{aligned} I(Z; S) &= -H(Z | S) + H(Z) \\ &= \mathbb{E}_{z, s \sim p(z), p^\pi(s|z)} [\log p(z | s) - \log p(z)] \\ &\geq \mathbb{E}_{z, s} [\log q(z | s)] + (\text{const}) \end{aligned} \quad (3)$$

where a skill  $z$  is sampled from a fixed prior distribution  $p(z)$ , and the skill discriminator  $q(z | s)$  is a variational approximation of the posterior  $p(z | s)$ . As a result, optimizing the MI objective, which quantifies the reduction in uncertainty about the states given the skill, leads to discovering diverse and distinguishable behaviors. These methods encourage  $z$  to be maximally informative of states or trajectories obtained from a skill conditioned policy  $\pi(a | s, z)$ , executed under the environment dynamics.

For example, Diversity is All You Need (DIAYN) [8] optimizes the MI between individual states and skills. Variation Option Discovery (VALOR) [1] also takes a similar approach but considers the whole trajectories instead of states. The objective of Variational Intrinsic Control (VIC) [10] maximizes the MI between the last states and skills given the initial state. Some other methods are based on a conditional form of mutual information [25].

### 3.4 Model Reinforcement Learning Methods

Temporal difference learning (TD) [12] and policy gradient based methods (PG) [18] are two types of popular model free reinforcement learning algorithms.

For instances, Deep Q-learning (DQN) [19], which is a Q-learning extension in which a deep neural network is trained to predict Q-values from states rather than manually tabulating Q-values. The loss function of a neural network trained for DQN is defined as follows:

$$\mathcal{L}^{TD} = \mathbb{E}_{s, a, r, s'} \left[ \left( r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)^2 \right] \quad (4)$$

In our framework, we can update low-level conditioned policy with both policy gradient and temporal difference methods to maximize the intrinsic reward. For both methods, we gather experiences using an epsilon-greedy technique in our implementations.

## 4 METHOD

In this section, we present our proposed approach Hierarchical Monte Carlo Tree Search for Latent Skill Planning. We describe an outline of H-MCTS in Algorithm 1, which begins by initializing replay buffers at both levels of the hierarchy for off-policy updates, and by initializing a dataset  $\mathcal{D}$  for the discriminator. The policies are continually updated depending on online interactions.

Specifically, H-MCTS acquires the following neural network distributions: A low-level policy  $\pi_\theta$  trained by environmental rollouts through intrinsic rewards; A discriminator  $q_v$  for the intrinsic reward learning; A high-level policy  $\mu_\phi$  that generates curriculums for skill training.

Figure 1 depicts our proposed framework, which includes a two-level hierarchy, applying skill-space planning. The high-level process equips the agent with the reasoning capacity to anticipate and seek potential outcomes. The low-level policy determines which action to perform given the local state  $s$  and the active latent variable

### Algorithm 1 Hierarchical Monte Carlo Planning

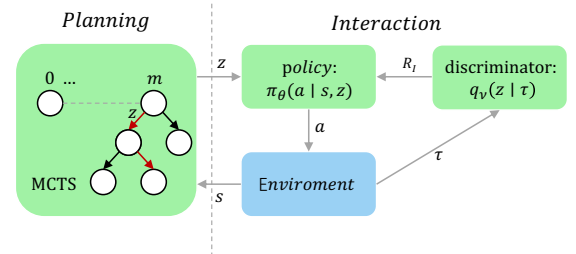
---

**Initialize:** Buffers, high-level  $Q_\phi$ , low-level  $Q_\theta$  trajectory-skill dataset  $\mathcal{D}$ , discriminator  $q_v$

- 1: **for**  $idx = 0, 1, 2, \dots$  **do**
- 2:   Initialize trajectory storage  $\tau$  of max length  $K$
- 3:   Begin episode at  $s$ .
- 4:   **while** not done **do**
- 5:     **if** high-level process **then**
- 6:       Compute  $Q_{\text{MCTS}}(s, \cdot) \leftarrow \text{MCTS}(s, Q_\phi)$
- 7:       Select  $z$  with epsilon-greedy over  $Q_{\text{MCTS}}(s, \cdot)$
- 8:       Store high-level transitions to high-level buffer.
- 9:       Update  $Q_\phi(s, z)$  with high-level buffer
- 10:     **end if**
- 11:     Update skill conditioned policy with  $\text{TrainPolicy}(\mathcal{D}, Q_\theta, z, q_v)$
- 12:   **end while**
- 13: **end for**
- 14: **return** solution

---

$z$ , which is responsible for interacting with the environment. The intuition for this design choice is that skill can assist the planner in exploring faster and planning more efficiently. The hierarchical structure decomposes the original task into smaller ones, and the state space in the tree search is correspondingly reduced to the sub-state space, thus the planner does not need to simulate each local state, but skips some states.



**Figure 1: Structure of H-MCTS.** In the MCTS procedure, the linkages within the tree search indicate skills, the root node represents the current state, and the remaining nodes represent states attained when skills are executed, which are equivalent to subgoals. As indicated above, each  $z$  is retained for a defined number of  $K$  low-level steps.

### 4.1 Skill Discovery via Intrinsic Reward

Our objective is to learn a low-level policy in which action distributions are conditioned on both the current state  $s$  and a latent variable  $z$ , which is most useful from the current state, instead of arbitrary skills. The latent variable  $z$  should be able to specify a specific mode of behavior in an unambiguous manner.

Following VALOR [1], we aim to assign or learn the meaning of a latent skill arbitrarily during training. Thus, we define the low-level reward in this section by introducing a discriminator  $q_v(z | \tau)$  that predicts the ground truth latent skill  $z$  applied in the

**Algorithm 2** Training Latent Skills

---

**Function:** *TrainPolicy*(trajectory-skill dataset  $\mathcal{D}$ , low-level  $Q_\theta$ , selected skill  $z$ , discriminator  $q_v$ )

- 1: **for**  $idx = 0, 1, 2, \dots, K$  **do**
- 2:   Initialize trajectory storage  $\tau$  of max length  $K$
- 3:   Select  $a$  with epsilon-greedy from  $Q_\theta(s, z, a)$
- 4:   Execute  $a$  in environment and receive  $s', r$
- 5:   Compute  $R_L$  using Equation 5
- 6:   Store low-level transitions to buffer
- 7:   Add  $s'$  to  $\tau$
- 8:   Update  $Q_\theta(s, z, a)$  with low-level buffer
- 9: **end for**
- 10: Store  $(z, \tau)$  into  $\mathcal{D}$
- 11: Calculate intrinsic reward  $R_I$  for  $\tau$  with  $q_v$
- 12: **if** discriminator training **then**
- 13:   Update discriminator  $q_v$  using  $\mathcal{D}$  then empty  $\mathcal{D}$
- 14: **end if**
- 15: **return** solution

---

low-level policy  $\pi_\theta$  that generated the trajectory  $\tau$ , indicating that  $z$  is encoded in the trajectory. If the trajectory is unique to  $z$ , the decoder will assign it a high probability, and the policy should be reinforced accordingly.

We train the discriminator using a dataset  $\mathcal{D} = \{z, \tau\}$  of skill trajectory pairs which are accrued online throughout training, where each pair consists of a high-level policy’s choice of  $z$  and the matching trajectory created by the low-level policy given  $z$ . Since we have access to the ground truth label  $z$  associated with each trajectory, training  $q_v$  can be viewed as a supervised learning task [1, 10, 34].

Furthermore, we define the intrinsic reward  $R_I(z_k, \tau_k)$  for the  $k$ -th trajectory segment  $\tau_k$  of the agent using the discriminator’s prediction performance on the tuple  $(z_k, \tau_k)$ . The underlying assumption is that a skill may be deduced from the history of states trajectory, thus the design of intrinsic reward promotes the generation of distinguishable behavior for distinct skills. The low-level reward can be defined as:

$$R_L(z, \tau) = \sum_{s_t, a_t \in \tau} \gamma^t R(s_t, a_t) + R_I(z, \tau) \quad (5)$$

Agent takes actions at each low-level time step based on skill selection, employing policy  $\pi$  induced by low-level Q-function  $Q_\theta(s_t, z_t, a)$ . We optimize the low-level objective by using DQN to optimize  $Q_\theta$  via minimizing the loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{z \sim \mu, a \sim \pi} \left[ \frac{1}{2} (y_t - Q_\theta(s_t, z, a_t))^2 \right]$$

$$y_t = R_L(z, \tau) + \gamma \max_a \hat{Q}_\theta(s_{t+1}, z, a_{t+1}) \quad (6)$$

The low-level process is summarized in Algorithm 2. Only at the final time step of each trajectory fragment does the low-level reward  $R_L$  include the intrinsic reward  $R_I$ . Once a batch of  $(z, \tau)$  is gathered into the dataset  $\mathcal{D}$ , the skill discriminator  $q_v(z | \tau)$  is trained to predict  $z$  given  $\tau$  by minimizing a typical cross-entropy loss using supervised learning on  $\mathcal{D}$ . Each selected  $z$  serves as the class label for the associated trajectory  $\tau$ .

**Algorithm 3** MCTS Procedure

---

**Function:** *MCTS*(initial state  $s_0$ , high level  $Q_\phi$ )

- 1: Initialize edge statics  $Q_0(s, a)$ ,  $N_0(s, a)$  for all  $s, a$
- 2: **for**  $idx = 0, 1, 2, \dots, SearchBudget$  **do**
- 3:   Traverse the search tree with  $\pi_m$  using Equation 7
- 4:   Search and expand new state  $s_H$
- 5:   Calculate  $\max_z Q_\phi(s, z)$
- 6:   Backup returns
- 7: **end for**
- 8: **return**  $Q_{MCTS}(s_0, \cdot)$

---

**4.2 Integrating Skill into MCTS**

We propose using model-based planning via MCTS as a policy improvement operator for high-level policy in our approach. We differ from previous MCTS papers in that we plan with the skill set. When executing a skill, we employ the skill-conditioned policy as a generator and roll out for  $K$  steps in order to obtain the states trajectory.

Consequently, in comparison to the original MCTS, the search state space is reduced to the sub-goal space, and the action space is transformed into the skill space. In our approach, skills serve as the preferred choice of interesting options available to planners to seek meaningful changes in state. While utilizing model-free RL, it is still possible to obtain the benefits of model-based planning, including exploration and long-term reasoning [5]. We summarize this subroutine in Algorithm 3.

In practice, in order to reduce planning cost, we follow Combining Q-learning and Search (SAVE) [11], an Alpha Zero like algorithm for single-player games.

For skill selection in the  $m^{\text{th}}$  iteration of Monte Carlo tree search, we descend the tree and take skills from the root node according to:

$$\pi_m(s) = \arg\max_z \left( Q_m(s, z) + c_{UCT} \sqrt{\frac{\log(\sum_z N_m(s, z))}{N_m(s, z)}} \right) \quad (7)$$

Where  $c_{UCT}$  is a constant that supports UCT term exploration [14].  $N(s, z)$  is the number of times we have explored taking  $z$  from state  $s$ .  $Q_m$  is the currently estimated value of taking  $z$  while in state  $s$ .

The selection procedure is repeated for  $H$  times. In the expansion phase,  $z_H$  is used in the low-level policy that executed in the simulator, resulting in a new state  $s_{H+1}$  and a reward  $r_H$  that can be obtained by summing the cumulative rewards of  $K$  steps. The new state is then added to the search tree and its value is estimated via a state-value function  $V(s) = \max_z Q_\phi(s, z)$ . After  $M$  iterations, we return  $Q_{MCTS}(s, z)$  to select one of the skills over  $Q_{MCTS}(s_0, z)$  [11]. And high-level Q-network  $Q_\phi(s, z)$  can be trained also by minimizing the Q loss function using DQN.

**5 EXPERIMENT**

In our experiments, the proposed H-MCTS method is evaluated on grid-world maze [38], a well-known navigation task, and compared to the standard MCTS algorithm. We assess the planning performance of H-MCTS and demonstrate that outperforms its standard counterparts.

As a baseline, we employ standard MCTS algorithm, which is implemented by limiting the H-MCTS to being a single-player implementation of an Alpha Zero like algorithm. For H-MCTS implementation, initial pretraining episodes are used to warm up the skill-conditioned policy learning, followed by training of the entire hierarchical structure until convergence. Both the algorithms utilize policy networks with the same fundamental architecture as heuristics for MCTS, as a result, their complexity is comparable. Except when specifically stated otherwise, all remaining parameters and design approach were the same for both planners. We assess the performance of H-MCTS versus regular MCTS as described in further detail below.

### 5.1 Environment Description

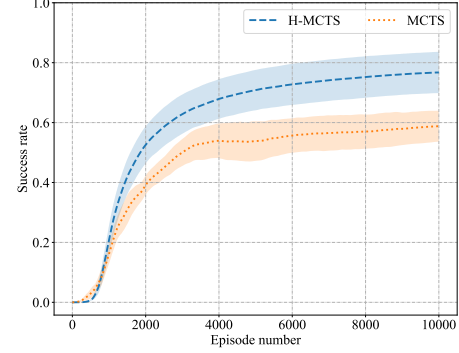
In the Grid-World Maze domain [21, 38], each task consists of a procedurally randomly generated maze. The structure of the maze is represented as a feature map of categorical variables, each of which has four categories, including empty, wall, start location, and goal position. The underlying MDP has four basic actions: up, down, left, and right. At each timesteps, the agent receives just its current coordinates and outputs an action that controls its position change, which is actually impacted by its collision with the wall. The density of walls,  $d$ , determines the complexity of the maze challenge, where  $d=0.0$  corresponds to the easiest setting, with no walls, and  $d=1.0$  corresponds to so-called perfect or single-connected mazes. In our experiments, we fixed the layout of the maze and  $d$  was adjusted to 0.5 to guarantee that there was sufficient space for skills to be randomly explored in order to get randomly assigned behavioral meanings. We constructed a maze on a  $20 \times 20$  grid that was initiated within the specified range. Both algorithms use policy neural networks as search heuristics and were given a maximum exploration budget of 200, which is inadequate for unguided planners to locate a feasible path in the majority of mazes [21].

### 5.2 Results

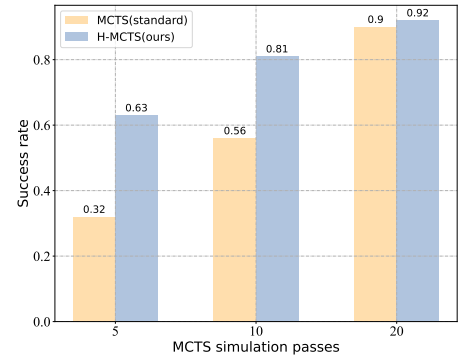
In this part, we report our finding: H-MCTS enables more efficient searching and better performance, which outperform the standard MCTS methods that do not incorporate skills.

Figure 2 depicts the learning curves of both approaches. The success rate is measured as a function of the training episodes. H-MCTS has improved performance in comparison to MCTS. This result follows the previous paper’s findings. Hierarchical planning can ease the issue that the policy cannot be reinforced proportionately due to the difficulty of receiving the incentive signal in a reward-sparse environment, and ultimately lead to improved performance. The hierarchical nature of our system decomposes the original work into smaller tasks to be solved. Based on skill planning, the planner is capable of searching the objective more quickly with expansion of the original single-step actions, while the intrinsic reward function gives a denser reinforcement signal to the policy.

In addition, we report MCTS simulation pass, which is a common metric for MCTS training. As shown in Figure 3, under the same simulation budgets, the performance of H-MCTS exceeds that of standard MCTS, especially when compared to a small search budget, such as a simulation pass equal to 5, where the advantage of H-MCTS is particularly evident. Standard MCTS gradually achieves



**Figure 2: Performance of H-MCTS and standard MCTS on gridworld maze navigation. Lines show rate of goal completion averaged and standard deviations over several experiments (shaded area shows mean $\pm$ std).**



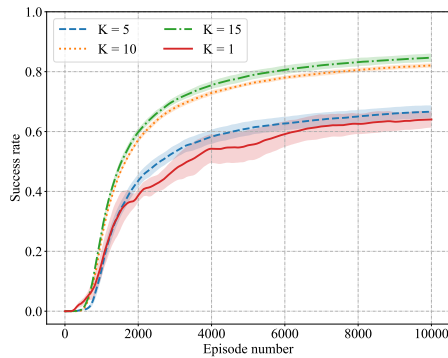
**Figure 3: The performance of Skill Search. Comparison on maze  $20 \times 20$  to standard MCTS. H-MCTS regularly delivers a high level of performance, even with limited computational resources. The advantage of H-MCTS for small budgets is clear.**

a higher level of prediction accuracy as the number of simulations passes increases, whereas H-MCTS consistently achieves significantly better performance even when there are fewer simulation passes. In our technique, the original state space in the tree search is drastically reduced to the sub-goal space. As a result, the H-MCTS planner does not need to simulate in every local state, but rather bypasses some states, requiring a search that is reasonably small or equal. Thus, skills can enable a high-level planner to explore more rapidly and assign credit more easily, significantly reducing the complexity of planning.

### 5.3 Analysis of the Skills’ Length

In this part, we analyzed the impact of altering the hyper-parameter  $K$  of H-MCTS, which represents the execution time steps sustained by the low-level policy conditioned on the current latent variable.

Figure 4 demonstrates that agents that maintain their skill for 10 or 15 time steps outperform those who maintain their skills for just 5 time steps, which is consistent with prior results that



**Figure 4: The success rates of H-MCTS at various  $K$  values. The orange curves reflect the  $K$  values utilized in the primary experiments, i.e.  $K = 10$ .**

higher  $K$  should make planning simpler since the search state space is smaller. Moreover, a lower  $K$  indicates that agents make more frequent decisions to maintain or alter their choice of skill, which allows for more flexible policies but increases the complexity of learning, as the fitting and convergence of discriminator will become extremely challenging. An assumption in our framework is that we rely on some pretraining episodes to warm up low-level conditioned policy learning, since skill acquisition is notoriously tough, particularly as the  $K$  size gets lower, training the discriminator becomes problematic.

## 6 CONCLUSION

We proposed an approach to implement hierarchical planning with MCTS, which incorporates skills with a conditioned policy. Our experiments show that H-MCTS using skills is competitive against standard MCTS planning with original actions in terms of performance and planning efficiency.

For future work, we may consider training a world model before the head instead of a real simulator. Moreover, owing to the behavioral meaning of skills being arbitrarily allocated, the way of examine the quality of skills remains an open challenge. In addition, for this study, we acquire skills in the raw state space, which may be insufficient for tasks requiring a high-dimensional visual input, thus we would like to test H-MCTS in more environments in the future as well.

## REFERENCES

- [1] Joshua Achiam, Harrison Edwards, Dario Amodei, and Pieter Abbeel. 2018. Variational option discovery algorithms. *arXiv preprint arXiv:1807.10299* (2018).
- [2] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. *Advances in neural information processing systems* 30 (2017).
- [3] Andrew G Barto and Sridhar Mahadevan. 2003. Recent advances in hierarchical reinforcement learning. *Discrete event dynamic systems* 13, 1 (2003), 41–77.
- [4] Andres Campero, Roberta Raileanu, Heinrich Küttler, Joshua B Tenenbaum, Tim Rocktäschel, and Edward Grefenstette. 2020. Learning with amigo: Adversarially motivated intrinsic goals. *arXiv preprint arXiv:2006.12122* (2020).
- [5] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. 2018. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems* 31 (2018).
- [6] Konrad Czechowski, Tomasz Odrzygóźdź, Marek Zbysiński, Michał Zawalski, Krzysztof Olejnik, Yuhuai Wu, Łukasz Kuciński, and Piotr Miłoś. 2021. Subgoal search for complex reasoning tasks. *Advances in Neural Information Processing Systems* 34 (2021), 624–638.
- [7] Thomas G Dietterich. 2000. Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of artificial intelligence research* 13 (2000), 227–303.
- [8] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070* (2018).
- [9] Thomas Gabor, Jan Peter, Thomy Phan, Christian Meyer, and Claudia Linnhoff-Popien. 2019. Subgoal-Based Temporal Abstraction in Monte-Carlo Tree Search. In *IJCAI*. 5562–5568.
- [10] Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. 2016. Variational intrinsic control. *arXiv preprint arXiv:1611.07507* (2016).
- [11] Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Tobias Pfaff, Theophane Weber, Lars Buesing, and Peter W Battaglia. 2019. Combining q-learning and search with amortized value estimates. *arXiv preprint arXiv:1912.02807* (2019).
- [12] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. 2018. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*.
- [13] Bilal Kartal, Pablo Hernandez-Leal, and Matthew E Taylor. 2019. Action guidance with MCTS for deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 15. 153–159.
- [14] Levente Kocsis and Csaba Szepesvári. 2006. Bandit based monte-carlo planning. In *European conference on machine learning*. Springer, 282–293.
- [15] Andrey Kolobov. 2012. Planning with Markov decision processes: An AI perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6, 1 (2012), 1–210.
- [16] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. 2019. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274* (2019).
- [17] Andrew Levy, Robert Platt, and Kate Saenko. 2017. Hierarchical actor-critic. *arXiv preprint arXiv:1712.00948* 12 (2017).
- [18] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.
- [19] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.
- [20] Soroush Nasiriany, Vitchyr Pong, Steven Lin, and Sergey Levine. 2019. Planning with goal-conditioned policies. *Advances in Neural Information Processing Systems* 32 (2019).
- [21] Giambattista Parascandolo, Lars Buesing, Josh Merel, Leonard Hasenclever, John Aslanides, Jessica B Hamrick, Nicolas Heess, Alexander Neitz, and Theophane Weber. 2020. Divide-and-conquer monte carlo tree search for goal-directed planning. *arXiv preprint arXiv:2004.11410* (2020).
- [22] Ronald Parr and Stuart Russell. 1997. Reinforcement learning with hierarchies of machines. *Advances in neural information processing systems* 10 (1997).
- [23] Karl Pertsch, Oleh Rybkin, Frederik Ebert, Shenghao Zhou, Dinesh Jayaraman, Chelsea Finn, and Sergey Levine. 2020. Long-horizon visual planning with goal-conditioned hierarchical predictors. *Advances in Neural Information Processing Systems* 33 (2020), 17321–17333.
- [24] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [25] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. 2019. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657* (2019).
- [26] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhruv Kumar, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.
- [27] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
- [28] David Silver, Richard S Sutton, and Martin Müller. 2008. Sample-based learning and search with permanent and transient memories. In *Proceedings of the 25th international conference on Machine learning*. 968–975.
- [29] Richard S Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* 112, 1–2 (1999), 181–211.
- [30] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. Feudal networks for hierarchical reinforcement learning. In *International Conference on Machine Learning*.

- PMLR, 3540–3549.
- [31] Ngo Anh Vien and Marc Toussaint. 2015. Hierarchical monte-carlo planning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
  - [32] Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. 2019. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057* (2019).
  - [33] Kevin Xie, Homanga Bharadhwaj, Danijar Hafner, Animesh Garg, and Florian Shkurti. 2020. Latent skill planning for exploration and transfer. *arXiv preprint arXiv:2011.13897* (2020).
  - [34] Jiachen Yang, Igor Borovikov, and Hongyuan Zha. 2019. Hierarchical cooperative multi-agent reinforcement learning with skill discovery. *arXiv preprint arXiv:1912.03558* (2019).
  - [35] Weirui Ye, Shaohuai Liu, Thanard Kurutach, Pieter Abbeel, and Yang Gao. 2021. Mastering atari games with limited data. *Advances in Neural Information Processing Systems* 34 (2021), 25476–25488.
  - [36] Tianhe Yu, Aviral Kumar, Rafael Rafailov, Aravind Rajeswaran, Sergey Levine, and Chelsea Finn. 2021. Combo: Conservative offline model-based policy optimization. *Advances in neural information processing systems* 34 (2021), 28954–28967.
  - [37] Jesse Zhang, Haonan Yu, and Wei Xu. 2021. Hierarchical reinforcement learning by discovering intrinsic options. *arXiv preprint arXiv:2101.06521* (2021).
  - [38] Xingdong Zuo. 2018. mazelab: A customizable framework to create maze and gridworld environments. <https://github.com/zuoxingdong/mazelab>.

# Adaptive model fusion algorithm for decision trees and association rules

Hui Zhang,H,Zhang  
China FAW Group Corporation CAV  
Development Research Institute  
wenzhangjida2022@163.com

Zhiling Nie ,ZL,Nie  
China FAW Group Corporation CAV  
Development Research Institute  
niezhiling@faw.com.cn

Hongwei Xiao,Hw,Xiao\*  
College of Automotive Engineering ,  
Jilin university  
xiaohw@jlu.edu.cn

## CCS CONCEPTS

• This paper proposes an adaptive model fusion algorithm based on decision trees and association rules. The decision trees are fused with association rules, the results of the decision trees are used as a priori conditions for the calculation of association rules, and the results of the model fusion are further fused with the results of the association rules to obtain the results of the algorithm. To determine the effectiveness of the algorithm, this paper collects data from 828 participants and applies the algorithm to obtain the results of the algorithm to effectively mine the relationships and rules that exist between real data, which has certain guiding significance in practical applications.; • CCS CONCEPTS; • Mathematics of computing → Probability and statistics;

## KEYWORDS

decision trees, association rules, model fusion, adaptive, data mining

### ACM Reference Format:

Hui Zhang,H,Zhang, Zhiling Nie ,ZL,Nie, and Hongwei Xiao,Hw,Xiao\*. 2023. Adaptive model fusion algorithm for decision trees and association rules. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590006>

## 1 INTRODUCTION

In recent years, with the continuous development of machine learning algorithms, the application scenarios have been expanded and decision tree algorithms and association rule algorithms have been updated among them. Decision tree algorithms are used in the education industry for course grade prediction [1], in the medical industry for inferring disease rules [2], and in the financial industry for financial self-service terminal diagnosis [3], decision trees have an extremely wide range of applications. Association rules are also used in a wide range of scenarios in the education industry for mining information related to education management [4], in the medical industry for data analysis of heart diseases [5], and in the financial industry for observation and early warning of key sectors and functions [6]. Although decision trees and association rules

have certain application scenarios, since decision trees can only give subgroup results, i.e., a certain group of people is classified into a specific category based on the characteristics of the data present, but cannot give a specific relationship, association rules can solve the problem of decision trees that cannot give rules. This paper addresses the above problem by developing an adaptive data mining algorithm combining decision trees and association rules and applying the algorithm to car color schemes to find the data rules that exist for car color scheme designs.

### 1.1 Decision Trees

The decision tree algorithm is a method for approximating a discrete-valued objective function that represents the classification law as a tree structure [7]. The algorithm ID3 [8] in decision tree was proposed in 1986. The ID3 algorithm assumes that the sample set  $E$  has a training set of  $C$  classes of samples, with the number of samples in each class being  $p_i, i = 1, 2, \dots, C$ . If attribute  $A$  is used as the test attribute, and the  $v$  different values of attribute  $A$  are  $\{v_1, v_2, \dots, v_n\}$ , one can use attribute  $A$  to partition  $E$  into  $v$  subsets  $\{E_1, E_2, \dots, E_v\}$ , and assume that the number of samples containing the  $j$ th class in  $E_i$  is  $p_{ij}, j = 1, 2, \dots, C$ , then the entropy [9] of subset  $E_i$  is

$$\text{Infor\_Entropy}(E_i) = - \sum_{j=1}^C \frac{p_{ij}}{|E_i|} \log_2 \frac{p_{ij}}{|E_i|}$$

The information entropy of attribute  $A$  is

$$\text{Infor\_Entropy}(A) = \sum_{i=1}^v \frac{E_i}{|E|} \text{Infor\_Entropy}(E_i)$$

The information required for a decision tree to make a correct category judgement for the same example is  $\text{Infor\_Entropy}(E) = - \sum_{i=1}^C \frac{|C_i|}{|E|} \log_2 \frac{|C_i|}{|E|} \quad i = 1, 2, \dots, C$ . Information gain:

$$\text{InforGain}(A) = \text{InforEntropy}(E) - \text{Infor\_Entropy}(A)$$

The first decision tree algorithm is CLS (Concept Learning System), the algorithm that brought attention to decision trees and made them a mainstream machine learning technique is ID3, the most used decision tree algorithm is C4.5, the decision tree algorithm that can be used for regression tasks is CART (Classification and Regression Tree), and the most powerful algorithm based on decision trees is RF (Random Forest). The most powerful algorithm based on decision trees is RF (Random Forest), and this paper uses the ID3 algorithm for algorithm development.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590006>

## 1.2 Association rules <sup>[10]</sup>

Association is a reflection that an event is dependent or related to other events to a certain extent and can be predicted according to the corresponding rules. Association rules [11] are a more widely used pattern recognition method and generally use three metrics to measure association rules, namely confidence, support, and boost. Support indicates the probability of both appearing in the rule at the same time, in no order, confidence indicates the probability of A appearing while B appears, and lift describes the correlation between A and B in the association rule.

Support is the probability of both occurring at the same time in the association rule. If the probability of occurring at the same time is small, there is little relationship, and if the probability of occurring at the same time is very frequent, the two are related, i.e.

$$\text{support}(A \Rightarrow B) = P(A \cup B)$$

Confidence is the probability that when A occurs B also occurs. If the confidence level is 100%, then AB can be bundled and introduced, otherwise AB will be disregarded as defined as relational intimacy, i.e.

$$\text{confidence}(A \Rightarrow B) = P(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

The lift is the ratio of the proportion containing A and containing B, to the proportion containing B. The higher the lift is above 1 and the higher the positive correlation, the lower the lift is below 1 and the opposite, i.e.

$$\text{lift}(A \Rightarrow B) = P(B|A) / P(B)$$

These three indicators are usually used to measure a correlation rule, according to the three indicators to filter the correlation rule to meet the conditions, to meet the minimum support and minimum confidence rule is called a strong correlation rule, if at the same time the lift is greater than 1 is called a valid strong correlation rule, the lift is less than 1 is invalid strong correlation rule, the lift is equal to 1 means that the two are independent of each other no relationship.

## 2 ADAPTIVE DATA MINING ALGORITHMS COMBINING DECISION TREES AND ASSOCIATION RULES

The problem with decision tree algorithms for data mining is that they can only classify subgroups in the traditional sense, but they cannot explain the internal relationships between subgroups very well. The problem with association rule algorithms for data mining is that first we need to find frequent item sets and then set support and confidence intervals for the frequent item sets to find rules with strong correlations. Association rules were first used to determine the association between products purchased in the purchasing process, for example, by buying milk and presumably buying bread, but the problem with association rules is that when the data contains too much information, the association rules will generate more frequent items and correlations, but when there are more weak correlations, resulting in lower computational efficiency, the decision tree and association rules combined with adaptive data mining algorithms can The adaptive data mining algorithm combining decision trees and association rules can solve this problem

well. At the same time, the results of the joint decision tree and association rule model are fused with the results of the association rule to avoid information loss to a certain extent.

The basic idea of the algorithm is as follows: firstly, the ID3 algorithm is used to classify the multi-level data into decision trees, the multi-level data will exist at a certain level where the decision trees cannot be classified, the algorithm will discard the data at this level and keep the data where the decision trees exist, and the decision trees will be divided into subgroups according to the results of the decision trees, the subgroups will be analyzed separately using association rules to find the relationships with strong correlation, and then the original data will be The top 10 support results obtained from the association rules are fused with the decision tree and association rule results, which is the result of data mining. This algorithm can achieve the input data directly output that there is a correlation relationship and can be divided into subgroups of the results, but also in the adaptive link to add human intervention information, proposed decision tree analysis results, or add some data for association rule analysis, the pseudo code of this algorithm is as follows:

## 3 EXPERIMENTAL PROCESS AND ANALYSIS OF RESULTS

In this paper, a questionnaire survey was used to collect a total of 828 participants from various provinces in China, collecting basic information on participants' age, gender, city, and education. The city is divided into two regions according to the south and the north, and the education level is divided into three levels: specialist, undergraduate and graduate. The colors were divided into four levels according to the red, yellow, green and blue tones, and the color brightness was divided into five levels, and the participants' color choices for instrument backlights, road displays, fuel signs and background colors were collected, and the instrument backlights, road displays, fuel signs and background colors were set to four levels, and the four levels were analyzed, and due to the large volume of data, the method of bar charts was used to show separately the situation of color and brightness, the basic situation of the data is as follows, combined with the results of the histogram can be found that the data distribution is balanced, conducive to the next step of analysis.

To facilitate visualization of the whole process, the output of the decision tree results is visualized in this paper. The visualization results of the color decision tree (Figure 3) show that there are no nodes for the background color, indicating that there is no difference in the color scheme of the background color between people of different ages, geographical regions, educational backgrounds, and genders. By analyzing the raw data it will be found that most of the investigators tend to choose the blue color scheme, from a psychological point of view, blue can bring people the intuitive feeling of composure, and in the process of driving a vehicle, the background color As a foil for other colors, it is necessary to maintain a relatively subdued color that can better highlight the role of other parts.

Combining the results of the color decision tree (Figure 3) with the color association rule (Table 1), it can be found that for the

Input	Training set: $D = \{(x_1, y_1, z_1, p_1, \dots), (x_2, y_2, z_2, p_2, \dots), \dots, (x_n, y_n, z_n, p_n, \dots)\}$ Property set: $A = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$
Process	$f(D, A)$
1:	Generate nodes;
2:	If the samples in D all belong to the same category C then;
3:	mark the node as a class C leaf node; return
4:	end if
5:	If $A = \emptyset$ or samples in D take the same value on A then
6:	mark the node as a leaf node and its class as the class with the most sample books in D; return
7:	end if
8:	Select the optimal division attribute $\alpha_*$ from A
9:	for each value of $\alpha_*$ $\alpha_*^V$ do
10:	generate a branch for the node; let $D_V$ denote the subset of samples in D that take the value $\alpha_*^V$ on $\alpha_*$ .
11:	If $D_V$ is empty, then
12:	mark the branch node as a leaf node and its class as the class with the most samples in D; return
13:	else
14:	with $\{D_V, A/(\alpha_*)\}$ as the branch node
15:	If $\alpha_*$ is not equal to empty, then
16:	Generate a new training set $M = \{(x_1, y_1, z_1, p_1, \dots), (x_2, y_2, z_2, p_2, \dots), \dots, (x_n, y_n, z_n, p_n, \dots)\}$
17:	Property set: $B = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$
18:	Retrieve data separately for each group of data in the dataset N then
19:	For each candidate set Q
20:	Check if Q is a subset of N
21:	If so, increase the count of Q
22:	For each candidate set
23:	If its support is not less than the minimum, then keep the set
24:	Return a list of all frequent sets
25:	End



Figure 1: Basic color selection

choice of instrument backlight, the decision tree shows that different regions and educational backgrounds affect the choice of instrument backlight. When the results obtained from the decision tree are subjected to the adaptive association rule, the results of the first 10 degrees of support are retained, and it can be found

that only regions are retained, and then combined with the results of the association rule, it can finally be found that there is a certain correlation between the south, men and young people in the choice of yellow and blue tones, and the probability that those

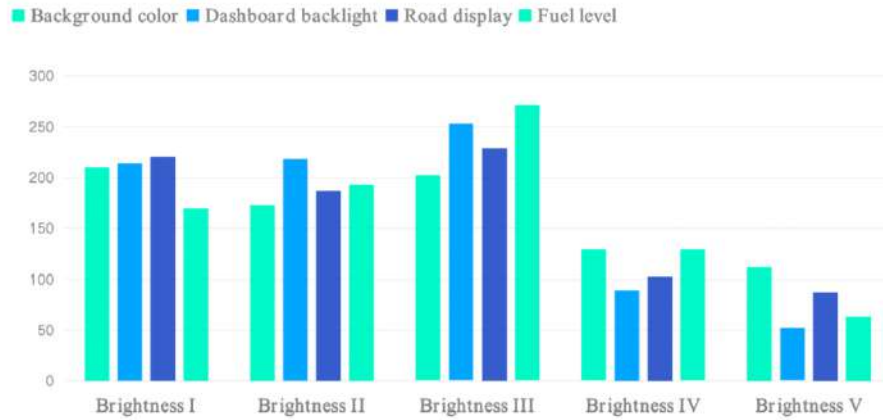


Figure 2: Basic brightness selection

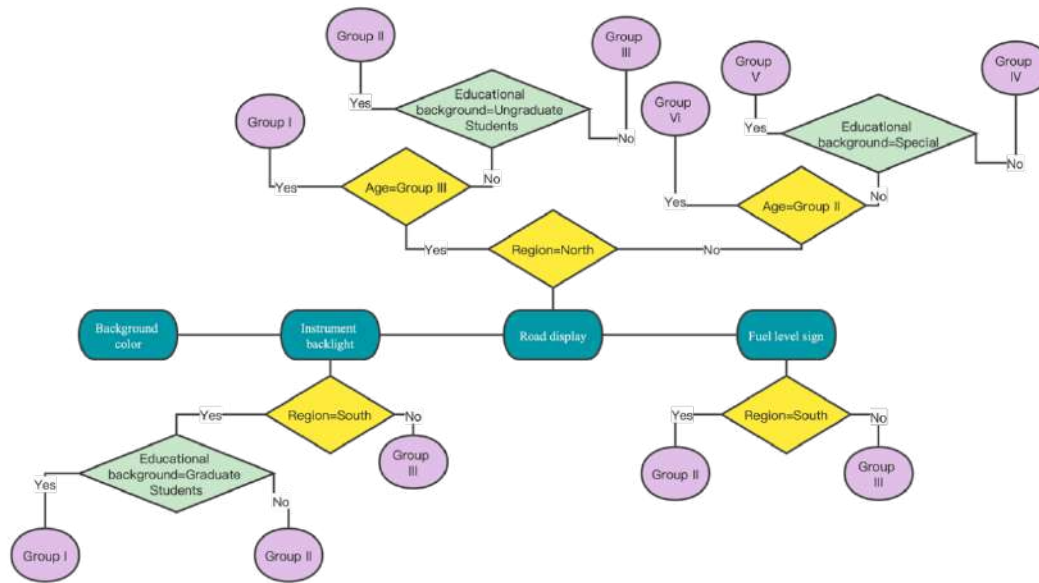


Figure 3: Decision tree results for color scheme selection

who choose yellow and blue tones are from the south amounts to 83.1%, and the probability that they are men amounts to 80%.

For the selection of road signs, the results of the decision tree show that region, age and educational background affect the group, and after fusing the adaptive association rule with the association rule model, it is found that gender is also an important factor, and the probability of choosing yellow and blue colors for road signs is 88% for people from the south, 87.5% for men and 87.5% for young people. For the selection of the fuel sign, the results of the decision tree show that region is an influential factor, and after fusing the adaptive association rules with the association rule

model the results show that the probability of choosing yellow and red colors is 61% for people from the south.

Combining the luminance analysis decision diagram with the association rule fusion results it can be found that the oil marker does not find a good subgroup for representation in the decision tree. For the choice of background color, the results of the decision tree show that region, education and age are the nodes, and the results of the association rule fusion can be found that region has a certain correlation for the choice of luminance, and the probability of choosing background colors luminance 3 and luminance 5 from the south is about 61.8%. For the choice of dashboard backlight,

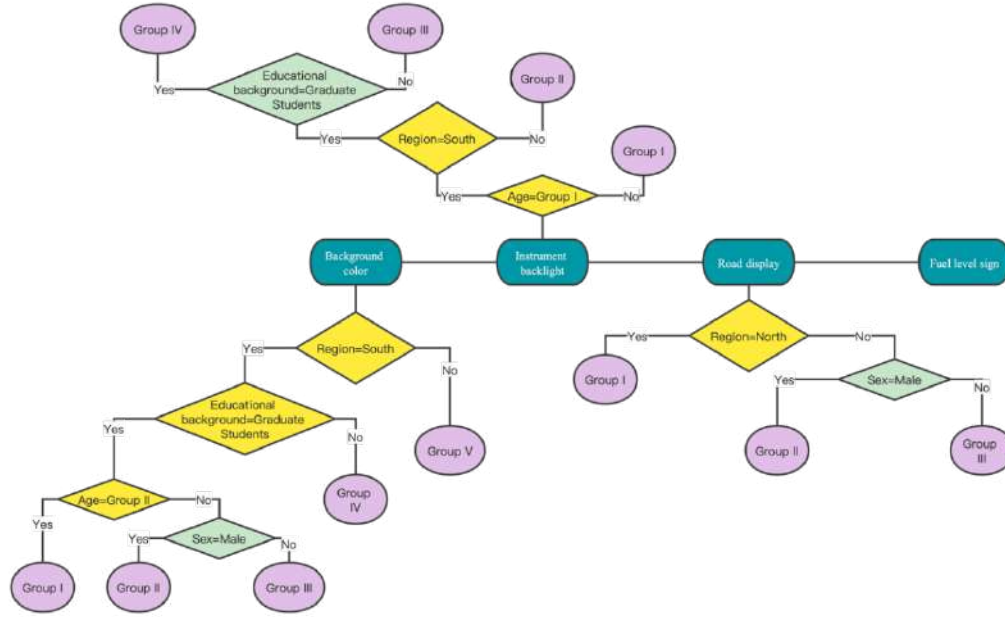


Figure 4: Decision tree results for brightness selection

Table 1: Association rule model fusion results

Instrument backlight			Support	Probability
Region = South	→	Instrument backlight = [2, 4]	0.523	0.638
Instrument backlight = [2, 4]	←	Region = South	0.523	0.831
Sex = male	→	Instrument backlight = [2, 4]	0.437	0.533
Instrument backlight = [2, 4]	←	Sex = male	0.437	0.805
Age = [1, 2]	→	Instrument backlight = [2, 4]	0.420	0.512
Instrument backlight = [2, 4]	←	Age = [1, 2]	0.420	0.724
Fuel level				
Fuel level = [3, 4]	→	Region = South	0.331	0.610
Region = South	←	Fuel level = [3, 4]	0.331	0.527
Road display				
Region = South	→	Road display = [2, 4]	0.553	0.638
Road display = [2, 4]	←	Region = South	0.553	0.880
Sex = male	→	Road display = [2, 4]	0.475	0.548
Road display = [2, 4]	←	Sex = male	0.475	0.875
Age = [1, 2]	→	Road display = [2, 4]	0.446	0.807
Road display = [2, 4]	←	Age = [1, 2]	0.446	0.878

the decision tree results show that age, region, and educational background have some influence. Combining the results of association rule fusion can reveal that age and gender have some rule, and the probability of choosing dashboard backlight luminance 3 and luminance 5 for participants who are male is about 41.5%, and the probability of being young is about 50.1%. For the choice of road display, the results of the decision tree show that region and gender have a certain influence, and the results of the fusion of association rules can find that region has a certain influence, and

the probability of participants from the south choosing road display luminance 3 and luminance 5 is about 61.3%.

## 4 RESULTS AND DISCUSSION

In this paper, we designed an adaptive fusion algorithm of decision trees and association rules and used the results of the decision trees to construct decisions for association rule data. After the results of the adaptive algorithm were fused with the results of the association rules themselves, we can find that the adaptive fusion algorithm,

**Table 2: Association rule model brightness results**

Background color			Support	Probability
Background color = [3, 5]	→	Region = South	0.332	0.535
Region = South	←	Background color = [3, 5]	0.332	0.618
Instrument backlight				
Instrument backlight = [3, 5]	→	Sex = male	0.225	0.415
Instrument backlight = [3, 5]	→	Age = [1, 2]	0.254	0.501
Road display				
Road display = [3, 5]	→	Region = South	0.311	0.501
Region = South	←	Road display = [3, 5]	0.311	0.613

to a certain extent, can be very good at mining the information and association relationships that exist in the data itself, and the visualization of the decision trees can help to find The results of the adaptive fusion algorithm, which are based on association rules, can, to a certain extent, explain the relationships between variables and the possible scenarios that can be inferred from the underlying indicators. During the experiments, the training time for the association rules was accelerated to a certain extent due to the embedded decision tree algorithm, which facilitated the practical application of the algorithm. However, there are still problems with the algorithm in this paper, for example, only the ID3 algorithm is used for decision tree analysis, and other algorithms will be tried in subsequent research; other algorithms will also be used for data mining experiments for association rules; for model fusion this algorithm uses the concept of late fusion, and if the model fusion can be made to generate new modules for algorithm cycle training, it can make the algorithm more optimized.

## REFERENCES

- [1] Yang L.P., Guo H.S.. Application of decision tree classification algorithm in course grade prediction [J]. Electronic Testing, 2022, 36(17):3.
- [2] Qiang Wu, Dingwei Wu, Xicheng Fu, *et al.* Decision tree based medical data analysis[J]. Computer CD-ROM Software and Applications, 2014, 17(1):2.
- [3] Du Miao, Tao Shengqing, Tang Song, *et al.* A method and system for fault diagnosis of financial self-service terminals based on decision tree learning algorithm., CN106600163A [P]. 2017.
- [4] Wu Xiu-Guo, Xing Ao-Lin. Research on education big data mining based on fuzzy association rules [J]. 2021(2020-21):67-71.
- [5] Li Hong, Cai Zhihua. Application of association rules in medical data analysis [J]. Microcomputer Development, 2003, 13(6):4.
- [6] Liu, L. F.. A financial data mining algorithm based on cleaning association rules[J]. Microelectronics and Computers, 2012, 29(5):4.
- [7] 12: Synthesizing and presenting findings using other methods | Cochrane Training
- [8] MANNLA H, SCRIBEANT R, et al.Fast discovery of association rule[C].Advances in Knowledge Discovery and data Mining.AAAI Press/The MIT Press, 1996:307-328.
- [9] R QUINLAN.Induction of decision trees[J].Machine Learning,1986,1(1):81-106.
- [10] Cai WJ, Zhang XH, Zhu JQ, *et al.* A review of association rule mining[J]. Computer Engineering, 2001, 27(5):4.
- [11] Ji Wenlu, Wang Hailong, Su Guibin, Liu Lin. A review of recommendation methods based on association rule algorithms [J]. Computer Engineering and Applications, 2020, 56(22):9.

# Speech image data mining algorithm based on multimodal decision fusion

Cong Lu,C,Lu\*  
China FAW Group Corporation CAV  
Development Research Institute  
lucong@faw.com.cn

Danxing Wang,Dx,Wang  
China FAW Group Corporation CAV  
Development Research Institute  
wangdanxing@faw.com.cn

Daquan Zhang,Dq,Zhang  
China FAW Group Corporation CAV  
Development Research Institute,  
zhangdaquan@faw.comChina FAW  
Group Corporation CAV  
Development Research Institute  
yuaqun@faw.com.cn

## ABSTRACT

This paper proposes a data mining algorithm based on multimodal decision fusion, which is mainly used to solve the correlation relationship of multi-level and multi-level multimodal data, the algorithm combines the methods of statistics, queueing study, machine learning and Bayesian decision fusion, compared with the results obtained by single modality, single data and single method, the algorithm proposed in this paper retains the information contained in the data to the maximum extent, and the algorithm is applied to the analysis of both numerical and text-based data. The proposed algorithm can be further extended by modifying the data types and methods to form new methods.

## CCS CONCEPTS

• **AiQun Yu,AQ,YU**; • **CCS CONCEPTS**; • **Mathematics of computing** → Probability and statistics; • **Keywords**: multimodality; textual data; decision fusion; speech image; data mining;

### ACM Reference Format:

Cong Lu,C,Lu\*, Danxing Wang,Dx,Wang, and Daquan Zhang,Dq,Zhang. 2023. Speech image data mining algorithm based on multimodal decision fusion. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590007>

## 1 INTRODUCTION

In today's era of explosive data growth, the era of big data is moving towards the era of 'many'. Data types have expanded from the original numerical data to text, image, audio, video, and hybrid data [1], and multimodal data refers to data collected from different perspectives or domains for the same description or the same problem [2]. target localization [3]. Due to the differences in the information contained in the different modalities, multimodal data analysis can, to a certain extent, achieve a more complete presentation of information.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590007>

Multimodal data processing methods can be broadly divided into model-independent fusion methods and model-based fusion methods, with model-independent fusion methods containing early fusion, late fusion, and hybrid fusion. Early fusion, also known as feature fusion, has relatively low information loss, but is more difficult to fuse and requires higher quality of features [4], while late fusion, also known as decision fusion, refers to the decision making of each modality and the subsequent fusion of the results. Hybrid fusion is a combination of early and late fusion, but also makes the model more complex. Common multimodal fusion methods include multicore-based learning methods [5], dynamic Bayesian networks [6], Hidden Markov Models [7], etc.

The analysis of speech image in recent years includes research based on speech-driven virtual image synthesis system [8], through subjective scoring and questionnaire to achieve speech image evaluation; elaborate the necessity of speech image research [9]; data perspective of the development of AI speech communication exploration [10], etc. In the related field there are less research from the perspective of the comprehensive effect of data, this paper addresses the problems existing in this field to be addressed.

The problems with the current algorithms are mainly focused on the lack of adaptation of the methods in conjunction with practical task-oriented approaches, the lack of intuitive fusion concepts for multimodal data, and the existence of certain technical barriers for non-specialist researchers. To address these problems, this paper proposes a data mining algorithm for multimodal decision fusion that can address multimodal data analysis combining textual and numerical data and consider the interaction of different methods for the same analysis problem. The method and framework can be further extended to other data types, and the ideas in this paper can be further extended to other research areas. The algorithm combines statistical, cohort studies, machine learning and Bayesian decision fusion methods, combining different types of data, different methods of analysis and fusion to form an algorithm that is instructive for practical implications and applied to speech image data mining.

## 2 METHODS

In response to the problems of current research, the following innovations exist in this paper.

①.The algorithm proposed in this paper can effectively deal with multi-source heterogeneous data, including textual data and numerical data, and the algorithm can modify the data type according

Input:	Number of clusters <b>K</b> and a database containing N of objects in the database.
1:	K objects are randomly selected from all data samples as the initial clustering centers
2:	The distance of the other data objects to the center of each cluster is calculated according to the principle of closest proximity to the center, and they are assigned to the respective classes
3:	For each class, the mean of all its pairs is calculated and used as the new cluster center
4:	Reallocation of data pairs based on closest proximity to the center
5:	Return to step 3 and the algorithm ends when the target functions no longer change
Output:	K clusters such that the squared error criterion is minimized

to the actual problem to be solved, or add multiple data types to increase the richness of information.

②.The algorithm in this paper can combine the advantages of different methods, with less information loss and higher accuracy of results compared to the results obtained by a single algorithm. (ii) The algorithm in this paper can combine the advantages of different methods and obtain results with less information loss and higher accuracy than a single algorithm.

③.The algorithm in this paper can automatically identify and categories data types, changing the traditional way of processing low-dimensional data to high-dimensional data, which can, to a certain extent, reduce the problem of abnormal results due to the non-existence of hierarchical relationships.

④.The algorithm can provide a research paradigm for the study of standards in areas where no such standards exist, and the framework modules can be modified or added in the process of research.

⑤.In this paper, the algorithm is applied to an actual problem that exists, and the algorithm is used to solve the current problem and retain comprehensive data information.

## 2.1 K-Means Algorithm

The K-Means algorithm<sup>[11]</sup> first needs to determine the initial clustering center, then classify all data points, calculate the average value of each cluster to adjust the clustering center, and continuously iterate, when the intra-class similarity reaches the maximum and inter-class similarity reaches the minimum, the clustering ends. the K-Means algorithm uses the error sum of squares criterion function as the most clustering criterion function, and the error sum of squares criterion function is defined<sup>[12]</sup> as:

$$J_e = \sum_{j=1}^K \sum_{i=1}^{n_j} x_i^j - m_j^2$$

The K-Means algorithm pseudo-code is as follows:

Table 1 K-Means algorithm pseudo-code

The illustration of the K-Means algorithm reveals that if the number of categories is large and the gap between categories is not obvious, the effect of clustering is often poor; the initial values of the categories of clustering will affect the effect of the iteration of the algorithm to a certain extent, and when the deviation of the initial value selection is large relative to the real situation, it will lead to the solution tending to be locally optimal and the global optimum cannot be found.

## 2.2 Data mining algorithm based on multimodal decision fusion

This paper proposes a new data mining algorithm based on multimodal decision fusion, which can solve data types of multi-level, multi-group data and contains both numerical and text-based data. The algorithm addresses the problem of evaluating the relationship between grouping variables and groupings under different groupings. The multilevel data is first stratified for numerical data and the dominance ratio<sup>[13]</sup> (OR) between variables and outcomes under different groupings is calculated. The OR value, traditionally conceptualized as the ratio of the number of exposed to the number of non-exposed in the case group to the number of exposed to the number of non-exposed in the kitchen art control group, is calculated as follows.

OR =

$$\frac{ad}{bc}$$

For this algorithm the results of the OR values indicate the relationship between the variables and the outcome of the effect, OR>1 means that an increase in the variable causes an increase in the outcome indicator i.e. a positive relationship, OR<1 means that an increase in the variable causes a decrease in the outcome indicator i.e. an inverse relationship and OR=1 means that a change in the variable does not affect a change in the outcome i.e. there is no relationship.

Indicators with positive and negative relationships were calculated from the OR values, and the positive and directional relationships were divided into two groups to categories the results. For text-based data, textual information is encoded using One-hot coding<sup>[14]</sup>, which uses N-bit status registers to encode N states, each with a separate register and only one valid at any given time. One-hot coding is a representation of categorical variables as binary vectors, which first requires mapping categorical values to integer values and then mapping each One-hot coding is the representation of categorical variables as binary vectors. In this algorithm, due to the presence of multi-level and multi-group data, One-hot coding is used to generate a high-dimensional matrix, for which K-Means algorithm based on prior information is used.

The traditional K-Means algorithm needs to set the initial clustering center based on experience. In this algorithm, the number of groups of grouped data can be set as a priori information, and the initial value of the K-Means algorithm is set as the number of groups, and the result of the K-Means algorithm is calculated as the true positive rate, i.e. the proportion of the group that originally belongs to a certain group and the K-Means result is also a certain group is calculated, and the group with the largest proportion is sorted according to the proportion, and the text data contained in

Driving experience	1-3 years 74 (44.6%)	4-6 years 8 (4.82%)	7-9 years 19 (11.4%)	10 years or more 65 (39.2%)	
Usage & Gender	Regularly used 81 (48.8%)	Not used 17 (10.2%)	Occasionally 68 (41.0%)	Male 89 (53.6%)	Female 77 (46.4%)
Age	< 22 34 (20.5%)	23-27 30 (18.1%)	28-32 30 (18.72%)	33-37 12 (7.23%)	>37 59 (35.5%)
Occupation	Service worker 25 (10.5%)	Career 29 (17.5%)	Students 48 (28.9%)	Government agencies 8 (4.82%)	Technical staff 56 (33.7%)
Educational level	Undergraduate 73 (44.0%)	Doctorate 23 (13.9%)	Master's degree 37 (22.3%)	Below secondary school 13 (7.83%)	Specialist 20 (12.0%)

the group with the largest proportion is taken as the text information representing the association strength of that group. Finally, the idea of decision fusion<sup>[15]</sup> is used to extract the key data information and text information that can be used to describe the different groups, and the flow chart of the algorithm is as follows.

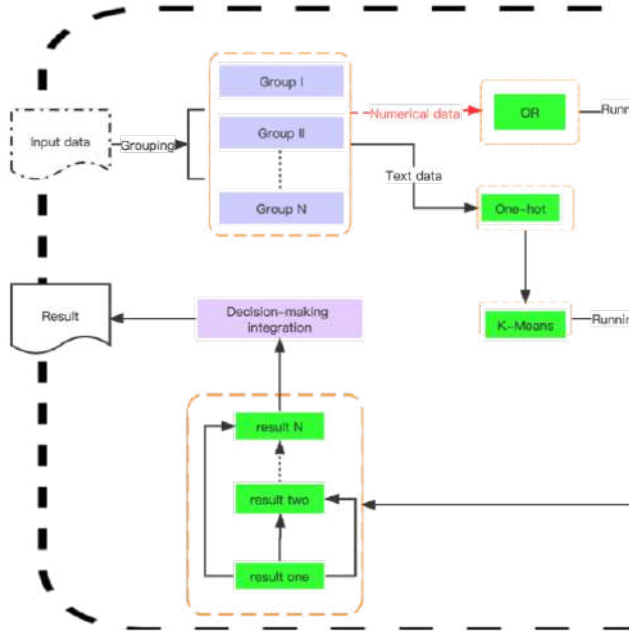


Figure 1 Basic framework of the algorithm

### 3 EXPERIMENTAL ANALYSES

In this paper, we collected 166 participants' evaluations of the voice character image and voice audio of three automobile brands. The participants evaluated the relevance of the automobile brands, image attractiveness, image correctness, image creativity, image vividness, image identity, image quality, the scoring of voice audio, the first feeling of voice, audio age, audio occupation, audio character temperament, audio conforming to the temperament, audio conforming to the tone. The participants' basic information was also collected (Table 2). Through Table 2, it can be found that the participants were concentrated in two categories of people with low driving experience and high driving experience, which can be a

good analysis of the evaluation of the in-car intelligent voice image by people with different driving experience, and the gender, age, occupation, and education. The balanced distribution of gender, age, occupation, and education level is conducive to better analysis.

Table 2 Baseline distribution of data

The data in this paper contains a total of two levels, the first level is the intelligent voice image, and the second level is the intelligent voice audio. Firstly, we analyze the intelligent voice image, divide the data into three groups according to brands, and calculate the OR and p-values between the evaluation indexes of voice image: image attractiveness, image modernity, image creativity, image vividness, image identity, image quality and the correlation degree (Table 3). This means that there is a positive relationship between image attractiveness, image era sense and brand one, and to a certain extent, the intelligent voice image of brand one can make drivers develop an association with the brand.

In the second brand image attractiveness, image era sense, image creativity and image identity are associated with the brand, and the OR values are all greater than 1. In other words, for the association analysis of the second brand, image attractiveness, image era sense, image creativity and image identity can be used, and they show a positive relationship, and when evaluating the second brand, image attractiveness, image era sense, image creativity and four indicators of image identity. In the third brand, image attractiveness, image creativity and the brand have a certain correlation, and the OR values are all greater than 1, that is, when evaluating the second brand, image attractiveness and image creativity can be used.

Table 3 Analysis of smart voice image evaluation indicators and brand correlation

The second level of evaluating intelligent voice image in this paper is the analysis of intelligent voice audio, this paper uses the first feeling of voice, audio age, audio occupation, audio character temperament, audio conforming to temperament, audio conforming to tone, the degree of tone ups and downs, the degree of speech speed, the degree of speech strength and weakness to analyze voice audio, where the first feeling of voice, audio age, audio occupation, the degree of tone ups and downs, the degree of speech speed. The OR values were calculated for the numerical data (Table 4), and for Brand 1, the first perception of voice, speed of speech and strength of speech triggered the association with Brand 1, and the OR values were all greater than 1, i.e. the higher the first perception of voice, the stronger the association with the brand. The analysis of the voice audio for brand two found that there was an association between the age of the voice, the occupation of the voice, the degree

	Group I		Group II		Group III	
	OR (95%CI)	P value	OR (95%CI)	P value	OR (95%CI)	P value
Image appeal	1.856 (1.509 -2.283)	0.000	1.267 (1.069 -1.502)	0.007	2.184 (1.753 -2.721)	0.000
Image contemporary	1.163 (0.974 -1.390)	0.097	1.185 (1.032 -1.359)	0.017	0.927 (0.756 -1.138)	0.470
Image Creativity	1.135 (0.904 -1.423)	0.277	1.271 (1.068 -1.512)	0.008	1.248 (1.025 -1.519)	0.029
Vivid image	0.997 (0.782 -1.273)	0.983	1.068 (0.882 -1.294)	0.502	0.882 (0.693 -1.123)	0.310
Image identity	1.061 (0.840 -1.339)	0.620	1.524 (1.231 -1.887)	0.000	0.959 (0.715 -1.287)	0.780
Sense of image quality	0.909 (0.726 -1.139)	0.409	0.855 (0.694 -1.053)	0.143	1.039 (0.820 -1.317)	0.750

	Group I		Group II		Group III	
	OR (95%CI)	P value	OR (95%CI)	P value	OR (95%CI)	P value
Sound First Feel	1.118 (1.042 -1.200)	0.002	1.049 (0.980 -1.122)	0.170	1.265 (0.776 -2.078 )	0.348
Audio age under 18	0.640 (0.144 -2.839 )	0.558	0.834 (0.459 -1.514 )	0.551	1.593 (1.001 -2.572 )	0.152
Audio age 23-27	0.791 (0.473 -1.322 )	0.372	0.590 (0.378 -0.920 )	0.021	1.659 (1.050 -2.665 )	0.033
Audio age 28-32	0.932 (0.508 -1.707 )	0.819	0.969 (0.602 -1.559 )	0.896	0.829 (0.474 -1.390 )	0.968
Audio age 33-37	1.052 (0.514 -2.153 )	0.891	0.736 (0.332 -1.631 )	0.452	0.453 (0.172 -0.995 )	0.493
Audio age 38-42	0.801 (0.322 -1.993 )	0.634	1.126 (0.486 -2.611 )	0.782	0.897 (0.337 -1.994 )	0.172
Audio age 42+	0.585 (0.266 -1.290 )	0.186	1.133 (0.489 -2.622 )	0.772	0.418 (0.123 -1.062 )	0.807
Audio Career Account Manager	1.428 (0.390 -5.225 )	0.591	1.525 (0.576 -4.039 )	0.397	0.258 (0.041 -0.860 )	0.103
Audio Career Flight Attendant	2.381 (0.620 -9.145 )	0.208	2.926 (1.117 -7.669 )	0.030	0.942 (0.613 -1.447 )	0.065
Audio Secretary	1.928 (0.543 -6.849 )	0.312	1.420 (0.580 -3.478 )	0.444	1.003 (1.000 -1.005 )	0.787
Audio Other	0.893 (0.251 -3.182 )	0.862	1.182 (0.461 -3.033 )	0.728	2.885 (1.344 -5.597 )	0.124
Audio Service	0.998 (0.278 -3.586 )	0.998	1.274 (0.485 -3.345 )	0.623	1.331(1.000 -1.564)	0.003
Audio Host / Hostess	1.655 (0.434 -6.301 )	0.462	1.522 (0.580 -3.993 )	0.395	1.005 (1.000 -1.132 )	0.207
How fast or slow you speak	1.611 (1.305 -1.987 )	0.000	1.530 (1.228 -1.907 )	0.000	1.744 (1.354 -1.900)	0.638
Strength of speech	1.184 (0.973 -1.440 )	0.093	1.395 (1.128 -1.726 )	0.003	1.000 (1.000 -1.030 )	0.425

of speed of speech, the degree of strength of speech and brand two, and for brand two, participants felt that the voice audio was more in line with 23–27 year-olds, which to some extent indicates that brand two could represent the choice of young people. The analysis of the voice audio for brand three found a correlation between audio age, audio occupation and brand three. There was some similarity to brand two in terms of audio age, but there were some differences for audio occupation, with brand three believing that the voice audio was more in line with service workers.

Table 4 Analysis of smart voice audio evaluation indicators and brand correlation

The K-Means algorithm was used to analyze the data for the text data, and the results obtained through the K-Means algorithm (Table 5) showed that for the evaluation of the words of audio personality temperament, the third group had a higher true positive rate, and by counting the words of specific audio personality temperament (Table 6) it could be found that there were more choices of stable, calm and cheerful, i.e. when the participants needed a stable, calm and cheerful voice audio assistant, the third group is a good choice. For the evaluation of the vocabulary of audio conforming temperament, again the third group had the highest true positive rate, and by counting the vocabulary of specific audio personality temperaments it can be found to be consistent with the vocabulary of audio personality temperaments, but the second group differed greatly, and the participants who chose the second group for the vocabulary of audio conforming temperament indicated being lively and cheerful, and by combining the results of the

numerical type data analysis it can be found that there is There is a certain similarity between the second and third groups. The first group had the highest rate of true positives for the evaluation of words that fit the tone of the audio, and combined with the results of the statistical analysis, it was found that the participants tended to evaluate the words in the first group as cordial and easy-going.

Table 5 Smart voice audio vocabulary evaluation and brand correlation analysis

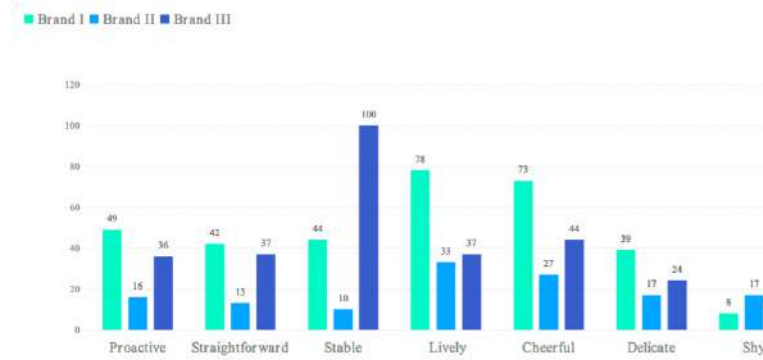


Figure 2 Audio personality temperament vocabulary selection

	Group I	Group II	Group III
Audio Personality Temperament			
Group I	54	56	35
Group II	70	81	34
Group III	42	29	97
Audio to match temperament			
Group I	81	105	42
Group II	43	28	24
Group III	42	33	100
Audio matches tone			
Group I	108	86	86
Group II	33	35	66
Group III	25	45	14

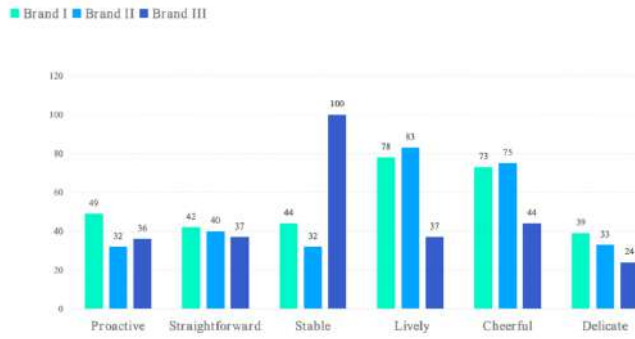


Figure 3 Audio matching temperament vocabulary selection

## 4 RESULTS

Through the analysis of the algorithm framework and application results, it can be found that the algorithm can solve the correlation analysis between multi-level and multi-level multimodal data to a certain extent. Relative to the correlation rules, the algorithm gives different processing methods from numerical data and textual data, combining the relationship under the same level, increasing the difference between levels, and at different levels, the analysis can be carried out to obtain results that are easy to visualize and understand the results can be visualized and understood. The results are more favorable for analysis than for traditional classification problems. The results are fused at the decision level based on a priori information (Table 6). Due to the instability and bias of the data in the actual problem, the use of concurrent decision fusion can retain more information to a certain extent relative to the data.

For unimodal data the results obtained in the above graphs can be obtained. The results obtained under unimodal data can only illustrate the direction of association between indicators and outcomes, while the results obtained under multimodal data can illustrate the direction of association between indicators and outcomes and the vocabulary that specifically describes the direction of association, with more informative results. The results obtained with the unimodal model do not consider the interaction between the different

outcomes and the a priori information. Decision fusion can compensate for the loss of information due to this drawback and increase the reliability, richness, and interpretability of the decisions.

### Table 6 Decision-level fusion results

For the real data collected in this paper, it can be found that the algorithm can effectively explore the relationship existing between the data and solve the technical barriers existing in the analysis process for numerical data and text-based data, and it can also be found through the results that the evaluation between different brands has certain differences. The results also show that there is some variation in the evaluation of different brands. The methodology of this paper makes better use of the data and retains more information than if only a single method had been used in the relevant field. In the area of research covered in this paper, evaluation criteria are time-sensitive and developments in the times lead to changes in evaluation criteria, so it is not possible to assess the relationship between variables and outcomes using a single or fixed criterion, and the data mining algorithm based on multimodal decision fusion proposed in this paper provides a framework for solving this type of problem to some extent.

In the future, we will conduct research such as replacing different classification algorithms as a method of processing textual data; the results of OR values can be replaced by forest graph visualization; replacing other application scenarios to determine the effectiveness of the algorithm; expanding the computational sample size given the evaluation criteria in the relevant field, etc.

## REFERENCES

- [1] Jia L, Tlab C, Peng X, *et al.* Urban big data fusion based on deep learning: An overview[J]. *Information Fusion*, 2020, 53:123-133.
- [2] ZHAO L. Research on multimodal data fusion methods[D]. Dalian: Dalian University of Technology, 2018.
- [3] Zhang XY, Li J, Wang L, *et al.* An attention mechanism-based multimodal data fusion method for vision and LIDAR
- [4] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. *Science*, 2006, 313(5786):504-507.
- [5] YEH Y R, LIN T C, CHUNG Y Y, *et al.* A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection[J]. *IEEE Transactions on multimedia*, 2012, 14(3):563-574.
- [6] SUTTON C, MCCALLUM A. An introduction to conditional random fields for relational learning[J]. *Introduction to Statistical Relational Learning*, 2006, 2:93-128.
- [7] FRIEDMAN N, MURPHY K, RUSSELL S. Learning the structure of dynamic probabilistic networks[J]. *arXiv 1301.7374*, 2013.
- [8] He Xiaoguang. Speech-driven virtual image synthesis system [J]. *Journal of Anhui Institute of Electronics and Information Technology*, 2021, 20(1):4.

	Evaluation methods
Group I	Image appeal; Image contemporary; Sound First Feel; Audio age; How fast or slow you speak; Strength of speech; Approachable; easy-going
Group II	Image appeal; Image contemporary; Image Creativity; Image identity; Audio age; Audio Career; How fast or slow you speak; Strength of speech; Lively; cheerful
Group III	Image appeal; Image Creativity; Audio age; Audio Career; Steady; calm; cheerful

- [9] Liu Bin, Ouyang Ye. Method and device for voice interaction based on virtual robot image, intelligent control system for in-vehicle devices; CN111124123A [P]. 2020.
- [10] Chunhui. AI voice perspective on foreign communication[J]. Journalism Research Guide, 2019(14):2.
- [11] Zhu Ming. Data Mining [M]. Hefei: China University of Science and Technology Publishing House. 2002.
- [12] Douglass F, Ousterhout J K. Transparent process migration: design alternatives and the Sprite implementation. ACM Press/Addison-Wesley Publishing Co. 1999.
- [13] Zhang XF, Pei GJ, Xu ZY, *et al.* Meta-analysis of risk factors for the development of esophageal cancer[J]. Modern preventive medicine, 2009, 36(5):4.
- [14] Buckman J, Roy A, Raffel C, *et al.* Thermometer Encoding: One Hot Way To Resist Adversarial Examples[C]// 2018.
- [15] Ruohong H, Ping Z, Yun P, *et al.* SAR target recognition using PCA, ICA and Gabor wavelet decision fusion PCA \ ICA and Gabor Wavelet decision fusion for SAR target identification[J]. Journal of Remote Sensing, 2012, 16(2):262-274.

# Optimized model analysis of blockchain PoW protocol under long delay attack

Tao Feng\*

School of Computer and Communication, Lanzhou  
University of Technology, Lanzhou, 730050, China  
fengt@lut.edu.cn

Yufeng Liu

School of Computer and Communication, Lanzhou  
University of  
Technology, Lanzhou, 730050, China, 1017373220@qq.com

## ABSTRACT

**Abstract:** Proof of work (POW) is one of the most widely used consensus method of bitcoin. In some chains, because of the large number of users, the huge amount of information interaction data, equipment hardware failure or malicious attacks on some nodes may cause communication delay. All of these may engender forks on the blockchain resulting data loss. Therefore, whether the blockchain protocol can achieve sufficient security in the asynchronous network environment with long delay, so how to reduce fork and user's data loss caused by long delay attack are important issue related to blockchain protocol. In this study, we optimized the existing model and proposed  $T_{OD}$  to describe the evolution state of the main chain accurately.

## CCS CONCEPTS

• **Security and privacy** → Cryptography.

## KEYWORDS

Blockchain; delay, bitcoin, security

### ACM Reference Format:

Tao Feng and Yufeng Liu. 2023. Optimized model analysis of blockchain PoW protocol under long delay attack. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590008>

## 1 INTRODUCTION

Proof of work (PoW) has been applied to produce new block since 2008 the bitcoin introduced [1], however, with the development of bitcoin, PoW has attracted large number of the computing power around the world. The emergence of large-scale mining rigs contradicts the idea of "decentralization" at the beginning of design, and the research on its improvement and optimization has never stopped because of long consensus period and large resource consumption. At the same time, the security of blockchains has also attracted considerable attention. "Selfish Mining" was originally

proposed by Sirer and Eyal [2]. This strategy is an attack against incentive mechanisms and bitcoin mining, which intentionally delays the publication of new calculated blocks causing "forks" on the blockchain, so that honest miners make invalid calculations and they obtain additional rewards, where the adversary just need to control about 33% of the total computing power to make an attack by delaying the public of blocks. Sompolinsky et al. [3] also proved that the adversary will be easier to attack in the case of high throughput and proposed the Greedy Heaviest-Observed Sub-Tree (GHOST) algorithm to improve the security of the blockchain under high throughput, this algorithm doesn't follow the longest chain principle, but uses the heaviest-observed Sub-Tree principle to alleviate the problems caused by forks. In 2015, Tromp proposed a graph-based PoW algorithm [4]. This algorithm finds a minimum cycle in the established graph by establishing a large random graph data structure to complete the proof of work. Since PoW reaches 50 % fault tolerance and become the highest among all consensus mechanisms, Gervais and Karame et al. [5] stated that if a node or mining pool has more than half of the computing power, he may launch a 51% attack. Bastiaan further analyzed the PoW consensus mechanism in Bitcoin in 2015 [6]. In 2016, Nayak et al. proposed "stubborn" in the literature [7]. This mining strategy extends the original "selfish mining" strategy. Based on this strategy, the income of malicious mining pools will increase by 13.94%, in the text, the author further optimized the "stubborn" strategy and proposed two new strategies, "the EqualFork Stubborn" and "Trail Stubborn". In the abstract model established by Garay, Kiayias, etc., they proved that if the adversary controls a certain proportion of computing power, interfering with the communication between miners can launch a delay attack, and they pointed out that the adversary can still cause delay attack when only the communication is controlled, at the same time they proposed three security properties such like chain growth, common prefix and chain quality [8] [9], they also analyzes the security in the application process. Pass et al. [10] proved that if the adversary performs a delay attack on the chain to cause a fork, the number of message delay rounds is not more than  $\Delta$  rounds, where  $\Delta \ll 1/n_p$  ( $n$  represents the number of miners, represents the mining probability), considering the actual situation due to the influence of the network, hardware equipment, etc., this situation is a small probability event, on this basis, Wei and Yuan simulated the evolution process of the main chain under PoW through a tree structure in [11], they proved that blockchain is secure under long delay attack of  $\Delta > 1/n_p$  in asynchronous network, and the proposed  $Tree_{MC}$  model idealized the evolution state of the chain in PoW. Due to delay, mining difficulty and other factors in the real world, we made some improvements and adjustments to this model,

\*Place the footnote text for the author (if applicable) here.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590008>

so as to better simulate the real situation and verify the feasibility of the new model.

## 2 PRELIMINARIES

### 2.1 Tree<sub>MC</sub> Model

The Tree<sub>MC</sub> model proposed in [11] represents the changing state of miner chain in each round of protocol through two steps of AddBlock and DeleteBlock, so as to simulate the evolution of PoW main chain.

When adding nodes, after receive the chain C with n nodes, the chain is added to the Tree<sub>MC</sub> branch on the basis of ensuring the common prefix to form a new sub-branch. If multiple broadcast chains are received, after a leaf node may be multiple sub-branches on the Tree<sub>MC</sub>. For each new block node on the tree, the tree depth increases by 1. After AddBlock operation, if the current tree depth is d, other invalid branches on Tree<sub>MC</sub> such as length less than d will be deleted immediately.

It can be seen that the delayed standby chain will be deleted immediately, and its current depth is less than d, but it may still become the main chain in the subsequent consensus round, thus causing recording errors.

### 2.2 Security Properties of the blockchain model

**2.2.1 Chain growth.** In [11], the author use Nakamoto model ( $\Pi^V, C$ ) and related environment oracle  $Z(\cdot)$ , security parameters k defines the view( $\Pi, C, A, Z, K$ ) to represent the status of each miner, and  $|\text{view}(\Pi, C, A, Z, K)|$  represents the number of rounds during the execution of ( $\Pi^V, C$ ), then that is:

Given  $\text{view}(\Pi, C, A, Z, K)$ , the blockchain grows by at least n blocks with majority  $\lambda$  ( $\lambda > 1/2$ ) from round r1 to r2, if  $\Pr_{i,j}[|C_j^{r2}| - |C_i^{r1}| \geq t] \geq \lambda$ , n represents the maximum value that satisfies the condition, so

$$\text{chain-increase}_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) = \max\{t | \Pr_{i,j}[|C_j^{r2}| - |C_i^{r1}| \geq t] \geq \lambda\} \quad (1)$$

**2.2.2 Common prefix.** If  $\lambda$  ( $\lambda > 1/2$ ), we can define the oracle that common - prefix $_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) = 1$ , so can get the following inequality

$$\Pr_{i,j}[(C_i^{rk} \leq C_j^r) \wedge (C_j^{rk} \leq C_i^r)] \geq \lambda \quad (2)$$

**2.2.3 Chain quality.** If there are n blocks out of t consecutive blocks are obtained by malicious miners in chain C, and we define chain - quality $_t(C, \rho) = 1$ , for given view( $\Pi, C, A, Z, K$ ), if

$$\Pr[\text{chain - quality}_t(C_i^r, \rho) = 1] \geq \lambda \quad (3)$$

## 3 OUR BLOCKCHAIN MODEL

### 3.1 Record the state of main chain

In the PoW consensus algorithm, the main chain is the blockchain that has accumulated the most difficulty, so it is also the chain that contains the most blocks. During the mining process, each node always tries to extend the main chain and solve the puzzles through cryptographic algorithms. The successor blocks of the current block are found, and after enough rounds of consensus, they will be unified into one or more main chains. If two miners

solve the puzzle in almost the same time during the mining process and broadcast the message, because the time when their nodes receive the information are different, the two candidate blocks will be selected as an extension of the main chain. The more workload is used as the main chain, and the other is saved as a backup chain. The backup chain is saved because it may exceed the difficulty of the main chain and become the new main chain.

Since the DeleteBlock operation is performed after each round of consensus process to delete the shorter chain in each round, the backup chain is deleted at the beginning of its formation, so the evolution process of the main chain and the backup chain cannot be reflected in the model Tree<sub>MC</sub>. Therefore, we propose an improved tree structure T<sub>OD</sub> to record the evolution of the main chain. Let B<sub>0</sub> be the genesis block and record it as the root node. Each node on the tree represents a block. As shown below, T<sub>OD</sub> mainly has three operations to execute the main chain. chain evolution.

1. Addblock: When the node broadcast the chain  $C' = (B_0, B_2, \dots, B_n)$ , Assume that there is a branch C<sub>0</sub> on TOD and C<sub>0</sub> is completely consistent with the first k nodes in  $C'$  ( $0 < k < n$ ), satisfy  $C_0 = (B_0, B_2, \dots, B_k)$ , and k reaches its maximum value, then add blocks in  $C'$  from k+1 to n to branch C<sub>0</sub>, and after a round of consensus algorithm is over, new blocks may be added on different branches due to the non-uniqueness of the node broadcast chain.

2. Boolturn: The return value determines whether the DeleteChain operation needs to be performed in the following operations

3. DeteleteChain: Branches that do not satisfy the current depth are deleted while satisfying the corresponding value of Boolturn.

Let B<sub>0</sub> be the initial root node on T<sub>OD</sub>. After AddBlock in round r ( $r > 0$ ), determine whether a new fork is generated on the current model and whether the fork depth on the original fork node increases, if a new branch is created or the depth of difference between the longest branch and the shortest branch on the current branch is not greater than 1, set Boolturn equal to 1, and set Boolturn to 0 in the next round, if there is no new forks generated or the difference between the longest branch and the shortest branch is greater than 1, set Boolturn to 0. After the Boolturn setting is completed in each round, if Boolturn equals 0, the DeleteChain operation is executed immediately; otherwise, the DeleteChain operation is not executed in the current round.

It can be seen that in the process of each node receiving the broadcast, each honest miner will take the longest chain for recording, which improves the problem that the shorter chain is deleted immediately in each round in the Tree<sub>MC</sub> model, which is quite different from the evolution of the chain in the consensus protocol. And if an honest miner chain longer than the main chain is maliciously delayed by the adversary, it can not be recorded in the previous model. Here, when the number of delay rounds is 1, it can also be recorded on B and will not be deleted, simulating the state of main chain more effectively.

### 3.2 Properties of our model

T<sub>OD</sub> can expand or modify the node branch according to the broadcast chain, so as to achieve the effect of better simulating the evolution of the main chain. In an ideal environment without considering

the delay of hardware devices such as the network,  $T_{OD}$  has the following properties:

**3.2.1 Properties of  $T_{OD}$ .** 1. If a new block is added to TOD after the round ends, but the tree depth  $d$  remains unchanged, the newly added blockchain will delay the broadcast for malicious miners.

2. If a new block is added to the branch after the DeleteChain operation is performed, and the depth of each chain in the tree is the same, the depth of tree  $A$  is increased by one.

3. If the tree depth  $d$  changes from  $r$  round to  $r+1$  round to  $d+t$ , where  $t>1$ , then all nodes after depth  $d+1$  in this round are produced by malicious miners.

4. If there are no consecutive forks, at the end of every two rounds, all forks have the same depth.

**3.2.2 Proof.** 1. If a new block is added to the tree after the round  $r$  ( $r>2$ ), and the tree depth  $d$  is consistent with the round  $r-1$ , it is because a fork occurs in the round  $r-2$ , BoolTurn is set to 1, and the DeleteChain operation was not performed for 1 round, resulting in inconsistent lengths of the two branches on the fork, and the miner node continued to work on the backup chain with the depth of  $d-1$ . After successful mining, the length  $d$  was delayed by the adversary, so that the tree depth is unchanged in the round  $r$ .

2. After a round of consensus protocol ends, a new block is added to  $T_{OD}$ , and the DeleteChain operation is performed, indicating that the backup chain and useless blocks have been deleted, so the added block is not on the backup chain, and the tree depth is increased by one.

3. The tree with a depth of  $d$  becomes the depth of  $d+t$  after one round, that is the chain received the new chain with a depth of  $d+t$ , and no miners have mined on the chain with a depth of  $d+2$  in the current round, so the blocks after the depth  $d+2$  are all mined by adversary miners.

4. If there is a fork in the current round, BoolTurn is set to 1, and the shorter backup chain will not be deleted in the next round, but in the subsequent round, after the DeleteChain operation is performed, all chains and some nodes whose depth is not  $d$  will be deleted, so if there is no continuous forks, all forks length on  $T_{OD}$  in this round are equal.

## 4 FEASIBILITY ANALYSIS

In [11], the author gives the relationship between the three properties of the PoW protocol under the Tree<sub>MC</sub> model. In our model, although there are still a little chains that will not be recorded in the delay environment, which may be different from the evolution of the main chain of the blockchain. But it is exactly the same, and this model has a wider range of recording chains, more flexible simulation, and still closely link to the properties of the PoW consensus. We will give relevant proofs below.

### 4.1 Chain growth

**Lemma 1.** In [11] call the event of  $m_{\text{delay}}^r > \frac{(1-\lambda)n}{4}$  as *Over-delay*, where  $m_{\text{delay}}^r$  represents the number of honest miners delayed in the  $r$ th round, here we make the following modifications

$$\Pr[m_{\text{delay}}^r > \frac{(1-\lambda)n}{4}] < e^{-u(k)} \quad (4)$$

where  $n=n(k)$  is a polynomial function of  $k$ .

**Proof:** In the  $r$  round of mining, at most  $\Delta$  rounds can be delayed, so let  $B_{\text{delay}}^r$  denote the number of delayable blocks obtained in the current  $r$  round. For  $n$  miners and in the mining process, the probability of reaching the delayable block is  $\alpha p$ , so we can get  $B_{\text{delay}}^r \sim B(n\Delta, \alpha p)$ , which can be obtained from the Hofding bound

$$\Pr[B_{\text{delay}}^r \geq (\alpha p + \varepsilon)n\Delta] < e^{-2\varepsilon^2 n}$$

Let  $(\alpha p + \varepsilon)n\Delta = \frac{(1-\lambda)n}{4}$ ,  $\frac{1}{2} < \lambda < 1 - 8\alpha p\Delta$ , we can get  $\varepsilon = \frac{1-\lambda}{4\Delta} - \alpha p$ , that is

$$\Pr[B_{\text{delay}}^r \geq \frac{(1-\lambda)n}{4}] \leq e^{-2\left(\frac{1-\lambda}{4\Delta} - \alpha p\right)^2 n} < e^{-\frac{(1-\lambda-4\Delta)^2}{8\Delta^2} n}$$

Let  $u(K) = \frac{(1-\lambda-4\Delta)^2}{8\Delta} n(K)$ , obviously  $m_{\text{delay}}^r < B_{\text{delay}}^r$ , so we can get

$$\Pr[m_{\text{delay}}^r > \frac{(1-\lambda)n}{4}] < \Pr[B_{\text{delay}}^r > \frac{(1-\lambda)n}{4}] < e^{-u(k)}$$

**Lemma 2.** Assume that  $d_{\text{tree}}^r$  is the depth of round  $r$ , and  $\frac{1}{2} < \lambda < 1 - 8\alpha p\Delta$ , so we can get

$$\Pr[\text{chain-increase}_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) \geq d_{\text{tree}}^{r_2} - d_{\text{tree}}^{r_1}] \geq 1 - 2e^{-u(k)} \quad (5)$$

**Proof:** We consider the case of  $|C_i^r| < d_{\text{tree}}^r$  for the chain  $C_i^r$  with the length  $|C_i^r|$  of miner  $i$  in the round  $r$ , since miners will make adjustments to their own chain after each round of receiving broadcasts, and mine on this basis, the situation of  $|C_i^r| < d_{\text{tree}}^r$  after modifying their own chain is strictly not true. Consider that if node  $i$  is delayed causing a fork, the  $|C_i^r|$  record in the current backup chain has  $|C_i^r| < d_{\text{tree}}^r$ , but the current node is still mining at the length of  $|C_i^r|$ , in the next round, if the mining is successful,  $|C_i^{r+1}| = d_{\text{tree}}^{r+1}$ , otherwise the DeleteChain operation will be performed to delete it, so the case of  $|C_i^r| < d_{\text{tree}}^r$  is not considered. If the deletion operation is performed in each round, that is, the  $d_{\text{tree}}^r$  branch nodes and illegal nodes on the tree are deleted immediately after each round, we can get

$$\Pr[C_i^r \neq d_{\text{tree}}^r] = \frac{m_{\text{delay}}^r}{n} \leq \frac{1-\lambda}{n}$$

In  $T_{OD}$ , the BoolTurn value needs to be judged after each round of adding nodes and then the deletion operation is performed. Therefore, without considering the continuous bifurcation, we can obtain

$$\Pr[C_i^r \neq d_{\text{tree}}^r] = \frac{m_{\text{delay}}^{r-1}}{n} \leq \frac{m_{\text{delay}}^r}{n} \leq \frac{1-\lambda}{4}$$

then

$$\begin{aligned} & \Pr_{i,j} \left[ |C_j^{r2}| - |C_i^{r1}| \geq d_{\text{tree}}^{r2} - d_{\text{tree}}^{r1} \right] \\ & \geq 1 - \Pr_j \left[ |C_j^{r2}| \neq d_{\text{tree}}^{r2} \right] - \Pr_i \left[ |C_i^{r1}| \neq d_{\text{tree}}^{r1} \right] \\ & \geq 1 - \frac{2(1-\lambda)}{4} \\ & \geq \lambda \end{aligned}$$

## 4.2 Common prefix

**Lemma 3.** In Assume that  $d_{\text{tree}}^r$  is the depth of round  $r$ , and  $\frac{1}{2} < \lambda < 1 - 8\alpha p\Delta$ , exist  $N > 0$ , if there is a common prefix of length  $d_{\text{tree}}^r - N$  on all branches of tree  $T_{\text{OD}}$  then

$$\Pr[\text{common - prefix}_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) = 1] \geq 1 - 2e^{-u(k)} - \varepsilon_1 \quad (6)$$

**Proof:** Assuming that there are branches  $C_i^r$  and  $C_j^r$  on a node, if the current depth is  $d_{\text{tree}}^r$ , then there is  $K > 0$  so that at the depth  $d_{\text{tree}}^r - K$ , there is a chain  $C^*$  that is the common prefix of branches  $C_i^r$  and  $C_j^r$  (denoted as  $C^* < C_i^r, C^* < C_j^r$ ). Since the chain fork is caused in the delay environment, it will also cause forks on  $T_{\text{OD}}$ , which is inconsistent with the actual consensus process. BoolTurn is introduced, so that when  $\Delta=1$ , it is completely consistent with  $C_i^r$ , so the continuous split when the number of forks or delay rounds is large, the main chain will not be completely consistent with  $T_{\text{OD}}$ . Let  $S_{\text{delay}}^r$  represent the set of all delayed miners in round  $r$ , then it can be seen from 4.4.1, when  $C_i^r \in S_{\text{delay}}^r$  and defined Over-delay does not occur, we can get

$$\Pr_i \left[ C_i^r \in S_{\text{delay}}^r \overline{\text{Over-delay}_r} \right] \leq \frac{|S_{\text{delay}}^r|}{n} = \frac{1-\lambda}{4}$$

If we consider the case of continuous forks, forks will be caused because of delay, and the probability of mining a delayed block and forking in one round is  $\alpha p$ , and the probability of mining a delayable block continuously in  $n$  rounds and delaying is  $(\alpha p)^n$ . Obviously, its probability decreases exponentially as  $n$  increases, and the direct impact of this situation is that the actual chain is inconsistent with the record branch, before the delay attack occurs, if the chain  $C'$  is not recorded in the round  $r'$  ( $r' < r$ ) because of delay, and the depth  $|C'| > d'$ , the adversary broadcasts it in the round  $r''$  ( $r' < r'' < r$ ) and the chain is updated, after that because of the delay, the  $C^*$  chain has the same length as the  $C'$  chain, so that the current chain cannot be recorded in the tree and the current tree depth is  $d_r = |C'|$ , thus causing a delay in the round  $r'$ , and the probability of continuous delay fork is not over  $\varepsilon_1$ , where  $\varepsilon_1$  is a negligible function of  $\lambda$ , and the probability of continuous bifurcation is extremely low in the actual situation, so Over-delay $_{r'}$  does not occur in the round  $r'$ , we can get

$$\Pr_i \left[ C_i^r \in S_{\text{delay}}^r \overline{\text{Over-delay}_{r'}} \right] \leq \frac{|S_{\text{delay}}^r|}{n} = \frac{1-\lambda}{4}$$

obviously we can get

$$\begin{aligned} & \Pr_{i,j} \left[ (C^* < C_i^r) \wedge (C^* < C_j^r) \right] \\ & \geq \Pr_{i,j} \left[ C_i^r \in \text{Tree}_{\text{OD}}^r \wedge C_j^r \in \text{Tree}_{\text{OD}}^r \right] \\ & \geq 1 - \Pr_i \left[ C_i^r \notin \text{Tree}_{\text{OD}}^r \right] - \Pr_i \left[ C_j^r \notin \text{Tree}_{\text{OD}}^r \right] - \varepsilon_1 \\ & \geq 1 - \frac{1-\lambda}{2} - \frac{1-\lambda}{2} - \varepsilon_1 \\ & > \lambda \end{aligned}$$

so it equivalent to common-prefix $_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) = 1$ , thus

$$\begin{aligned} & \Pr \left[ \text{common - prefix}_{A,Z,K}^{(\Pi,C)}(r_1, r_2, \lambda) = 1 \right] \\ & \geq 1 - \Pr [\text{Over-delay}_r] - \varepsilon_1 \\ & \geq 1 - 2e^{-u(k)} - \varepsilon_1 \end{aligned}$$

## 4.3 Chain quality

**Lemma 4.** Assume that  $\frac{1}{2} < \lambda < 1 - 8\alpha p\Delta$ , on consecutive  $t$  nodes in the round  $r(t > 0)$ , at most  $N$  nodes are mined by malicious miners, then

$$\Pr \left[ \text{chain - quality}_{A,Z,K}^{(\Pi,C)}(r, \rho, k, \lambda) = 1 \right] \geq 1 - 2e^{-u(k)} - \varepsilon_1 \quad (7)$$

**Proof:** if chain-quality $_{A,Z,K}^{(\Pi,C)}(r, \rho, k, \lambda) = 1$ , then the chain  $C$  satisfies the chain quality security attribute, that is, the number of malicious nodes on the consecutive  $t$  nodes in the round  $r$  is less than  $\rho t$ , and most of its nodes are mined by honest miners, so we can obtain  $\Pr[C^r \notin T_{\text{OD}}^r] < \frac{1-\lambda}{2}$ . And without considering the Over-delay event, to make sure that chain-quality $_{A,Z,K}^{(\Pi,C)}(r, \rho, k, \lambda) = 1$  cannot cause continuous forks, so

$$\Pr \left[ \text{chain - quality}_{A,Z,K}^{(\Pi,C)}(r, \rho, k, \lambda) = 1 \right] \geq 1 - 2e^{-u(k)} - \varepsilon_1$$

## 5 EXPERIMENT ANALYSIS

In this section, We use the model to simulate and analyze the potential threat of long-latency attacks to the growth rate of the blockchain chain, and compare the results of different nodes in the same round by comparing with the original Tree $_{\text{MC}}$  model.

### 5.1 Long delay attack on chain growth

We use the same attack model as [11] to test the attack results of the improved model, we use the model to simulate and analyze the potential threat of long-delay attacks to the chain growth. Assuming  $\mu = 0$ , that is, the adversary does not control any miners' computing power, the miners are divided into  $H_A$  and  $H_B$  to represent the two branches on  $T_{\text{OD}}$ , and the number of miners in these two sets are dynamically equal  $H_A = H_B = \frac{n}{2}$ , if  $n$  miners mined successfully, the probability of successful mining is  $\eta(n, p) = 1 - (1 - p)^n$ , where we can consider  $\eta(n, p) \approx np$ . Here we consider the change of the chain per unit time. If the adversary successfully prevents the new block from being added to  $T_{\text{OD}}$ , we called the adversary successful. After analysis, the probability of the adversary's success in consecutive  $T$  rounds is as follows

$$S = \left( \frac{(2 - np)^2 \alpha p}{4np} + \frac{np - 2\alpha(np - 2)}{4} \times \frac{np(1 - P_{\text{onedelay}}^\Delta)}{2 - 2P_{\text{onedelay}}} \right)^T$$

Here we have  $\alpha = 0.8$ ,  $T = 5$ , we can see that when  $\Delta = 60$ , the adversary's attack success rate begins to increase gradually, and increases with the increase of the number of delay rounds. It can be seen that the number of delay rounds is an intuitive factor affecting the delay attack. Here we choose the delay probability  $\alpha$  when the number of delay rounds is 60 (10min) to observe the influence of the delay probability  $\alpha$  on the success rate of the delay attack.

It can be seen that when the delay probability reaches 0.8, the delay success rate will increase significantly.

Therefore, we set  $\alpha = 0.8$  and  $\Delta = 60$ . It can be observed that although the number of consecutive attacks increases, the success rate of delayed attacks is greatly reduced. Therefore, the results reflected by this model show that in the face of long-delayed attacks, POW has good security.

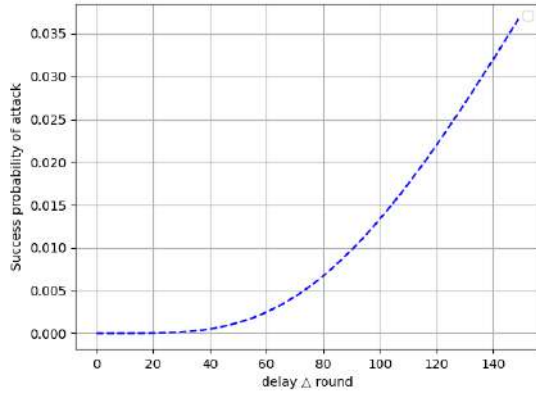


Figure 1: xxxxx

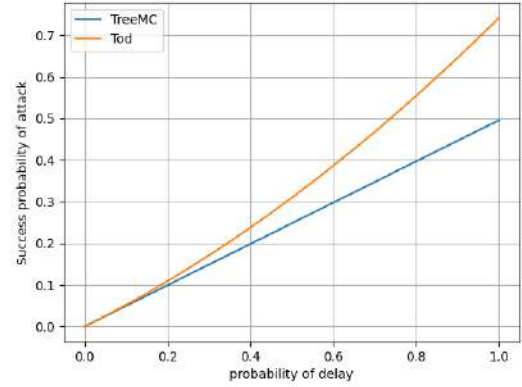


Figure 4: xxxxx

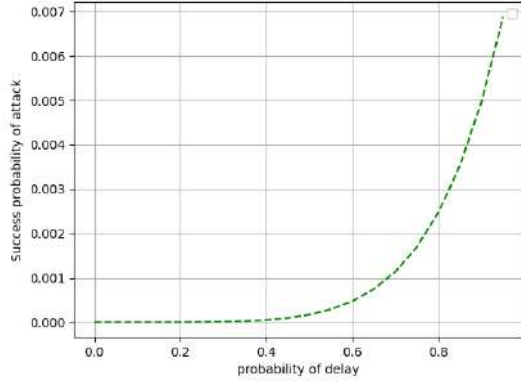


Figure 2: : xxxxx

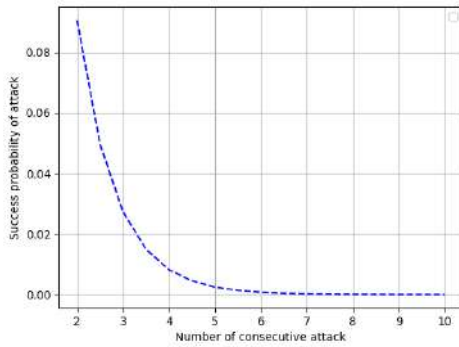


Figure 3: xxxxx

## 5.2 Comparative analysis

We compared and analyzed the adversary's attack on the security property of chain growth after the two models  $T_{OD}$  and  $Tree_{MC}$  mined delayable blocks on the two branches of  $H_A$  and  $H_A$  in one round at long delay attack. Similarly,  $np=1/60$ . It can be seen from the figure that when set  $\alpha$  is unchanged, the optimized  $T_{OD}$  model shows that the adversary's attack success probability is higher than that of the  $Tree_{MC}$  model, and with the increase of  $\alpha$ , the adversary's attack success probability on the chain growth rate increases in one round, but not over 0.8. Combining with the previous experimental data, although the success probability of the attack may increase after the adversary mines the delayable block, with the increase of the number of consecutive rounds  $T$  and the change of the number of delayable blocks and the number of delay rounds  $\Delta$ , its success probability will be greatly reduced, so PoW consensus still has good security in the face of long-delay attacks.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grant No. 62162039, 61762060), Foundation for the Key Research and Development Program of Gansu Province, China (Grant No.20YF3GA016).

## REFERENCES

- [1] Nakamoto S. Bitcoin: A peer-to-peer electronic cash system[J]. Consulted, 2008.
- [2] Ittay Eyal and Emin G.Sirer. Majority is not enough: Bitcoin mining is vulnerable. In Financial Cryptography and Data Security, volume 8437 of Lecture Notes in Computer Science, Pages 436-454. Springer-Verlag, 2014.
- [3] Yonatan Sompolinsky and Aviv Zohar. Secure high-rate transaction processing in bitcoin. In Financial Cryptography and Data Security, volume 8975 of Lecture Notes in Computer Science, pages 507-527.
- [4] Trom J. Cuckoo cycle: a memory bound graph-theoretic proof-of-work[C]/International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2015: 49-62.
- [5] Arthur Gervais, Ghassan O. Karame, Karl Wust, Vasileios Glykantzis, Hubert Ritzdorf, and Srdjan Capkun. On the security and performance of proof of work blockchains. In ACM SIGSAC Conference on Computer and Communications Security, Press, 2016.
- [6] Bastiaan M. Preventing the 51%-attack: a stochastic analysis of two phase proof of work in bitcoin. 2015.
- [7] Nayak K, Kumar S, Miller A, *et al*. Stubborn mining: generalizing selfish mining and combining with an eclipse attack. In: Proceedings of IEEE European Symposium on Security and Privacy, 2016. 305–320.

- [8] Aggelos Kiayas and Giorgos Panagiotakos. Speed-Security Tradeoffs in Blockchain Protocols. In IACR ePrint Archive Report:2015/1019.2016.
- [9] Juan Garay, Aggelos Kiayas, and Nikos Leonardos. The bitcoin backbone protocol with chains of variable difficulty. In Advances in Cryptology - CRYPT. TO 2017, volume 10401 of Lecture Notes in Computer Science, pages 291–323. Springer-Verlag, 2017.
- [10] Pass R, Seeman L, shelat A. Analysis of the blockchain protocol in asynchronous networks. In: Proceedings of Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2017. 643–673.
- [11] Wei P W, Yuan Q, Zheng Y L. Security of the blockchain protocol against long delay attack. In: Proceedings of International Conference on the Theory and Application of Cryptology and Information Security, 2018. 250–275.
- [12] Fork Rate-based Analysis of the Longest Chain Growth Time Interval of a PoW Blockchain, International Conference on Blockchain (Blockchain), 2019 IEEE DOI 10.1109/Blockchain.2019.00040.
- [13] Delay Analysis of Consensus Communication for Blockchain-Based Applications Using Network Calculus .IEEE Wireless Communications Letters. DOI 10.1109/LWC.2022.3183197.
- [14] Moustapha BA. The effect of propagation delay on the dynamic evolution of the Bitcoin blockchain. In: Digital Communications and Networks Volume 6, Issue 2. 2020. PP 157-166.
- [15] Ziyu Zhou, Zongyang Zhang, Jianwei Liu .Research Method of Nakamoto Consensus Security Properties. SCIENTIA SINICA Informationis. 2022.53(05)

# ADCapsNet: An Efficient and Robust Capsule Network Model for Anomaly Detection

Xiangyu Cai  
caixiangyu\_0205@qq.com  
Fujian Normal University  
Fuzhou, Fujian, China

Ruliang Xiao  
Fujian Normal University  
Fuzhou, China  
xiaoruliang@fjnu.edu.cn

Zhixia Zeng  
Fujian Normal University  
Fuzhou, China

Ping Gong  
Fujian Normal University  
Fuzhou, China

Shi Zhang  
Fujian Normal University  
Fuzhou, China

## ABSTRACT

With the rapid development of the industrial internet of things(IIoT), the anomalies will cause significant damage to the ordinary operation of the industry. Anomaly detection work has increasingly become a hot spot. Although many related kinds of research exist, some problems still need to be solved. This paper proposes an efficient and robust semi-supervised capsule network (ADCapsNet) for anomaly detection by changing the convolution structure to better extract the features of the data and adding a new SecondaryCaps layer to better extract spatial relationships. Besides, we optimize the vector selection for dynamic anomaly detection routing and propose the scoring operation, the modified probability mechanism. The modified probability mechanism can widen the score gap between positive and negative samples. This model can accurately identify and output the spatial relationships. Extensive experiments on four datasets show that the ADCapsNet has good performance in anomaly detection.

## CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; Scene anomaly detection; • **Security and privacy** → *Intrusion/anomaly detection and malware mitigation.*

## KEYWORDS

anomaly detection; capsule network; modified probability

### ACM Reference Format:

Xiangyu Cai, Ruliang Xiao, Zhixia Zeng, Ping Gong, and Shi Zhang. 2023. ADCapsNet: An Efficient and Robust Capsule Network Model for Anomaly Detection. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3590003.3590009>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590009>

## 1 INTRODUCTION

The IIoT system is widely used in national infrastructure. It integrates various types of sensors or controllers with sensing and monitoring capabilities into all aspects of industrial production[8]. With the rapid development of the IIoT, the anomalies generated in industrial production will cause significant damage to the ordinary operation of the industry. Nowadays, anomaly detection has increasingly become a hot spot.

So far, there has been much research on anomaly detection algorithms in academia and industry. These works of anomaly detection can be divided into the following four categories: statistics-based methods[11, 12], distance-based methods[13, 15], density-based methods[17] and deep learning methods[14, 24, 26]. Before using deep learning for anomaly detection, traditional methods had good results when dealing with some data from the past. However, due to the modern large scale and high dimensional data, applying a traditional anomaly detection model becomes increasingly challenging. Current researches show that deep learning models have successfully improved anomaly detection performance in the face of these challenges[23]. However, the above anomaly detection methods still have the following problems:

- The problem of view angle invariance: Due to the existence of pooling, the neural network has the problem of invariance of view angle. The rotated and the original images are considered two different images, which will cause significant interference in anomaly detection.
- The problem of position transformation relation: Traditional neural networks are unable to construct spatial relationships. For example, traditional neural networks are likely to mistake position reversed samples for normal samples.

To solve the above problems, we propose a novel capsule network for anomaly detection(ADCapsNet). The main contributions of this paper are as follows:

- In the proposed ADCapsNet, we reconstruct CapsNet by adding a convolution layer and SecondaryCaps connected to PrimaryCaps. These allow our proposed model to better extract features of the input.
- This paper proposes a modified probability mechanism, combining the proposed modified probability with the reconstruction loss to widen the score gap between positive and

negative samples. It has good robustness and spatial sensitivity.

- This paper proposes the dynamic anomaly detection routing to optimize the vector selection for dynamic routing so that it can mask vectors with large gaps from the current vector.
- We often do not know the characteristics and forms of abnormal samples, so it is difficult to learn the characteristics of anomalies. The ADCapsNet is a semi-supervised method that only needs to learn the characteristics of normal samples, which is very important to deal with anomalies in the real world.

The proposed ADCapsNet can effectively solve the anomaly detection difficulties caused by the mentioned problems. We have constructed sufficient experiments, and the results show that our proposed ADCapsNet method has a better result than the traditional anomaly detection methods and the current deep learning anomaly detection models.

The rest of this paper is organized as follows: Section 2 introduces the related work. Section 3 focuses on the structure and advantages of the ADCapsNet. Section 4 conducts sufficient experiments to give the results obtained from four experiments on four different datasets. Finally, the conclusions are given in Section 5.

## 2 RELATED WORK

Anomaly detection plays a vital role in various fields and is particularly important in data centers, logistics centers, monitoring centers, and life science centers for diagnosing system failures, detecting foreign intrusions, and discovering new knowledge. In general, anomaly detection can be divided into the following categories.

- **Statistics-Based Methods:** These methods are based on the assumption that the normal data follow a specific distribution and account for a large proportion. The problem with using this method is that both mean and variance are sensitive to outliers. The most representative methods are RHF[18] and COPOD[11].
- **Distance-Based Methods:** It calculates the distance between each point and the surrounding point. The distance between the normal point and the surrounding point is very close, while the distance between the abnormal and surrounding points is far. The classical algorithms are KNN[3], K-Means[15], and so on.
- **Density-Based Methods:** This method calculates the points' density and its neighboring points' density. Then calculates the relative density based on these two density values as the anomaly score. The classical algorithm is LOF[17].
- **Deep Learning Methods:** The core is to learn hierarchical features from data, then computes each sample's score to determine whether it is normal or not, such as Ganomaly[1] and Mkd4AD[20]. When it encounters data with ambiguous abnormal statuses, such as numbers 2 and 7 in the MNIST dataset, which we will introduce in section 4.1, the score difference is often small, so it is very likely to judge normal data as abnormal.

## 3 METHODOLOGY

In this section, we will present the structure and operations of ADCapsNet.

### 3.1 The Structure of ADCapsNet

The proposed ADCapsNet is mainly divided into two parts: encoder and decoder. The former five layers are encoders, and the last three layers are decoders. The main works are carried out in the encoder: we add a convolution layer to extract the data's features better, and a SecondaryCaps connected to PrimaryCaps by dynamic anomaly detection routing to make our proposed method more spatially sensitive and robust. We also propose the modified probability mechanism applied in DigitCaps. The main parameters and structure of each layer are shown in Fig. 1:

Before introducing the structure of ADCapsNet, we first introduce a concept, anomaly detection capsule (ADCapsule). It is a group of anomaly detection vector neurons in which all the important information (including position information, spatial relationship, etc.) is stored. The information will be encapsulated in the form of vector  $\vec{X}_i = (x_{i1}, x_{i2}, \dots, x_{id})$  by ADCapsules. When multiple lower-level predictions are consistent, higher-level ADCapsules become more active.

The first and second layers are convolution layers. The inputs are processed by 64 and 128 5×5 filters, respectively. We add an additional convolution layer to extract the features of the input data better.

The third layer is PrimaryCaps. It uses 128 5×5 convolution filters to manipulate the output of the second layer. Then all output feature maps are reshaped to 64 ADCapsule channels, and each channel contains 2 instantiation parameters.

The fourth layer is an additionally added SecondaryCaps. It uses a 3×3 slide window with a stride of 1. It uses the dynamic anomaly detection routing, which we will introduce in the next part, to fully connect 3×3×64=576 ADCapsules in PrimaryCaps to 32 ADCapsules in this layer. Therefore, the output of this layer is 4×4×32=512 ADCapsules. Each ADCapsule has 4 instantiation parameters.

The last layer is DigitCaps. It also uses dynamic anomaly detection routing to fully connect the 512 ADCapsules in the SecondaryCaps to 10 ADCapsules in this layer. Each ADCapsule has 8 instantiation parameters, and the length of each ADCapsule represents the probability that the input data belongs to each class.

In this case, the lower-level ADCapsule must decide how to distribute its output to the higher-level ADCapsule. The weight assignment mechanism is shown in Fig. 2:

We use margin loss as part of the loss function in the ADCapsNet. It defines an edge loss for the output of each ADCapsule and then calculates the total loss of all ADCapsules to optimize. The formula is as follows:

$$L_c = \varphi(n \cap c) \max(0, 0.9 - \|h_c\|_2)^2 + 0.5 \cdot (1 - \phi(n \cap c)) \max(0, \|h_c\|_2 - 0.1)^2 \quad (1)$$

Where  $\varphi(n \cap c) = 1$  if and only when the predicted class  $z$  corresponds to the real class  $n$  else  $\varphi(n \cap c) = 0$ .  $\|h_c\|_2$  represents the probability that the current sample belongs to category  $c$ .

The reconstruction loss is also part of the loss function, but it occupies a small proportion of the total loss function. The formula

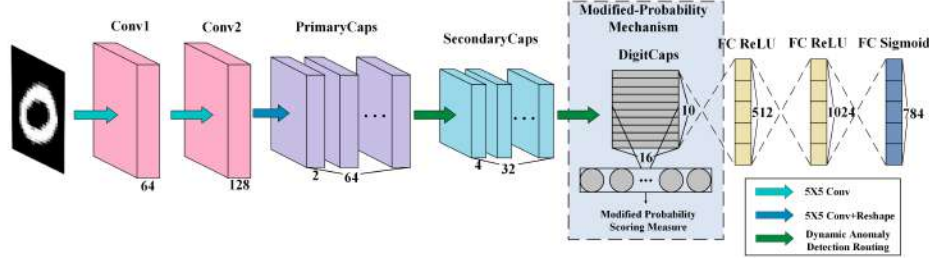


Figure 1: The structure of ADCapsNet: The first four layers are mainly used to extract features. The next layer, DigitCaps, applies the modified probability mechanism to score the input data. The last three layers are mainly used to reconstruct the images.

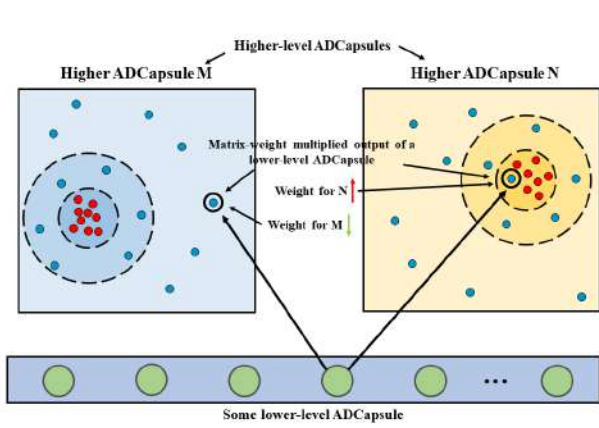


Figure 2: In the above figure, the prediction vector of the low-level ADCapsule is far away from the red cluster predicted by "correct" in ADCapsule M and close to the red cluster in ADCapsule N. Then the low-level capsule will increase the weight corresponding to ADCapsule N and decrease the weight corresponding to ADCapsule M.

is as follows:

$$L_r = \|X_{reconstruction} - X_{real}\|_2 \quad (2)$$

where  $X_{reconstruction}$  refers to the reconstruction image and  $X_{real}$  refers to the real image.

The total loss is as follows:

$$L_{total} = \sum_{c=1}^C L_c + 0.0005 \cdot L_r \quad (3)$$

### 3.2 Vector Selection of Dynamic Anomaly Detection Routing

The ADCapsNet uses dynamic anomaly detection routing to find a set of coefficients to predict the input vectors that best match the output vectors. We optimize the vector selection by shielding the input vectors that are too far away from the output vectors from participating in the iterative process, thus improving the accuracy of the predicted vectors and speeding up the iteration and convergence process. The demonstration process of dynamic anomaly detection routing is shown in Figure 5:

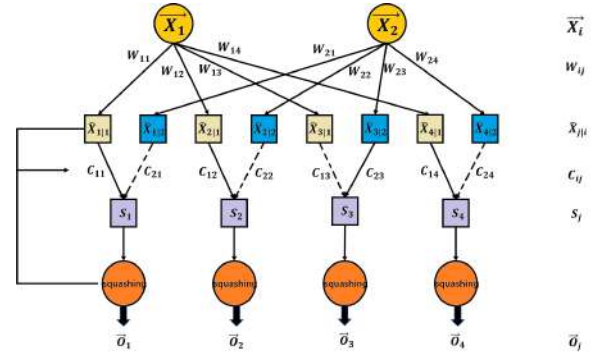


Figure 3: The procedure of dynamic anomaly detection routing: The dashed lines represent  $c_{21}, c_{22}, c_{13}$  and  $c_{24}$ , which are shielded from the iterative process.

The weights updating process of dynamic anomaly detection routing is explained as follows:

For two input vectors  $\vec{X}_1$  and  $\vec{X}_2$ , multiply the matrices  $W_{11}, W_{12}, \hat{x}_{1|1}, \hat{x}_{1|2}, \hat{x}_{2|1}, \hat{x}_{2|2}$  are obtained. Then  $\vec{s}_j$ , a weighted sum of  $\hat{x}_{i|j}$  is obtained.

$$\vec{s}_j = \sum_i c_{ij} \hat{x}_{i|j} \quad (4)$$

After that,  $\vec{s}_j$  is fed into the squashing function, and  $\vec{o}_j$  is obtained.

$$\vec{o}_j = \frac{\|\vec{s}_j\|_2}{1 + \|\vec{s}_j\|_2} \cdot \frac{\vec{s}_j}{\|\vec{s}_j\|_2} \quad (5)$$

What we should pay attention to here is the determination of the parameter  $c_{ij}$  in Eq.4. First, set parameters  $b_{ij} = 0$ , then perform the Softmax operation on  $b_{ij}$  to get the corresponding  $c_{ij}$ . Then we filter  $c_{ij}$ . Mask  $c_{ij}$  less than the threshold and keep  $c_{ij}$  greater than the threshold. Finally we can get the  $\vec{s}_j$  from the Eq.4, then get the  $\vec{o}_j$  from Eq.5. We can update  $b_{ij}$  after get the  $\vec{o}_j$ .

$$b_{ij} = b_{ij} + \hat{x}_{i|j} \cdot \vec{o}_j \quad (6)$$

$\hat{x}_{i|j}$  is equivalent to the personal prediction of ADCapsule  $i$ , and  $\vec{o}_j$  is equivalent to the result of the combined action of all ADCapsules. So the magnitude of the vector product represents the coupling degree of the ADCapsule to the final result. The more similar the output vectors are to the input vectors, the smaller the

angle between them, the larger the vector product, and the higher the degree of coupling.

### 3.3 Modified Probability Mechanism for Data Scoring

In ADCapsNet, the outputs of DigitCaps represent the probability that the input samples belong to these classes. However, the sum of all ADCapsules' activation probability is not necessarily 1. If the network training is good enough, there should be only one probability close to 1, indicating the possibility that the object belongs to this class. When an image is abnormal and cannot be interpreted by the network, the activation probability of each Digit ADCapsule will be very low.

This unique feature inspires us to propose the modified probability mechanism to judge the outlier degree of data. By using this mechanism, the score gap between abnormal and normal images is widened so we can find a threshold to distinguish positive and negative samples more easily. We extract the maximum prediction probability of the DigitCaps and then subtract the sum of the prediction probabilities of other classes. In this way, the score gap between normal data and abnormal data will be very large under the unique characteristics of ADCapsNet after applying the modified probability mechanism, so as to improve the performance of anomaly detection.

In addition, we add reconstruction loss as a correction to the score. It accounts for a small fraction of the total score. The score function formula based on modified probability is as follows:

$$score(x) = \max_{i \in M} \|h_i\|_2 - \sum_{j \in M \wedge j \neq i} \|h_j\|_2 + \gamma \frac{\max_{r \in N} L_r - L_x}{\max_{r \in N} L_r - \min_{r \in N} L_r} \quad (7)$$

Where  $x$  represents the input image,  $\|h_i\|_2$  represents the probability that image  $x$  belongs to the  $i$ -th class.  $L_x$  represents the reconstruction anomaly loss of sample  $x$ .  $M$  represents the total categories and  $N$  represents the total samples.  $\gamma$  is a coefficient that represents the weight of reconstruction error in the score.

Since the reconstruction loss is distance-preserving, we do not need to bring it into the analysis of the score distance preservation in Section 3.3. In this paper, we make  $\gamma = 0.1$ , which controls the score range of reconstruction error between 0.01 and 0.1, to modify the total score finely.

### 3.4 Score Distance Preservation Analysis of ADCapsNet

In this section, we analyze the proposed ADCapsNet to prove that it not only theoretically ensures the score gap between positive and negative samples but also widens the score gap between positive and negative samples. It is easier to determine the appropriate threshold to judge the positive and negative samples.

The proof is as follows:

For normal sample  $X \in U_{normal}$ , there exists  $\alpha, \varepsilon \in (0, 1)$ , such that:

$$0 < \|X_j\|_2 < \alpha < \varepsilon \leq \max \|X_i\|_2 < 1 \quad (8)$$

For abnormal sample  $Y \in U_{abnormal}$ , there exists  $\gamma \in (0, 1)$ , such that:

$$0 < \alpha < \|Y_j\|_2 < \max \|Y_i\|_2 \leq \gamma < 1 \quad (9)$$

Where  $i \in M, j \in M \wedge j \neq i$ ,  $\|X_i\|_2$  represents the probability that sample  $X$  belongs to the  $i$ -th class.  $M$  represents the total classes.

From Eq.8, we can get the score range of normal sample  $X$ :

$$\varepsilon - (C - 1)\alpha \leq Score(X) = \max_{i \in M} \|X_i\|_2 - \sum_{j \in M \wedge j \neq i} \|X_j\|_2 < 1 \quad (10)$$

From Eq.9, we can get the score range of abnormal sample  $Y$ :

$$-(C-2)\gamma \leq Score(Y) = \max_{i \in M} \|Y_i\|_2 - \sum_{j \in M \wedge j \neq i} \|Y_j\|_2 \leq -(C-2)\alpha \leq 0 \quad (11)$$

We can use Eq.10,11 to compute the score gap  $Gap1$  through our proposed method between normal sample  $X$  and abnormal sample  $Y$ :

$$\varepsilon - \alpha \leq Gap1 = Score(X) - Score(Y) \leq 1 + (C - 2)\gamma$$

The score gap  $Gap2$  through usual normal method between normal sample  $X$  and abnormal sample  $Y$  is:

$$\varepsilon - \gamma \leq Gap2 = \max_{i \in M} \|X_i\|_2 - \max_{j \in M} \|Y_j\|_2 \leq 1$$

Cause  $Gap1 \in [\varepsilon - \alpha, 1 + (C - 2)\gamma]$ ,  $Gap2 \in [\varepsilon - \gamma, 1]$  and  $\varepsilon - \alpha < \varepsilon - \gamma, 1 + (C - 2)\gamma > 1$ . So the range of  $Gap1$  is larger than that of  $Gap2$ . Thus, it is proved that the method proposed by us can guarantee the score gap between positive and negative samples and enlarge the score gap and the method proposed by us has better score distance preservation.

## 4 EXPERIMENT AND RESULT ANALYSIS

In this part, we introduce the datasets, the evaluation indicators, and experimental setups. We will conduct four sets of simulation experiments on four groups of large scale and high dimensional datasets to prove the superiority of our proposed method. The **Experiment 1** is to prove that the performance of our proposed method is better than the current method, and the **Experiment 2** is to prove that our method still has good performance in the case of uneven data. The purpose of **Experiment 3** is to find a suitable threshold to determine anomalies, and the purpose of **Experiment 4** is to prove that our proposed method has good spatial robustness.

### 4.1 Datasets

We use three datasets, MNIST<sup>1</sup>, Fashion-MNIST<sup>2</sup> and CIFAR-10<sup>3</sup> to evaluate the performance of ADCapsNet. In addition, we also used affNIST<sup>4</sup> to prove that ADCapsNet has good spatial sensitivity and robustness. We follow the common practice of anomaly detection and define one of the classes as an abnormal class and the others as normal classes for anomaly detection.

<sup>1</sup><http://www.research.att.com/~yann/ocr/mnist/>

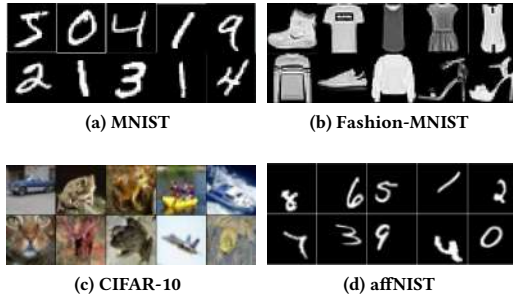
<sup>2</sup><https://github.com/zalandoresearch/fashion-mnist>

<sup>3</sup><http://www.cs.toronto.edu/~kriz/cifar.html>

<sup>4</sup><http://www.cs.toronto.edu/~tijmen/affNIST/>

- **MNIST**: The training set consists of 50000 28×28 pixels handwritten digital images, and the test set includes 10000 28×28 pixels handwritten digital images.
- **Fashion-MNIST** The training set consists of 50000 28×28 pixels images and the test set includes 10000 28×28 pixels images. These images are divided into ten categories.
- **CIFAR-10** There are 10 categories of RGB color pictures: airplanes, cars, birds, cats, deer, dogs, frogs, horses, boats and trucks. The size of each picture is 32×32 pixels, and each category has 60000 images. There are 50,000 32×32 pixels training pictures and 10,000 32×32 pixels test pictures.
- **affNIST** This is the dataset extended from MNIST. The training and test set consists of 50000 and 10000 images of 40×40 pixels, respectively. The training set is expanded to 40×40 pixels without any spatial variation. The test dataset is randomly placed anywhere in the 40×40 pixels after random rotation transformation.

Parts of the four datasets are shown as Figure 4:



**Figure 4:** Parts of the four datasets:(a)MNIST,(b)Fashion-MNIST,(c)CIFAR-10,(d)affNIST.

## 4.2 Evaluation indicators

In order to verify the effectiveness and superiority of ADCapsNet, we use AUROC, AUPRC, and other indicators to evaluate the test results. In the real world, category imbalance often occurs. That is, the imbalance between positive and negative samples and the distribution of positive and negative samples may change over time. In this case, the ROC curve can remain unchanged. AUROC is the area under the ROC curve. Its value range is 0-1. The closer it is to 1, the better the effect of the model.

The formula of AUROC is as follows:

$$AUROC = \frac{\sum_i rank_i - \frac{M(1+M)}{2}}{M \times N} \quad (12)$$

Where  $M$  represents the number of negative samples.  $N$  represents the normal samples.  $rank_i$  represents the rank value of data score sorted from small to large.  $rank_i$  is from 0- $(M + N)$ .

However, there is still a problem with the ROC curve. When there are more negative samples than positive samples, even if many negative samples are misjudged as positive samples due to the characteristics of the ROC curve, the value of AUROC will still be maintained at a relatively high level. The PRC[21] curve can

better reflect the essence of the data, and the value of AURRC can better reflect the effect of the model.

The PRC curve integrates two evaluation indicators of *precision* and *recall*. The *precision* is the  $y$ -axis and the *recall* is the  $x$ -axis. *Precision* evaluates the accuracy of anomaly detection results, and the *recall* rate evaluates the completeness of the experiment. The formula is as follows:

$$precision = \frac{TP}{TP + FP} \quad (13)$$

$$recall = \frac{TP}{TP + FN} \quad (14)$$

Where  $TP$  is true positive.  $FP$  is false positive.  $FN$  is false negative.  $TN$  is true negative.

In addition, we also use the F1-Score[6] as a performance indicator to judge our proposed model. The formula is as follows:

$$F1 - Score = \frac{2 \times precision \times recall}{precision + recall} \quad (15)$$

## 4.3 Experimental setups

During training, only normal samples were used. During testing, a mixture of normal and abnormal samples was fed into the trained model. We used python 3.6 and TensorFlow 1.15 and trained on an Nvidia Quadro P5000 only.

The specific training parameters are shown in Table 1:

## 4.4 Simulation experiments

**Experiment 1:** The purpose of this experiment is to prove the superiority of our proposed method. The above three datasets and AUROC were used as important indicators to evaluate anomaly detection performance and compared with existing methods, including generative, self-supervised, and autoencoder-based methods. We select one of the classes as the abnormal class and the rest of the classes as normal classes for anomaly detection. As is shown in Table 2:

Experiments show that the performance of our proposed method is better than that of the existing methods in the above three datasets. However, with the increase in image complexity, the performance of all kinds of anomaly detection models decreases in varying degrees, which is inevitable.

**Experiment 2:** This experiment aims to prove that our proposed method still has good performance in the case of uneven data, using MNIST dataset and AUPRC as the evaluation indicator. As in Experiment 1, we also select one class as the abnormal class and the rest as the normal classes. The results are shown in Fig. 5.

The results show that the proposed method performs well in the case of unbalanced data. It is worth mentioning that our method is tested on the data with a positive-negative sample ratio of 1:10, while the other methods are tested on the normal test set. If we test on the normal test set, the AUPRC score will improve further.

**Experiment 3:** The purpose of this experiment is to find a suitable threshold and use it to detect the anomaly. Use the Fashion-MNIST dataset and 2, 5, and 8 as the abnormal categories.

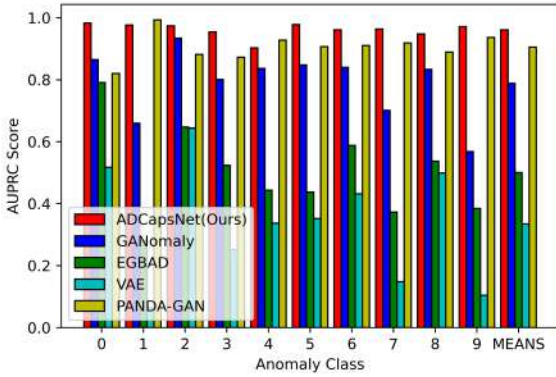
First, we use the t-SNE[16] algorithm to reduce the dimensionality of the data and show the spatial distribution of normal and abnormal data in 3-D space, as is shown in Fig. 6.

**Table 1: The experimental setups of ADCapsNet.**

Setups	Optimizer	Learning Rate	Batch Size	Weight Decay	Epochs
Value	Adam	0.0001	128	cosine warmup	2000

**Table 2: Performance comparison for different algorithms on three different datasets: where 0 column represents that we select the first class as the abnormal class and the other classes as the normal classes.**

Dataset	Method	0	1	2	3	4	5	6	7	8	9	Mean
MNIST	OCSVM[4]	0.948	0.970	0.901	0.913	0.921	0.920	0.889	0.891	0.908	0.912	0.917
	Ganomaly[1]	0.890	0.231	0.94	0.771	0.720	0.800	0.880	0.784	0.864	0.58	0.746
	EGBAD[25]	0.750	0.323	0.800	0.694	0.652	0.715	0.684	0.794	0.678	0.491	0.658
	CapsNet <sub>pp</sub> [10]	0.994	0.997	0.965	0.976	0.945	0.974	0.980	0.975	<b>0.989</b>	<b>0.989</b>	0.978
	DAGPR[5]	0.993	0.993	0.979	0.959	0.949	0.934	0.970	0.951	0.943	0.938	0.960
	Skip-Ganomaly[2]	0.781	<b>0.997</b>	0.589	0.657	0.387	0.698	0.552	0.758	0.874	0.856	0.715
	PUMAD[7]	<b>0.995</b>	0.992	0.964	0.966	<b>0.981</b>	0.945	<b>0.988</b>	0.977	0.870	0.933	0.961
	<b>ADCapsNet</b>	0.993	0.995	<b>0.992</b>	<b>0.986</b>	0.973	<b>0.990</b>	0.983	<b>0.988</b>	0.986	0.983	<b>0.987</b>
Fashion-MNIST	OCSVM[4]	0.699	0.487	0.743	0.624	0.712	0.903	0.698	0.446	0.932	0.672	0.692
	Skip-Ganomaly[2]	0.653	0.785	0.823	<b>0.913</b>	<b>0.855</b>	0.542	0.555	<b>0.875</b>	0.722	<b>0.927</b>	0.765
	CapsNet <sub>RE</sub> [10]	0.456	0.878	0.479	0.693	0.397	0.903	0.464	0.781	0.865	0.762	0.669
	CapsNet <sub>pp</sub> [10]	0.609	0.721	0.814	0.881	0.758	0.720	0.762	0.626	0.940	0.670	0.750
	DAGMM[27]	0.303	0.311	0.475	0.481	0.499	0.413	0.42	0.374	0.518	0.378	0.417
	<b>ADCapsNet</b>	<b>0.759</b>	<b>0.948</b>	<b>0.827</b>	0.897	0.809	<b>0.915</b>	<b>0.789</b>	0.729	<b>0.962</b>	0.683	<b>0.832</b>
Cifar-10	OCSVM[4]	0.630	0.440	0.649	0.487	0.735	0.500	0.725	0.533	0.679	0.508	0.587
	ARAE[19]	<b>0.722</b>	0.431	0.690	0.550	<b>0.752</b>	0.547	0.701	0.510	0.722	0.400	0.602
	VAE <sub>Gradient</sub> [9]	0.658	0.543	0.632	0.461	0.725	0.493	0.699	0.490	0.641	0.477	0.582
	Capsnet <sub>RE</sub> [10]	0.377	<b>0.736</b>	0.413	0.597	0.390	0.590	0.486	0.628	0.403	0.688	0.531
	CapsNet <sub>pp</sub> [10]	0.622	0.455	0.671	0.675	0.683	0.635	<b>0.727</b>	0.673	0.710	0.466	0.612
	AnoGAN[22]	0.671	0.547	0.529	0.545	0.651	0.603	0.585	0.625	<b>0.758</b>	0.665	0.618
	<b>ADCapsNet</b>	0.702	0.687	<b>0.747</b>	<b>0.732</b>	0.653	<b>0.736</b>	0.696	<b>0.721</b>	0.645	<b>0.698</b>	<b>0.702</b>

**Figure 5: The AUPRC results of different anomaly detection methods in the case of data imbalance: It can be seen that the performance of our proposed method has good performance in most cases.**

Then we feed the test data into our anomaly detection model to get the anomaly scores. Fig. 7 shows the distribution of scores after the data is processed by our model.

Through the experiment, it is found that the score gap between the normal samples and the abnormal samples is very large. The score of the normal samples mainly accumulates between 0 and 0.8, and the score of the abnormal sample mainly accumulates between -3 and 0.

Finally, we take the threshold values -0.7, -0.6, -0.5, -0.4, -0.3, -0.2, respectively, for testing. Four evaluation indicators are used: Precision, Recall, F1-Score, and Accuracy. The results are shown in Fig. 8.

It can be seen that when the threshold is -0.5, all the performance indicators of the model achieve the optimal effect.

**Experiment 4:** The images received through all kinds of sensors are likely to be affine transformed images, so the anomaly detection model needs to be sensitive and robust to space relationships. This experiment aims to prove that our proposed method has good spatial sensitivity and robustness using the affnIST dataset. The result is shown in Fig. 9.

It can be seen that our proposed method still has remarkable performance even in the position of affine transformed images. This shows that our proposed method has good spatial robustness and sensitivity.

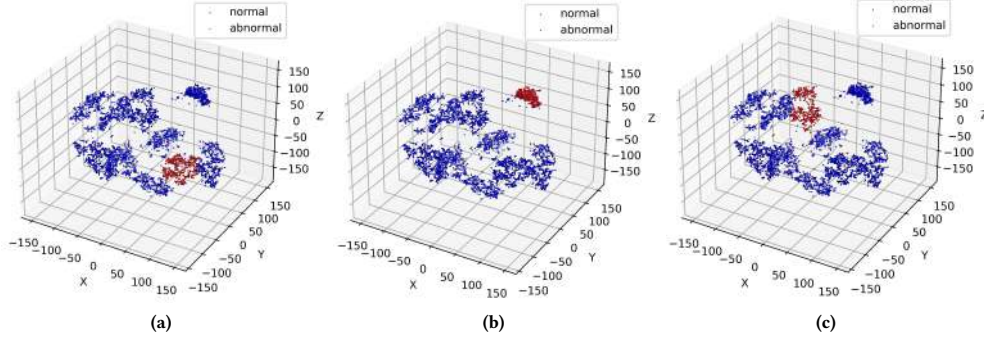


Figure 6: These three pictures show the spatial distribution of normal and abnormal data processed by the t-SNE algorithm. The red dots represent the anomaly, and the blue dots represent normal data.

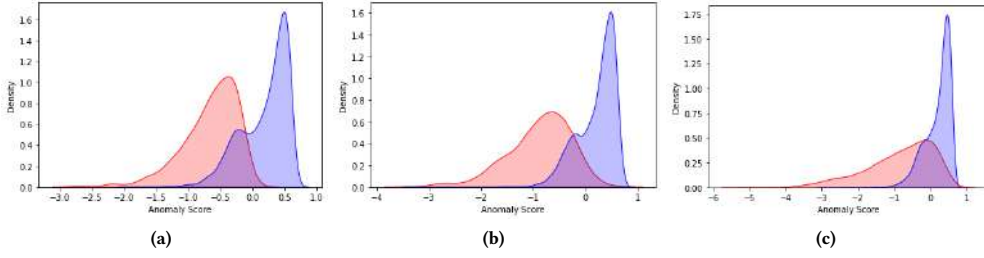


Figure 7: These three images show the distribution of the scores of the data. The red part indicates the density of abnormal samples, and the blue part is the density of normal samples.

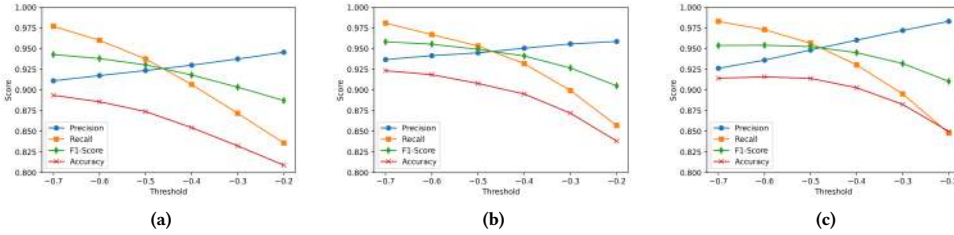


Figure 8: These three pictures shows the performance of the model under different thresholds. It can be seen that when we select -0.5 as threshold, all the indicators achieve the optimal effect.

#### 4.5 Ablation studies

The ADCapsNet proposed in this paper is composed of several components. We perform sufficient ablation studies to demonstrate the effectiveness of each component and our proposed modified probability mechanism. We will study the impact of SecondaryCaps, dynamic anomaly detection routing, and the modification probability mechanism on the performance of the ADCapsNet.

**SecondaryCaps and dynamic anomaly detection routing:** In this subsection, we will study the impact of SecondaryCaps and dynamic anomaly detection routing (DADR) on the ADCapsNet. We chose AUROC(%) as the evaluation metric. The results are shown in Table 3:

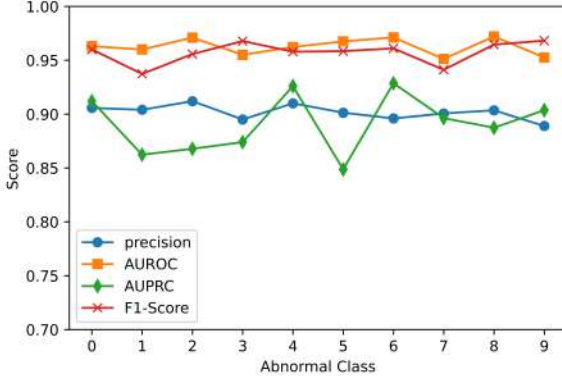
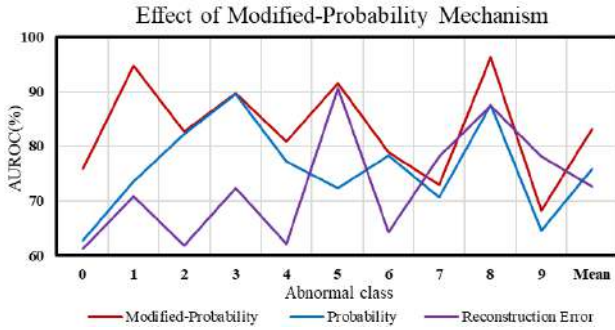
It can be seen that our proposed SecondaryCaps and dynamic anomaly detection routing both improve the performance of the proposed method. The best results are achieved when the two components are combined.

**Modified Probability Mechanism:** In this subsection, we will study the impact of the modified probability mechanism on the performance of our proposed method. We use the Fashion-MNIST dataset to compare with probability-based and reconstruction-based scoring mechanisms. The results are shown in Figure 10:

It can be seen that the modified probability mechanism has a great degree of improvement for the model proposed in this paper. The modified probability mechanism outperforms both the ordinary

**Table 3: The effect of SecondaryCaps and dynamic anomaly detection routing on ADCapsNet, where DADR indicates Dynamic Anomaly Detection Routing.**

Method	SecondaryCaps	DADR	Ordinary Routing	MNIST	F-MNIST
ADCapsNet	×	×	✓	96.8	75.0
	×	✓	×	97.6	77.3
	✓	×	✓	97.6	79.2
	✓	✓	×	98.7	82.1

**Figure 9: The results of each evaluation indicators on the affNIST dataset.****Figure 10: Impact of different scoring mechanisms on the model.**

probability-based and reconstruction-based approaches for almost every anomaly class.

## 5 CONCLUSIONS

The large scale and high dimensional data have brought great difficulties to the current anomaly detection work, so it is significant to put forward an anomaly detection method with better performance. In this paper, we propose a novel anomaly detection method AD-CapsNet and the modified probability mechanism to improve the efficiency of anomaly detection. It is better than most of the current advanced anomaly detection methods in each evaluation indicator. Besides, the method has good spatial sensitivity and robustness.

The anomaly detection of the affine transformed images also has a good effect, which is of significant advantage in dealing with the data in the real world.

In future work, we will further improve the vector selection for dynamic anomaly detection routing to enhance the effectiveness of our proposed method on complex image datasets.

## REFERENCES

- [1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. 2018. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *Asian conference on computer vision*. Springer, Cham, 622–637.
- [2] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. 2019. Skip-ganomaly: Skip connected and adversarially trained encoder-decoder anomaly detection. In *2019 International Joint Conference on Neural Networks (IJCNN)*. 1–8.
- [3] Yewang Chen, Xiaoliang Hu, Wentao Fan, Lianlian Shen, Zheng Zhang, Xin Liu, Jixiang Du, Haibo Li, Yi Chen, and Hailin Li. 2020. Fast density peak clustering for large scale data based on kNN. *Knowledge-Based Systems* 187 (2020), 104824.
- [4] Yunqiang Chen, Xiang Sean Zhou, and Thomas S Huang. 2001. One-class SVM for learning in image retrieval. In *Proceedings 2001 International Conference on Image Processing (ICIP)*, Vol. 1. IEEE, 34–37.
- [5] Jinan Fan, Qianru Zhang, Jialei Zhu, Meng Zhang, Zhou Yang, and Hanxiang Cao. 2020. Robust deep auto-encoding Gaussian process regression for unsupervised anomaly detection. *Neurocomputing* 376 (2020), 180–190.
- [6] Damien Fourure, Muhammad Usama Javaid, Nicolas Posocco, and Simon Tihon. 2021. Anomaly Detection: How to Artificially Increase Your F1-Score with a Biased Evaluation Protocol. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 3–18.
- [7] Hyunjun Ju, Dongha Lee, Junyoung Hwang, Junghyun Namkung, and Hwanjo Yu. 2020. PUMAD: PU metric learning for anomaly detection. *Information Sciences* 523 (2020), 167–183.
- [8] W.Z. Khan, M.H. Rehman, H.M. Zangoti, M.K. Afzal, N. Armi, and K. Salah. 2020. Industrial internet of things: Recent advances, enabling technologies and open challenges. *Computers & Electrical Engineering* 81 (2020), 106522.
- [9] Gukyeon Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. 2020. Novelty detection through model-based characterization of neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3179–3183.
- [10] Xiaoyan Li, Iluju Kiringa, Tet Yeap, Xiaodan Zhu, and Yifeng Li. 2020. Exploring deep anomaly detection methods based on capsule net. In *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*. Springer, 375–387.
- [11] Zheng Li, Yue Zhao, N Botta, C Ionescu, and X COPOD Hu. 2020. Copula-based outlier detection. In *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*, Sorrento, Italy. 17–20.
- [12] Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. 2022. ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering* (2022), 1–1. <https://doi.org/10.1109/TKDE.2022.3159580>
- [13] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11474–11481.
- [14] Tao Liu, Meiqian Duan, Luyang Sun, and Bo Zhang. 2022. An Auditory Measure for Anomaly Detection based on Auto-encoders. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. 109–114. <https://doi.org/10.1109/CACML55074.2022.00026>
- [15] Konstantin Makarychev and Liren Shan. 2021. Near-Optimal Algorithms for Explainable k-Medians and k-Means. In *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 7358–7367.
- [16] Bruno Henrique Meyer, Aurora Trinidad Ramirez Pozo, and Wagner M. Nunan Zola. 2022. Global and local structure preserving GPU t-SNE methods for large-scale applications. *Expert Systems with Applications* (2022), 116918. <https://doi.org/10.1016/j.eswa.2022.116918>

- [org/10.1016/j.eswa.2022.116918](https://doi.org/10.1016/j.eswa.2022.116918)
- [17] Tshepiso Mokoena, Turgay Celik, and Vukosi Marivate. 2022. Why is this an anomaly? Explaining anomalies using sequential explanations. *Pattern Recognition* 121 (2022), 108227. <https://doi.org/10.1016/j.patcog.2021.108227>
  - [18] Andrian Putina, Mauro Sozio, Dario Rossi, and José Manuel Navarro. 2020. Random Histogram Forest for Unsupervised Anomaly Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*. 1226–1231. <https://doi.org/10.1109/ICDM50108.2020.00154>
  - [19] Mohammadreza Salehi, Atrin Arya, Barbod Pajoum, Mohammad Otoofi, Amirreza Shaeiri, Mohammad Hossein Rohban, and Hamid R Rabiee. 2021. Arae: Adversarially robust training of autoencoders improves novelty detection. *Neural Networks* (2021).
  - [20] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H. Rohban, and Hamid R. Rabiee. 2021. Multiresolution Knowledge Distillation for Anomaly Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14902–14912.
  - [21] Niels Ole Salscheider. 2021. FeatureNMS: Non-Maximum Suppression by Learning Feature Embeddings. In *2020 25th International Conference on Pattern Recognition (ICPR)*. 7848–7854. <https://doi.org/10.1109/ICPR48806.2021.9412930>
  - [22] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*. Springer, Cham, 146–157.
  - [23] Ruoying Wang, Kexin Nie, Tie Wang, Yang Yang, and Bo Long. 2020. Deep Learning for Anomaly Detection. In *Proceedings of the 13th International Conference on Web Search and Data Mining (Houston, TX, USA) (WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 894–896.
  - [24] Jie Xie, Qing Cheng, Guangquan Cheng, and Jincai Huang. 2022. Detection of anomalies in key performance indicator data by a convolutional long short-term memory prediction model. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. 328–336. <https://doi.org/10.1109/CACML55074.2022.00062>
  - [25] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. 2018. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222* (2018).
  - [26] Xianchao Zhang, Jie Mu, Xiaotong Zhang, Han Liu, Linlin Zong, and Yuqiang Li. 2022. Deep anomaly detection with self-supervised learning and adversarial training. *Pattern Recognition* 121 (2022), 108234. <https://doi.org/10.1016/j.patcog.2021.108234>
  - [27] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *International Conference on Learning Representations*.

# Customer Service Hot event Discovery Based on Dynamic Dialogue Embedding

Fei Li

School of Computer Science and  
Technology, University of Science and  
Technology of China and Innovation  
+ Research Institute, GuoChuang  
Cloud Technology Co., Ltd., China  
fli312@mail.ustc.edu.cn

Yanyan Wang\*

Innovation + Research Institute,  
GuoChuang Cloud Technology Co.,  
Ltd., China  
ywang0619@163.com

Ying Feng

Innovation + Research Institute,  
GuoChuang Cloud Technology Co.,  
Ltd., China  
feng.ying@ustcinfo.com

Qiangzhong Feng

Innovation + Research Institute,  
GuoChuang Cloud Technology Co.,  
Ltd., China  
qzfeng@ustcinfo.com

Yuan Zhou

Innovation + Research Institute,  
GuoChuang Cloud Technology Co.,  
Ltd., China  
zhou.yuan@ustcinfo.com

Dexuan Wang

Innovation + Research Institute,  
GuoChuang Cloud Technology Co.,  
Ltd., China  
wang.dexuan@ustcinfo.com

## ABSTRACT

Frequent customer service conversations focus on hot topics of communication users, and automatic hot topic discovery is critical to improving user experience. Traditionally, Customer service relies on operator to write traffic summaries. It leads to the source of the conversation difficult to analyze, which makes difficult to spot aggregated hotspot events. In this paper, we propose a Customer Service hot event Discovery based on dynamic dialogue embedding (CShe-D). This model includes dynamic semantic representation of customer service dialogue, clustering-based customer service hot event discovery and new hot event prediction. In the dialogue semantic embedding module, we obtain the dynamic embedding of each dialogue with combining word importance and word length based on the pre-trained language model to capture richer semantic information in different contexts. We further apply a clustering iterative algorithm with dynamic dialogue embedding to discover customer service hotspots. It can monitor the change trend of events in real time, optimize the accuracy of hot event discovery in operator customer service. Finally, the effectiveness of our CShe-D model is verified by experiments on real dialogue data in the field of customer service.

## CCS CONCEPTS

• **Computing methodologies**; • **Information extraction**;

\*The first and primary author of this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590011>

## KEYWORDS

Customer service, Dialogue embedding, Semantic representation, Hot event discovery

### ACM Reference Format:

Fei Li, Yanyan Wang, Ying Feng, Qiangzhong Feng, Yuan Zhou, and Dexuan Wang. 2023. Customer Service Hot event Discovery Based on Dynamic Dialogue Embedding. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3590003.3590011>

## 1 INTRODUCTION

The average daily traffic volume in the field of operator customer service has been large in recent year. As an example, an operator in certain province in China, gains 4.2 million monthly incoming call volumes, and the average daily traffic volume is about 130 times per person. The traffic volume has grown rapidly, with an average increase of about 17% over the same period last year. In the field of operator customer service, there is an urgent need to analyze traffic sources and take inventory of problems to reduce traffic volume. However, in the field of customer service, the analysis of traffic sources is basically carried out through the experience of the operators. It is relies entirely on manual subjectivity, and there is a certain risk of error. In addition, there are also ways to rely on experience to solidify the types of traffic source. When using supervised algorithms, it is difficult to find the causes of emergencies and abnormal fluctuations.

The dynamic dialogue semantic representation can improve the accuracy of clustering algorithms and realize the discovery of internal relationships between dialogues. Therefore, we propose a customer service hot event discovery based on dynamic dialogue embedding. In this work, we propose a improved semantic representation method based on the dynamic TF-IDF for customer service to obtain semantically rich dialogue embedding. Then we utilize the K-Means algorithm with dynamic dialogue semantic representation to quickly and efficiently detect customer service hotspot events. It can help to analyze the source of customer service problems, and assist in manual quality inspection.

The main contributions of this paper are as follows:

- Our model uses unsupervised algorithm to realize automatic hot event discovery and subsequent hot event prediction integration. Existing methods are mostly based on word frequency, our model considers the length of the word and the context word order of the texts to obtain dynamic dialogue semantic representation. Therefore, in the paper, we propose a dynamic semantic representation model of customer service dialogue based on improved TF-IDF and can obtain rich semantic information.
- For the field of operator customer service, we design three algorithm models: hierarchical clustering, density clustering and K-Means clustering, respectively, combined with domain semantic representation for comparative experiments. The customer service hotspot event discovery model is more effective, and the performance is improved by 13.83% compared with the best results in other clustering methods, which can effectively improve the accuracy of clustered hotspot event discovery.

## 2 RELATED WORK

### 2.1 Semantic Representation

Semantic representation utilizes sparse or dense matrix to generate a space embedding for text. In the field of customer service, semantic representation model can learn to gather customer service professional field information. It captures rich semantic information, which will greatly affect the effect of subsequent text analysis.

The TF-IDF proposed by Jones [11] is the most common semantic representation method based on word frequency statistics. But this method ignores the importance of word length and context word order, and cannot obtain accurate text semantic representation. After that, Tomas Mikolov [17] considered the meaning of words and the relationship between words, and proposed Word2vec for word vector representation. This method has efficient word vector representation. Because the word and vector are in a one-to-one relationship, it is impossible to solve the problem of polysemy. And it is difficult to dynamically optimize for specific task scenarios. These approaches provide mainly word-level information, while our aim is to capture a higher level of semantics [12]. In addition, phrase-level or sentence-level embedding has been used to encode text into vector images that can be trained using an unmarked body [15]. In recent years, a series of large-scale language models have been proposed, including ELMo [19] (Embeddings from Language Models), GPT [20], BERT [4] and ERNIE [27]. ELMo [19] uses a linear combination of layers to represent word vectors. GPT [20] builds with one-way Transformer Decoder module. BERT [4] trains deep bidirectional representations according to the context of all layers (Bidirectional Encoder Representations from Transformer). ERNIE [27] improves on knowledge mapping based on BERT. These models highlight the characteristics of using large-scale corpora to train word vectors, but there are also problems such as poor interpretability of results and over-reliance on corpus.

In order to obtain semantic information with high accuracy, we propose a conversation semantic embedding method based on word length, which utilizes improved TF-IDF to obtain embedding

weights. This will dynamically strengthen the weight information of domain words.

### 2.2 Hot Event Discovery

Hot event discovery is to mine some frequently occurring events from massive texts. Currently, it is mainly implemented using clustering algorithms. As an unsupervised machine learning method, clustering algorithm does not require a lot of unlabeled data. It is mainly based on the text of the same category has greater text similarity, and different types of text have less text similarity [3]. At present, clustering has become an important means for intelligent recommendation [18] [28] and summarization [2] [8] of text, which has high flexibility and has been widely studied by scholars [5].

Dai et al. [24] proposed to use the hierarchical clustering algorithm to perform topic detection, and obtain subtopics, so as to reflect the characteristics of hot topics in a deeper level. Jiang et al. [13] used the peak density clustering algorithm to detect hot topics of complaints in the telecommunications industry. According to the similarity and density of text segmentation, they obtained cluster division, and mined keywords to form hot topic descriptions. However, the convergence time of density clustering for large sample sets is long. When the density of the sample set is not uniform, the clustering quality will be relatively poor. Zhang [6] proposed to use the K-Means algorithm to optimize the initial clustering center. They found potential hot topics in massive texts, and designed a hot topic ranking strategy to present topics in an orderly manner. The principle of the algorithm is simple and highly interpretable, which has efficient for large-scale data sets [9].

K-Means clustering algorithm is widely used in data operation, cluster analysis and other fields, but it is rarely used in the field of operator customer service. In this paper, we utilize K-Means to cluster customer service text, which can effectively improve the accuracy of hot event discovery in the field of customer service.

## 3 THE CUSTOMER SERVICE HOTSPOT EVENT DISCOVERY MODEL

### 3.1 Overall

In view of the large average daily traffic volume in the customer service, it is difficult to analyze the hot event sources and other problems. In this paper, we propose a Customer Service hot event Discovery model based on dynamic dialogue embedding, namely CSHe-D model, including customer service semantic representation module, hot event discovery module based on K-Means clustering and new event category prediction module. The specific framework is shown in Figure 1.

Firstly, the dialogue semantic representation of customer service utilize new words to build a domain lexicon, and then calculates the dynamic sentence embeddings based on the improved TF-IDF algorithm. After that, it obtains the dialogue semantic representation through semantic compression, which can realize multi-dimensional aggregation of semantics to capture rich conversation information. Secondly, in customer service hot event discovery module, we utilize K-Means algorithm to cluster problems in different customer service scenarios, by divided customer service text into K sample categories. Then, we extract the hot word combinations from the K categories separately, which can realize the

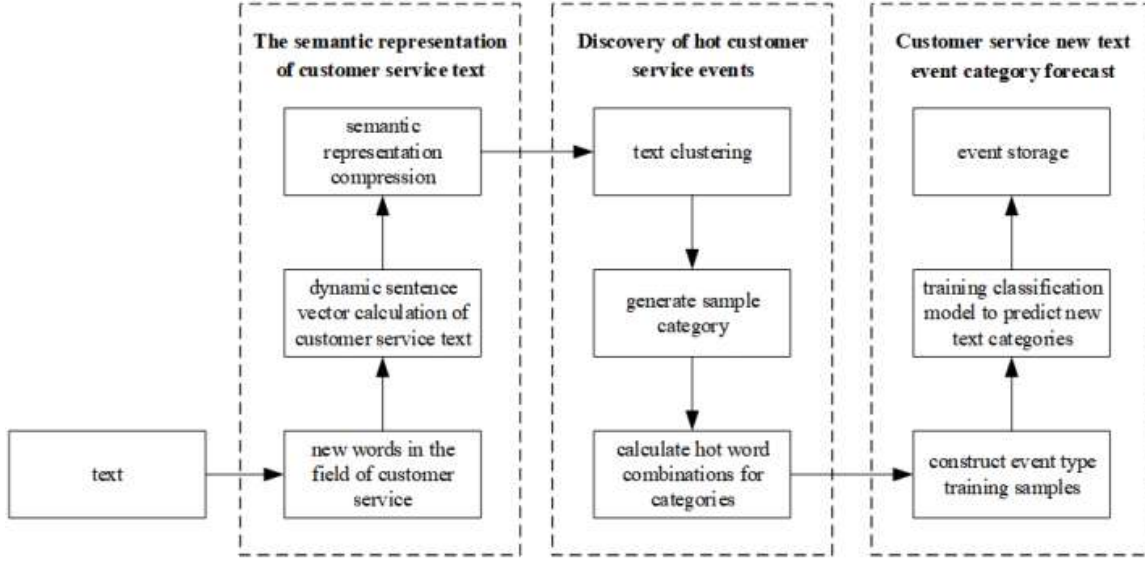


Figure 1: Overall architecture of the CSHe-D model

discovery of customer service text hot events. Thirdly, customer service new event category prediction is trained based on the supervised Fasttext, which can realize the integration of event discovery of new texts. Finally, our CSHe-D model solves the problem that it is difficult to determine the source of traffic by relying on manual statistical analysis, and optimize the accuracy of finding hot events in the field of customer service.

### 3.2 The Semantic Representation Of Customer Service

Most of the dialogue representations in customer service are based on word frequency, ignoring the length of the words themselves and the word order of the context. Hence, the obtained semantic representation of customer service texts are biased. In order to obtain rich semantic information, we propose a semantic representation of customer service text based on dynamic conversation embedding with improved TF-IDF.

**3.2.1 New Word Discovery In Customer Service.** Due to severe spoken dialogue, we preprocess the customer service text. Then, it is necessary to discover new words in the customer service field, because the current general thesaurus does not have some related terms of the operator's customer service. Specific steps are as follows:

- Step 1: The word embedding model Word2vec are used to train customer service corpus. We generate a domain basic thesaurus for customer service.
- Step 2: If the words after the text segmentation are not in domain the basic thesaurus, we perform word frequency statistics.
- Step 3: Set the new word discovery threshold  $R$  and record the words that exceed the threshold as candidate domain words.

- Step 4: Integrate candidate domain words with the domain basic thesaurus to establish a domain thesaurus in the field of customer service.

**3.2.2 Dynamic Sentence Vector Of Customer Service.** In order to consider the importance of each word and the influence of context words on the current word, we propose a dynamic sentence vector calculation for customer service. For the customer service text sequence, firstly, we use the operator customer service domain thesaurus to remove the words in the sequence  $s$  that are not related to the customer service, hence the sequence is updated to  $s' = \{c_1, \dots, c_i, \dots, c_N\}$ , where  $c_i$  means the  $i$ 'th word in the sequence,  $N$  means the text length of the new sequence. Then we obtain the initial vector of each word in the sequence  $s$  based on the Word2vec method,  $e_s = \{e_1, \dots, e_i, \dots, e_N\}$ , which uses CBOW (Continuous Bag Of Words) to map the word into a fixed-dimensional vector, which can contain the contextual semantic information.

To reflect the importance of different words, we consider that the length of words. Therefore, we adds weight of word length on the basis of TF-IDF algorithm to improve semantic representation. The specific definition of weight is:

$$w_{(c_i)} = \text{len}(c_i) * \frac{m}{T} * \log\left(\frac{H}{1+h}\right), \quad (1)$$

where  $w_{(c_i)}$  is the weight of the  $i$ 'th word,  $\text{len}(c_i)$  is the length of entry  $c_i$ ,  $m$  is the frequency of  $c_i$ ,  $T$  is the total number of entries in the text,  $H$  is the total number of texts,  $h$  is the number of texts which contains  $c_i$ . The sentence vector  $V_s$  of  $s$  is dynamically calculated with combining the vector  $e_{s'} = \{e_1, \dots, e_i, \dots, e_N\}$  of each word in the sequence  $s'$  and its corresponding word weight. It completes the vectorization of the customer service dialogue. The  $V_s$  is defined as:

$$V_s = w_{(c_1)} * e_1 + \dots + w_{(c_i)} * e_i + \dots + w_{(c_N)} * e_N. \quad (2)$$

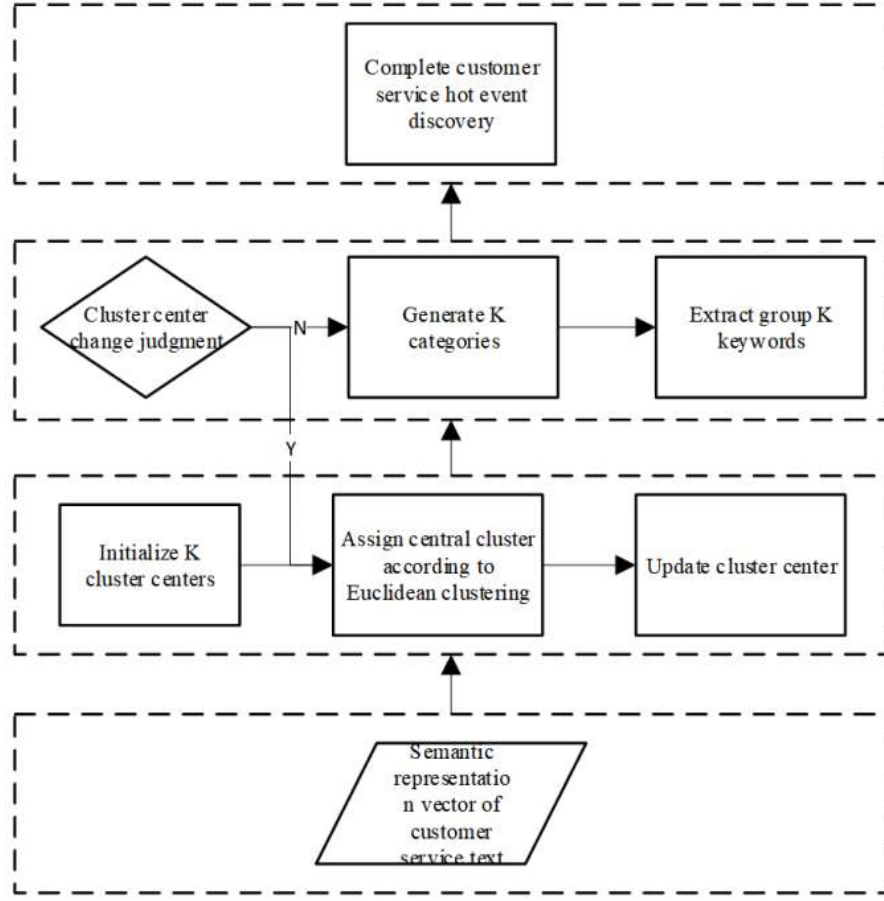


Figure 2: Customer service hot event discovery process

where  $w_{(c_i)}$  is the weight of the  $i'$ th entry, and  $e_i$  is the word vector of the  $i'$ th entry.

**3.2.3 Semantic Representation Compression.** We normalize the generated sentence vector to reduce the influence of the features with large variance in the sentence vector. This makes the features of different dimensions in the same numerical order and speeds up the convergence speed of the algorithm. Then, we adopt the PCA [1] (Principal Component Analysis) dimensionality reduction method to reduce the dimension of the original sentence vector to the specified dimension. At the same time, the loss of customer service text information is guaranteed to be minimized. The sentence vector  $V_{s'}$  after compression is defined as:

$$V_{s'} = F\left(\frac{X_i - X_{\min}}{X_{\max} - X_{\min}}, d\right), \quad (3)$$

where  $X_i$  is the value of the first dimension in the original sentence vector, and  $X_{\max}$ ,  $X_{\min}$  are the maximum and minimum values in the original sentence vector separately,  $F$  is the PCA (Principal Component Analysis) dimension reduction function, and  $d$  is the specified dimension.

### 3.3 Hot Events Discovery Based On K-Means Clustering

Since the source of customer service traffic currently mainly relies on the traffic summary of the operator, there are problems of difficult clustering and inaccurate division of various scenarios. Hence, it is difficult to find clustered hot events. In this paper, we propose a hot event discovery method of customer service based on K-Means clustering. The dialogues are clustered in different customer service scenarios with the semantic representation vector of customer service. The specific process is shown in Figure 2.

Firstly,  $K$  points are randomly selected as the initial clustering center, and the Euclidean distance is chosen as the clustering method. The data point closest to the initial clustering center is divided into the same cluster. This iteration continues, and finally the cluster center is updated according to the mean value of each cluster data point until the cluster center does not change. Then the customer service text is divided into  $K$  sample categories.

Finally, the category title is formed using hot word combinations of  $K$  categories to realize the discovery of customer service hot events.

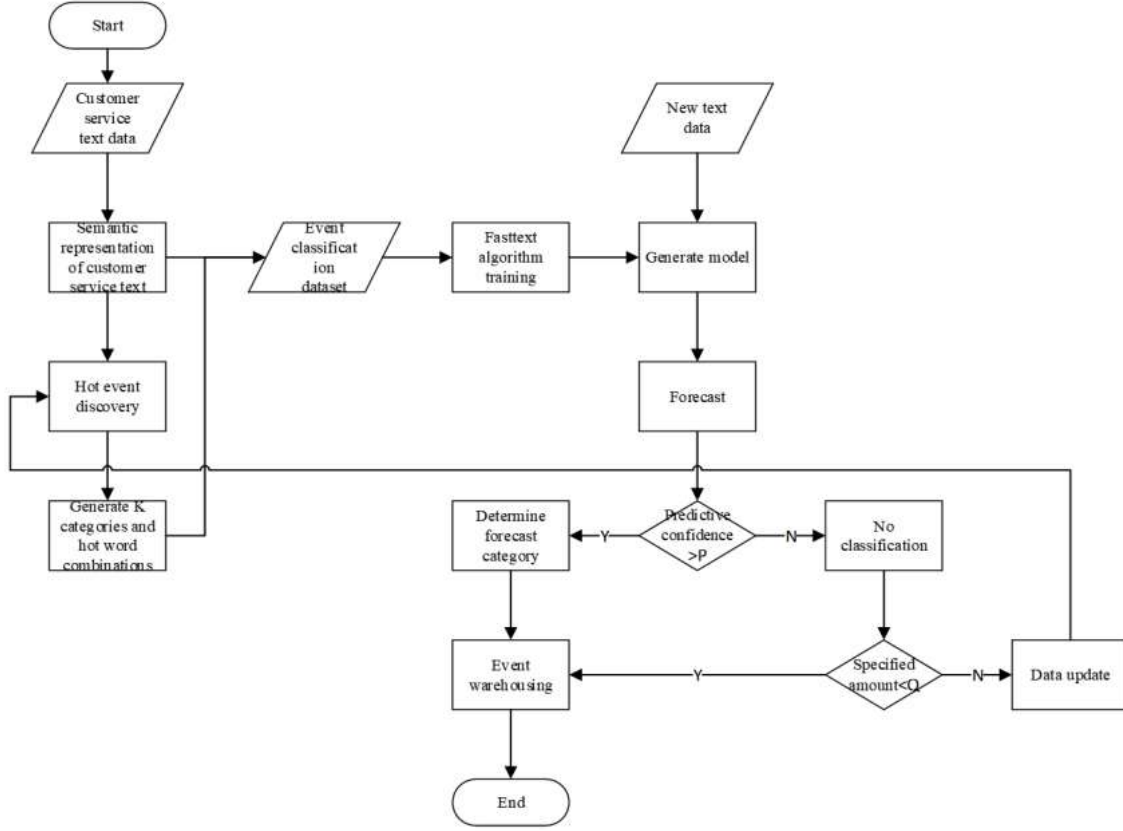


Figure 3: Forecast process of new text event category of customer service

### 3.4 Prediction Of New Hot Event

According to the  $K$  sample categories of the above events and the hot word combinations of each category, we construct the training sample to predict customer service text event type. Then, the supervised Fasttext algorithm [25] is used to train the customer service text event samples, and the text event classification  $model_f$  based on the Fasttext algorithm is obtained. For the new sequence  $x$ , we obtain the probability  $y_1, \dots, y_i, \dots, y_k$  that the sequence  $x$  belongs to  $K$  types through the trained  $model_f$ . And then, the max function are used to obtain the probability  $\hat{y}$  of a sequence  $x$  belonging to a possible class:

$$\hat{y} = \max(y_1, \dots, y_i, \dots, y_k) \quad (4)$$

Considering that the new customer service dialogue may be a new type and does not belong to the existing  $K$  categories, we set the prediction confidence  $P$  to predict the new text event category. The specific process is shown in Figure 3. If the probability  $\hat{y}$  reaches the prediction confidence, it will be classified into the similar category with the highest probability among the  $K$  categories, otherwise the item will not be classified until the number of texts meets the specified amount  $Q$ . Then all data are re-clustered to integrate subsequent events and complete event storage. It could monitor the trend of hot events in real time, and assist manual quality inspection.

## 4 EXPERIMENTAL RESULTS AND ANALYSIS

### 4.1 Experimental Setup

**4.1.1 Data Processing.** We collect dialogue data of user and customer service during 30 days. And a total of 41,607 pieces of data on the first day are used for clustering algorithm. The data involves events such as "point exchange", "call fee inquiry", "broadband consultation" and so on. In order to verify the accuracy of the algorithm, a total of 279 pieces of "broadband" data were manually marked, including 148 pieces of "broadband report" and 131 pieces of "broadband query".

**4.1.2 Parameter Setting.** In this work, we set the new word discovery threshold  $R$  to 2, and the compression vector dimension  $d$  is 20 dimensions. In order to verify the accuracy of the algorithm, the initial  $K$  value of the experiment is 2, the learning rate in the Fasttext model is 0.05, and the number of epochs is 10. Through multiple experimental adjustments, the prediction confidence is set to  $P$  is 0.8, and the specified amount  $Q$  is 100.

**4.1.3 Evaluation Index.** In order to evaluate the effect of the clustering algorithm, we adopted multiple indicators, including accuracy, mutual information score [10] and adjusted Rand coefficient [23], to measure the performance between the clustering results and the true label distribution.

**Table 1: Comparison of experimental effects of word vector models**

	Broadband failure report counts	Broadband Inquiry counts	Accuracy
BERT	109	67	0.6308
ERNIE	89	98	0.6703
Word2vec	102	104	<b>0.7384</b>

The accuracy is defined as follows:

$$Acc = \frac{TP+TN}{P+N}, \quad (5)$$

where  $Acc$  is the accuracy of the clustering evaluation index,  $P$  and  $N$  are counted as positive and negative examples,  $TP$  is the number of correctly classified as positive examples,  $TN$  is the number of correctly classified as negative examples.

The mutual information is defined as follows:

$$MI(X, Y) = \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} P(i, j) \log \left( \frac{P(i, j)}{P(i)P(j)} \right) \quad (6)$$

where  $MI(X, Y)$  is the mutual information score of the clustering evaluation index, the value is between  $[0, 1]$ , the closer to 1, the better the clustering effect is,  $X$  and  $Y$  are the clustering and the real results,  $i$  and  $j$  are the values in the set  $X$  and  $Y$  respectively,  $P(i, j)$  are the joint probability distribution of  $i$  and  $j$ ,  $P(i)$  and  $P(j)$  are the probability distribution functions of  $i$  and  $j$  respectively.

The Rand coefficient is defined as follows:

$$RI = a + b + c + d \quad (7)$$

where  $RI$  is the Rand coefficient,  $U$  and  $V$  are the real labels and clustering results,  $a$  is the number of the data points logarithms that belong to the same class in  $V$  and  $U$ ,  $b$  is the number of the data points logarithms that have same class in  $U$ , but not same in  $V$ . To the opposite, we let  $c$  is the number of the data points logarithms that are same in  $V$ , but not same in  $U$ ,  $d$  is the number of the data points logarithm which are not same in both  $U$  and  $V$ .

In order to avoid the situation that the Rand coefficient may be close to zero when the clustering results are randomly generated, we adopt the adjusted Rand coefficient, and the specific Rand coefficient  $ARI$  is defined as:

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}, \quad (8)$$

where  $ARI$  is adjust the Rand coefficient, which takes the value between  $[-1, 1]$ . The larger the value, the more consistent the clustering results are with the real situation,  $E[RI]$  is the expectation of the Rand coefficient, and  $\max(RI)$  is the maximum value of the Rand coefficient.

## 4.2 Experimental Comparison

In order to explore the effect of different word vector pre-training models, in this paper, we select BERT, ERNIE and Word2vec for experimental comparison, and verifies the effect of the word vector model based on the labeled data of the "broadband" class. The greater the degree of distinction, the better the model effect. The experimental results are shown in Table 1:

The above experimental results show that BERT and ERNIR are not ideal for the representation of long texts with less corpus, and the semantic distinction of words with the same sentence pattern

is not obvious. The Word2vec selected in this paper has better performance on small corpus, high accuracy, and can obtain similar words vector.

In this paper, hierarchical clustering, density clustering and K-Means clustering are used to train Word2vec-enhanced text semantic representation. The experimental results are shown in Table 2:

It can be seen from the Table 2 that some of the existing methods have poor results. Our CSHe-D model has the best overall effect, which is shown in the following:

The evaluation indicators of the traditional density clustering algorithm are all 0. This is because the density clustering assumes that the cluster structure can be determined by the tightness of the spatial distribution of the samples, and the algorithm itself is difficult to analyze high-dimensional data. The data is included in discrete points. K-Means clustering algorithm has an evaluation index accuracy rate of 0.6487, a mutual information score  $MI$  of 0.0660, and the adjusted Rand coefficient  $ARI$  of 0.0852, which is significantly better than density and hierarchical clustering.

After adding the improved semantic representation, the evaluation indicators of the three clustering algorithms are improved to a certain extent. The K-Means clustering effect is very significant, the mutual information score  $MI$  is 0.1614, and the adjusted Rand coefficient  $ARI$  is 0.2110. This shows that our proposed CSHe-D model for finding customer service hotspots based on the improved TF-IDF algorithm works well, and the accuracy is increased by 13.83%. And it can capture richer semantic information.

## 4.3 Case Analysis

We take the actual dialogue scene in the field of customer service as an example, and select the dialogue text instance for analysis by following the method of customer service hotspot event discovery based on our CSHe-D model.

**4.3.1 Example Of Semantic Representation Of Customer Service.** In order to verify the performance of the semantic representation of the model, we take two dialogues as examples in customer service, as shown in Table 3:

First, establish an industry business thesaurus, use Word2vec to train customer service corpus, and generate a vocabulary as the basic thesaurus of the customer service industry. On the basis of the basic thesaurus, new words in the customer service field are discovered.

Then, words such as "customer" and "agent" that are meaningless in business analysis are removed. After that, the Word2vec is used to map the word into a 100-dimensional vector, and the word vector is generated into a sentence vector by combining the maintained business words in the customer service field of the operator and

**Table 2: Comparison of experimental effects of different clustering models**

No.	Model	Acc	MI	ARI
1	Hierarchical Clustering [26]	0.5090	0.00008	-0.0033
2	Our CShe-D+Hierarchical Clustering	<b>0.5161</b>	<b>0.0037</b>	<b>-0.0026</b>
3	Density Clustering [21]	0	0	0
4	Our CShe-D+Density Clustering	<b>0.5305</b>	<b>0.0326</b>	<b>0.0030</b>
5	K-Means Clustering [16]	0.6487	0.0660	0.0852
6	Our CShe-D+K-Means Clustering	<b>0.7384</b>	<b>0.1614</b>	<b>0.2110</b>
7	Enhanced effect	<b>13.83%</b>	\	\

**Table 3: Texts of two customer service industries**

Index	Texts
text1	# Customer: Hello, I'd like to check the number of traffic packets? If it exceeds the limit, you want to add traffic packets. # Agent: OK, I'll check it first. # Customer: Yes.
text2	# Agent: There's another g, no super # Customer: Please help me check how many points I have? Can I exchange the phone bill. # Agent: OK, your current score query is 215, which can be exchanged. # Customer: OK, please exchange it for me.

**Table 4: New word weight table**

Index	Chinese word	Weight
1	Super	3.3862
2	Flow packet	6.2958
3	Integral query	6.7724
4	Convertible	8.3944
5	Add flow package	8.4655

**Table 5: Example of clustering results**

Texts	Cluster labels	Class keyword
text1	1	super, check flow, check first, month, flow
text2	0	integral, check it out, convertible, for a long time, consumption

their weight information. Through Eq.(1), the weight calculation results are shown in Table 4:

It can be seen from the Table 4 that the weight of the new word "Super" is relatively small, and there is no obvious semantic information in the field of customer service; the weight of the new word "Add flow package" is larger than that of "SFlow packet", which is mainly due to the high amount of semantic information it contains, it can better distinguish event types.

**4.3.2 Example Of Customer Service Hot Event Discovery.** Use K-Means clustering and specify the initial K parameter as 2, that is, cluster the sample texts into two categories. Then the hot word

combinations of the above two categories are extracted, namely keyword texts. These combinations can automatically generate category titles to realize hot event discovery in customer service. For example, in this paper, the customer service dialogues of the sample text involving "Flow" and "integral" are classified respectively, and the category keywords are extracted. The clustering results are shown in Table 5:

From the Table 5, the customer service texts of "Flow" and "integral" are clustered in different categories, respectively. We can interpret specific events from the categories keywords, which assist

**Table 6: Example of prediction results**

Texts	True labels	Predict labels	Probability
# Customer: I didn't run out of traffic last month. Can I use it this month? Check the flow. # Agent: Hello, no, the traffic will be cleared at the end of the month. # Customer: Oh, OK.	1	1	0.8535
# Customer: Hello, how can I exchange my points for the phone bill? # Agent: Hello, here you have 300 points to redeem. Here you can redeem them for you # Customer: OK, thanks.	0	1	0.8125

manual quality inspection. It also generates user group characteristics to predict potential users with similar characteristics and push customized service solutions to users.

**4.3.3 New Text Event Category Prediction.** The new text event category prediction can monitor the trend of hot events in real time and integrate subsequent events. Based on the customer service dialogues and their clustering labels, we construct a training sample for customer service event type prediction. The supervised Fast-text algorithm is used to train dialogues samples to obtain a event classification model. Taking new texts in the customer service as examples, the classification model prediction results is shown in Table 6.

From the Table 6, the customer service event classification model trained based on the Fasttext algorithm can well predict new customer service text events, detect the emergence of new hot events in time, and complete the process of subsequent event storage.

## 5 CONCLUSION

Massive dialogues in customer service contain rich user needs. Through data analysis and mining, it can be found that these needs are mainly concentrated on some specific hot events. In order to reduce manual participation and discover hot events, we propose a Customer Service hot event Discovery based on dynamic dialogue embedding (CSHe-D). This model evaluates the effects of hierarchical clustering, density clustering and K-Means clustering through the optimized text semantic representation in customer service. And it improves the accuracy of hotspot event discovery. As an AI scenario of intelligent customer service, this method can assist operators in predicting potential user problems before calling and identifying intentions during real-time conversations during conversations. Thereby, operators could reduce call volume, improve hotline service satisfaction, and accelerate the intelligent and digital transformation of customer service.

## REFERENCES

- [1] Abdi.H and Williams. 2010. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010), 433–459.
- [2] Akter.S, Asa.A.S, and Uddin.M.P. 2017. An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm. ICIWPR IEEE (2017), 1–6.
- [3] ALLAN.J. 2012. Introduction to topic detection and tracking. Topic detection and tracking Springer, Boston (2012), 1–16.
- [4] Devlin.J, Chang.M.W, and Lee.K. 2019. Pre-training of Deep Bidirectional Transformers for Language Understanding. Human Language Technologies 1 (2019), 4171–4186.
- [5] XU Feifei and CHEN Saihong. 2021. Summary of Research on Chinese Text Topic Clustering Algorithms. Shanghai University of Electric Power 37, 06 (2021), 613–619.
- [6] ZHANG Guofeng. 2019. Research and Implementation of Topic Hot Sorting in Article Clustering. Donghua University (2019).
- [7] Hadifar.A, Sterckx.L, and Demeester.T. 2019. A self-training approach for short text clustering. Proceedings of the 4th Workshop on Representation Learning for NLP RepL4NLP (2019), 194–199.
- [8] Haider.M.M, Hossin.M.A, and Mahi.H.R. 2020. Automatic text summarization using gensim word2vec and k-means clustering algorithm. TENSYPM IEEE (2020), 283–286.
- [9] SHAO Jie and ZHAO Qian. 2017. Hierarchical Adaptive Clustering Based Group Detection in the Crowd. Journal of Shanghai University of Electric Power 33, 1 (2017), 91–96.
- [10] CAI Jincheng and SUN Haojun. 2018. Relation Hierarchical Distance Clustering of Mixed Data Based on Mutual Information and Bayesian Networks. Journal of Shantou University 33, 02 (2018), 3–12+2.
- [11] JONES.K.S. 1972. A statistical interpretation of term specificity and its application in retrieval. Journal of documentation 28, 1 (1972), 11–21.
- [12] J.Pennington, R.Socher, and C.Manning. 2014. Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (2014), 1532–1543.
- [13] JIANG Jun, HUANG Hua, REN Tiaojuan, and ZHANG Denghui. 2019. Telecom complaint hot topic detection method based on density peaks clustering. Telecommunication Science 35, 05 (2019), 97–103.
- [14] Li.Q, Li.S, and Zhang.S. 2019. A review of text corpus-based tourism big data mining. Applied Sciences 9, 16 (2019), 3300.
- [15] L.Logeswaran and H.Lee. 2018. An efficient framework for learning sentence representations. ICLR (2018).
- [16] Macqueen.J. 1965. Some Methods for Classification and Analysis of Multi-Variate Observations. Proc of Berkeley Symposium on Mathematical Statistics Probability 344 (1965), 281–297.
- [17] Mikolov.T, Chen.K, and Corrado.G. 2013. Efficient estimation of word representations in vector space. International Conference on Learning Representations (2013).
- [18] Moradi.P, Ahmadian.S, and Akhlaghian.F. 2015. An effective trust-based recommendation method using a novel graph clustering algorithm. Statistical mechanics and its applications 436 (2015), 462–481.
- [19] Peters.M, Neumann.M, and Iyyer.M. 2018. Deep Contextualized Word Representations. Human Language Technologies 1 (2018), 2227–2237.
- [20] Radford. 2018. Improving Language Understanding by Generative Pre-Training. (2018).
- [21] Rodriguez.A and Laio.A. 2014. Clustering by Fast Search and Find of Density Peaks. Science 344 (2014), 1492–1496.
- [22] WeiBer.T, SaBmannshausen.T, and Ohrndorf.D. 2020. A cluster- ing approach for topic filtering within systematic literature reviews. MethodsX 7 (2020), 100831.
- [23] GUO Wenjuan. 2022. K-means Clustering Algorithm Based on Optimized Initial Clustering Center. Technology Wind 04 (2022), 63–65.
- [24] DAI Xiang, HUANG Xifeng, TANG Rui, JIANG Mengting, CHEN Xingshu, WANG Haizhou, and LUO Liang. 2019. Sub Topic Detection Algorithm Based on Hierarchical Clustering. Journal of South China University of Technology 47, 08 (2019), 84–95.
- [25] ZHANG Yanbo and GUO Kai. 2021. Text Classification Model Based on Fasttext and Multi-Feature Fusion. Computer simulation 38, 07 (2021), 461–466.

- [26] Zhang.T, Ramakrishnan.R, and Livny.M. 1996. An Efficient Data Clustering Method for Very Large Databases. Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data (Jun 1996), 103–114.
- [27] Zhang.Z, Han.X, and Liu.Z. 2019. Enhanced Language Representation with Informative Entities. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), 1441–1451.
- [28] Zhu.X, Li.Y, and Wang.J. 2020. Automatic Recommendation of a Distance Measure for Clustering Algorithms. ACM Transactions on Knowledge Discovery from

Data 15, 1 (2020), 1–22.

## A APPENDICES

The author(s) declare(s) that there is no conflict of interest regarding the publication of this paper.

# MergeTree: a Tree Model with Merged Nodes for Threat Induction

Ping Chen

China Electronic Product Reliability  
and Environmental Testing Research  
Institute  
Guangzhou, Guangdong, China

Jingjing Hu

China Electronic Product Reliability  
and Environmental Testing Research  
Institute  
Guangzhou, Guangdong, China

Zhitao Wu

China Electronic Product Reliability  
and Environmental Testing Research  
Institute  
Guangzhou, Guangdong, China

Ruoting Xiong

School of Computing Sciences,  
University of East Anglia  
Norwich, Norfolk, UK

Wei Ren

School of Computer Science, China  
University of Geosciences (Wuhan)  
Wuhan, Hubei, China  
weirencs@cug.edu.cn

## ABSTRACT

Threat tree model can clearly organize threat induction information and thus is widely used for risk analysis in software assurance. Threat tree will grow to complicated structures, e.g., the number of nodes and branches, when the threat information grows to a huge volume. To extend the scalability of the threat tree model, we propose a tree model with merged nodes so as to largely decrease the number of nodes and branches. The formal model and dedicated algorithms are proposed in details. The experimental results show the practicality of MergeTree. We also formally analyze the soundness and completeness of the proposed model.

## CCS CONCEPTS

• Security and privacy → Formal security models.

## KEYWORDS

Threat Tree, Semantics, Risk Analysis, Software Assurance.

### ACM Reference Format:

Ping Chen, Jingjing Hu, Zhitao Wu, Ruoting Xiong, and Wei Ren. 2023. MergeTree: a Tree Model with Merged Nodes for Threat Induction. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590012>

## 1 INTRODUCTION

Threat tree model is a modeling method for threat representation and risks, which is widely used in software assurance [1]. The conditions for a threat are organized as a branch in the tree, which consists of several brothers with logical relations either “AND” or “OR” and a father who is the result of the conditions. This branch representation is iterative so that can represent complicated threat conditions like a tree [2].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590012>

The tree model presents simplicity in terms of semantics as threat information is always provided as a list of inductions, each of which is composed of certain conditions and one conclusion. Thus, tree model can smoothly represent the induction list, and afterward tree structure can be operated by available tree algorithms.

With the growing volume of induction list, however, tree structure will be complicated to due the increasing of the number of nodes [3, 4]. Simply speaking, the number of nodes is proportional to the number of inductions. Hence, how to decrease the number of nodes yet maintaining the semantics of threat tree is of great importance.

The problem seems to be straightforward, intuitively, because the number of nodes can only be decreased by merging. However, if the merge is not proper, the overhead of merging will be great. The challenge of the problem is that merging must not damage the semantics of the threat and merging should maintain the tree structure, so that legacy algorithms for tree processing can be remained.

The contribution of the paper is as follows:

- (1) We formally model the semantics of the threat tree with merged nodes.
- (2) We propose several algorithms for threat tree construction from the threat database.
- (3) We formally prove the equivalence in terms of semantics between induction set and threat tree.

The rest of the paper is organized as follows: Related work is reviewed in Section 2. Section 3 describes the proposed model. We propose key algorithms for threat tree in Section 4. The experimental results are shown in 5. Then analysis is given in 6. Finally, Section 7 concludes the paper.

## 2 RELATED WORK

Attack tree or their variation trees are good tools for security assessment and threat defense. Many scholars also focus on the studies for attack tree synthesis and refinement [5]. Attack trees have been applied in many fields, such as e-mail systems [6] and Cyber Physical Systems (CPS) security [7]. Other applications can be seen in [8–10]. Generally, the studies for attack trees divide into these three categories: security assessment and threat defense, tree synthesis, and large-scale nodes of attack tree management.

- (1) Attack trees for security assessment and threat defense

The main functions of attack trees can be security assessment and threat defense. For example, ApproxTree+ [11] analysis tool is proposed for attacker profiling and enhanced it by incorporating the attacker's capabilities into it. Wouter et al. [12] introduced a methodology to evaluate the security of CSP system, which generates attack trees based on the system architecture automatically. The generated attack trees can provide both technical and non-technical feedback. Nishihara et al. [13] focused on refinement scenarios for attack trees which enables the assessment of the validity of attack decomposition systematically. That is, the attributes that sub-attacks refine an attack are described by the relationship among their effects. The proposed ideas are applied to the case study of a vehicular network system that is well-behaved.

#### (2) Attack tree synthesis

Sophie et al. put forward ATSyRa [14], which provides a user-friendly environment for attack tree synthesis. In the ATSyRa system, users can define a structured attack tree by high-level description, and refine the synthesis interactively with the system. The relations of the nodes can be three types, namely AND, OR, and SAND. Nonetheless, the refinement analysis is still done by humans, not automatically and intelligently compared with [16]. In 2020, a Library-Based Attack Tree Synthesis scheme is proposed [17], where the inputs of the system are a library and a trace. Whereas, the proposed algorithm is only polynomial in the size of the trace. Olga et al. [15] proposed a refinement-aware method for attack tree generation by labeling technique. It solves the problem of constructing a correct attack tree while maintaining a predefined refinement relationship. Nevertheless, it has not addressed the challenge of the growing volume of nodes.

#### (3) Large-scale nodes of attack tree management

With the increasing number of attacks and the complex operations between the nodes, the management of attack trees with complex structures has become a challenge [18]. Paul et al. [19] suggested that traditional risk assessment schemes are now reaching their limit and they proposed graphical extensions to deal with scalability issues, like chain diagrams. The method has been applied in the Galileo risk management program and the results show that the system can deal with both software-intensive situations and a large number of small problems. Fila et al. [20] focused on finding the best series of Attack Defense Trees (ADTrees) from the directed acyclic graphs. Experiment results show that the countermeasures can block numerical ways of attacking and a wide class of optimization problems can be solved despite the growing volume of nodes. Vigo et al. [21] put forward a static analysis method implemented by Java to avoid the exponential explosion of analysis for attack trees, where they are automatically deduced from an algebraic specification in a syntax-directed manner. The study of a national-scale authentication system has proved the flexibility and effectiveness of the scheme.

As we can see, there is no research on the induction method to manage the attack trees with large-scale nodes. Therefore, it is of great importance to reduce the complexity of the tree model when the threat information grows to a huge volume while maintaining the semantics.

## 3 PROPOSED MODEL

### 3.1 Induction Model

To explore the atom induction, it can be classified into four types as follows:

$F1 : A_1 \wedge A_2 \Rightarrow B$ . We denote it as  $A_1, A_2 \Rightarrow B$ .

$F2 : A_1 \vee A_2 \Rightarrow B$ . We rephrase it as two inductions:  $A_1 \Rightarrow B$ ;  $A_2 \Rightarrow B$ .

$F3 : (A_1 \wedge A_2) \vee A_3 \Rightarrow B$ . It can be denoted as two inductions:  $A_1, A_2 \Rightarrow B$ ;  $A_3 \Rightarrow B$ .

$F4 : (A_1 \vee A_2) \wedge A_3 \Rightarrow B$ . It can be denoted as two inductions:  $A_1, A_3 \Rightarrow B$ ;  $A_2, A_3 \Rightarrow B$ .

**DEFINITION 3.1.** *Simplex Form.* Induction  $A_1, \dots, A_n \Rightarrow B$  is called *simplex form* where all conditions (namely,  $A_i$ ,  $i = 1, \dots, n$ ) have one relation "AND" for a conclusion (namely,  $B$ ).

**PROPOSITION 3.2.** *All inductions can be regulated into simplex form.*

**PROOF.** Given any inductions, a method can change it into simplex form as follows:

(1) Split all  $\vee$  at the outlier layer by  $F2$  and  $F3$ , and then obtain an induction set.

(2) For any induction in the set, it must be in the form either  $F1$  or  $F4$ . Both can be changed into simplex form.

(3) If any induction is not simplex form, go to (1).

Later, we will propose a tuple model for representing a simplex form, and further a tree model for representing all simplex form.

### 3.2 Tree Model

**DEFINITION 3.3.** *SimplexInduction* ::=  $\langle \{condition \in Label\}, conclusion \in Label \rangle$ , where *condition* are one label or multiple labels; *conclusion* is one label; *Label* is a set of strings in context, e.g.,  $A_1, \dots, A_n, B$ .

For example, simplex induction for  $A_1 \Rightarrow B$  is  $\langle \{A_1\}, B \rangle$ , induction for  $A_1, A_2 \Rightarrow B$  is  $\langle \{A_1, A_2\}, B \rangle$ .

As an assumption, we suppose  $\forall \langle condition, conclusion \rangle \in SimplexInduction$ ,  $condition \times conclusion$  is a one-to-one mapping and onto.

**DEFINITION 3.4.** *NODE* ::=  $\langle \{label \in Label\}, to \in Label \rangle$ , where *label* denotes one condition or multiple conditions; *to* denotes a conclusion.

For example, the node for  $A_1 \Rightarrow B$  is  $\langle \{A_1\}, B \rangle$ , induction for  $A_1, A_2 \Rightarrow B$  is  $\langle \{A_1, A_2\}, B \rangle$ . For visualization, *label* can be looked as a node, and *to* can be looked as an edge that starts from the node and with a label *to* at the end.

**PROPOSITION 3.5.** *SimplexInduction is equivalent to Node.*

**PROOF.** Straightforward. The definition is identical for the type of the components, although the names of the components are distinct.

**DEFINITION 3.6.** *EDGE* ::=  $\langle from \in NODE, to \in NODE \rangle$ .

**PROPOSITION 3.7.** *If  $\exists node_a, node_b \in NODE$ ,  $node_a.to \in node_b.label$ , then  $\exists edge \in EDGE$ ,  $edge.from = node_a$  and  $edge.to = node_b$ .*

PROOF. Straightforward.

DEFINITION 3.8.  $ROOT ::= \{node | node \in NODE, \nexists m \in EDGE, m.from = node\}$ .

DEFINITION 3.9.  $LEAF ::= \{node | node \in NODE, \nexists m \in EDGE, m.to = node\}$ .

DEFINITION 3.10.  $Tree ::= \langle \{node \in NODE\}, \{edge \in EDGE\} \rangle$ .

For example, suppose an induction set consisting of simplex forms is as follows:

- (1)  $A_1 \wedge A_2 \wedge A_3 \Rightarrow B_1$ ;
- (2)  $A_2 \wedge A_3 \Rightarrow B_2$ ;
- (3)  $A_3 \wedge A_4 \Rightarrow B_4$ ;
- (4)  $B_1 \wedge B_2 \wedge B_3 \Rightarrow Root$ ;
- (5)  $B_3 \wedge B_4 \wedge B_5 \Rightarrow Root$ .

Above induction set can be converted into a threat tree with merged nodes in Fig. 1. Note that, at the end of the edge there exists a label to denote the conclusion. Besides, if the union of edge labels is not equal to the label of the “father” node, that “father” will not be reached indeed. In that words, the threat path to the root is broken at this “father”.

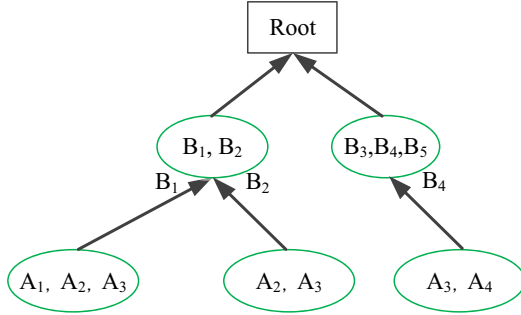


Figure 1: A threat tree with merged nodes.

## 4 PROPOSED ALGORITHMS

PROPOSITION 4.1. *There exists an algorithm that can convert a simplex form to a threat branch.*

PROOF. Given any induction (e.g.,  $I$ ) in the set, if there does not exist any node whose label is equal to the condition of the induction (i.e.,  $\nexists n \in NODE, n.label = I.condition$ ), then create a node  $node$  and set the label of node as that condition (i.e.,  $node.label = I.condition$ ). Create an edge  $e$  who starts from the node ( $e.from = node$ ). The conclusion of the induction is denoted it at the end of the edge (i.e.,  $node.to = conclusion$ ). If there exists a node (e.g.,  $m$ ) whose label includes the label of this edge (i.e.,  $node.to \in m.label$ ), then set the edge point to this node (i.e.,  $e.to = m$ ).

PROPOSITION 4.2. *There exists an algorithm that can convert an induction set with simplex forms into a threat tree with merged nodes.*

PROOF. It is accomplished by conducting all inductions in the set, due to Proposition 4.1.

The algorithm that can convert an induction set with simplex forms into a tree with merged nodes, can be proposed as follows (see Algorithm 1):

```

Data: A set of SimplexInduction - set.
Result: A threat tree T.
while (set! = Null) do
  Fetch a SimplexInduction  $\in$  set;
  if ( $\forall node \in T$ ,
    SimplexInduction.condition! = node.label) then
    Add node to T;
    node.label  $\leftarrow$  condition, node.to  $\leftarrow$  conclusion;
    Add edge to T;
    edge.from  $\leftarrow$  node;
    Find n  $\in T$  such that conclusion  $\in$  n.label;
    edge.to  $\leftarrow$  n;
  end
  Delete SimplexInduction from set;
end
return 1;

```

Algorithm 1: Threat tree construction algorithm.

PROPOSITION 4.3. *There exists an algorithm that can convert a threat branch into a simplex form.*

PROOF. Given a branch, e.g.,  $node_1$  and  $node_2$  where  $edge.from = node_1$  and  $edge.to = node_2$ . The simplex form is as follows:  $node_1.label \Rightarrow node_1.to \in node_2.label$ .

PROPOSITION 4.4. *There exists an algorithm that can convert a threat tree with merged nodes into an induction set with simplex forms*

PROOF. It is accomplished by conducting all tree branches into inductions, due to Proposition 4.3.

The algorithm that can convert a tree with merged nodes into an induction set with simplex forms, can be proposed as follows (see Algorithm 2):

```

Data: A threat tree T.
Result: A set of SimplexInduction - set.
while (T! = Null) do
  Fetch a node  $\in T$ ;
  Find edge  $\in T$  such that edge.from = node;
  if ( $\forall SimplexInduction \in set$ ,
    SimplexInduction.condition! = node.label) then
    | Add node.label  $\Rightarrow$  node.to to set.
  end
  Delete node and edge from T;
end
return 1;

```

Algorithm 2: Induction construction algorithm.

## 5 PERFORMANCE EVALUATION

**Time consumption.** We set up an experiment to calculate the time consumption of these two functions: induction sets convert to threat tree, and threat tree convert to induction sets. We quantitative measure the benefits of the proposed solution. In our experiment, we set the number of induction sets are 50, 100, and 150. The results are shown in Fig. 2.

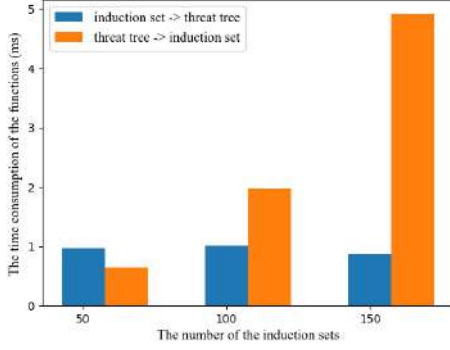


Figure 2: The time consumption of the algorithms.

As we can see, when the number of induction sets are 50, the time of converting them to threat tree is 0.96ms, while threat tree to induction set is 0.65ms. When the number of induction set reaches to 100, the time of these two functions are 0.997ms and 1.993ms, separately. When the number of induction set is 150, which means the tree is large enough, the time of these two functions are only 1.06ms and 4.92ms, respectively. Therefore, the time consumptions of the algorithms are within 'ms' level, which means MergeTree is practical to merge the nodes to build a threat tree from induction sets.

Besides, we calculate the edge decrease ratio for the MergeTree. We find that when the average number of nodes merged in one Node is 2 to 3, the edge decrease ratio is about 50% to 60%. When the number of nodes that merge in one Node increases, the edge decrease ratio is going up, which means the MergeTree can load a large induction set with low edge degree. To notice, the MergeTree can build the tree of induction sets without much limit on node and edge degree, paving a practical and flexible way for building attack trees.

**Algorithm complexity analysis.** Given that the total number of nodes in MergeTree is  $n$ , and the number of edge is  $m$ . We can see that the complexity of MergeTree construction is  $O(m)$ . When we leverage the MergeTree for attack analysis, the complexity is  $O(\log_m)$ . Usually,  $m$  is smaller than  $n$ , and it is proportional to  $n$ , which depends on the number of nodes in one Node. When compared with other methods, they do not list the complexity of the algorithm. Most of them focus on risk assessment. However, we focus on the tree construction algorithm and we prove that our MergeTree construction algorithm is applicable in terms of complexity and time consumption.

## 6 ANALYSIS

**PROPOSITION 6.1.** *The threat tree with merged nodes is equivalent to a set of simplex induction in terms of semantics.*

**PROOF.** It is due to Proposition 4.2 and Proposition 4.4. We need to prove that the induction set can be transformed into threat tree. In the threat tree, there are multiple nodes (node.label, node.to) connected by edges (edge.from, edge.to). Given a simplex induction set, the context is a list of  $\langle \text{conditions}, \text{conclusion} \rangle$ . Each condition and conclusion can be considered as nodes and they are connected by edges. That is, the value of node.label and edge.from is condition, and the value of node.to and edge.to is conclusion. Besides, we need to prove that the threat tree can be converted to simplex induction set. Obviously, in a threat tree, each edge is linked with two nodes, which represent the edge.from and edge.to. In simplex induction set, the conditions are edge.from and the conclusions are edge.to.

**REMARK 6.2.** *A simplex induction is equivalent to a node and an edge indeed. The visualization in the tree provides a better understanding for induction.*

*A tree model will facilitate the manipulation of threat information processing, e.g., searching a threat route, detecting the existence of a threat, listing possible threat methods, and so on, by off-the-shelf tree algorithms. That is the reason of tree representation of threat induction information.*

*The simplex form simplifies the structure of induction, so as to simplify the structure of threat tree, with manageable overhead in the number of induction numbers.*

## 7 CONCLUSION

In this paper, we study how to decrease the number of nodes (also branches) for threat model when the number of threat information keeps on increasing. The threat model with merged nodes is proposed and respective critical algorithms are provided. The presentation is formal, e.g., the equivalence between simplex induction and tree branch, as well as between threat tree and induction set, so as to derive the respective key conversion algorithms. The experiment results and analysis justifies the soundness and completeness of the proposed model. Besides, the tree structure has almost remained so that original tree algorithms for threat analysis can still work.

## ACKNOWLEDGMENTS

The research was financially supported by the Science and Technology Program of Guangzhou, China (No. 202102021216).

## REFERENCES

- [1] B. Schneier, "Attack Trees: Modeling Security Threats", *Dr.Dobbs Journal*, vol.24, no.12, pp. 21-29, 1999.
- [2] A.T. Ali, D.P. Gruska, "Attack Trees with Time Constraints", in *Proc. of the 28th International Workshop on Concurrency, Specification and Programming (CS&P2021)*, 2021, pp. 27-28.
- [3] Asif, Waqar, Indranil Ghosh Ray, and Muttukrishnan Rajarajan. "An attack tree based risk evaluation approach for the internet of things," in *Proc. of the 8th International Conference on the Internet of Things*, 2018, pp. 1-8.
- [4] Schiele, Nathan Daniel, and Olga Gadyatskaya. "A Novel Approach for Attack Tree to Attack Graph Transformation," *International Conference on Risks and Security of Internet and Systems*. Springer, 2022, pp. 1-8.

- [5] H. Mantel, C. W. Probst, "On the Meaning and Purpose of Attack Trees", in *Proc. of 2019 IEEE 32nd Computer Security Foundations Symposium (CSF2019)*, 2019, pp. 184-18415.
- [6] Scala, Natalie M., et al. "Evaluating mail-based security for electoral processes using attack trees." *Risk Analysis* (2022).
- [7] Ji, Xiang, et al. "Attack-defense trees based cyber security analysis for CPSs." in *Proc. of 2016 17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, 2016, pp. 693-698.
- [8] Kammüller, Florian. "Attack trees in Isabelle extended with probabilities for quantum cryptography." *Computers and Security*, vol. 87, pp: 101572, 2019.
- [9] Sen He, Wei Ren, Tianqing Zhu and Kim-Kwang Raymond Choo, BoSMoS, "A Blockchain-Based Status Monitoring System for Defending Against Unauthorized Software Updating Industrial Internet of Things", *IEEE Internet of Things Journal*, IEEE, pp. 948-959, 2020.
- [10] A.T. Ali, D.P. Gruska, "Attack Protection Tree", in *Proc. of the 28th International Workshop on Concurrency, Specification and Programming (CS&P2019)*, 2019, pp. 1-6.
- [11] Lenin, A., Willemson, J., Sari, D.P., "Attacker Profiling in Quantitative Security Assessment Based on Attack Trees", Bernsmed, K., Fischer-HÄEbner, S. (eds) *Secure IT Systems. NordSec 2014. Lecture Notes in Computer Science()*, vol 8788, pp. 199-212, 2014.
- [12] Depamelaere, Wouter, et al. "CPS security assessment using automatically generated attack trees," in *Proc. of the 5th international symposium for ICS & SCADA cyber security research 2018. British Computer Society (BCS)*, 2018, pp. 1-10.
- [13] Nishihara, Hideaki, et al. "On Validating Attack Trees with Attack Effects: An Approach from Barwise-Seligman's Channel Theory," arXiv preprint arXiv:2204.06223 (2022).
- [14] Pinchinat, Sophie, Mathieu Acher, and Didier Vojtisek. "ATSyRa: an integrated environment for synthesizing attack trees," *International Workshop on Graphical Models for Security*, 2015, pp. 97-101.
- [15] Gadyatskaya, Olga, et al. "Refinement-aware generation of attack trees." in *Proc. of International Workshop on Security and Trust Management*, 2017, pp. 164-179.
- [16] Ali, Aliyu Tanko, and Damas Gruska. "Dynamic Attack Trees Methodology," in *Proc. of 2022 Interdisciplinary Research in Technology and Management (IRTM)*, 2022, pp. 1-9.
- [17] Pinchinat, Sophie, Francois Schwarzentruher, and Sebastien Le Cong. "Library-Based Attack Tree Synthesis," in *Proc. of International Workshop on Graphical Models for Security*, 2020, pp. 24-44.
- [18] Yaocheng Zhang, Wei Ren, Tianqing Zhu, Yi Ren, SaaS, A Situational Awareness and Analysis System for Massive Android Malware Detection, Future Generation Computer Systems, Volume 95, June 2019, 548-559.
- [19] Paul, Stephane, and Raphael Vignon-Davillier. "Unifying traditional risk assessment approaches with attack trees," *Journal of Information Security and Applications*, vol. 19, no. 3, pp. 165-181, 2014.
- [20] Fila, Barbara, and Wojciech Wide. "Exploiting attack defense trees to find an optimal set of countermeasures," in *Proc. of 2020 IEEE 33rd Computer Security Foundations Symposium (CSF)*, 2020, pp. 395-410.
- [21] Vigo, Roberto, Flemming Nielson, and Hanne Riis Nielson. "Automated generation of attack trees." in *Proc. of 2014 IEEE 27th computer security foundations symposium*, 2014, pp. 337-350.

# Heart Sound Classification Algorithm Based on Sub-band Statistics and Time-frequency Fusion Features

Xiaoqing, Zhang\*

Department of Information, Yunnan University  
2662013019@qq.com

Weilian, Wang

Department of Information, Yunnan University  
wlwang\_47@126.com

## ABSTRACT

The clinically acquired heart sound signals always have inevitable noise, and the statistical features of these noises are different from heart sounds, so a heart sound classification algorithm based on sub-band statistics and time-frequency fusion features is proposed. Firstly, the statistical moments (mean, variance, skewness and kurtosis), normalized correlation coefficients between sub-band and sub-band modulation spectrum are extracted from each sub-band envelope of the heart sound signal, and these three features are fused into fusion features by Z-score normalization method. Finally, a convolutional neural network classification model is constructed, which are used for training and testing. The experimental results showed that the accuracy, sensitivity, specificity and F1 score of the algorithm were 95.12%, 92.27%, 97.93% and 94.95%, respectively. It has great potential in machine-aided diagnosis of precordial diseases.

## CCS CONCEPTS

• Theory of computation; • Theory and algorithms for application domains; • Machine learning theory;

## KEYWORDS

Heart sounds, Congenital heart disease, Fusion features, Mel filter banks, Z-score normalization, Convolutional neural networks

## ACM Reference Format:

Xiaoqing, Zhang and Weilian, Wang. 2023. Heart Sound Classification Algorithm Based on Sub-band Statistics and Time-frequency Fusion Features. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590013>

\*Supported by the National Natural Science Foundation of China (81960067); Major Science and Technology Special Project of Yunnan Province (2018ZF017). Corresponding author: Xiao-qing, Zhang(1996-), F, Yunnan, China, M.S. student, mainly researching on biomedical signal processing. Email: 2662013019 @qq.com. Author: Weilian Wang(1947-), Male, Yunnan, China, Professor, Main research on signal processing and pattern recognition, biomedical signal processing, digital-analog hybrid IC and ASIC design. Email: wlwang\_47@126.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590013>

## 1 INTRODUCTION

“The automatic analysis and processing of heart sounds consists of three main steps: pre-processing, feature extraction and classification recognition.” [1]. “Since the heart sound signal is inevitably disturbed by environmental noise, noise from the acquisition equipment and noise from other human organs during the acquisition process.” [2]. Therefore, a large number of researchers perform noise reduction and segmentation on the heart sound signal in the pre-processing stage. “In conventional heart sound classification, pre-processing requires synchronized electrocardiogram (ECG) to segment the phonocardiogram (PCG).” [3]. Nogueira et al. [4] after segmentation of the PCG, time series of single cardiac cycles and Mel-frequency cepstral coefficients (MFCC) features were extracted for classification with a sensitivity of 0.8737, specificity of 0.7907 and accuracy of 0.8322. Abduh et al. [5] extracted segmented heart sounds based on the features of log Mel-frequency spectral coefficients (MFSC) of fractional Fourier transform for classification with a sensitivity of 0.8735, specificity of 0.9666, and accuracy of 0.9200. Although the accuracy obtained by this method is high, the difference between sensitivity and specificity is too large to be applied to large area screening.

Since simultaneous acquisition of PCG and ECG is inconvenient in CHD screening, a number of studies have been conducted in recent years to segment PCG directly without the help of ECG. Yin et al. [6] proposed an algorithm for direct heart sound segmentation by combining the hidden Markov model and temporal convolutional network, and tested 403 heart sound samples with an accuracy of 91.64% and an F1 score of 97.02%. However, the number of tested samples is small and the generalization ability of the method needs to be verified. Since the segmentation effect directly affects the effectiveness of feature extraction, Wang et al. [7] proposed a heart sound classification algorithm based on sub-band envelope features and convolutional neural network without relying on accurate segmentation, which was tested on 1000 heart sound sample data with sensitivity, specificity, accuracy and F1 score of 0.938, 0.940, 0.939 and 0.939, respectively. But the method has high complexity and is not suitable for large area application.

Feature extraction is the key to heart sound classification. Recent researchers have achieved good results in extracting heart sound features in time domain, frequency domain and time-frequency domain. Karan et al. [8] extracted wavelet packet energy-based features from PCG signals for classification and finally achieved 84.88% accuracy on the public dataset. Xu et al. [9] extracted a total of 84 features containing time and frequency domains from the segmented PCG, and achieved accuracy, sensitivity, specificity and F1 scores of 0.953, 0.946, 0.961 and 0.953, respectively, on a self-developed database of 941 CHD heart sounds in children. Although better results were achieved, it still relies on accurate segmentation, and the

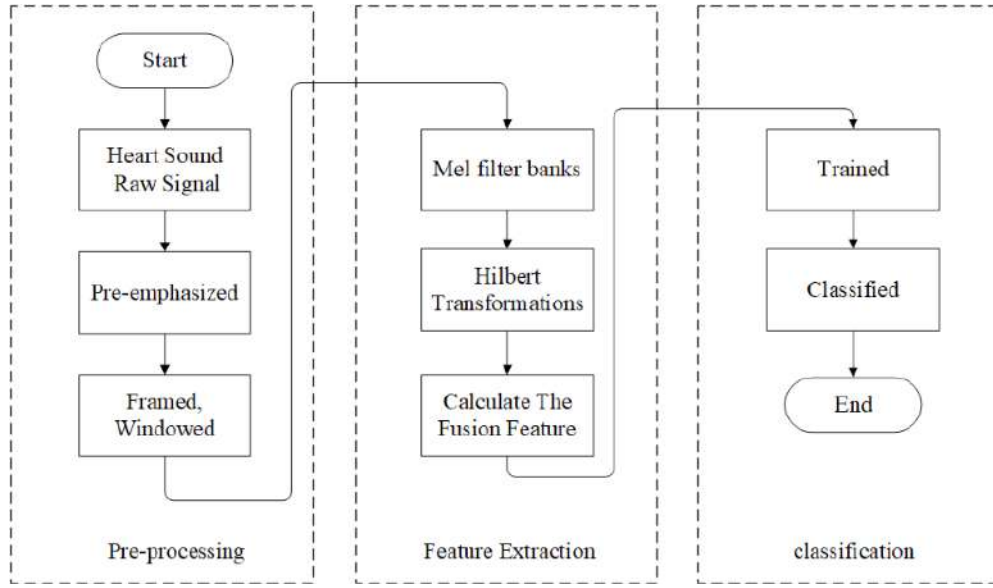


Figure 1: Algorithm flow chart

study was based on a small sample dataset, and its generalizability is difficult to guarantee.

“Considering that the statistical characteristics of the heart sound signal are always different from the background noise.”[10] Therefore, this paper proposes a heart sound classification algorithm based on sub-band statistics and time-frequency fusion features. This method can effectively explore the depth features of heart sounds without noise reduction and segmentation, which simplifies the algorithm steps and effectively avoids the removal of pathological features due to noise reduction.

## 2 METHODOLOGY

### 2.1 Dataset

Two heart sound datasets were used in this study. One dataset was the “Precocious Heart Sound Sample Dataset” (the subject dataset), which was jointly collected by Yunnan University and Affiliated Cardiovascular Hospital of Kunming Medical University, and was collected from clinical patients at Affiliated Cardiovascular Hospital of Kunming Medical University and children who were screened for precocious heart disease in various states of Yunnan Province. All signals were confirmed by clinical experts and diagnosed using echocardiography. A total of 5000 heart sound samples were used in this study, including 2500 normal heart sound samples and 2500 abnormal (with precordial disease) heart sound samples in the subject dataset. The other dataset used was the “Heart Sound Challenge PhysioNet/CinC 2016 public dataset.”[11], which has a total of 3240 heart sound samples.

### 2.2 Models and Methods

The flow of the proposed heart sound classification algorithm for precordial disease based on sub-band statistics and time-frequency fusion features is shown in Figure 1.

**2.2.1 Pre-processing.** Ambient sounds, lung sounds and cough sounds are the main interfering factors in the recording and analysis of the heart sound signal. The main purpose of preprocessing is to enhance the useful information of the heart sound signal. In this study, 5s of each heart sound signal was randomly intercepted as the heart sound sample to be analyzed, and it was pre-emphasized, framed, windowed, and normalized. The heart sound signal is a typical non-stationary signal, which can be regarded as locally stationary when the window is small enough. In this study, a fixed frame length of 0.1 s and a step shift of 0.05 s were used, and a Hamming window was added to the signal to reduce frequency leakage and the effect of partials.

**2.2.2 Feature Extraction.** The heart sound signal is a low-frequency signal, and “the Mel filter set has the characteristics of dense low frequency and sparse high frequency.”[12] The pre-processed signal is passed through the Mel filter set to capture as much information as possible at low frequencies. As the signal undergoes Fourier transform, there are positive and negative frequencies in the frequency spectrum. “Hilbert transform of the filtered signal can play a role in eliminating redundant frequencies.”[13]

“Sub-band envelopes differ in the presence of width, mean and long tail. These properties can be captured by statistical moments (mean, variance, skewness and kurtosis, respectively).”[14] “The envelope energy of the heart tone signal will be concentrated in the low frequency band, while the energy of the noise envelope is mainly uniformly distributed.”[15] “The modulated spectral bands can therefore be used as one of the features. Considering that there is an obvious correlation between the coefficients of the corresponding spatial positions of each sub-band.”[16] Therefore, the correlation is also used as one of the features. As shown in Figure 2. The specific feature extraction steps are as follows.

(1) The pre-processed one-dimensional heart sound signal is decomposed by a Mel-scale auditory filter set.

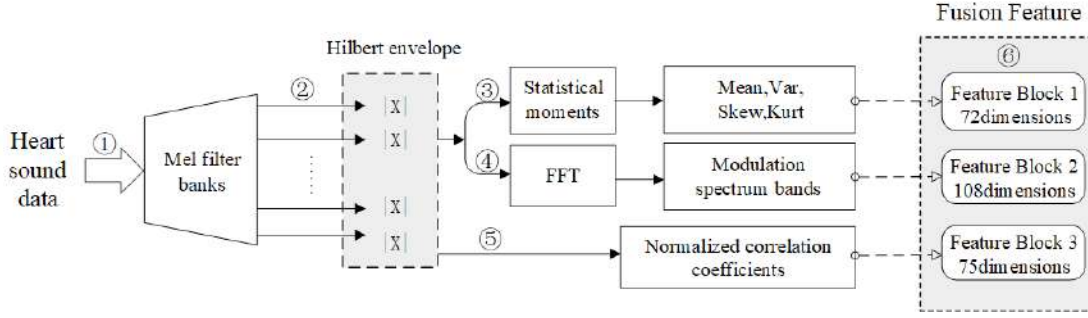


Figure 2: Algorithm flow chart

(2) Calculate the Hilbert envelope of the sub-band signal.

(3) Extracting feature block one of the sub-band envelope. Four statistical moments of mean, variance, skewness and kurtosis are obtained for the sub-band envelope to obtain a 72 (18×4) dimensional feature block. The formulas for calculating the mean and variance are shown in equations 1) and (2).

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (2)$$

Where  $X_i$  denotes the first  $i$  subband envelope signal,  $\bar{X}$  denotes the mean of the subband envelope signal, and  $S^2$  denotes the variance of the subband envelope signal. The variance statistics can represent the sparsity of the subband envelope signal well, but the variance contains limited information, so this study incorporates the higher order statistics (skewness and kurtosis) into the characteristic statistics in conjunction with the characteristics of the filter. The formulas are shown in equations 3) and (4).

$$Skew(X) = E \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] \quad (3)$$

$$Kurt(X) = E \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] \quad (4)$$

Where,  $X$  denotes the subband envelope signal,  $\mu$  denotes the mean of the subband envelope signal,  $\sigma$  denotes the standard deviation of the sub-band envelope signal,  $E[\cdot]$  denotes the expectation operator, and  $Skew(X)$  and  $Kurt(X)$  denote the skewness and kurtosis of the subband envelope signal, respectively.

(4) Extraction of feature block two of the sub-band envelope. Fast Fourier transform is performed on the sub-band envelope signal, and its spectrum is divided into 6 spectral bands according to the corresponding digital corner frequencies, and each spectral band is normalized by its sub-band variance to obtain 6 modulated spectral bands, and finally a 108 (18×6) dimensional feature block is obtained.

(5) Extracting the feature block III of the sub-band envelope. This study reflects the characteristics of this relationship by calculating the Pearson correlation coefficient matrix (18 × 18 matrix) between

the sub-band signals and grabbing the diagonal from it. The correlation coefficient matrix is calculated as shown in equation 5).

$$R(E, Y) = \frac{Cov(X, Y)}{\sqrt{Var|X| Var|Y|}} \quad (5)$$

Where,  $Cov(X, Y)$  is the covariance of the sub-band signal  $X$  and the sub-band signal  $Y$ ,  $Var|X|$  is the variance of  $X$ , and  $Var|Y|$  is the variance of  $Y$ . Since the coefficients on the main diagonal of the correlation matrix are all correlation coefficients of two identical variables, the correlation between the sub-band signals is characterized by ignoring the main diagonal when capturing the diagonal and capturing the longest 5 diagonals to form a 75 (17+16+15+14+13=75) dimensional feature block.

(6) The 3 feature blocks are fused into one-dimensional fused features. Considering the influence of the difference in the magnitude of different feature blocks on the sample distribution, the Z-score normalization method is used to normalize the 3 feature data, which can achieve the elimination of the magnitude of the three feature blocks while reducing the error and preserving the distribution of the original feature data.

**2.2.3 Classifier.** Since the fused features extracted in this paper are of low complexity and the use of a higher complexity model can easily lead to overfitting phenomena, this study improves on the traditional one-dimensional CNN, and the specific model structure is shown in Figure 3.

Compared with the traditional CNN model, the model has made the following main improvements.

The first convolutional layer of the model adopts a large size 64×1 convolutional kernel, which can increase the perceptual field of the model input and obtain more effective information for the subsequent layers of the network. At the same time, the second convolutional layer adopts a 2×1 small-sized convolutional kernel, which can compress the feature channels and achieve cross-channel feature fusion.

Replacing the dense fully connected layers (and spreading layers) with GAP layer can reduce the model parameters exponentially and effectively prevent the overfitting phenomenon.

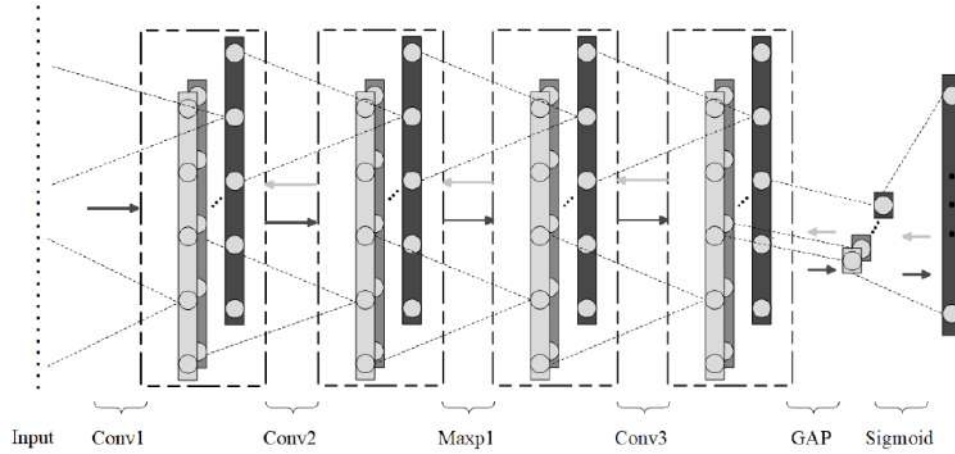


Figure 3: Structure of the improved CNN-1D model

Table 1: Classification results of different feature extraction methods

Analytical method	Evaluation Criteria			
	Accuracy/%	Sensitivity/%	Specificity/%	F1 /%
MFCC+CNN	91.01	85.99	96.03	90.51
MFSC+CNN	86.10	89.37	82.83	86.45
Wavelet based +CNN	83.38	87.60	79.17	83.95
Our method +CNN	95.12	92.27	97.93	94.95

### 3 EXPERIMENTAL RESULTS AND ANALYSIS

#### 3.1 Comparison of feature extraction methods

To evaluate the effectiveness of the extracted fusion features, different feature extraction methods are compared with different classifier combinations in this paper. Three feature extraction methods commonly used for heart sound classification are included: MFCC, MFSC and Wavelet-based features, under the subject dataset, and the results are shown in Table 1. Both MFCC and MFSC can simulate the auditory system, but MFCC compresses the features compared to MFSC, which has less data and fits the classification model of this paper better.

#### 3.2 Classifier performance comparison

To verify the generalizability of the feature extraction algorithm in this paper, the algorithm in this paper is compared with three classifier models commonly used in classification recognition, K-Nearest Neighbor (KNN), SVM and RF and under the same dataset of the subject group for experiments, and the results are shown in Table 2. From Table 2, we can see that the fused features combined with KNN, RF and SVM classifiers can all perform the heart sound classification task well, and the best performance among the CNN classifiers built in this paper. The worst performance is KNN, which may be due to the fact that KNN is more dependent on the training set than the other three classifiers. The use of fusion feature extraction method

with CNN classifier is more suitable for the application of heart sound classification.

Figure 4 shows the ROC curves of the experimental results of different feature extraction methods combined with different classification models. the closer the ROC curve is to the upper left, the better the performance of the system. From Fig.4, it can be seen that different classification models with the algorithms in this paper have good performance in terms of the overall evaluation index.

#### 3.3 Validation of different data sets

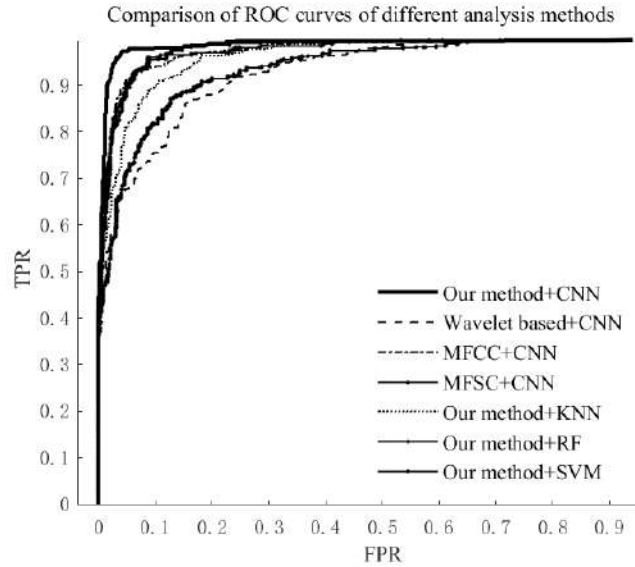
Table 5 shows the comparison of the experimental results of this paper's method on the two datasets introduced in Section 2.1 under the same experimental environment. From the table, it can be seen that the specificity index of this paper's method is lower on the public dataset, probably due to the unbalanced number of normal and abnormal heart sound samples in the public dataset, but the overall performance is better when all the indexes are considered together, which proves that this paper's method has strong noise immunity and generalization ability.

### 4 CONCLUSION

In this paper, we propose a heart sound classification algorithm based on sub-band statistical and time-frequency fusion features, which removes the complex denoising and segmentation steps and fuses the computationally simple statistical features with time-frequency features to reduce the computational effort while improving the classification accuracy compared with previous studies.

**Table 2: Feature extraction methods and recognition results of different classifiers in this paper**

Analytical method	Evaluation Criteria			
	Accuracy/%	Sensitivity/%	Specificity/%	F1/%
Our method +KNN	87.73	95.33	80.13	88.49
Our method +RF	90.92	96.62	85.21	91.32
Our method +SVM	91.68	85.19	98.09	91.05
Our method +CNN	95.12	92.27	97.93	94.95

**Figure 4: Comparison of ROC curves of algorithms****Table 3: Identification results of our method on different datasets**

Dataset	Evaluation Criteria			
	Accuracy/%	Sensitivity/%	Specificity/%	F1 /%
Public data sets	91.72	94.56	88.89	91.75
Subject dataset	95.12	92.27	97.93	94.95

The proposed algorithm is evaluated using four different evaluation metrics, and the experimental results show that the comprehensive performance of the feature extraction algorithm in this paper is better than other methods in comparison with other feature extraction algorithms, with an accuracy of 95.12% and a better fit with different classification models. In addition, this paper uses the subject dataset and public dataset for testing, which further shows that the proposed algorithm has better generalization ability and is more suitable for the actual CHD screening environment.

## ACKNOWLEDGMENTS

This work was funded by the Major Science and Technology Projects of Yunnan Province under Grants 2018ZF017 and the National Natural Science Foundation of China under Grants 81960067.

Thanks to Affiliated Cardiovascular Hospital of Kunming Medical University for providing a clinical research environment and medical guidance for this study.

## REFERENCES

- [1] YANG Yang, GUO Xingming, ZHENG Yineng, WANG Hui. Study on the identification of heart sound signals of left ventricular diastolic dysfunction based on ICEEMDAN-MSE[J]. Journal of Instrumentation, 2022, 43(01): 274-281.
- [2] XU Chun-dong, ZHOU Jing, YING Dong-wen, XIN Peng-li. Heart sound noise reduction under dynamic estimation of noise[J]. Journal of Biomedical Engineering, 2020, 37(05): 775-785.

- [3] Radia F, Zine-Eddine H S. A new heart sounds segmentation approach based on the correlation between ECG and PCG signals[J]. *International Journal of Biomedical Engineering and Technology*, 2019, 29(2): 174-185.
- [4] Nogueira D M, Ferreira C A, Gomes E F, *et al*. Classifying heart sounds using images of motifs, MFCC and temporal features[J]. *Journal of medical systems*, 2019, 43(6): 1-13.
- [5] Abduh Z, Nehary E A, Wahed M A, *et al*. Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers[J]. *Biomedical Signal Processing and Control*, 2020, 57: 101788.
- [6] Yin Y, Ma K, Liu M. Temporal convolutional network connected with an anti-arrhythmia hidden semi-Markov model for heart sound segmentation[J]. *Applied Sciences*, 2020, 10(20): 7049.
- [7] Wang, Xingzhi, Yang, Hongbo, Zong, Rong, Pan, Jiahua, Wang, William. A heart sound classification algorithm based on subband envelope and convolutional neural network[J]. *Journal of Biomedical Engineering*, 2021, 38(05): 969-978.
- [8] Karan B, Thakur G, Rath A, *et al*. Heart Sound Abnormality Detection using Wavelet Packet Features and Machine Learning[C]//2021 International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA). IEEE, 2021: 310-314.
- [9] Xu W, Yu K, Ye J, *et al*. Automatic pediatric congenital heart disease classification based on heart sound signal[J]. *Artificial Intelligence in Medicine*, 2022, 126: 102257.
- [10] WU Jiangbo, JIA Yunwei, YAO Chengbin, HAO Chenxiang, WANG Kun. Algorithm for spectral signal extraction based on statistical features and saliency[J]. *Spectroscopy and Spectral Analysis*, 2021, 41(07): 2294-2300.
- [11] Liu C, Springer D, Li Q, *et al*. An open access database for the evaluation of heart sound algorithms[J]. *Physiological measurement*, 2016, 37(12): 2181.
- [12] Gao W, Kan Y, Zha F. Filter algorithm based on cochlear mechanics and neuron filter mechanism and application on enhancement of audio signals[J]. *Journal of Central South University*, 2021, 28(6): 1813-1828.
- [13] KANG Jia-Fang, WANG Hong-Xing, ZHONG Pei-Lin, LIU Chuan-Hui. Baseband DSSS modulation method based on Hilbert shaped waveform transform[J]. *Journal of the University of Electronic Science and Technology*, 2020, 49(02): 201-205.
- [14] Saint-Arnaud N, Popat K. Analysis and synthesis of sound textures[M]//*Computational auditory scene analysis*. CRC Press, 2021: 293-308.
- [15] Lan T, Peng Chuan, Li Sen, Ye Wenzheng, Li Meng, Hui Guoqiang, Lu Yilan, Qian Yuxin, Liu Highest. A review of monophonic speech noise reduction and de-reverberation research[J]. *Computer Research and Development*, 2020, 57(05): 928-953.
- [16] Cheng Xie-Feng, Wang Jing, Wang Yue. Design and application of a multi-threshold fused heart sound recurrence map[J]. *Vibration and Shock*, 2019, 38(16): 108-114. DOI:10.13465/j.cnki.jvs.2019.16.016.

# Garment Metaverse: Parametric Digital Human and Dynamic Scene Try-on

Hua Wang  
Wuhan Textile University  
wanghua@yeah.net

Minghua Jiang  
Wuhan Textile University  
minghua.jiang@wtu.edu.cn

Xiaoxiao Liu  
Wuhan Textile University  
liuxiaoxiaolxx@yeah.net

Changlong Zhou\*  
Wuhan Textile University  
zcl@wtu.edu.cn

## ABSTRACT

As a new concept, the metaverse has been widely concerned by the industry, academia, media and the public. Many domestic and foreign companies have also set up in the field of the metaverse. The traditional 2D and 3D virtual fitting has not achieved breakthrough and development because of the technical problems of authenticity and timeliness. With this in mind, we developed a virtual fitting system based on the metaverse community, which includes two modules: parameterized virtual digital human modeling and multi-scene and multi-action fitting. The system realizes the construction of individualized virtual digital person. As the medium between the metaverse clothing community and the real world, it is used to achieve multi-category dynamic fitting display actions in the metaverse clothing community platform for multi-category scenes. The system integrates the high simulation and synchronization of the metaverse with the virtual fitting system, to break the barriers of traditional virtual fitting technology and realize the combination of the garment industry and the metaverse. The experimental results show that the system can realize the construction of virtual digital human in 2.1ms at the fastest and realize the dynamic display of multi-action and multi-scene fitting.

## KEYWORDS

Metaverse, Digital human, 3D face reconstruction, Virtual try-on

### ACM Reference Format:

Hua Wang, Xiaoxiao Liu, Minghua Jiang, and Changlong Zhou\*. 2023. Garment Metaverse: Parametric Digital Human and Dynamic Scene Try-on. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590014>

## 1 INTRODUCTION

Driven by the wave of new technologies such as artificial intelligence and virtual reality, the concept of the metaverse has been put forward and constantly given new meaning. 2021 is also known

as the first year of the metaverse. The term "metaverse" comes from the science fiction novel *Snow Crash* by Neal Stephenson [1]. In the novel, humans live in a virtual three-dimensional space through avatars, which the author calls the metaverse. It is born out of and parallel to the real world. Un-Kon Lee [2] proposes that metaverse is the permanent, immersive mixed-reality world, where people and objects can synchronously interact, collaborate, and live within the limitation of time and space, using avatars and immersion-supporting devices, platforms, and infrastructures. For instance, in the field of urban transportation, smart city services provide effective solutions to urban problems. Vinayakumar R et.al [3] propose a visualized botnet detection system based deep learning, it can monitor all Internet-connected IoT devices online in real time. Shuai Wang et.al [4] propose a novel human Short-Long Cognitive Memory mechanism for video surveillance in smart cities. It can realize visualization, real-time automatic and effective target monitoring and tracking in complex monitoring environment with limited resources; In the medical field, the technology of metaverse is used as assistive rehabilitation tools for patients who suffer from stroke [5, 6], cerebral palsy [7] and severe burns [8, 9]. Immersive technology-based rehabilitation allows patients to obtain a more intensive learning and rehabilitation experience by immersing themselves in an enriched practice environment specifically designed for rehabilitation purposes [10]. Apart from rehabilitation, immersive technologies also are used in telehealth and virtual care [11]; In the field of education, Currently, scholars often apply technology of metaverse to medical training and education as teaching aids [12, 13] to support, supplement, or even replace traditional teaching methods. At present, Metaverse has become a hot topic in the digital economy, but the related research in academic circles lacks depth, especially the specific implementation cases of Metaverse. In the field of clothing, there is no specific implementation case of Metaverse. According to the framework of Metaverse, this paper builds the first Metaverse clothing community.

## 2 METHODOLOGY

The system is divided into two modules, which are parametric virtual digital human modeling and dynamic fitting under multiple scenes and actions. The parametric virtual digital human modeling, which is divided into parametric human modeling unit and 3D face reconstruction unit, builds the virtual digital human model that is closest to users through the human-computer interaction between the user and the system, which reflects the high fidelity of Metaverse. The dynamic fitting module under multiple scenes

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590014>

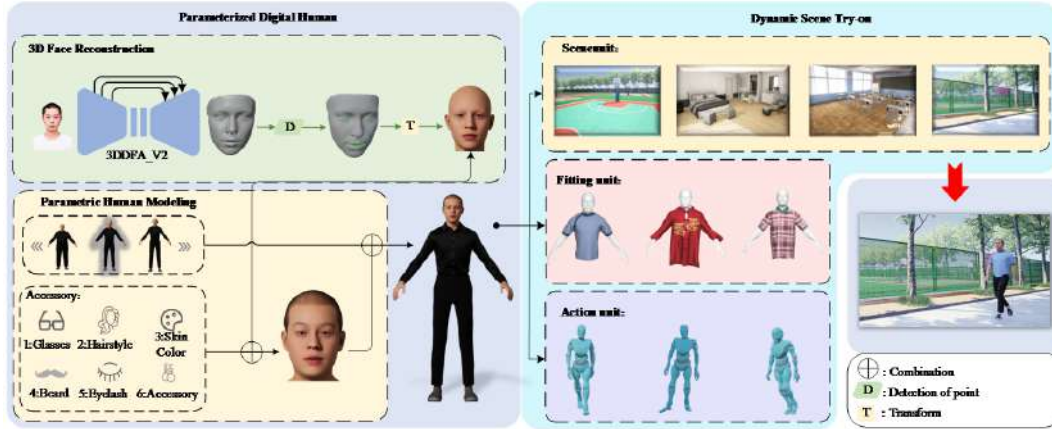


Figure 1: System frame diagram.



Figure 2: Human model diagram.

and actions includes scene unit, action unit and fitting unit. The dynamic virtual fitting effect under multiple scenes and actions is realized through the integration of the construction of fitting scenes, fitting actions and fabric simulation, which reflects the synchronicity of Metaverse. The overall framework of the system is shown in Figure 1.

## 2.1 Parametric Virtual Digital Human

In order to ensure the high fidelity of digital human, we use Epic's metahuman as the basic digital human model of the system. Parametric virtual digital human consists of digital human body model and digital human head model. After the basic body model and the reconstructed face model are fused, the accessories such as hairstyles, beards and the like are selected through human computer interaction, so as to realize the personalized needs of users. Finally the parameterized virtual digital person is fused. The self-defined parameterized virtual digital person in this system has the characteristic of high-fidelity texture mapping, personalized accessories selection, multicategory body shape selection, and image-based face reconstruction.

**2.1.1 3D face reconstruction.** Realistic face models can improve the user's fitting experience, and higher fidelity of virtual digital people can also be one of the necessary conditions of the Metaverse. In this paper, the 3D face reconstruction algorithm based on a single photo is implemented in the system, which makes the system more realistic and personalized. In order to speed up the running speed of the algorithm and keep the real-time calculation, 3DDFA-V2[14] also introduces a lightweight network structure and designs a meta joint optimization strategy to regress the parameters of the 3D morphable face model, which further improved the accuracy and running speed. At present, 3DDFA-V2 is the fastest 3D face reconstruction algorithm, which only takes 2.1ms on CPU.

**2.1.2 Parametric human model.** In this system the virtual digital human model is divided into 9 types of male and female models according to metahuman body database. The initial virtual digital human models are shown in Figure 2.

The system selects a suitable character model as the user's virtual digital human identity in the metaverse based on the human-computer interaction information, where the human-computer interaction information includes the user's gender and the size of the user's height, head, neck, chest, waist, hip, thigh, and calf



Figure 3: Scene diagram.

circumference. First, the system determines the male or female digital human model by the input of gender information, respectively, with  $l_i$  indicating the data of the size of height, head, neck, chest, waist, hip, thigh, and calf of the digital human model, with  $l_i^j$  ( $i \in [1, 8]$ ) denoting the data of first  $j$  height, head, neck, chest, waist, hip, thigh, and calf circumference, and  $j$  denoting the number of the digital human model in the manikin database. The system automatically calculates the relationship between the input size and the size deviation of the corresponding part of the digital human model in the human model database, and the calculation formula is shown as follows.

$$p_j = \sum_{i=1}^8 \sum_{i=1}^8 w_i (l_i - l_i^j)^2 \quad (1)$$

where  $p_j$  indicates the deviation of the body dimension data which the user inputs from the digital human model dimension data, and  $w_i$  indicates the influence weight value of part of the body dimension data in the model, and the system selects the smallest  $p_j$  corresponding to the number  $j$  of the digital human model as the user's personalized digital human model

## 2.2 Dynamic fitting

In order to simulate the real-world clothing shopping scene more realistically in the Metaverse clothing community, the system constructs the real-world try-on scene and try-on action in the Metaverse clothing community, realizes the effect of dynamic fitting, and meets the personalized fitting needs of users.

**2.2.1 Scene unit.** The method of generating the environment and objects in Metaverse is divided into the method of depicting by reflecting the real world and the method of creating a new imaginary environment. A realistic way to reflect the real world environment is to reproduce famous places (such as museums and Eiffel Tower) and places familiar to individuals (such as home and school) in the real world. Alternatively, it creates a hard-to-reach environment (such as underwater and Mars) to provide a surreal experience. Just like the virtual digital human, the Metaverse not only needs to construct the same scenes and characters as the real world, but also needs to construct the imaginary scenes and characters, so as to conform to the characteristics of individuality and pluralism of the Metaverse.

In this paper, modeling tools such as 3ds Max are used to build multiple types of fitting scenes with high fidelity, including catwalk, classroom, playground, street, bedroom and other scenes as shown in Figure 3, and the model is imported into the system database for multi-scene fitting, so as to meet the diversified display needs of user's fitting and make the Metaverse more diversified.

**2.2.2 Action unit.** In order to show the synchronization of the Metaverse, the action display part after the virtual fitting is added in the system. With some of the actions in the public action library Mixamo as the basis, and through the skeletal redirection technology, the virtual digital human can synchronize the actions in the Mixamo action library. The redirection relationship of the two types of skeletal mesh bodies is shown in Figure 4. Bone redirection is to let the same animation reuse among similar bones. The system maps the skeletal levels and names of the skeletal mesh of the person model in the person model library to the ones in the Mixamo action library one by one, so that the skeletal actions in Mixamo can be reused on the human model skeletons in the system to meet the demand of diversified display of user's fitting and the requirement of synchronization of the Metaverse.

**2.2.3 Fitting unit.** In the fitting process, in order to improve the authenticity and reduction degree of the clothing model, so that the clothing can simulate the morphological changes caused by air resistance, gravity, etc, as well as the wrinkles and subtle shaking effects caused by the human body traction process, it is necessary to carry out dynamic simulation of clothing fabric.

This system uses 3D Draper plug-in from TriMirror Company as the fitting unit of the system. 3D Draper is the world's first real-time cloth simulation and multi-platform 3D virtual fitting solution. Through 3D mathematics and algorithms, as well as graphics, network, cloud, server, mobile, desktop programming techniques. And by using the actual CAD patterns and grading charts used in the design and manufacture of the real garments, as well as their true physical properties to capture their accurate fit and behavior.

The process of the unit is shown in Figure 5. The fashion designer provides a two-dimensional garment pattern, matches it to the virtual digital human model, sews the garment and simulates the fabric, and finally tries on the garment. On the right of Figure 5 is the display of virtual digital people after trying on different clothes. 3D Draper combines with the parameterized virtual digital human

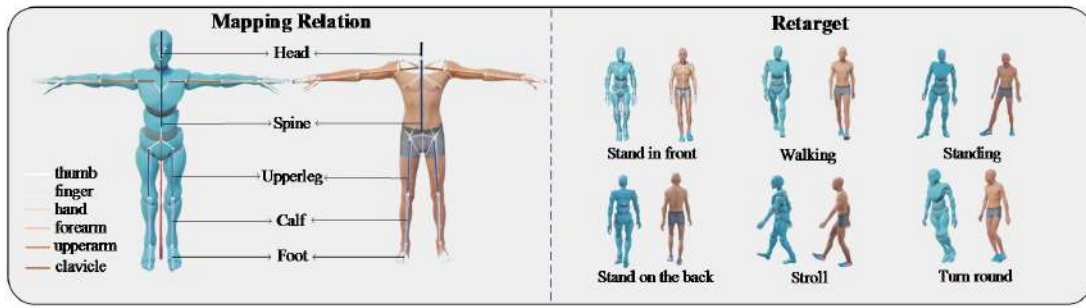


Figure 4: Skeletal redirection.



Figure 5: Clothing fitting module.

template in the system can realize the real-time fitting requirements of virtual digital human.

### 3 RESULTS

Through the 3D face reconstruction algorithm, the corresponding 3D face mesh can be obtained by inputting 2D face images. Metahuman provides the module of mesh transformation. By tracking the key points of the face of the mesh, it solves the head model of metahuman, realizes the process of mesh transformation to metahuman, and finally gets the head model corresponding to the input image. Then, it gives the skin texture, hairstyle, eyelashes and other accessories to the texture database built in the system, and gets the parameterized virtual digital person.

The system selects four face images from AFLW2000- 3D data set and one of our testers' face images as input, and uses DECA, LAP, 3DDFA, 3DDFA-V2 algorithms to output them respectively, thus obtaining a three-dimensional reconstructed face model. Then, the parameterized head model is obtained by tracking and solving the key points of metahuman, and finally, it is combined with the body part of the initial virtual digital human to form a parameterized virtual digital human. Figure 7 shows the reconstruction effect of an example face picture. Because the Morphable Face Model used by LAP, DECA and 3DDFA algorithms are different, the results of the reconstructed models are also different. Finally, the mesh volume conversion in metahuman will also be affected by the results of 3D face reconstruction. The reconstructed face model generated by

LAP algorithm in Figure 6 finally affects the deformed face shape of metahuman, so that the faces of the virtual digital people after the final deformation all incline to the square outline of LAP's Morphable Face Model. And because the 3D face model obtained by 3DDFA has no basic texture information, metahuman cannot detect the key point information of its face, hence the final face model of the virtual digital human cannot be obtained.

The 3D face reconstruction algorithm in the system gets the reconstructed face model according to the user's input face image, and the system detects the key points, and then deforms the initial virtual digital person's head model to get a customized virtual human head model. Through human-computer interaction, the user selects the appropriate hairstyle accessories and body model, and finally gets the parameterized virtual digital person model. Then, try-on scenes and try-on actions in the meta-universe clothing community are selected for virtual digital people, and the effect of dynamic fitting is realized by combining virtual fitting module. The final display effect is shown in Figure 6.

In this system, the main clothing model, character model resources, action resources, mapping resources, user data and other data are encrypted and then placed in the data server, which effectively reduces the storage space of the mobile terminal and protects the security of users and system data. The calculation unit of single-picture face reconstruction algorithm is directly placed in the client. In the client, there are still a few functional modules such as clothing simulation, data decryption, etc. that need real-time operation,

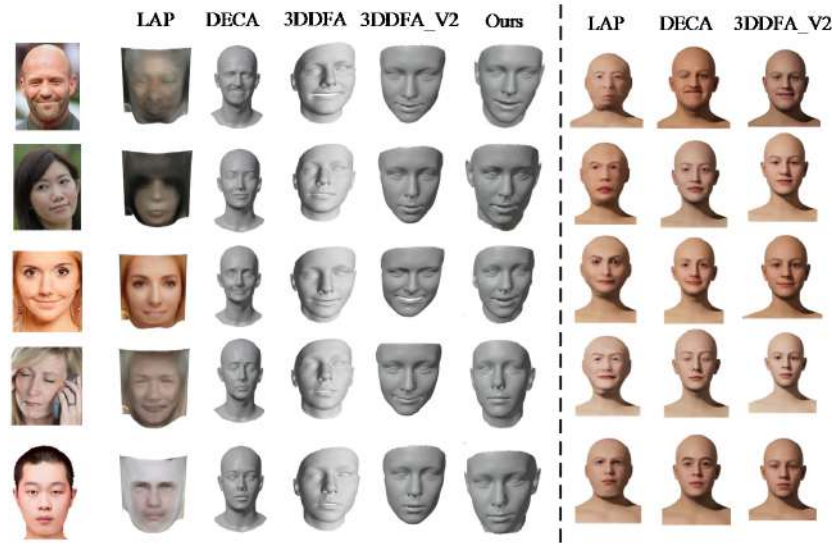


Figure 6: Face reconstruction results.

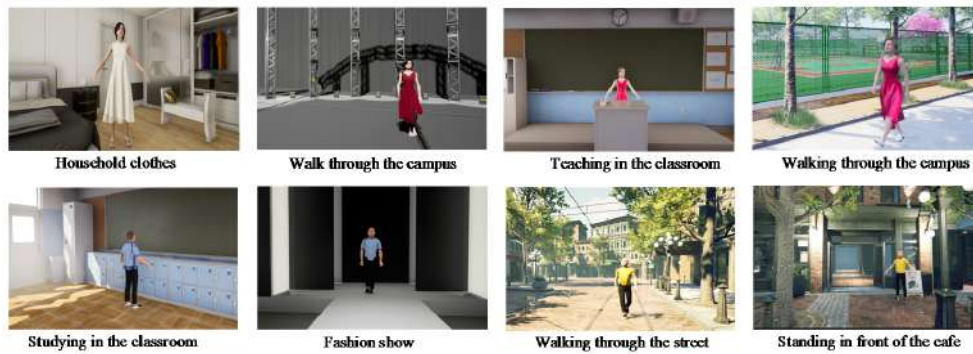


Figure 7: Virtual fitting results.

which is conducive to the development and implementation of the system on different mobile platforms and reduces the amount of system calculation. The system uses Intel Core i7-10700, NVIDIA GeForce RTX3070 8GB as the development platform, Unreal Engine as the development tool, combining mapping, model and material database and through man-machine interaction, it completes the communication between the client and the storage database, builds the Metaverse clothing community, and realizes the dynamic fitting of parameterized virtual digital people in Metaverse clothing community. The fitting results of parameterized virtual digital people are shown in Figure 7.

#### 4 CONCLUSION

This paper mainly puts forward the methods of parametric reconstruction of human body with high fidelity and multi-scene and multi-action fitting. In this system, the parametric digital human body model is constructed by parametric digital human body model and 3D face reconstruction algorithm, and the carrier construction

of metaverse and real world is realized by parametric parametric digital human model. Using the technology of skeletal redirection, the garment try-on display actions from the public multi-type character action library are transplanted to the parametric models, aiding the self-modeling try-on scenes of multiple categories in the system as well as fabric. For future work, we need to collect more data, especially adding special face and costume datasets for users in different regions and under different customs and cultures, to ensure that the parametric digital human as a medium can be more anthropomorphic and personalized.

#### REFERENCES

- [1] N. Stephenson. 2003. Snow crash: A novel. Spectra.
- [2] U.-K. Lee and H. Kim. 2022. Utaut in metaverse: An “ifland” case. Journal of Theoretical and Applied Electronic Commerce Research, 17(2):613–635. <https://doi.org/10.25147/ijcsr.2017.001.1.122>
- [3] R. Vinayakumar, M. Alazab, S. Srinivasan, Q.-V. Pham, S. K. Padannayil, and K. Simran. 2020. A visualized botnet detection system based deep learning for the internet of things networks of smart cities. IEEE Transactions on Industry Applications, vol. 56, no. 4, pp. 4436–4456, <https://doi.org/10.1109/TIA.2020.2971952>

- [4] S. Wang, X. Liu, S. Liu, K. Muhammad, A. A. Heidari, J. Del Ser, and V. H. C. de Albuquerque, 2021. Human short long-term cognitive memory mechanism for visual monitoring in iot-assisted smart cities. *IEEE Internet of Things Journal*, vol. 9, no. 10, pp. 7128–7139, <https://doi.org/10.1109/JIOT.2021.3077600>
- [5] K. N. Fong, Y. M. Tang, K. Sie, A. K. Yu, C. C. Lo, and Y. W. Ma, 2022. Task-specific virtual reality training on hemiparetic upper extremity in patients with stroke. *Virtual Reality*, vol. 26, no. 2, pp. 453–464, <https://doi.org/10.1007/s10055-021-00583-6>
- [6] Gorman and L. Gustafsson, 2022. The use of augmented reality for rehabilitation after stroke: a narrative review. *Disability and rehabilitation: assistive technology*, vol. 17, no. 4, pp. 409–417, <https://doi.org/10.1080/17483107.2020.1791264>
- [7] Choi, J. Y., Yi, S. H., Ao, L., Tang, X., Xu, X., Shim, D., Yoo, B., Park, E. S., & Rha, D. W. 2021. Virtual reality rehabilitation in children with brain injury: a randomized controlled trial. *Developmental medicine and child neurology*, 63(4), 480–487. <https://doi.org/10.1111/dmcn.14762>
- [8] . A. H. Kamel and M. A. Basha, 2021. Effects of virtual reality and task-oriented training on hand function and activity performance in pediatric hand burns: A randomized controlled trial. *Archives of Physical Medicine and Rehabilitation*, vol. 102, no. 6, pp. 1059–1066, <https://doi.org/10.1016/j.apmr.2021.01.087>.
- [9] Scapin, S., Echevarría-Guanilo, M. E., Boeira Fuculo Junior, P. R., Gonçalves, N., Rocha, P. K., & Coimbra, R. 2018. Virtual Reality in the treatment of burn patients: A systematic review. *Burns: journal of the International Society for Burn Injuries*, 44(6), 1403–1416. <https://doi.org/10.1016/j.burns.2017.11.002>
- [10] Levin, M. F., & Demers, M. 2021. Motor learning in neurological rehabilitation. *Disability and rehabilitation*, 43(24), 3445–3453. <https://doi.org/10.1080/09638288.2020.1752317>
- [11] Rutkowski S. 2021. Management Challenges in Chronic Obstructive Pulmonary Disease in the COVID-19 Pandemic: Telehealth and Virtual Reality. *Journal of clinical medicine*, 10(6), 1261. <https://doi.org/10.3390/jcm10061261>
- [12] D. Chytas, E. O. Johnson, M. Piagkou, A. Mazarakis, G. C. Babis, E. Chronopoulos, V. S. Nikolaou, N. Lazaridis, and K. Natsis, 2020. The role of augmented reality in anatomical education: An overview. *Annals of Anatomy-Anatomischer Anzeiger*, vol. 229, p. 151463, <https://doi.org/10.1016/j.aanat.2020.151463>
- [13] Maniam, P., Schnell, P., Dan, L., Portelli, R., Erolin, C., Mountain, R., & Wilkinson, T. 2020. Exploration of temporal bone anatomy using mixed reality (HoloLens): development of a mixed reality anatomy teaching resource prototype. *Journal of visual communication in medicine*, 43(1), 17–26. <https://doi.org/10.1080/17453054.2019.1671813>
- [14] Guo, J., Zhu, X., Yang, Y., Yang, F., Lei, Z., Li, S. Z. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. *ECCV 2020*. vol 12364. [https://doi.org/10.1007/978-3-030-58529-7\\_10](https://doi.org/10.1007/978-3-030-58529-7_10)

# Multi-dimensional analysis of urban shrinkage problem in Liaoning Province based on multi-index system, grey correlation analysis and BP neural network with particle swarm optimization

Zhenyu Fang

Houston International Institute, Dalian Maritime University, Dalian, Liaoning 116000, China  
fzy@dlmu.edu.cn

Junyu Xiong

School of Information Science and Technology, Dalian Maritime University, Dalian, Liaoning 116000, China  
xjy@dlmu.edu.cn

Junpeng Li

School of Science, Dalian Maritime University, Dalian, Liaoning 116000, China  
172069005qq@dlmu.edu.cn

Xin Wang\*

School of Science, Dalian Maritime University, Dalian, Liaoning 116000, China  
xenawang@dlmu.edu.cn

## ABSTRACT

The rapid development of urbanization in modern China is accompanied by the increasingly serious problem of urban shrinkage. To provide an effective analytical model for the urban shrinkage problem, this paper takes Liaoning Province, which is one of the typical provinces with a serious urban shrinkage issue in China, as an example. Based on the data from 30 cities in Liaoning Province in recent years, this paper constructs a multi-index system for shrinking cities to evaluate and classify the shrinkage degree of 30 cities. The grey relation analysis model is also used to quantitatively analyze the influence of various factors on the shrinking city population, while the back-propagation neural network algorithm model optimized with particle swarm optimization is also applied to predict the development trend of shrinking cities. The results present the shrinking properties of 30 cities and correlations between different city indicators, as well as the predictive development trend of the shrinking city.

## CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation; Model development and analysis; Modeling methodologies; Machine learning; Machine learning approaches; Neural networks.

## KEYWORDS

Grey relation analysis, Back-propagation neural network, Particle swarm optimization, A multi-index system for shrinking cities

### ACM Reference Format:

Zhenyu Fang, Junpeng Li, Junyu Xiong, and Xin Wang. 2023. Multi-dimensional analysis of urban shrinkage problem in Liaoning Province

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590016>

based on multi-index system, grey correlation analysis and BP neural network with particle swarm optimization. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590016>

## 1 INTRODUCTION

Over the years, some issues that appeared in the development process of the developed countries have gradually emerged in Chinese cities. One of the most typical issues is urban shrinkage, mainly characterized by population loss, economic decline, and other negative effects of development. Moreover, the development process of urban shrinkage in many developed countries has shown that urban shrinkage is a comprehensive concept [1]. Recent census results of China show that although the Chinese urbanization rate still maintains a rapid growth trend, some cities have been experiencing obvious population loss, economic decline, and other forms of urban shrinkage [2, 3]. Similarly, urban shrinkage issues also have occurred in many American and European cities [4], including many regions in Japan [5, 6]. The causes of the urban shrinkage to these cities are various, e.g., deindustrialization, low birth rate, political effects, and long-term industrial transformation. Thus, it is of great significance to establish a reasonable analysis model to judge and classify the shrinking characteristics of cities, so as to grasp the current situation and trend of urban development and provide feasible suggestions for the implementation and formulation of local policies. Therefore, this paper puts forward a set of the analytical model, including the judgment and classification of the cities in shrinking degree, as well as further prediction of the development of shrinking cities, and carries out an empirical analysis of urban shrinkage in Liaoning Province, a typical and representative province in China.

Liaoning Province is located in northeast China and used to be a famous heavy industry base in the middle of the 19th century when the population and the social economy rapidly grew. Nevertheless, affected by policy shifts and natural resources drying up, the conventional economic structure was pushed to be transformed. As a result, the economic downturn started to emerge in Liaoning Province while the development rate slowed down, and the urban shrinkage problem gradually spread to most cities. There are 3

types of urban shrinkage classification methods in current mainstream researches [1, 7, 8]: population index changes, economic index changes, and multi-index comprehensive changes. However, except for population index changes, there is no unified standard for the remaining two classification methods [1]. Therefore, this paper innovates evaluation criteria in terms of economic and social indicators to help discriminate urban shrinkage. But different from some previous studies which focused on the classification and qualitative analysis of the shrinking cities [1, 9], this paper proposes a method to quantitatively forecast the development of shrinking cities, helping to have a more specific grasp of the future development of shrinking cities.

Based on the economic and social data of 30 cities in Liaoning Province collected in recent years, a multi-index system is established for shrinking cities, to better identify and classify the shrinking degree of these cities. According to the properties of collected data of the city, the types of data are generally divided into four parts: population, economy, society, and policy. Then, the grey relation analysis (GRA) is applied to quantitatively analyze and discuss the influential degree of different factors on the population change of identified shrinking cities. Besides, this paper uses the back-propagation neural network (BPNN) optimized with the particle swarm optimization (PSO), which is termed PSO-BPNN here, to predict the development trend of two selected iconic composite shrinking cities. By comparing the predicted results with the general evolution characteristics of shrinking cities, it is found that our final predicted results have a better interpretation, indicating that the mathematical model established by this research can provide an effective analytical tool for the classification, discrimination, and prediction of urban shrinkage.

## 2 METHODOLOGY

### 2.1 The multi-index system for shrinking cities

The multi-index system for shrinking cities consists of the single index and multi-index evaluations. Among the types of collected data, some are public data types and others are specific data for some cities. In order to ensure that the final classification results of shrinking cities are comparable and universal, we choose to use the public data while constructing the multi-index system for shrinking cities. In public data types, there are two categories related to the population index, namely the number of employed people in the secondary and tertiary industries (EP23) and the registered population (RP), and three categories related to the economic index, namely urban GDP (UGDP), total retail sales of consumer goods (TRSCG) and public budget expenditure (PBE).

The single index evaluation is to judge that whether the city is shrinking in one aspect (economy, population, or society) or not. We select the RP change as the quantitative factor of the population index, UGDP change as the quantitative factor of the economic index, and EP23 and TRSCG changes as the quantitative factors of the social index. For the multi-index evaluation, we give different weights to those indicators used in the single index evaluation based on three standards, i.e., population, economy, and society. We choose the RP change as the sole index of population standard because the RP shrinkage is an important indicator of urban shrinkage. As for the economic standard, we choose the UGDP and PBE

changes as indexes. Since the UGDP reflects the overall economic development of the city, and the PBE reflects how much convenience the city provides for citizens' public services. In terms of the social standard, we choose the TRSCG and EP23 as the indicators.

After we specify the concrete classification criteria for shrinking cities herein, we use the weights given in the previous literature [10] for population, economic, and social standards, which respectively are 0.5, 0.3, and 0.2, as shown in Table 1.

**2.1.1 Formulas for the single index analysis.** The population index is based on the RP change. By referring to the theory proposed by scholar P.Oswalt, we identify cities with a total registered population loss rate of more than 1% as population shrinkage cities [11]. The calculation formula for the registered population loss rate is:

$$P_{di} = \frac{P_{i-1} - P_i}{P_i} \quad (2009 \leq i \leq 2020), \quad (1)$$

where  $p_i$  represents the RP in  $i^{th}$  year,  $P_{di}$  is the loss rate in  $i^{th}$  year. The formula for the population index is:

$$P_d = \frac{\sum P_{di}}{n} \quad n = 11, \quad (2)$$

where  $n$  is the total number of years,  $P_d$  represents the population index.

Therefore, to judge whether a city is the population shrinking city, the population index of 1% is taken as the dividing line, namely:

$$P_d = \begin{cases} \geq 1\% & \text{Shrinking City} \\ < 1\% & \text{Unshrinking City} \end{cases} \quad (3)$$

The economic index is mainly based on urban economic change, which is usually defined by indicators such as economic development speed and quality [12]. However, due to the complexity of social and economic factors, there is no fixed and unified quantitative standard. Here, due to the limitation of data types, the UGDP is selected and relevant research [11] (the total economic growth rate is less than 10%) is referred to define the economic index formula:

$$E_{di} = \frac{e_{i-1} - e_i}{e_i} \quad (2009 \leq i \leq 2020), \quad (4)$$

where  $e_i$  represents the UGDP in the  $i^{th}$  year,  $E_{di}$  represents the growth rate of UGDP in the  $i^{th}$  year. The formula for the economic index is:

$$E_d = \frac{\sum E_{di}}{n} \quad n = 11, \quad (5)$$

where  $n$  is the total number of years,  $E_d$  represents the economic index.

Therefore, to judge whether a city is the economically shrinking city, the economic index of 10% is taken as the dividing line:

$$E_d = \begin{cases} \geq -10\% & \text{Shrinking City} \\ < -10\% & \text{Unshrinking City} \end{cases} \quad (6)$$

Social shrinkage is also one of the important characteristics of urban shrinkage. The social index is mainly based on the change index of urban social services, which usually indicates the intensity of social service functions. Here, we select the EP23 and the TRSCG as factors, and calculate the average annual growth rate of TRSCG by analogy with the calculation method of population index:

$$CS_{di} = \frac{CS_{i-1} - CS_i}{CS_i} \quad (2009 \leq i \leq 2020), \quad (7)$$

**Table 1: Weights for different standards in the multi-index analysis [10]**

Objective Layer	Criterion Layer	Weight	Index Layer
Shrinking City	Population standard	0.5	RP
	Economic standard	0.3	UGDP
	Social standard	0.2	EP23
			TRSCG

**Table 2: Classification criteria for urban shrinkage**

Shrinkage Classification	Criticality or Growth	Mild Shrinkage	Moderate Shrinkage	Severe Shrinkage
Composite Shrinking Degree	$US \leq 0$	$0 < US \leq 0.2$	$0.2 < US \leq 0.5$	$0.5 < US$

where  $cs_i$  represents the TRSCG in the  $i^{th}$  year,  $CS_{di}$  represents the growth rate of TRSCG in the  $i^{th}$  year. The formula for calculating the annual growth rate of TRSCG is:

$$CS_d = \frac{\sum CS_{di}}{n} \quad n = 11, \quad (8)$$

where,  $n$  is the total number of years,  $CS_d$  represents the social index.

Similarly, we calculate the loss rate of EP23, namely:

$$RS_{di} = \frac{rs_{i-1} - rs_i}{rs_i}; \quad RS_d = \frac{\sum RS_{di}}{n}, \quad (9)$$

where  $RS_{di}$  represents the loss rate of EP23 in the  $i^{th}$  year,  $rs_i$  represents the EP23 in the  $i^{th}$  year. Finally, the social index of each city is calculated by:

$$S_d = \frac{RS_d + CS_d}{2}, \quad (10)$$

where  $S_d$  represents the social index.

Therefore, to judge whether a city is the socially shrinking city, the social index of 1% is taken as the dividing line:

$$S_d = \begin{cases} \geq 1\% & \text{Shrinking City} \\ < 1\% & \text{Unshrinking City} \end{cases} \quad (11)$$

**2.1.2 Z-score normalization.** In the multi-index analysis, we use the weights of the different standards (cf., Table 1). To eliminate the dimensional effects, reduce the differences among indicators in terms of magnitude, and weaken the mutual influence of indicators, normalized operations need to be conducted on each index data. Since there are no certain maximum and minimum values for the indicators selected in this question, Z-score normalization is used to normalize the data of the four indicators. Combining the normalized data matrices with the weights of each standard under the specific formula, we can get the composite shrinking degree of the city:

$$US = \sum_{j=1}^m (US_c)_j W_j, \quad (12)$$

where  $US$  represents the composite shrinking degree,  $(US_c)_j$  represents the normalized index layer values,  $W_j$  represents weights in criterion layer and  $m$  represents the serial number of the criterion layer.

**2.1.3 Classification criteria for urban shrinkage.** To further standard classification of urban shrinkage types, cities are divided into four grades according to the composite shrinking degree, i.e., criticality or growing, mild shrinkage, moderate shrinkage, and severe shrinkage (cf., Table 2) based on the existing research [10].

After computing the composite shrinking degree  $US$  of each city by formula (12), We can categorize the city by referring to Table 2. For example, If the computed composite shrinking degree of a city is 0.3, the shrinking extent of this city belongs to the moderate shrinkage under the multi-index analysis.

## 2.2 Grey relation analysis (GRA)

In view of the small amount of collected data and the unknown data distribution, we choose grey relation analysis (GRA). Since GRA is relatively reliable in the calculation of small data set and famous for its convenience and accuracy, even when the data distribution is unknown [13]. With the aid of GRA, we quantitatively analyze the influence of different factors on the population of each shrinking city.

**2.2.1 Determine analysis sequence and normalize data.** Here the reference sequence is the RP, and the comparison sequences are the remaining indicators. To prevent different units of indicators from affecting the final results of the model, we need to normalize all variable data to eliminate dimensional effects and reduce the difference in absolute values of the data.

**2.2.2 Solve correlation coefficient and define grey relational degree.** After normalization, the reference sequence is vector  $Y$ , and the comparison sequence is vector  $X_m$ . The two-level minimum difference  $a$  of the reference sequence and comparison sequence is:

$$a = \min (\min (|Y(k) - X_i(k)|)) \quad i = 1, 2, \dots, m \quad k = 1, 2, \dots, n. \quad (13)$$

The two-level maximum difference  $b$  is:

$$b = \max (\max (|Y(k) - X_i(k)|)) \quad i = 1, 2, \dots, m \quad k = 1, 2, \dots, n. \quad (14)$$

Then, we can calculate the correlation coefficients  $\gamma$  between the reference sequence and each comparison sequence by the following

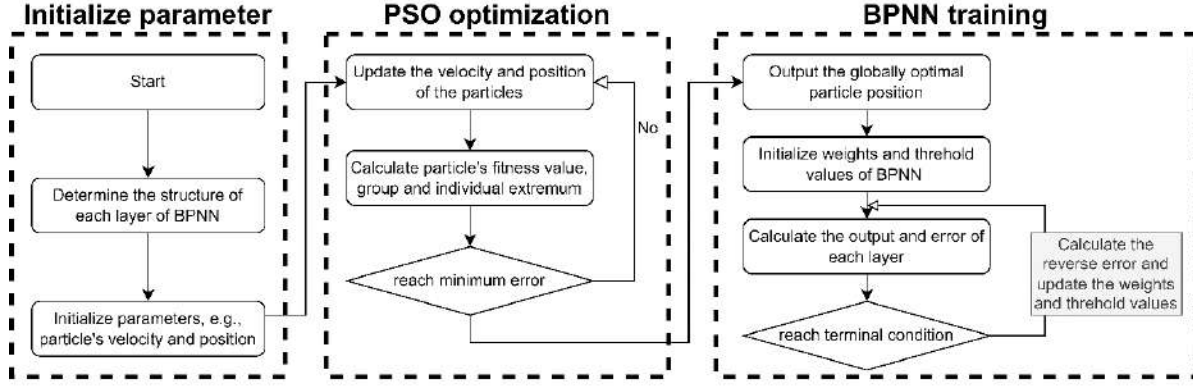


Figure 1: Working flow chart of PSO-BPNN.

calculation formula:

$$y(Y(k), X_i(k)) = a + ro \times \frac{b}{(|Y(k) - X_i(k) + ro \times b|)} \quad k = 1, 2, \dots, n, \quad (15)$$

where  $ro$  is the resolution coefficient, which is generally 0.5.

After calculating the correlation coefficient, the grey relation degree  $\bar{y}$  is defined as:

$$\bar{y}(Y(k), X_i(k)) = \frac{1}{n} \sum_{k=1}^n y(Y(k), X_i(k)). \quad (16)$$

### 2.3 Back-propagation neural network optimized with the particle swarm optimization

PSO-BPNN algorithm has shown some advantages in predicting data containing time series [14, 15]. Therefore, in order to predict the development trend of shrinking cities in the next year, we choose PSO-BPNN algorithm for prediction. Existing data have different dimensions, which will affect the speed of neural network training and learning. Therefore, it is necessary to normalize the data set first.

**2.3.1 Back-propagation neural network (BPNN).** BPNN is a classical neural network algorithm trained by an error back propagation algorithm [16]. The most common method to minimize the sum of error squares of the neural networks is gradient descent. The structure of the BPNN consists of an input layer, a hidden layer, and an output layer. The training and learning process are divided into two stages: forward propagation and backpropagation.

**2.3.2 Particle swarm optimization (PSO).** PSO is a kind of intelligent optimization algorithm based on the concept of birds' flock's looking for food [17]. Its core idea is to use information shared between individuals and groups, which is robust and not easy to fall into the local optimal solution. Update of particle velocity and position in iteration, and the specific formula is:

$$V_{iD}^{n+1} = V_{iD}^n + c_1 r_1 (P_{iD}^n - X_{iD}^n) + c_2 r_2 (P_{gD}^n - X_{iD}^n), \quad (17)$$

$$X_{iD}^{n+1} = X_{iD}^n + V_{iD}^{n+1}, \quad (18)$$

where,  $c_1$  and  $c_2$  represents the learning rate;  $r_1$  and  $r_2$  represents the random numbers distributed in range of  $(0, 1)$ ;  $n$  is the number of Iterations;  $X_{iD}^n$  represents the position of the particle  $i$  after

$n$  iterations;  $V_{iD}^n$  represents the velocity of the particle  $i$  after  $n$  iterations;  $P_{iD}^n$  represents the global extremum found by the search from the start to the current iteration.

**2.3.3 Neural network optimization based on PSO.** The solution of traditional BPNN is easy to fall into the local optimal. Therefore, PSO can effectively optimize the algorithm [14, 15]. The process is shown in Figure 1. The fundamental idea of this algorithm is to employ the weights and threshold values optimized by PSO as the initial weights and threshold values of BPNN for training and prediction.

Specifically, the group extremum of each particle group is set as the optimal value of the target function, and calculate the fitness value of each particle. When fitness greater than group extremum, fitness value will be replaced with the individual extremum and optimize group extremum. According to the formula (17) and (18), the particle's speed and position will be updated in each iteration. When PSO optimization is terminated, we can get the optimal weights and threshold values, which can be substituted into in to BPNN. Also, the initial weights and threshold values of BPNN are initialized. Then, the sample data will be input into the BPNN, and the model will be trained in the way of backpropagation. When the error between output value and direct value is large, the reverse error will be recalculated, accompanied by updating the connection weight and threshold between each layer, until the terminal condition is reached.

## 3 RESULTS

### 3.1 Classification results under the multi-index system for shrinking cities

After the data processing and calculation with proposed formulas, relevant results can be obtained. In terms of the single index analysis, we obtain 10 population-shrinking cities, 16 economically shrinking cities, 5 social-shrinking cities. For the multi-index analysis, we obtained 17 composite shrinking cities. The number of economically shrinking cities is very close to the number of composite shrinking cities. Since the weight of the economic standard is the largest when the multi-index analysis is established, this result is in line with expectations.

**Table 3: Shrinking indexes calculated under a multi-index system (partial)**

City	Population	Society	Economy	Composite	City	Population	Society	Economy	Composite
Benxi	<b>1.04%</b>	<b>1.25%</b>	-12.31%	<b>0.43</b>	Lingyuan	0.42%	-3.01%	-22.13%	-0.8
Chaoyang	<b>2.21%</b>	-5.12%	-11.78%	<b>0.75</b>	Xingcheng	0.96%	-7.21%	<b>-8.13%</b>	<b>0.08</b>
...	...	...	...	...	...	...	...	...	...
Donggang	0.72%	-0.18%	-11.80%	<b>0.17</b>	Anshan	-0.37%	-0.86%	<b>-7.73%</b>	-0.29

\* The bold numbers indicate the city is shrinking under the index

**Table 4: Grey relation degree between indicators and the population of PPL shrinking cities (partial)**

City	GDP	TRSCG	FI	PBE	GIOVADS	EP23	WSDU	IREDD
Fushun	0.7805	0.6576	0.6032	0.6694	0.7151	0.8740	0.8389	0.6558
Benxi	0.7633	0.6491	0.5678	0.7446	0.7089	0.8476	0.8182	0.5807
...	...	...	...	...	...	...	...	...
Huludao	0.8330	0.5833	0.6891	0.6790	0.7353	0.6884	0.6379	0.5778

**Table 5: Grey relation degree between indicators and population of CL shrinking cities (partial)**

City	GDP	TRSCG	PBE	EB23	HB
Diaobinshan	0.5443	0.5895	0.6343	0.6123	0.5805
Linghai	0.5446	0.5388	0.6766	0.6929	0.5361
...	...	...	...	...	...
Gaizhou	0.7345	0.5958	0.5964	0.7502	0.6195

With formulas proposed in section 2, the shrinking indexes of all cities can be computed, as shown in Table 3. It can be seen that the index values with respect to the population, society and economy are listed under the label “Population” “Society” “Economy” respectively. The column labelled “Composite” is the city’s composite shrinking degree. Compared with the prescribed dividing lines (single index evaluation) and classification criteria (multi-index analysis), it is easy to judge whether the city is shrinking in certain aspect or not.

### 3.2 Results of the grey relation analysis

Through the grey relation analysis, we can quantitatively obtain the impact of various indicators on the RP of 17 composite shrinking cities. Since the data types of provincial and prefecture-level (PPL) shrinking cities and county-level (CL) shrinking cities (cf., Table 5) are not the same, the data of these two types of cities are solved separately. For convenience, we define abbreviations for some indexes: fixed investments (FI), water supply for domestic use (WSDU), investment in real estate development (IREDD), gross industrial output value above designated size (GIOVADS), and hospital bed (HB). Table 4 and Table 5 respectively present the grey relation degrees between indicators and the population of PPL and CL shrinking cities.

From Table 4, it can be obtained that for PPL shrinking cities, the correlation between FI and RP is weak, but that between UGDP is strong. Also, the correlation intensity in the EP23 and WSDU in Huludao is quite different from that of other PPL shrinking cities.

For county-level shrinking cities, the PBE and EP23 have a greater impact on the RP. Different from PPL shrinking cities, half

of the CL shrinking cities’ UGDP is not strongly correlated with the RP. The rest of the indicators in the mutual influence is not strong commonality.

### 3.3 Prediction of shrinking city development trend by PSO-BPNN

First set the parameters of the prediction model:

1. PSO part: the number of population size is 5; learning rate  $c_1 = c_2 = 2.4$ ; iterative number  $n = 30$ ; particle position  $X_{max} = 5$ ,  $X_{min} = -5$ ; particle velocity  $V_{max} = 2$ ,  $V_{min} = -2$ . The fitness function of the particle is equal to the mean square error of the true and predicted values.
2. BPNN part: The whole network is composed of three layers, the number of nodes in the input layer is 14, the main input data is the data of various economic and social factors, and the number of nodes in the hidden layer is 10. Transfer function Sigmoid, training times of 1000, training target is 0.00001.

To evaluate the model’s prediction accuracy more accurately, the root mean square error is selected as the indicator. The evaluation results show that the predicted results have a good fit with the measured values.

Since the number of shrinking cities is large, Fushun and Beipiao are taken as examples of the development trend forecast of PPL and CL shrinking cities respectively (cf. Table 6).

Beipiao belongs to a population and mild composite shrinking cities. According to the forecast data, the total UGDP and RP of Beipiao in 2021, including the EP23 in 2020, will all decline, while

**Table 6: Results of city development forecast by PSO-BPNN (partial)**

CL shrinking cities: take Beipiao as an example				
Beipiao	UGDP	TRSCG	EP23	RP
	Unit / 10000¥	Unit / 10000¥	Unit / 10000 people	Unit / 10000 people
2013	2305298	545500	17.89	57.56
...	...	...	...	...
2020	1245000	349000	<b>7.34</b>	54.05
2021	<b>1198682</b>	<b>423541</b>	/	<b>53.98</b>
PPL shrinking cities: take Fushun as an example				
Fushun	UGDP	TRSCG	EP23	RP
	Unit / 10000¥	Unit / 10000¥	Unit / 10000 people	Unit / 10000 people
2007	4296407	2030220	24.01	140.07
...	...	...	...	...
2020	6870000	1521347	20.32	132.3
2021	<b>6750000</b>	<b>5407623</b>	<b>18.89</b>	<b>131.8</b>

\* The bold numbers indicate the predicted values

only the TRSCG in 2021 will increase. In 2020, the social index will be 10.63% (socially shrinking city). In 2021 the population and economic indexes of Beipiao will be 0.1297% and 3.86% (economically shrinking city), respectively.

Fushun is a shrinking city in terms of population, economy, and society, belonging to the severe composite shrinking city. According to the forecast data, both the UGDP and RP of Fushun will decline in 2021, but the TRSCG will show a sharp upward trend. Moreover, the population, economic, and social indexes of Fushun in 2021 are 0.3794%, 1.78%, and -64.29%, respectively, indicating that Fushun is an economically shrinking city.

## 4 CONCLUSIONS

By comparing the computed results, the following conclusions are seen:

1. The classification results of shrinking cities in this paper are highly consistent with the previous survey [1, 9].
2. There are 17 composite shrinking cities (56.67% of all) under the multi-index analysis. This indicates that half of the cities in Liaoning province experienced a decline in development level these years, and the decline in urban development quality is a serious problem.
3. The causes of population loss in shrinking cities are complex and related to factors which have different impacts in different cities.
4. The predicted data indicate that the development of shrinking cities in the short term still has shrinking characteristics.

On the whole, the prospects of shrinking cities are generally grim when the factors that may lead to dramatic changes in urban development are not considered. As the characteristics of urban shrinkage strengthen, negative feedback effect is formed, which is not conducive to the long-term development of a city. Liaoning Province is experiencing severe urban shrinkage issue, where the traditional urban planning method based on growth expectation has been difficult to meet the needs of current development [1]. These cities should recognize the shrinking situation and take some measures like giving full play to industrial restructuring, in order

to help themselves quickly get out of the "negative feedback effect of development" cycle.

## ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (Grant No. 61803065).

## REFERENCES

- [1] Sun P, Wang K. Identification and stage division of urban shrinkage in the three provinces of Northeast China[J]. *Acta Geographica Sinica*, 2021,76(06):1366-1379. DOI: 10.11821/dlxb202106004.
- [2] Liu J, Sun P, Luo N, Peng Y. Research Progress of Urban Shrinkage and Its Thoughts on Localization in China[J]. *Areal Research and Development*, 2022,41(03):55-60. DOI:10.3969/j.issn.1003-2363.2022.03.010.
- [3] Gao X, Zhao M, Shen W, Song Z, Zhang M. Study on Spatial Distribution Characteristics and Development Prediction of Shrinking Cities in Three Provinces of Northeast China[J]. *Global Cities Research*, 2021,2(02):72-83+191-192. [https://kns.cnki.net/kcms/detail/detail.aspx?FileName=\\$%QQCS202102007&DbName=\\$%CJFQ2021](https://kns.cnki.net/kcms/detail/detail.aspx?FileName=$%QQCS202102007&DbName=$%CJFQ2021)
- [4] Wiechmann T, Pallagst K M. Urban shrinkage in Germany and the USA: A comparison of transformation patterns and local strategies[J]. *International journal of urban and regional research*, 2012, 36(2): 261-280. DOI:10.1111/j.1468-2427.2011.01095.x
- [5] Buhnuk S. From shrinking cities to Toshi no Shukushō: Identifying patterns of urban shrinkage in the Osaka metropolitan area[J]. *Berkeley Planning Journal*, 2010, 23(1). DOI:10.5070/BP323111434.
- [6] Ducom E. Tama new town, west of Tokyo: Analysis of a shrinking suburb[J]. 2008. <https://shs.hal.science/halshs-00203107/>.
- [7] Mallach A, Haase A, Hattori K. The shrinking city in comparative perspective: Contrasting dynamics and responses to urban shrinkage[J]. *Cities*, 2017, 69: 102-108. DOI:10.1016/j.cities.2016.09.008.
- [8] Yang D, Long Y, Yang W, Sun H. Losing Population with Expanding Space: Paradox of Urban Shrinkage in China[J]. *Modern Urban Research*, 2015(09):20-25. DOI:10.3969/j.issn.1009-6000.2015.09.003
- [9] Wang G, Wang Y. Shrinkage city identification of prefecture-level cities in Liaoning province[J]. *Journal of Liaoning Normal University(Natural Science Edition)*, 2020,43(04):552-557. DOI:10.11679/lxblk2020040552.
- [10] ZHANG Shuai, WANG Chengxin, WANG Jing, *et al.* On the comprehensive measurement of urban shrink in China and its spatio-temporal differentiation[J]. *China population, resources and environment*, 2020,30(08):72-82. <http://ir.igsnrr.ac.cn/handle/311030/156123>.
- [11] Oswalt P. Shrinking cities, volume 1: International research[J]. Ostfildern-Ruit: Hatje Cantz, 2005.
- [12] Yang Z, Yang D. Exploring shrinking areas in China availing of city development index[J]. *Human Geography*, 2019, 34(4): 63-72. DOI: 10.13959/j.issn.1003-2398.2019.04.008.

- [13] Tan X, Deng J. Grey Relation Analysis:A new method for statistical analysis of multivariate[J]. Statistical Research,1995(03):46-48. DOI:CNKI:SUN:TJYJ.0.1995-03-010.
- [14] Niu M, Li X, Zhang Y, Lei X, Wang Y, Li W. Short-term Prediction Model of Soil Water Content Based on BP Neural Network Optimized by PSO[J]. Vegetables, 2020(08):24-30. [https://kns.cnki.net/kcms/detail/detail.aspx?FileName=\\$SCZZ202008007&DbName=\\$CJFQ2020](https://kns.cnki.net/kcms/detail/detail.aspx?FileName=$SCZZ202008007&DbName=$CJFQ2020)
- [15] Z. Lin, G. Chen, W. Guo and Y. Liu, "PSO-BPNN-Based Prediction of Network Security Situation," 2008 3rd International Conference on Innovative Computing Information and Control, 2008, pp. 37-37. DOI: 10.1109/ICICIC.2008.436.
- [16] Hecht-Nielsen R. Theory of the backpropagation neural network[M]//Neural networks for perception. Academic Press, 1992: 65-93. DOI:10.1016/B978-0-12-741252-8.50010-8.
- [17] Poli R, Kennedy J, Blackwell T. Particle swarm optimization[J]. Swarm intelligence, 2007, 1(1): 33-57. DOI: 10.1007/s11721-007-0002-0

# Helmet wear detection based on YOLOV5

Jun Liu  
Wuhan Textile University  
junliuwtu@yeah.net

Jiacheng Cao  
Wuhan Textile University  
caojiachengjc@gmail.com

Changlong Zhou\*  
Wuhan Textile University  
zcl@wtu.edu.cn

## ABSTRACT

Safety helmet wearing detection is an important safety inspection task with widespread applications in industries, construction, and transportation. Traditional safety helmet wearing detection methods typically use feature-based classifiers such as SVM and decision trees, but these methods often have low accuracy and poor adaptability. In this paper, we propose an improved helmet detection method that uses a combination of SPD Conv, ASPP and BiFPN structures to increase the perceptual field to ensure maximum feature extraction from the helmet, and can ensure fusion between different feature layers to pass semantic information to deeper neural networks, effectively avoiding information loss and improving the performance of detecting helmets. Experimental results show that our method has a 1% improvement in the average accuracy of detection in the public dataset VCO2007 set compared to YOLOv5, which still allows for real-time detection and meets the needs of industry with some practicality.

## CCS CONCEPTS

• **Software and its engineering**; • **Computing methodologies**  
→ Artificial intelligence; Computer vision; Computer vision tasks; Scene anomaly detection;

## KEYWORDS

Helmet Detection, SPD-Conv, ASPP, BiFPN, Feature extraction

### ACM Reference Format:

Jun Liu, Jiacheng Cao, and Changlong Zhou. 2023. Helmet wear detection based on YOLOV5. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590017>

## 1 INTRODUCTION

In infrastructure construction, the Work Safety Law clearly stipulates that all personnel must wear helmets when entering the workplace. Due to hot weather, workers' negligence, comfort of wearing helmets and other factors, entering the workplace without wearing helmets result in economic losses and casualties. Helmets are the most basic protective equipment for workers, which effectively protect the life safety of workers. It is of great research

significance to realize real-time detection of whether workers at the workplace wear helmets.

At present, helmet detection methods are divided into two types: traditional detection methods which require manual design features. This method does not have generalization ability and robustness. In practical applications, there are problems such as high time complexity, slow detection rate, low accuracy, and many false positives. Based on the detection method of depth learning, the weights used in the convolution layer can realize parameter sharing, sparse connection, feature extraction without manual design, and have the advantages of translation invariance.

Based on the idea of the one-stage [6] algorithm of the YOLO series [1–5], a "divide and conquer" strategy is used to divide the image into a number of grids, and if the center of a target falls on the grid, the grid is responsible for predicting this target. As input to the network, it only needs to go through a neural network to get the position of the Bounding box and its class. Moreover, the feature pyramid network BiFPN [7] has a feature information acquisition network structure, which can effectively improve the network's ability to obtain deep information about the target in the case of complex colors, shapes and backgrounds.

Li et al.'s [8] SSD MobileNet algorithm base on convolutional neural network enhances feature information extraction and improves detection accuracy. Cheng et al. [9, 14, 15] propose a helmet detection algorithm (SAS-YOLOv3 tiny), add an improved spatial pyramid pool (SPP) module to the feature extraction network, extract local and global features with rich semantic information, and improve the accuracy of helmet detection. Rattapoom Waranusast et al. [10] propose a system using a KNN classifier that recognises moving objects and classifies heads as wearing a helmet or not, and achieves correct detection rates of 84%, 68% and 74% for near lane, far lane and two lane situations. Kang Li et al. [11] propose a framework for detecting helmets, using the ViBe algorithm to identify moving objects in substations and the C4 algorithm to locate pedestrians. It uses head position and colour information to determine if a safety helmet is being worn. Madhuchanda Dasgupta et al. [12, 16] propose in YOLOv3 and CNN to detect if wearing motorcyclists helmets. It obtains good results on traffic videos. CA Rohith et al. [13] working on a system that would automatically distinguish between cyclists wearing helmets and fines for non-compliance as part of enforcement. This system is not yet used by the police or other agencies. We intend to use trained models and datasets to build a deep learning based helmet recognition system to help police departments enforce the law.

The above algorithms have made some contributions to helmet recognition and helmet feature extraction, but there are some difficulties in accurately identifying whether workers wear helmets in the complex actual operating environment: 1) Workers wear helmets. Helmets are relatively small targets, and it is difficult to obtain helmet features at a deeper level. 2) The actual environment is very complex, and the influence of light intensity, helmet overlap,

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590017>

color, reflection and other factors leads to the detection of missing. In order to solve the above problems, we propose a new object detection network algorithm, which effectively solves the problem of helmet missing detection under low accuracy and low resolution of helmet detection in practical applications. Our contribution is threefold as follows:

- 1) We propose the SPD-Conv module to address the loss of fine-grained information and the low feature extraction capability of convolutional neural networks.
- 2) We propose using cavity convolution, which expands the receptive field while maintaining resolution.
- 3) We propagate features of different sizes and scales in two-way propagation, helping to effectively reduce the loss of feature information, improves the network's extraction of target location information, and improves the accuracy of helmet detection in the training model.

## 2 METHODOLOGY

Object detection algorithms based on deep learning convolutional neural networks are divided into two categories. The two-stage, represented by Faster RCNN, has high detection accuracy. The images are input into the convolutional neural networks (CNN) for feature extraction. Regional Proposal Network (PRN) is used instead of the Selective Search method to generate a more accurate anchor box. After clipping and filtering, Softmax is used to judge whether the anchors belong to the foreground or the background[17]. The other branch carries out anchor box correction, and then maps to the feature map generated by the convolution layer of the last layer of CNN, and uses the RoI-Pooling layer to fix the feature map. Use Softmax Loss and Smooth L1 Loss to jointly train the classification probability with Bounding box regression. The other category, end-to-end detection, represented by YOLO, uses the "divide and conquer" strategy to divide the image into several grids. If the center of a target falls on the grid, this grid is responsible for predicting the target. YOLO turns object detection into a regression task, taking the entire image as the input of the network, and only needs to go through a neural network to get the location and category of the bounding box.

Xu uses increasing the number of anchor points on Faster RCNN to enhance the robustness of the model, and uses multi-scale training to increase the detection accuracy, but the detection speed cannot meet the application requirements[18]. Jiang et al. reconstruct a new backbone feature extraction network based on Inverted Res-block structure, and BN layer conducted sparse training to delete channels with smaller weights, reduce model parameters, make the model lightweight, greatly improve the detection speed, but the detection accuracy is low[19]. Shi et al. base on the image pyramid, fuse and connect the feature maps[20]. They use the K-means algorithm on the datasets to determine a prior box and multi-scale training. The input image pass through Darknet-53, followed by full convolution, and then output the results.

Convolutional neural network (CNN) achieve great success in the machine vision, such as image classification and object detection. However, in object detection, CNN's detection effect for small targets is always unsatisfactory, because it always uses pooling layer, up sampling and other aspects to obtain some features of

small targets in the image, which is a defect in the design of CNN framework, which will lead to the loss of fine-grained information and inefficient feature representation learning.

The loss of fine-grained information caused by convolution and pooling layer of convolutional neural network itself and the low feature extraction ability lead to the rapid decline of detection accuracy of the network in the detection task of low resolution images or small objects. This paper introduces a new convolutional neural module SPD-Conv, which is composed of space to depth (SPD) layer and non step convolution (Conv) layer. Space\_ to\_ Depth means that the dimension is superimposed on the length and width of the depth, which is equivalent to pooling layer, but pooling is to select one of all the dimensions, and this method uses one of the dimensions, and the remaining dimensions are superimposed on the depth direction, so as to retain the characteristics of low latitude.

Feature Pyramid Networks (FPN), because it is difficult to detect small objects in object detection, and in the process of convolution, there are many pixels of large objects and few pixels of small objects[21]. With the deepening of convolution, the features of large objects are easy to be retained, and the features of small objects are easier to be ignored later. After successive down-sampling of feature points, it has a bunch of feature layers with high semantic content, which are then resampled so that the length and width of the feature layers become large again, and a large size feature map is used to detect small targets. As the result of up-sampling is not clear about the features and information of the small target, the feature layers with the same length and width in down-sampling and up-sampling can be superimposed to ensure the features and information of the small target.

## 3 OUR APPROACH

### 3.1 SPD-Conv

The loss of fine-grained information caused by the convolution and pooling layers of convolutional neural networks (CNNs) and their low feature extraction ability leads to a rapid decline in the detection accuracy of the network when detecting low resolution images or small objects. To address this issue, we propose the SPD-Conv method Figure 1., which consists of a space-to-depth (SPD) layer followed by a non-stepped convolution (Conv) layer. The SPD layer down-samples the feature map while retaining all the information in the channel dimension, so no information is lost. The original image is scaled before it is input into the neural network and this scaling is applied throughout the network to down-sample the feature map. Additionally, a non-stepped convolution operation is added after each SPD layer to increase the number of channels using learnable parameters in the added convolution layer.

### 3.2 ASPP

In object detection algorithms, down-sampling is often used to increase the receptive field, but this method reduces the spatial resolution. Generally, feature extraction is top-down, transferring high-level strong semantic information throughout the network- Figure 2.. This enhances the semantic information, but does not transfer the target location information. In the case of detecting helmeted workers in a complex and cluttered environment, the

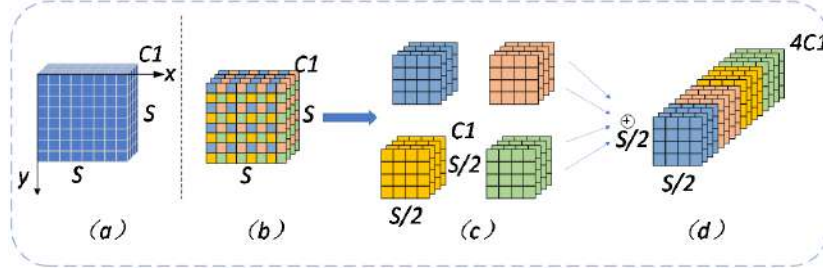


Figure 1: The illustration of SPD. The figure shows the feature map of  $s \times s \times C1$ . The original map is downsampled to scale, with a downsampling step of 2 in the figure, to obtain four sub-maps, before connecting these special wholes along the dimension of the channels, thus increasing the channel dimension of the feature map by a factor of 2.

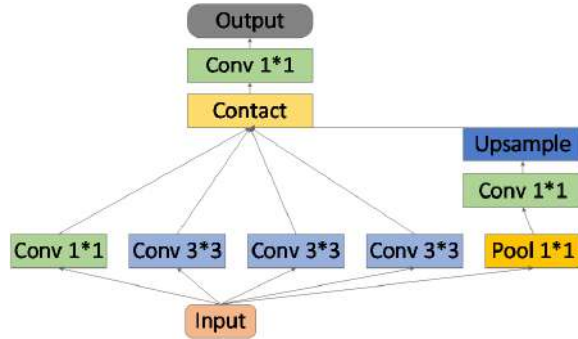


Figure 2: ASPP Structure. The input image is processed with  $1 \times 1$ ,  $3 \times 3$  (padding=6, dilation=6),  $3 \times 3$  (padding=12, dilation=12),  $3 \times 3$  (padding=18, dilation=18) operations, followed by a global average pooling operation. The output channel is then changed using a  $1 \times 1$  convolution and the image is upsampled to the original input size. Finally, feature fusion is performed. This effectively performs parallel sampling of the input with different sampling rates using the given input and cavity convolution, thus enhancing the network’s ability to extract features of safety helmets (color, shape, background, etc.).

helmets are often small and overlapped in the image, leading to low detection accuracy. To address this issue, we propose using cavity convolution, which expands the receptive field while maintaining resolution. This allows for the detection and segmentation of large targets, as well as accurate target positioning using high resolution. Additionally, cavity convolution can extract deep semantic information in the image, enhance the network’s sensitivity to the location information of the feature map, and avoid the loss of effective information for better positioning.

### 3.3 BiFPN

We propose using Bidirectional Feature Pyramid Network (BiFPN) to address the issue of unequal weights being generated during feature fusion between different scales of each feature layer during network training. This can occur when fusing features from different depths for the same target, leading to a loss of effective information and affecting the final training model. BiFPN solves

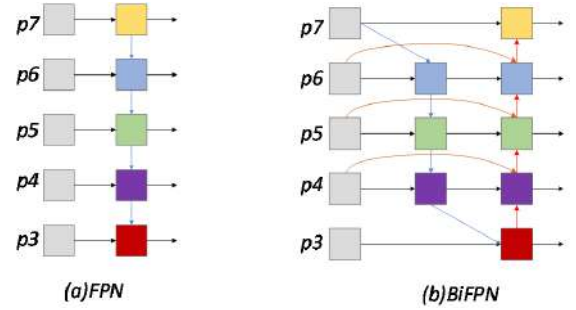


Figure 3: FPN compared with BiFPN. FPNs are self-directed upward network results in deep semantic information loss during the feature fusion phase. In this paper, the BiFPN structure is used to fuse different features of the feature network from top-down and bottom-up, adding jump connections to fuse more features without increasing the cost and enhancing the fusion of feature information obtained by the network. The BiFPN can be seen as a unit, then it can be repeated many times for superposition, thus obtaining higher level feature fusion.

this problem by allowing two-way propagation of features of different sizes, transferring strong semantic information from deep feature layers to shallow ones and strong positioning information from shallow layers to deep ones. This path fusion of different size feature layers allows for the acquisition of shallow strong target location information, as well as semantic information, avoiding the loss of effective information. The weighted two-way network helps to effectively reduce the loss of feature information, improves the network’s extraction of target location information, and improves the accuracy of helmet detection in the training model.

## 4 EXPERIMENT

We change the backbone and neck layers of YOLOv5 by innovatively adding the structure of ASPP, using convolution instead of maximum pooling down-sampling to increase the perceptual field and maximise the feature extraction of small target helmets, while adding the structure of BiFPN to the neck layer to ensure the fusion of features from different feature layers and to transfer the stronger semantic information to the shallow feature layer, avoiding the

loss of effective information and largely improving the detection efficiency of safety caps. Relative to YOLOv5's network, a 3% improvement in mAP is achieved on the official dataset.

#### 4.1 Experimental platform

The system configuration of our experimental platform is: the system version is Ubuntu 18.04.3 LTS, the processor is Intel(R) Xeon(R) Gold 5118 CPU @ 2.30GHz, the graphics processing unit is Tesla V100 16GB, it relies on python 3.8.13, numpy 1.23.3, opencv-python 4.6.0.66, torchvision 0.9.1+cu101, torch 1.8.1+cu101, pandas 1.5.1. The optimizer used is the SGD optimizer. The initial learning rate was 0.01. The input network was  $640 \times 640$  pixels, the batch size was 16, the number of model iterations was 300, and the learning rate was set to  $1e-4$ . YOLOv3, YOLOv3-tiny, YOLOv3-SSP, SSD, Faster-RCNN, YOLOv4, and YOLOX-S were chosen to perform cross-sectional comparison experiments to compare the detection effects of the models.

#### 4.2 Datasets

In our experiments, we use VOC datasets to evaluate the improved network model separately. VOC (Visual Object Classes) is a dataset that is widely used for object detection and classification. It contains a variety of object classes, including annotated images with bounding boxes and class labels for objects in the image. The ones we use in this experiment include "Person" and "Helmet" including 126,627 Person targets, and 8,213 Helmet targets, with a total of 7,851 images. The ratio of training and test sets we use for network training is 8:2.

#### 4.3 Evaluation Protocol

In our experiments, we use mAP (mean average precision) as the main evaluation metric. The definition is as follows:

$$mAP = \frac{1}{C} \sum_{i=1}^C AP_i \quad (1)$$

Where  $AP$ ,  $Precision$ ,  $Recall$  are calculated as (2) (3) (4) respectively:

$$AP = \int_0^1 P(R) \quad (2)$$

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

Of these, TP is actually the number of instances where positive samples are detected as positive samples. FP is actually the number of instances where negative samples are detected as positive samples. FN is actually the number of instances where positive samples are detected as negative samples. TN is actually the number of instances where negative samples are detected as negative samples. As shown in equation 2), AP (Average, Precision) is the average of the accuracy of each category. As shown in equation 3), Precision is the proportion of all classifiers that consider a positive class and are indeed positive as a percentage of all classifiers that consider a positive class. As shown in equation 4), Recall is how many positive proportions of the sample were correctly predicted.

#### 4.4 Comparative experiments

We first compare the performance of Helmet Detection on VOC 2007 train using our method with that of SOTA. As shown in columns 3, 4 and 5 of Table 1, we use the Person and Helmet labels as data benchmarks to calculate mAP to evaluate network performance. Our method narrows the gap between human and Helmet predictions and improves performance, resulting in a 1.5%, 6.7% and 8.6% improvement of mAP compared with YOLOvX-s, YOLOv4, Faster-RCNN-resnet50 respectively.

#### 4.5 Ablation experiments

To further validate the robustness and generalization of our improved neural network, we conduct ablation experiments on different datasets to demonstrate the superiority of our network over other networks by comparing the accuracy performance between different datasets. In table 2, we first show the Person, Helmets detection performance and mean average precision of the networks corresponding to the progressive addition of SPD, ASPP and BiFPN modules on the VOC dataset, using YOLOv5 as a baseline. In rows 2, 3 and 4, showing the mAP of YOLOv5 after adding just ASPP, SPD and BiFPN respectively, you can see that there is a 0.4% improvement in ASPP and SPD and a 0.2% drop in BiFPN, but this drop is not our final network. We further add 2 modules at once and see that adding ASPP+BiFPN improves YOLOv5's performance by 0.2%, adding ASPP+SPD improves YOLOv5's performance by 0.6% and adding SPD+BiFPN YOLOv5 improves performance by 0.7%. At this point, we combine all three modules simultaneously into YOLOv5 and we can see a 1% performance improvement. It shows that by adding the three modules, the detection accuracy for Helmet and Person on the VOC dataset gradually improves, finally reaching 96.3% of mAP.

### 5 CONCLUSION

In this work, we focus on improving YOLOv5 by adding SPD, ASPP and BiFPN, achieving a 1% improvement relative to YOLOv5. 1) SPD allows the network to sample information with increased accuracy and enhanced information retention. 2) ASPP discards the receptive field idea in favor of the cavity convolution idea, which retains more valid information. 3) BiFPN delivers more semantic information to the deeper layers of the neural network, it avoids the loss of effective feature information. Our dataset is broader in terms of scenarios, considering various application scenarios in practice, focusing more on small targets as well as dense helmet images, and using higher criteria for factors such as lighting, viewing angle, weather, occlusion and background interference, which largely improves the efficiency of helmet detection. In our next step of work, we will further improve the current network, enhance the network's ability to obtain features of small targets, and accurately detect safety helmets from more angles during detection. At the same time, we will conduct practical applications and form a monitoring system to monitor whether the workers wear safety helmets when entering the work area, reducing the occurrence of safety accidents and economic property damage.

**Table 1: A side-by-side comparison in terms of MAP (%) with the baseline of Helmet Detection on VOC 2007.**

Method	Person	Helmet	MAP
YOLOv5-ASPP+SPD+BiFPN	96.1	96.2	96.3
YOLOv3	87.1	84.5	85.8
SSD	77.5	75.3	76.4
Faster-RCNN-resnet50	89.1	86.3	87.7
YOLOv3-tiny	63.2	59.8	61.5
YOLOv3-SPP	85.9	86.5	86.2
YOLOv4	90.1	90.1	89.6
YOLOvX-s	96.4	93.4	94.8

Demonstrate the accuracy of different methods for datasets of VOC 2007.

**Table 2: Ablation experiment in terms of MAP (%) from Helmet Detection on VOC 2007.**

Method	ASPP	BiFPN	SPD	Person	Helmet	MAP
YOLOv5				95.4	95.2	95.3
YOLOv5-ASPP	✓			96	95.4	95.7
YOLOv5-BiFPN		✓		95.6	94.5	95.1
YOLOv5-SPD			✓	96.2	95.2	95.7
YOLOv5-ASPP+BiFPN	✓	✓		95.2	95.8	95.5
YOLOv5-ASPP+SPD	✓		✓	96.2	95.6	95.9
YOLOv5-SPD+BiFPN		✓	✓	96.2	95.8	96
YOLOv5-ASPP+SPD+BiFPN	✓	✓	✓	96.1	96.2	96.3

Based on YOLOv5, we show the detection accuracy of ASPP, SPD and BiFPN modules after adding them respectively

## REFERENCES

- [1] Redmon J, Divvala S, Girshick R, *et al.* You only look once: Unified, real-time object detection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [2] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271
- [3] Redmon, J.; Farhadi, A. YoloV3: An incremental improvement. arXiv 2018, arXiv: 1804.02767.
- [4] Bochkovskiy A, Wang C Y, Liao H Y M. YoloV4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
- [5] Ultralytics-Yolov5. Available online: <https://github.com/ultralytics/yolov5> (accessed on 5 June 2022).
- [6] Tian Z, Shen C, Chen H, *et al.* Fcos: Fully convolutional one-stage object detection[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.
- [7] Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June; pp. 10781–10790.
- [8] LI, Yiting, *et al.* Research on a surface defect detection algorithm based on MobileNet-SSD. Applied Sciences, 2018, 8,9: 1678.
- [9] CHENG, Rao, *et al.* Multi-scale safety helmet detection based on SAS-YOLOv3-tiny. Applied Sciences, 2021, 11,8: 3652.
- [10] WARANUSAST, Rattapoom, *et al.* Machine vision techniques for motorcycle safety helmet detection. In: 2013 28th International conference on image and vision computing New Zealand (IVCNZ 2013). IEEE, 2013. p. 35-40.
- [11] LI, Kang, *et al.* Automatic safety helmet wearing detection. arXiv preprint arXiv: 1802.00264, 2018.
- [12] DASGUPTA, Madhuchhanda; BANDYOPADHYAY, Oishila; CHATTERJI, Sanjay. Automated helmet detection for multiple motorcycle riders using CNN. In: 2019 IEEE Conference on Information and Communication Technology. IEEE, 2019. p. 1-4.
- [13] ROHITH, C. A., *et al.* An efficient helmet detection for MVD using deep learning. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2019. p. 282-286.
- [14] DENG, Lixia, *et al.* A lightweight YOLOv3 algorithm used for safety helmet detection. Scientific reports, 2022, 12,1: 10981.
- [15] HUANG, Li, *et al.* Detection algorithm of safety helmet wearing based on deep learning. Concurrency and Computation: Practice and Experience, 2021, 33,13: e6234.
- [16] MOHAMMADI, Esmail, *et al.* Barriers and factors associated with the use of helmets by Motorcyclists: A scoping review. Accident Analysis & Prevention, 2022, 171: 106667.
- [17] LIU, Weiyang, *et al.* Large-margin softmax loss for convolutional neural networks. arXiv preprint arXiv:1612.02295, 2016.
- [18] XU, Zhanwei; WU, Ziyi; FENG, Jianjiang. CFUN: Combining faster R-CNN and U-net network for efficient whole heart segmentation. arXiv preprint arXiv: 1812.04914, 2018.
- [19] JIANG, Zicong, *et al.* Real-time object detection method based on improved YOLOv4-tiny. arXiv preprint arXiv:2011.04244, 2020.
- [20] WANG, Rui; SHI, Yijie; CAO, Wenming. GA-SURF: A new speeded-up robust feature extraction algorithm for multispectral images based on geometric algebra. Pattern Recognition Letters, 2019, 127: 11-17.
- [21] LIN, Tsung-Yi, *et al.* Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 2117-2125.

# An Intrusion Detection Model With Attention and BiLSTM-DNN

Yongcai Tao  
School of Computer and Artificial  
Intelligence, Zhengzhou University,  
Zhengzhou, China  
ieyctao@zzu.edu.cn

Jitao Zhang  
School of Cyber Science and  
Engineering, Zhengzhou University,  
Zhengzhou, China  
597778912@qq.com

Lin Wei  
School of Cyber Science and  
Engineering, Zhengzhou University,  
Zhengzhou, China  
weilin@zzu.edu.cn

Yufei Gao  
School of Cyber Science and  
Engineering, Zhengzhou University,  
Zhengzhou, China  
yfgao@zzu.edu.cn

Lei Shi  
School of Cyber Science and  
Engineering, Zhengzhou University,  
Zhengzhou, China  
ielshi@zzu.edu.cn

## ABSTRACT

*Abstract*—At present, machine learning and deep learning are often used for network traffic intrusion detection. In order to solve the problem of unfocused feature extraction in these methods and improve the accuracy of network intrusion detection, this paper proposes an intrusion detection model that combines Attention and BiLSTM-DNN(ABD). The model uses Attention to perform preliminary feature extraction on input data, reads the relationship between different features, then uses BiLSTM to extract long-distance dependent features, uses DNN to further extract deep-level features, and finally obtains classification through SoftMax classifier. The comparison experiment uses the NSL\_KDD data set, and models such as BiLSTM-DNN, support vector machine, decision tree and random forest are selected as the comparison experiment model. The experimental results show that the accuracy of the ABD is improved by 1.0% and 2.0% on the two-category and five-category tasks, respectively, which verifies the effectiveness of the method.

## CCS CONCEPTS

• Security and privacy; • Intrusion/anomaly detection and malware mitigation; • Intrusion detection systems;

## KEYWORDS

Network intrusion detection, Multi-head attention, Bi-directional long short-term memory, Deep neural network

### ACM Reference Format:

Yongcai Tao, Jitao Zhang, Lin Wei, Yufei Gao, and Lei Shi. 2023. An Intrusion Detection Model With Attention and BiLSTM-DNN. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590018>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590018>

## 1 INTRODUCTION

Network intrusion detection technology can identify and classify network traffic and detect malicious traffic. However, due to the continuous expansion of the scale of the Internet, the explosive growth of network traffic and the variety of network attack methods have put forward higher requirements for intrusion detection technology. How to improve the accuracy of detecting malicious traffic and how to efficiently distinguish different types of malicious traffic has become a top priority.

Traditional computer network security defense methods mainly include firewalls, anti-virus software, and network security hardware products [1]. Traditional methods have limitations and cannot adapt to the increasingly complex network environment and attack methods. The industry has introduced machine learning methods, such as Bayesian, support vector machine, and logistic regression algorithms. The methods based on machine learning realize the intelligent detection of network attacks and improves the efficiency of intrusion detection [2]. With the development of computer hardware technology, deep learning algorithms have brought breakthrough results in multimedia processing. Compared with traditional Advanced machine learning algorithm, deep learning improves detection efficiency, reduces false alarm rate, can automatically and intelligently identify attack characteristics, and helps discover potential security threats.

Therefore, we propose a new intrusion detection model. By introducing the attention mechanism, pay more attention between different features and between local features and overall features, and the internal relationship between features is explored. At the same time, the bidirectional long-short-time neural network is used to further consider the correlation between different features before and after, and then DNN is used to extract deep-level features, and then the SoftMax classifier is used to classify the output. In summary, the contributions of this paper are as follows:

- Propose a new intrusion detection method to improve the accuracy of intrusion detection.
- Fusion of global features and local features, with excellent feature extraction ability.
- Extensive experiments were performed on NSL\_KDD and optimal parameters were selected for the model.

## 2 RELATED WORK

At present, machine learning is widely used in various disciplines [3], and has attracted much attention in the field of network intrusion detection. Literature [4] [5] [6] is based on the improvement and optimization of the decision tree to realize the classification and processing of network intrusion data. Literature [7] [8] is based on random forest, combined with artificial immune algorithm and MapReduce distributed framework respectively, which reduces the false alarm rate of intrusion detection and improves the accuracy of intrusion detection.

Qi Ming yu et al. [9] used Principal Component Analysis (PCA) method to filter and reduce the dimensionality of KDD99 data set, and then trained with support vector machines (SVM), which reduced the detection time and improved the detection efficiency. Literature [10] considers the characteristics of data in three dimensions of time, space and content, and proposes a multi-dimensional feature fusion and overlay integration mechanism. Zeng Guo bin et al. [11] proposed an attribute value algorithm NBC, which uses the maximum a posteriori probability MAP of the possible category of unclassified nodes to determine the last node in the network model, but when applied to actual experimental detection, due to too many Interference factors and different feature node samples are different, resulting in low classification accuracy.

Machine learning methods can analyze the surface characteristics of data. Deep learning can mine deeper features of data to achieve better detection results. Literature [12] uses a preprocessing method combining dimensionality reduction technology and feature engineering to solve the problem of data imbalance, generate meaningful features. Wang Zhen dong [13] proposed an improved gray wolf algorithm to optimize the intrusion detection model of BP neural network in view of the problem of large randomness in the initial value of BP neural network. In the literature [14], based on the dimensional features and time series features of the data, the author first uses PCA to simplify the data features, and uses the stacked GRU detection model based on transfer learning to perform intrusion detection on the simplified features. Literature [15] proposes a genetic convolutional neural network model. First, a genetic algorithm combining KNN fitness function and fuzzy C-means clustering is used for feature selection to obtain an improved feature subset. Literature [16] constructed an integrated model by integrating multiple neural networks including building blocks and residual blocks. Liu Yue feng et al. [17] proposed a network intrusion detection method that combines convolutional neural network and bidirectional long-short-term network. Literature [18] [19] constructed detection models based on deep belief networks.

Many scholars have applied the deep learning method combined with the attention mechanism to intrusion detection. In the literature [20], the method of combining BiLSTM with the attention mechanism is proposed to classify the dataset NSL\_KDD. The model considers the influence of feature attributes before and after. Cao Lei [21] and others proposed a two-layer attention neural network model, which directly extracts features from the collected raw traffic data and captures key byte information and packet information. These studies have shown that applying deep learning to intrusion detection can improve detection efficiency and accuracy, and at

the same time, the introduction of attention mechanism can establish the connection between features and make the model focus on more important features. Therefore, this paper proposes an intrusion detection model ABD that combines Attention and neural networks.

## 3 METHOD

The ABD inputs the preprocessed data into the Attention module to establish the connection between different features, and extract richer feature information through multi-head attention. Then the data is input into the Bi-LSTM neural network to obtain the connection between the front and rear features, and finally the features are further extracted through the DNN and the SoftMax classifier is used to classify and identify the features to obtain the result. The process framework of the ABD is shown in Figure 1.

The NSL\_KDD dataset [22] used in this experiment trains and verifies the model. There are 9 data features in the data set that are discrete data, which are encoded using one-hot and inserted into the initial features for training as part of the whole.

### 3.1 Attention

This module mainly uses the Multi-Head Attention part of the Transformer structure, and its calculation formula (1) is as follows.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Q, K, and V represent the three matrices of Query, Key, and Value respectively,  $d_k$  is the dimension of Key. First calculate the result of the dot product of Q and K, multiply the result by  $\frac{1}{\sqrt{d_k}}$  to scale, use the SoftMax function to obtain the weights assigned to these values, and then perform the dot product operation with V to calculate the weighted sum. The structure is shown in Figure 2.

Before the multi-head attention operation, a linear change is made to Q, K, and V, and the input is projected to different spaces to enhance the generalization ability of the expression. Using h parallel attention operations allows the model to focus on different subspace information and extract richer features. Afterwards, the information is spliced and then output through a linear map. Its structure is shown in Figure 3. The calculation formula (2) of multi-head attention is as follows.

$$\begin{cases} Q_i = QW_i^Q, K_i = KW_i^K, V_i = VW_i^V & i = 1, \dots, 8 \\ head_i = Attention(Q_i, K_i, V_i) & i = 1, \dots, 8 \\ MultiHead(Q, K, V) = Concat(head_1, \dots, head_8)W^O \end{cases} \quad (2)$$

### 3.2 BiLSTM

The long-short-term memory neural network LSTM [23] is a recurrent neural network RNN that solves long-sequence and long-distance information loss. Its structural unit is shown in Figure 4.

The LSTM model is composed of six parts: input, cell state, hidden layer state, forget gate, memory gate, and output gate at time t. Represented by  $X_t, C_t, h_t, f_t, i_t$  and  $o_t$  respectively. The calculation process is to forget the useless information in the cell state, add new useful information, pass the integrated information to the next unit, and output the hidden layer state at every moment. The

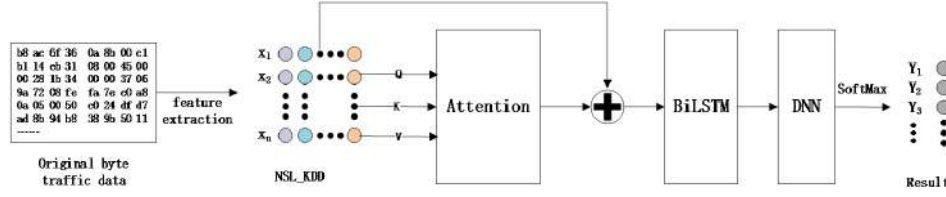


Figure 1: Process framework of ABD

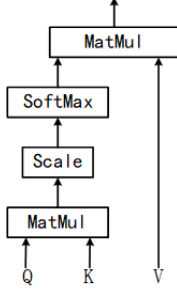


Figure 2: Attention mechanism

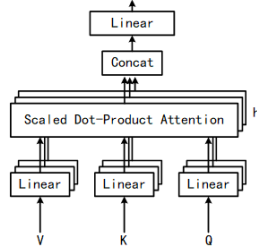


Figure 3: Multi-head attention

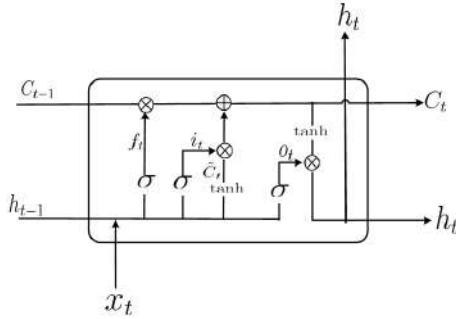


Figure 4: LSTM unit

corresponding calculation is shown in formula (3).

$$\begin{cases} f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t * \tanh(C_t) \end{cases} \quad (3)$$

Table 1: Details of Parameter initialization

Network Type	Parameter initialization distribution
Linear layer	Normal distribution $N(0,0.01)/N(0,0.1)$
LayerNorm layer	Fixed value 1
LSTM layer	Xavier uniform distribution

The bidirectional long short-term memory neural network is composed of a forward LSTM and a reverse LSTM. In this paper, this part splices the last of the output of the forward network and the backward network as the input of the next layer.

### 3.3 DNN

DNN is also commonly referred to as a multi-layer perceptron, which is a framework for deep learning. The network is divided into three parts: input layer, hidden layer and output layer. The DNN layer is fully connected between layers. The DNN in this model has two hidden layers, the input layer has 128 neurons, and the two hidden layers have 64 and 32 neurons respectively. Using RELU as the activation function, Dropout is set to 0.5. The general formula (4) for the calculation of DNN is as follows.

$$f(x) = \sigma(WX + b) \quad (4)$$

## 4 EXPERIMENTS

### 4.1 Parameter settings

The initialization of different experimental parameters will affect the model convergence speed and experimental results. In the parameter initialization of this experiment, the bias item initialization is set to 0, and the specific parameter initialization is shown in Table 1.

The weight parameters of the Linear layer in Table 1 are set to obey the normal distribution of  $N(0,0.01)$  in the binary classification task, and set to obey the normal distribution of  $N(0,0.1)$  in the five-classification experimental task. During the training process, the experiment uses the Adam optimizer to optimize the model parameters, which can automatically adjust the learning rate, but the initial learning rate still needs to be determined through experiments, otherwise it may directly converge to a poor local optimum.

**4.1.1 The effect of different learning rates on the classification results of the model.** The setting of the learning rate has an important impact on the classification results of the model. If the learning rate

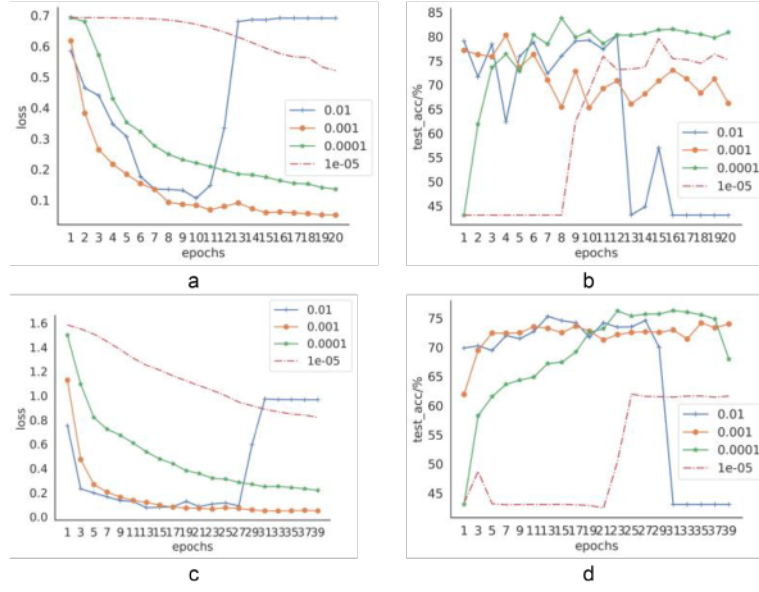


Figure 5: The influence of different learning rates on the effectiveness of experiments

is set too high, the model may not converge well, causing the loss value to oscillate back and forth on both sides of the optimal point, or even cross the optimal point. Setting the learning rate too small will reduce the speed of network optimization and increase the training time. In this experiment, 0.01, 0.001, 0.0001 and 0.00001 are selected as the candidate values of the learning rate respectively, and the optimal learning rate value is selected through experimental comparison. In Figure 5, a and b show the changes in the training set loss value and test set accuracy under different learning rates in the two-category task, and c and d show the changes in the training set loss value and test set accuracy rate under different learning rates in the five-category task.

Analyze a, b, c, and d. When the learning rate is set to 0.01, it can be seen from a and c that the loss value of the training set shows a downward trend at the beginning, whether it is a two-category task or a five-category task. When the epoch increases to a certain value, the loss value will rise linearly and remain stable, and the corresponding test set accuracy in b and d will also decrease linearly and then remain stable. When the learning rate is set to 0.0001, it can be seen from a and c that the loss value maintains a low rate of decline, and b and d show that the accuracy rate is poor. This shows that when the learning rate is set larger, the model may jump out of the optimal point at a certain moment, and when the learning rate is set smaller, the model training time will be prolonged and a better training effect cannot be achieved. When the learning rate is set to 0.001 and 0.0001, the loss value decreases steadily with the increase of epoch, but in terms of test set accuracy, the model with a learning rate of 0.0001 exceeds the model with a learning rate of 0.001. Therefore, 0.0001 is selected as the learning rate of the ABD in this paper.

**4.1.2 The effect of different weight parameter initialization settings on the model.** In the training process of the ABD, weight parameter

initialization has an important impact on the training of the model. The parameter initialization settings in this paper are shown in Table 1. The comparison experiment uses the default parameter initialization settings of PyTorch. That is, the initial weight parameters of the Linear layer obey the uniform distribution of Kaiming, the fixed value of the weight parameters of the LayerNorm layer is 1, and the weight parameters of the LSTM layer obey the uniform distribution. In Figure 6, a, b, c, and d represent the change curves of the loss value of the training set and the accuracy of the test set under different weight initialization settings in the two-category task and the five-category task, respectively.

Analyzing a, b, c, and d in Figure 6, it can be clearly seen that the weight parameters are initialized in the way of this paper, the loss value of the training set decreases more smoothly and stably, and the accuracy of the test set is also greatly improved. However, initializing the weight parameters in the default way may cause the model to converge to the wrong local optimum at the initial moment, resulting in large fluctuations in the accuracy of the test set.

## 4.2 Experiments result

This paper uses the commonly used machine learning methods Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) as comparative experimental methods. At the same time, the BiLSTM+DNN model is used as a comparative experiment to explore the role of the attention part in the model.

**4.2.1 Two classifications.** In this paper, 20 epochs are set in the experiment of two classifications. The test set accuracy results of each ABD training experiment are different, and the accuracy results are distributed between 80% and 84%. Therefore, this paper trains the ABD 10 times, and finally takes the average value of all evaluation indicators for 10 times as the final result, and uses

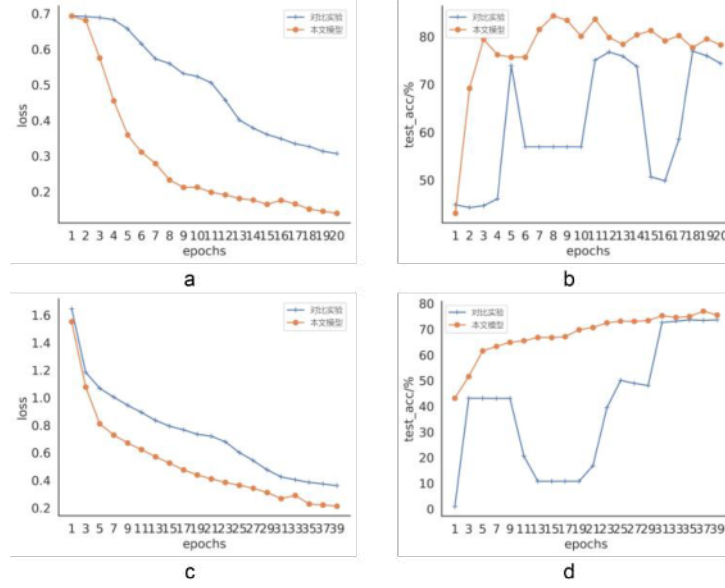


Figure 6: The influence of different weight initializations on the effect of the experiment

Table 2: Effect comparison of Two-class model

Model	Evaluation index			
	Accuracy	Precision	Recall	F1-score
SVM	0.76	0.96	0.61	0.75
RF	0.77	0.96	0.62	0.76
DT	0.81	0.97	0.69	0.80
BiLSTM-DNN	0.76	0.96	0.61	0.74
ABD	0.82	0.95	0.72	0.82

Table 3: Effect comparison of Five-class model

Model	Evaluation index			
	Accuracy	Precision	Recall	F1-score
SVM	0.70	0.42	0.43	0.42
RF	0.74	0.75	0.47	0.47
DT	0.74	0.71	0.51	0.51
BiLSTM-DNN	0.73	0.45	0.44	0.43
ABD	0.76	0.47	0.50	0.48

this result for comparative analysis with other models. The effect comparison table of the two classification models is shown in Table 2.

By comparing the indicators of the five model methods, it can be seen from Table 2 that the model in this paper has a better experimental effect in terms of accuracy. Compared with the BiLSTM+DNN model that removes the attention module, the experimental effect has been significantly improved.

**4.2.2 Five classifications.** In this paper, 40 epochs are set in the five-category experiment. Similar to the experimental process of

two classifications, the experimental results of five classifications are shown in Table 3 below.

According to Table 3, it can be seen that the model in this paper has significant advantages in terms of accuracy, but due to the unbalanced distribution of various data in the initial data set, the detection rates of the U2R and R2L attack categories are low, and the precision and recall indicators are relatively low. The performance of RF and DT on the precision index is significantly better than other models, mainly because the tree model is not sensitive to the problem of sample imbalance. As shown in the heat map of the confusion matrix in Figure 7, where a, b, and c respectively represent

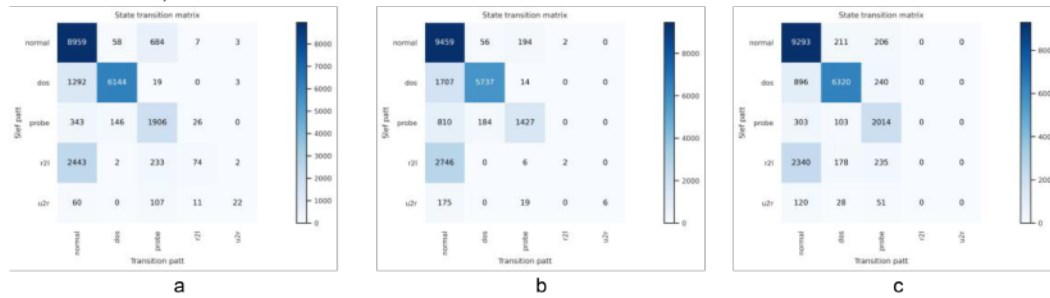


Figure 7: Heat map of confusion matrix

the DT, RF, and ABD, it can be seen that for the classification of r2l and u2r, both DT and RF can correctly classify some minority sample data. But taking RF as an example, its precision index can reach 0.5 and 1 in the classification of r2l and u2r, which directly leads to a high overall precision index, but this does not mean that it is better in multi-classification. The superiority of the ABD is proved by combining the results of the two-category experiment and the five-category experiment. However, due to the unbalanced distribution of various types of data in the initial data set, the detection rates of U2R and R2L attack classifications are low.

## 5 CONCLUSIONS

This paper proposes a model ABD that combines Attention and BiLSTM. The Attention part of the method focuses on the connection between different attribute features through a multi-head attention mechanism, and extracts more abundant features. The BiLSTM-DNN part has further deep-level features. Compared with traditional machine learning methods, the accuracy of the ABD has been improved. By manually setting the distribution of parameter initialization, the training process of the model can be made more stable. However, due to the uneven distribution of various sample data in the NSL\_KDD dataset, the detection rate of some classifications is low, and the ABD has a problem of long training time. In the future, we can carry out further research work on the model to address these issues.

## ACKNOWLEDGMENTS

This work was supported in part by the Key Project of the National R&D Program of China (2018YFB1701401, 2018YFB1701400, 2020YFB1712401)

## REFERENCES

- [1] Yang Yang. 2017. Analysis of Network Security Incident Association Analysis Technology[J]. Network Security Technology & Application, 2017(08):14+30.
- [2] Kun Zhu, Qi Zhang. 2017. Application of Machine Learning in Network Intrusion Detection, College of Computer Science and Technology, 2017, 32(03):479-488. DOI:10.16337/j.1004-9037.2017.03.006.
- [3] Qiao Wu. 2020. Application of machine learning in network security intrusion detection[J]. Technology Innovation and Application, 2020(25):1-4.
- [4] Guo quan Zhang, Wen li Li. 2009. Intrusion detection based on decision tree with mutual information, Journal of Liaoning Technical University(Natural Science), 2009, 2802:273-276.
- [5] Yuan fang Pu , Hong le Du. 2010. Research and Application of Decision Tree in Network Intrusion Detection, Computer Knowledge and Technology, 2010, 607:1560-1563.
- [6] Ming qiu Song, Yun Fu, Gui shi Deng. 2007. Intrusion detection via decision tree and protocol analysis, Application Research of Computers, 2007, 12:171-173, 176.
- [7] Ling Zhang, Jian wei Zhang, Yong xuan Sang, Bo Wang, Ze xiang Hou. 2020. Intrusion Detection Algorithm Based on Random Forest and Artificial Immunity[J]. Computer Engineering, 2020, 46(08):146-152. DOI:10.19678/j.issn.1000-3428.0057085.
- [8] Ji jun Guo , Jun hua LI , Chen Chen , Yiming Chen , Yida Lv. 2020. Network Intrusion Detection Method Based on Random Forest[J]. Computer Engineering and Applications, 2020, 56(02):82-88.
- [9] QI Ming-yu , LIU Ming , FU Yan-ming. Research on Network Intrusion Detection Using Support Vector Machines Based on Principal Component Analysis, [J]. Netinfo Security, 2015(02):15-18.
- [10] Hao Zhang, Jie-Ling Li, Xi-Meng Liu, Chen Dong, Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection, Future Generation Computer Systems, Volume 122, 2021, Pages 130-143.
- [11] Guo bin Zeng, Zhao chun Ran. 2017. Based on the study of improving the intrusion detection system of the simple Bayesian algorithm[J]. Computer Knowledge and Technology, 2017, 1328:28-29.
- [12] Al-Turaiki, Isra & Altwaijry, Najwa. 2021. A Convolutional Neural Network for Improved Anomaly-Based Network Intrusion Detection. Big Data. 9. 233-252. 10.1089/big.2020.0263.
- [13] Zhen dong Wang, Yao di Liu, Zhong dong Hu, Da hai LI, Jun ling Wang. 2021. Use Improved Grey Wolf Algorithm to Optimize BP Neural Network Intrusion Detection[J]. Journal of Chinese Computer Systems, 2021, 42(04):875-884.
- [14] Nongmeikapam Brajabidhu Singh, Moirangthem Marjit Singh, Arindam Sarkar, Jyotsna Kumar Mandal. 2021. A novel wide & deep transfer learning stacked GRU framework for network intrusion detection, Journal of Information Security and Applications, Volume 61, 2021, 102899.
- [15] Minh Tuan Nguyen, Kiseon Kim. 2020. Genetic convolutional neural network for intrusion detection systems, Future Generation Computer Systems, Volume 113, 2020, Pages 418-427.
- [16] F. Folino, G. Folino, M. Guarascio, F.S. Pisani, L. Pontieri, 2021. On learning effective ensembles of deep neural networks for intrusion detection, Information Fusion, Volume 72, 2021, Pages 48-69.
- [17] Yue feng Liu, Shuang Cai, Han xi Yang, Chen rong Zhang. 2019. Network Intrusion Detection Method Integrating CNN and BiLSTM[J]. Computer Engineering, 2019, 45(12):127-133. DOI:10.19678/j.issn.1000-3428.0053263.
- [18] Pan Wang, Xue hua Song , Chang da Wang, Feng Chen, Xia qiang Xu, Guan yu Cai. 2020. Intrusion Detection Method Based on Improved Deep Belief Network[J]. Computer Engineering and Applications, 2020, 56(20):87-92.
- [19] Jing ming Xia, Chun jian Ding, Ling Tan. 2020. Deep belief network intrusion detection method based on grey wolf algorithm[J]. Computer Engineering and Design, 2020, 41(06):1534-1539. DOI:10.16208/j.issn1000-7024.2020.06.006.
- [20] Hao Shu, Chen Wang, Yin. Shi. 2020. Intrusion detection based on BiLSTM and attention mechanism[J]. Computer Engineering and Design, 2020, 41(11):3042-3046. DOI:10.16208/j.issn1000-7024.2020.11.007.
- [21] Lei Cao, Zhan bin LI, Yong sheng Yang , Long fei Zhao. 2021. Intrusion Detection Method Based on Two-Layer Attention Networks[J]. Computer Engineering and Applications, 2021, 57(19):142-149.
- [22] Tavallae, Mahbod & Bagheri, Ebrahim & Lu, Wei & Ghorbani, Ali. 2009. A detailed analysis of the KDD CUP 99 data set. IEEE Symposium. Computational Intelligence for Security and Defense Applications, CISDA. 2. 10.1109/CISDA.2009.5356528.
- [23] Xiaodan Zhu, Parinaz Sobhani, and Hongyu Guo. 2015. Long short-term memory over recursive structures. In Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37 (ICML'15). JMLR.org, 1604-1612

# Deep Reinforcement Learning with Copy-oriented Context Awareness and Weighted Rewards for Abstractive Summarization

Caidong Tan

2020171143@mail.hfut.edu.cn

Hefei University of Technology

Hefei, China

## ABSTRACT

This paper presents a deep context-aware model with a copy mechanism based on reinforcement learning for abstractive text summarization. Our model is optimized using weighted ROUGE as global prediction-based rewards and the self-critical policy gradient training algorithm, which can reduce the inconsistency between training and testing by directly optimizing the evaluation metrics. To alleviate the lexical diversity and component diversity problems caused by global prediction rewards, we improve the richness of the multi-head self-attention mechanism to capture context through global deep context representation with copy mechanism. We conduct experiments and demonstrate that our model outperforms many existing benchmarks over the Gigaword, LCSTS, and CNN/DM datasets. The experimental results demonstrate that our model has a significant effect on improving the quality of summarization.

## CCS CONCEPTS

• Computing methodologies → Artificial intelligence; Natural language processing.

## KEYWORDS

abstractive summarization, reinforcement learning, deep context-aware, copy mechanism

## ACM Reference Format:

Caidong Tan . 2023. Deep Reinforcement Learning with Copy-oriented Context Awareness and Weighted Rewards for Abstractive Summarization. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590019>

## 1 INTRODUCTION

The goal of text summarization is to produce a concise, high-quality text snippet from the input document, and existing methods are mainly divided into abstractive and extractive. The extractive summaries can select keywords or key phrases from the original text and combine them into a summary, while the abstractive summaries can generate new words and both methods have their

advantages and disadvantages. In recent years, text summarization methods have been based on the Seq2Seq framework, incorporating attention mechanisms, such as efficient pointer generator [17] and multi-head self-attention mechanism [19]. However, there are some unavoidable problems with its use for summary generation, which generally uses cross-entropy (XENT) loss and is trained iteratively using the teacher’s forcing algorithm [2] for local and token-level. It introduces the mismatch between the objective function and the ROUGE [11] evaluation metrics, suffers from exposure bias, and is unfriendly to the sentence generation task [15].

The reinforcement learning (RL) paradigm can provide solutions for summary generation models based on the Seq2Seq frameworks, such as self-critical policy gradient training algorithms which add non-trivial ROUGE metrics as rewards to reinforcement learning losses to alleviate mismatch and exposure bias [7, 13]. Nevertheless, using the ROUGE metrics as a reward for joining the reinforcement learning (RL) paradigm for optimization is not robust to different words with similar meanings and does not capture the lexical diversity, and compositional diversity of natural language well [10].

The transformer is based on the self-attention mechanism, effectively modeling long-term and short-term dependencies. Existing approaches use the pointer generator mechanism combined with LSTM [4] or Transformer [19]. However, LSTM cannot capture long-term dependencies. The transformer needs to perceive better the mutual information between context and semantics (grammatical awareness), which does not consider context information when calculating the correlation between elements. Yang et al. [23] have shown that context information can enhance the ability of attentional models to model dependencies between neural representations.

Therefore, we propose a deep context-aware approach combined with a pointer mechanism, using ROUGE with weights as a reward using reinforcement learning to train a model that increases the diversity of words and composition of the generated summaries and improves the quality of summary generation. Second, our method can rely only on reinforcement learning loss for training instead of using a loss training model with a combination of cross-entropy and reinforcement learning, which can better alleviate the exposure bias problem. Finally, we experimentally explored the impact of different weighted ROUGE metric scores as a reward for training the model.

## 2 RELATED WORK

Most of the existing text summarization methods are based on the Seq2Seq structure and incorporate some different attention mechanisms [17, 20]. Encoder and Decoder are built using LSTM or Transformer model based entirely on a self-attention mechanism

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590019>

[9]. Compared with LSTM, the Transformer has several advantages, including efficient training with parallel computing and effective self-attention capturing long and short dependencies.

Ranzato et al. [15] proposed a hybrid training objective based on cross-entropy (XENT) and reinforcement learning (RL) in order to alleviate the training-test mismatch problem caused by the traditional Seq2Seq framework as a generative model. However, it introduces problems such as high variance. Bahdanau et al. [1] proposed the actor-critic method to generate sequences to solve the previous problem of high variance of mixed targets, which has been widely used. However, to ensure better convergence, other networks need to be constructed to assist. Kryscinski et al. [8] proposed that the self-critical policy gradient training algorithm can solve the previous problem very well.

### Background

Cross-entropy (XENT) loss is generally used when training the summary model with the Seq2Seq-based framework. The effectiveness of summary generation is measured in the testing phase using the non-trivial ROUGE as an evaluation metric.

$$L_{XENT} = - \sum_{t=1}^T \log P(\hat{y}_t | \hat{y}_{1 \sim t-1}, x) \quad (1)$$

where  $x$  denotes the input of source text,  $\hat{y}_{1 \sim t-1}$  denotes the sequence already decoded before time step  $t$ , and  $\hat{y}_t$  denotes the word to be decoded at time step  $t$ .

In RL methods, the goal is to find a set of parameters  $\theta$  that maximizes the returned reward value, which is expressed mathematically as the expected value. The specific form is defined as follows:

$$L_{RL} = - \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \log P(\hat{y}_t | \hat{y}_{1 \sim t-1}, x) \quad (2)$$

$$\times \{r_{ROUGE}(\hat{y}_{1 \sim t}) - r_{ROUGE}(\hat{y}_{1 \sim t}^g)\}$$

where  $\hat{y}_{1 \sim t}$  is a sequence of actions from the policy, and  $\hat{y}_{1 \sim t}^g$  denotes the sequence obtained using the greedy approach.  $r_{ROUGE}(\ast)$  denotes the calculation of ROUGE-L score.

The training will start with the cross-entropy (XENT) loss function to obtain a relatively good pretraining model, and then use the cross-entropy (XENT) and reinforcement learning (RL) mixed loss function for training. [13]:

$$L_{total} = \mu L_{XENT} + (1 - \mu) L_{RL} \quad (3)$$

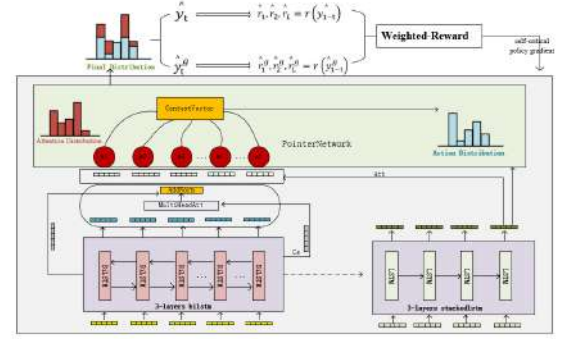
where  $\mu$  is a hyperparameter. This optimization objective combines XENT and RL mixed loss function, but this will also introduce mismatch and exposure bias problems.

Found in practice, using only ROUGE-L as the reward score depresses the ROUGE-2 score, so multiple measures were introduced for balance. Huang et al. [6] and Xiao et al. [22] raised ROUGE scores by combining them with multiple choice completion scores as a reward function, where ROUGE scores were also combined considering ROUGE-1, ROUGE-2, and ROUGE-L.

$$r_{ROUGE}(\ast) = \lambda r_{ROUGE-1}(\ast) + \eta r_{ROUGE-2}(\ast) + \varepsilon r_{ROUGE-L}(\ast) \quad (4)$$

where  $\lambda^2 + \eta^2 + \varepsilon^2 \neq 0$ , and  $\lambda + \eta + \varepsilon = 1$ . During the training phase, we can set the value of  $\lambda, \eta, \varepsilon$  to adjust reward.

## 3 METHODOLOGY



**Figure 1: The overview of our model. The model includes a deep LSTM with a multi-head self attention module and a pointer generator module, and uses Weighted-Rewards to optimize the model. The  $\hat{y}_{1 \sim t}$  is a sequence of actions from the policy, and  $\hat{y}_{1 \sim t}^g$  denotes the sequence obtained using the greedy approach.**

### 3.1 Deep copy-oriented context awareness

In order to solve the problem that the automatic evaluation metric of ROUGE is not robust to different words with similar meanings and cannot capture the lexical diversity and component diversity of natural language, we optimize the pointer network model. To enable the pointer generator to more accurately replicate the valid information in the source text, we chose a three-layer bidirectional LSTM as the encoder. We added a multi-head self-attention mechanism, combining both advantages to encode the contextual semantics at a deeper level and capture the mutual semantic information between words.

Overall, we define the summary problem as converting a long sequence to a short sequence which requires a high degree of semantic consistency. Specifically, we stipulate that the long text input sequence  $X$  of length  $N$ :  $\{x_1, \dots, x_N\}$ , each of  $x_i$  in brackets represents a specific word in the input text. The output summary  $Y$  of length  $N'$ :  $\{y_1, \dots, y_{N'}\}$ , and  $y_i$  is the word in the output. Our goal is to find a suitable set of parameters for the model to better map the input to the desired output:  $f_\theta : \{X_{1 \sim N} \xrightarrow{\theta} Y_{1 \sim N'}\}$ . Formally,

$$Y = \{y_1, \dots, y_{N'}\} = f_\theta(X_{1 \sim N}) = f_\theta(\{x_1, \dots, x_N\}) \quad (5)$$

#### Encoder-Decoder

Our model uses three layers of bidirectional LSTM encoders to construct a global representation of input sequences and encode deep semantics. The bidirectional LSTM can represent the relative position information of words in the input sequence, so it can make up for the defect that the multi-head self-attention mechanism cannot represent the position information. The long text input sequence passes through the bidirectional LSTM to generate the encoded

vector at the corresponding position. We use  $\{h_{f1}, \dots, h_{fn}\}$  to represent the set of forward coded implicit vectors, and  $\{h_{b1}, \dots, h_{bn}\}$  represents the set of backward coded implicit vectors. The whole is expressed as  $h_i = Bi \cdot LSTM(h_{i-1}, x_i)$ , each  $x_i$  in the sequence is represented by a series of forward and backward implicit vectors, eg  $h_i^e = [h_{fi}, h_{bi}]$ . The output context in the coding phase is represented by  $H_e = [h_1^e, \dots, h_N^e] = DeepLSTM(E_{x1}, \dots, E_{xN})$ . On the de-coder side, we also use three layers stacked LSTM, where the output of each time step  $t$  is represented as  $s_t^d = StackLSTM(s_{t-1}^d, y_{t-1})$ .

The LSTM can encode global and deep semantics of the input text well, but its ability to capture long-term dependencies is limited, especially for digest tasks. Therefore, the multi-head self-attention mechanism [19] remedies this defect. The multi-head self-attention mechanism is very effective for learning long-term dependence. It allows a direct connection between each pair of positions, and each position will pay attention to all past positions to improve its representation. Based on deep context encoding of the source text, we fuse syntactic and semantic information of different types with multi-head self-attention mechanisms and then calculate scores with the common attentional mechanism at each decoding step. We use  $u$  to represent the output of multi-head self-attention. The final formula of prior distributions is as follows:

$$P_{vocab}(w) = softmax(W^O[C_{contextVec}, s_t^d] + b) \quad (6)$$

part of the formula is explained as follows:

$$C_{contextVec} = \sum_{i=1}^N \frac{s_t^d W_a u_i}{\sum_{j=1}^N e_j^t} \cdot u_i \quad (7)$$

the output of the multi-head self-attention mechanism is expressed as:

$$\begin{aligned} u &= AddNorm(Z_u + H_e) \\ &= AddNorm(MultiHead(Q, K, V) + H_e) \\ &= AddNorm(Concat(head_{1 \sim k}) \cdot W^z + H_e) \end{aligned} \quad (8)$$

Where  $head_i = softmax(\frac{QW_i^Q \times (KW_i^K)^T}{\sqrt{d_k}})VW_i^V$ ,  $V = W^V \cdot H_e$  and  $\begin{bmatrix} Q \\ K \end{bmatrix} = H_e \begin{bmatrix} W^Q \\ W^K \end{bmatrix} + \bar{H}_e \begin{bmatrix} U^Q \\ U^K \end{bmatrix}$ . the  $\bar{H}_e$  represents the global context of the input layer function, which represents the global meaning of the sequence. We use mean arithmetic to summarize the representation of the input layer, which is commonly used in Seq2Seq models. The  $W^O, W^Q, W^K, W^V, U^Q, U^K$  are trainable hyperparameters.

#### Pointer Generator

Attention is used as a pointer to determine the probability of generating words from word distribution and source text. Traditional pointer generator mechanisms can only generate surface words from source documents. We consider the representation of contextual information for each word using the multi-head self-attention mechanism. Therefore, when generating the summary, the words generated by using the pointer generator are more meaningful, focusing not only on the words on the surface of the source document but also on capturing the lexical diversity and component diversity of natural language.

The pointer generation mechanism defines a pointer probability  $P_{gen}$ .

$$P_{gen} = \sigma(W^G[s_t^d, u_t, y_t^d] + b_u) \quad (9)$$

$$P_{final}(w) = P_{gen} \times P_{vocab}(w) + (1 - P_{gen}) \times \sum_{i=1}^N \frac{e_i^t}{\sum_{j=1}^N e_j^t} \quad (10)$$

where  $e_i^t = s_t^d W_a u_i$  in the part of Equation 10.

## 4 EXPERIMENTAL

### 1) Dataset

We use three common datasets to verify the effectiveness of our method. First, we choose the anonymous English Gigaword dataset and process it in the same way as [16], which contains 3.8 million text summary pairs with 189k validation sets and 1,951 text summary pairs after processing. Second, we chose A large Chinese short Text abstract (LCSTS) dataset collected and constructed by the Chinese microblogging site Sina Weibo. We refer to Hu et al. [5] and divide the dataset into training, validation, and test sets. We used the first part of the LCSTS dataset for training, which contained 2.4 million text summary pairs, and selected 725 text summary pairs with high annotation scores from the last part as our test dataset. Finally, we also chose the long-text English summary dataset CNN/DM. We obtained a non-anonymous version of the data using the data processing method provided by See et al. [17], which included 287,227 training pairs, 13,368 verification pairs, and 11,490 test pairs.

### 2) Implementation

We first pre-train the Seq2Seq model using the cross-entropy (XENT) loss function and then select the model parameters that perform best on the validation dataset to initialize the RL model. Our model can then be trained directly using equations (2) and (4) instead of equation (3), which does not introduce cross-entropy loss and thus reduces the impact of exposure bias. The model is built on the Pytorch framework with a word embedding dimension of 256 and a hidden embedding dimension of 512. Adam optimizer was selected and is trained on a 3090 GPU. For the  $\lambda$ ,  $\eta$ , and  $\epsilon$  we choose 0.6, 0.2, and 0.2, respectively, and we obtain more competitive results on this set of weights.

### 3) Evaluation Metrics

We chose the ROUGE score to evaluate the metrics for comparison [11]. The ROUGE score calculates the degree of overlap between generated summaries and references, including N-grams. The F1 scores of ROUGE-1, ROUGE-2, and ROUGE-L were used as evaluation metrics. ROUGE-1 and ROUGE-2 are used to evaluate the amount of information, while ROUGE-L is used to evaluate sentence fluency. In addition, we compare the average ROUGE F1 score for each model.

## 4.1 Results

To better verify the effectiveness of our proposed methods, the large-scale pretraining model was not introduced into our model. We introduce the seq2seq summary model based on the attention mechanism as the baseline model and also select some newer models for comparison. In the table of the paper, the Base (XENT) denotes training our model based on cross-entropy loss function, the Mixed denotes training our model based on mixed loss of cross-entropy in equation (3), and RL denotes training our model in the combination of equation (2) and (4).

**Table 1: Evaluation results on Gigaword test dataset based on ROUGE F1 metric.**

Models	ROUGE-1	ROUGE-2	ROUGE-L
ABS(beam) [16]	37.41	15.87	34.70
SEASS(beam) [24]	46.86	24.58	43.53
DSR+ROUGE [10]	-	-	42.95
DSR+XENT [10]	-	-	42.29
Reinforced-ConvS2S [21]	46.68	24.22	43.76
<b>Base(XENT)</b>	46.20	24.65	44.21
<b>Mixed</b> ( $\lambda = 0.0, \eta = 0.0, \varepsilon = 1.0$ )	46.28	24.52	44.91
<b>RL</b> ( $\lambda = 0.6, \eta = 0.2, \varepsilon = 0.2$ )	<b>47.10</b>	<b>25.05</b>	<b>45.64</b>

**Results on Gigaword.** The evaluation results of Gigaword’s 1,951 test dataset based on the ROUGE F1 metrics are compared with various generative text summary models. The baselines are as follows, and results are shown in Table 1.

- **ABS [16]**: This model introduces an attention mechanism to generate text summaries.
- **SEASS [24]**: This model expands the sequence to sequence structure by incorporating a selective encoder, and we choose the model using beam strategy for comparison.
- **Reinforced-ConvS2S [21]**: This model is a sequence-to-sequence convolution model based on the reinforcement learning paradigm.
- **DSR+ROUGE/XENT [10]**: This model adds distributed semantic reward to the reinforcement learning formula for training.

According to the experimental results in table 1, the proposed method is better than the previous method based on the reinforcement learning paradigm in improving the accuracy of text summarization. Our proposed model achieves the best performance in all three ROUGE scores, indicating that the generated summary is of higher quality.

**Results on LCSTS.** For the large-scale Chinese dataset LCSTS, Huet et al. [5] provide two preprocessing methods. We choose character-based processing and reference the experimental results in the original paper as the baseline. In addition, some representative algorithms are also selected for comparison. The baselines are as follows.

- **Transformer [3]**: This model first report on Transformer’s performance on the LCSTS dataset.
- **ProphetNet-Zh [14]**: This model is a Chinese model trained with ProphetNet as a pretraining method.

The experimental results are shown in Table 2, our approach achieves optimal results on two metrics (ROUGE-1 and ROUGE-L), and outperforms Transformer and ProphetNet-Zh.

**Results on CNN/DM.** For the CNN/DM dataset, in addition to the partial baselines mentioned in the Gigaword and the LCSTS, we also added baselines based on the transformer model correlation for comparison. The baselines are as follows, and the results are shown in Table 3, and our method achieves competitive results in ROUGE-1 and ROUGE-L.

- **Pointer-Generator [17]**: The model is based on the Seq2Seq framework and introduces the pointer network mechanism,

**Table 2: Evaluation results on LCSTS dataset based on ROUGE F1 metric.x**

Models	ROUGE-1	ROUGE-2	ROUGE-L
Transformer [3]	42.35	29.38	39.23
Reinforced-ConvS2S [21]	42.61	<b>29.79</b>	40.03
ProphetNet-Zh [14]	42.32	27.33	37.08
<b>Base(XENT)</b>	41.04	28.07	39.78
<b>Mixed</b> ( $\lambda = 0.0, \eta = 0.0, \varepsilon = 1.0$ )	41.47	28.11	40.05
<b>RL</b> ( $\lambda = 0.6, \eta = 0.2, \varepsilon = 0.2$ )	<b>42.82</b>	29.26	<b>40.99</b>

**Table 3: Evaluation results on CNN/DM dataset based on ROUGE F1 metric.**

Models	ROUGE-1	ROUGE-2	ROUGE-L
Pointer-Generator [17]	39.53	17.28	36.38
SENECA [18]	41.52	18.36	38.0
TransformerABS [12]	40.21	17.76	37.09
BERTSUMABS [12]	41.72	<b>19.39</b>	38.76
<b>Base(XENT)</b>	39.51	16.86	36.63
<b>Mixed</b> ( $\lambda = 0.0, \eta = 0.0, \varepsilon = 1.0$ )	40.00	17.42	37.03
<b>RL</b> ( $\lambda = 0.6, \eta = 0.2, \varepsilon = 0.2$ )	<b>42.25</b>	18.72	<b>39.03</b>

**Table 4: Comparison of experimental results with different weights on the Gigaword dataset.**

Hyperparameters	ROUGE-1	ROUGE-2	ROUGE-L
$\lambda = 1.0, \eta = 0.0, \varepsilon = 0.0$	48.32	<b>22.04</b>	47.13
$\lambda = 0.0, \eta = 1.0, \varepsilon = 0.0$	46.39	25.10	44.32
$\lambda = 0.0, \eta = 0.0, \varepsilon = 1.0$	47.90	<b>22.86</b>	45.95
$\lambda = 0.6, \eta = 0.0, \varepsilon = 0.4$	46.96	24.29	45.63
$\lambda = 0.6, \eta = 0.2, \varepsilon = 0.2$	47.10	24.52	45.64

so the model can replicate the source text to some extent and can solve the OOV problem well.

- **SENECA [18]**: A coherent abstraction framework system that generates informative and coherent summary using entity information and introduces RL for training.
- **TransformerABS [12]**: An abstractive Transformer baseline based on CNN/DM dataset, the experimental results were first announced by Liu et al [12].
- **BERTSUMABS [12]**: This model is a new document-level encoder based on BERT. It is able to encode the semantics of the document well and capture the semantic information of the document to obtain a better vector representation of the sentences.

#### Parameters Setting

We selected different parameters for validation on the Gigaword dataset to compare the effects of different parameters on the metrics, and the experimental results are shown in Table 4.

As seen from the data in table 4, the desired results cannot be obtained when training with just one metric as a reward. When training with a combination of all metrics as a reward, a better level can be achieved. When the scores of ROUGE-1 and ROUGE-L are

**Table 5: Ablation study on the LCSTS dataset and Gigaword dataset.**

	Gigaword				LCSTS			
	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-AVG	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-AVG
baseline	44.74	24.00	42.85	37.20	38.87	25.46	36.14	33.49
+deepaware (Ours)	46.20	24.65	44.21	38.69	41.04	28.07	39.78	36.30
+deepaware (Ours) +mixed	46.28	24.52	44.91	38.75	41.47	28.11	40.05	36.54
+deepaware +RL622 (Ours)	<b>47.10</b>	<b>25.05</b>	<b>45.64</b>	<b>39.09</b>	<b>42.82</b>	<b>29.26</b>	<b>40.99</b>	<b>37.69</b>

high and the scores of ROUGE-2 are relatively low, the generated summaries have duplicated words or phrases. Therefore, for both LCSTS and CNN/DM datasets, we also choose the optimal set of parameters in the table for the experiments. The optimal results are generated in this set of parameters.

#### Ablation Study

We also conduct ablation experiments on the Gigaword and the LCSTS datasets to analyze the gain impact of the deep copy-oriented context awareness module and the optimized reinforcement learning loss on system performance. We remove the deep copy-oriented context awareness module based on the entire model in the ablation experiments. Then we use the Mixed traditional loss and the optimized reinforcement learning loss to conduct multiple experiments and compare them to verify the effectiveness of our method. In Table 5, the first row represents the experimental results of the baseline model built without deep copy-oriented context awareness module based on XENT loss and Mixed loss. The RL in the table represents reinforcement learning loss. The optimized reinforcement learning loss compared to the Mixed loss, and +RL622 means the selection of parameters  $\lambda = 0.6$ ,  $\eta = 0.2$ ,  $\varepsilon = 0.2$  for RL training respectively.

#### Analysis

The effectiveness of our method can be seen from Table 1, Table 2 and Table 3 as a whole. In Table 5, by comparing the metrics on the LCSTS dataset, we can see the advantage of the deep copy-oriented context awareness module, which increases by 2.17, 2.61, and 3.64 for ROUGE-1, ROUGE-2 and ROUGE-L, respectively, based on the baseline. When deep copy-oriented context awareness module and RL622 are combined and compared to baseline, ROUGE-1, ROUGE-2 and ROUGE-L are improved by 3.95, 3.80 and 4.85, respectively. Comparing the Mixed and the RL622, the results show that the model improvement is more significant when combined with our RL622 training, which illustrates the effectiveness of our combined approach, better results with RL622 than Mixed. Further analysis, adding a depth-aware replication mechanism that can extract key information from the deep semantics of the original document using a pointer mechanism to include the information as part of the summary. Secondly, training using RL622 loss better reduces exposure bias because Mixed loss introduces cross entropy loss. Finally, training based entirely on RL622 loss is not robust to different words with similar meanings and does not capture the lexical diversity, which using deep copy-oriented context awareness can compensate for this deficiency.

## 5 CONCLUSION

In this study, we propose a deep context-aware approach combined with a pointer mechanism, using ROUGE with weights as a reward using reinforcement learning to train a model that increases the diversity of words and composition of the generated summaries and improves the quality of summary generation. Our method can rely only on reinforcement learning loss for training. Experiments show that this mechanism introduces some high-level contextual information for summarization. In addition, our model can produce summaries with good information, consistency, and diversity. We find that existing methods for evaluating summary quality have limitations and remain an open problem, and research on this issue is also of great importance. In the future, we aim to propose more representative evaluation metrics and try them out in conjunction with the reinforcement learning paradigm.

## REFERENCES

- [1] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2016. An actor-critic algorithm for sequence prediction. *arXiv preprint arXiv:1607.07086* (2016).
- [2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems* 28 (2015).
- [3] C Chang, C Huang, and Jane Yungjen Hsu. 2018. A hybrid word-character model for abstractive summarization. *arXiv preprint arXiv:1802.09968* (2018).
- [4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [5] Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. Lcsts: A large scale chinese short text summarization dataset. *arXiv preprint arXiv:1506.05865* (2015).
- [6] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *arXiv preprint arXiv:2005.01159* (2020).
- [7] Yaser Keneshloo, Tian Shi, Naren Ramakrishnan, and Chandan K Reddy. 2019. Deep reinforcement learning for sequence-to-sequence models. *IEEE transactions on neural networks and learning systems* 31, 7 (2019), 2469–2489.
- [8] Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. *arXiv preprint arXiv:1808.07913* (2018).
- [9] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems* 29 (2016).
- [10] Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. *arXiv preprint arXiv:1909.00141* (2019).
- [11] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [12] Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345* (2019).
- [13] Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304* (2017).
- [14] Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. ProphetNet-x: large-scale pre-training models for English, Chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006* (2021).

- [15] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732* (2015).
- [16] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685* (2015).
- [17] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368* (2017).
- [18] Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. 2019. An entity-driven framework for abstractive summarization. *arXiv preprint arXiv:1909.02059* (2019).
- [19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [20] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems* 28 (2015).
- [21] Li Wang, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for abstractive text summarization. *arXiv preprint arXiv:1805.03616* (2018).
- [22] Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. 2020. Copy or rewrite: Hybrid summarization with hierarchical reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9306–9313.
- [23] Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019. Context-aware self-attention networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 387–394.
- [24] Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. *arXiv preprint arXiv:1704.07073* (2017).

# An autoencoder-based fast online clustering algorithm for evolving data stream

Dazheng Gao

gaodazheng@mail.ustc.edu.cn

University of Science and Technology of China  
Hefei, China

## ABSTRACT

In the era of Big Data, more and more IoT devices are generating huge amounts of high-dimensional, real-time and dynamic data streams. As a result, there is a growing interest in how to cluster this data effectively and efficiently. Although a number of popular two-stage data stream clustering algorithms have been proposed, these algorithms still have some problems that are difficult to solve in the face of real-world data streams: poor handling of high-dimensional data streams and difficulty in effective dimensionality reduction; a slow clustering process that makes it difficult to meet real-time requirements; and too many manually defined parameters that make it difficult to cope with evolving data streams. This paper proposes an autoencoder-based fast online clustering algorithm for evolving data stream (AFOCEDS). The algorithm uses a stacked denoising autoencoder to reduce the dimensionality of the data, a multi-threaded approach to improve response speed, and a mechanism to automatically update parameters to cope with evolving data streams. The experiments on several realistic data streams show that AFOCEDS outperforms other algorithms in terms of effectiveness and speed.

## CCS CONCEPTS

• Information systems → Data stream mining; Clustering.

## KEYWORDS

evolving data stream, cluster, autoencoder

## ACM Reference Format:

Dazheng Gao. 2023. An autoencoder-based fast online clustering algorithm for evolving data stream. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590020>

## 1 INTRODUCTION

Recently, the devices in many areas, such as network monitoring, environmental inspection and satellite remote sensing, have generated data streams constantly and rapidly[2][4][6]. Data streams are characterized by large data volumes, sequential order in time

and constant evolution, which bring many challenges for effective data mining.

Clustering is an unsupervised learning method. When the data stream arrives, its distribution evolves over time. Then its clustering results in the data space change, but static clustering algorithm cannot handle its evolution. At the same time, online systems also have high requirements for real-time. Considering these problems, a good data stream clustering algorithm tries to realize the minimum processing delay, detecting noise and the evolution of the data stream without a predefined number of clusters[12]. Several data stream clustering algorithms have been proposed, including Denstream[5], DBstream[7], CEDAS[9], BOCEDS[10]. These algorithms are almost performed in two phases: the online phase (summarizes the data into micro-clusters) and the offline phase (re-clusters the micro-clusters to obtain the final clustering results). Re-clustering micro-clusters is time consuming, which makes it difficult to do quickly to deal with problems such as the evolution of data stream. At the same time, the algorithms also have difficulty in handling high-dimensional data because of the “Curse of Dimensionality”. To overcome these drawbacks, this paper proposes a density-based clustering algorithm, which can handle the problem of data stream evolution. The algorithm also uses an autoencoder for dimension reduction, which can effectively deal with high-dimensional data.

In summary, the main contributions of this paper are:

1. We propose a novel streaming data clustering method that can efficiently detect clusters of arbitrary shapes. The algorithm uses SDAE, which is capable of nonlinear dimensionality reduction and solves “Curse of Dimensionality”.
2. The algorithm can automatically modify the parameters during the clustering process and handle the evolution actively. It classifies micro-clusters into three classes according to their density which acts as pruning. In the process of re-clustering the micro-clusters, the algorithm uses a multi-threaded approach to speed up it.
3. Our experimental results show that AFOCEDS outperforms other algorithms for several popular real-world datasets.

The rest of the paper is constructed as follows: In Section 2, we review the existing stream clustering algorithms, as well as autoencoders and their use in clustering. Then, in Section 3, we describe the proposed algorithm AFOCEDS. Later, in Section 4, we describe our evaluation method and report our experimental results. Finally, in Section 5, we present the main conclusions of our study.

## 2 RELATED WORK

The clustering algorithms for data streams can be classified into the same five categories as the criteria for static clustering algorithms: hierarchy-based, division-based, model-based, grid-based

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590020>

and density-based. Among them, density-based and grid-based clustering algorithms are widely preferred because they can efficiently generate arbitrarily shaped clusters.

In the field of density-based data stream clustering, Clustream[1] is a two-stage model based on k-means, but it can only generate hyperspherical clusters. Based on Clustream, Denstream[5] is improved by dividing the clusters into potential micro-clusters and outlier micro-clusters according to their density. Denstream maintains these two types of micro-clusters separately and transforms them with the input of data and the decay of time. Though it can generate arbitrarily shaped clusters based on DBSCAN, it takes a lot of time to prune outlier clusters. Edwin et al.[13] investigates the merging and separation strategies of micro-clusters. DBstream[7] determines whether two micro-clusters belong to the same cluster by considering the shared density between them. In 2017, Hyde et al.[9] proposed CEDAS, a fully online data stream clustering algorithm, which introduces a simple linear aging process to handle the evolutionary features of the data streams and improves the clustering effectiveness. However, it still has the problem of too many manual parameters and can't handle high-dimensional data streams.

An autoencoder is a neural network in the field of unsupervised learning and is commonly used for dimensionality reduction and feature learning. In 2016, Peng et al.[14] used an autoencoder to learn a feature on which a global sparse prior constraint was imposed such that the learned subspace maintained the local and global structure of the original space. The clustering results are then obtained in the learned feature space by using the k-means algorithm. However, Peng's algorithm can only work on static data. Yang et al.[18] proposed a recurrent framework for joint unsupervised learning of deep representation and image clustering. There have been many clustering algorithms that use this approach. But most of them were used for image clustering and it is difficult to apply it for structured data stream clustering[17].

### 3 ALGORITHM

The AFOCEDS proposed in this paper consists of both offline and online phases. In the offline phase, AFOCEDS builds an encoder to learn the data in the training set, and this encoder generates a low-dimensional representation of each data. The online phase generates clusters continuously from the low-dimensional data stream. In this section, we first introduce the encoding phase of AFOCEDS. Then, we introduce the clustering phase of AFOCEDS and describe the process of data stream and clustering update.

#### 3.1 AFOCEDS's offline phase

Stacked denoising autoencoder (SDAE) is an important component in AFOCEDS, which acts as a nonlinear dimension reduction for the data, generating meaningful and well-separated representations that are more conducive for clustering.

Denoising autoencoders add noise to the input data on the basis of autoencoders. This approach can prevent overfitting and make the encoder obtained robust[16]. The stacked autoencoders are deeper than the single-layer autoencoder and better able to extract effective features, yielding better compression efficiency.

The SDAE model consists of the encoder and the decoder. The encoder is used to encode the high-dimensional input  $x$  into a low-dimensional representation  $h$ , where  $h$  is the most informative feature learned by the neural network. The decoder tries to reduce  $h$  to a result  $y$  that is consistent in the input dimension. Ideally, the input  $x$  and the reduced output  $y$  can be approximately equal or exactly the same. The output  $y$  is used to reconstruct the original data  $x$ . A suitable loss function is chosen to minimize the error of the reconstructed data from the original data, such that the potential representation  $h$  of the encoder can represent the original data  $x$ . Decoders are generally constructed in the reverse order of the encoder's structure.

The SDAE encoder is composed of multiple layers of denoising autoencoders. The structure of each denoising autoencoder consists of 3 layers, the dropout layer, the neurons and the activation function. The SDAE decoder is a symmetric structure of the encoder. In the process of training, the goal is to reduce the reconstruction error.

A model for dimensionality reduction of high-dimensional structured data is generated by training and fine-tuning the SDAE. The low-dimensional representation  $h$  output by the encoder is used in the subsequent clustering process.

#### 3.2 AFOCEDS's online phase

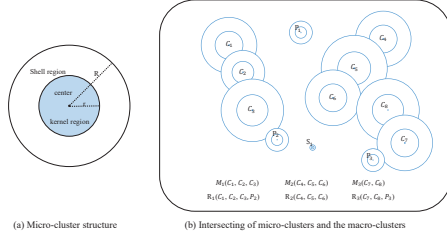
The goal of the AFOCEDS's online phase is to generate high-quality clusters that meets low memory requirements and low processing latency and effectively detects evolution in the data stream.

##### 3.2.1 Data structures of AFOCEDS.

Before introducing the algorithm, we need to define the relevant terms.

- a) Cluster graph: The cluster graph records the set of micro-clusters and the result of macro-clusters formed by micro-clusters.
- b) Core micro-cluster ( $MC_c$ ): The density in this micro-cluster is not less than the high-density threshold  $D_u$ , which is used to construct macro-clusters.
- c) Potential micro-cluster ( $MC_p$ ): The density in this micro-cluster is less than the high-density threshold  $D_u$ , but not less than the low-density threshold  $D_l$ . It may be shrunken from the core micro-cluster or extended from the sparse micro-cluster. In the process of clustering, it is only connected to macro-clusters and does not participate in the construction of macro-clusters. It is usually the boundary of the clustering.
- d) Sparse micro-cluster ( $MC_s$ ): The density in the micro-cluster is less than  $D_l$ . The sparse micro-clusters are stored in the buffer and are identified as noise and do not participate in the clustering process.
- e) Macro-clusters: Macro-clusters consist of several micro-clusters that intersect with each other. Macro-clusters are also the final clustering results.

The clustering phase of AFOCEDS is an online micro-cluster-based clustering algorithm. It only stores information related to micro-clusters and records the graph structure of micro-clusters connections. Fig. 1(a) shows the structure of micro-clusters, which is divided into core and shell regions with radius  $r$  for the core region and  $R$  for the shell region. In general,  $R=2r$  is set.



**Figure 1: Formation of macro-clusters in the proposed AFOCEDS.**

Fig. 1(b) shows the structure of micro-clusters and an example of the relationship between micro-clusters and macro-clusters. The input data is stored in the micro-clusters, and the subsequent clustering process operates only on the micro-clusters. When the shell region of a micro-cluster intersects with the core region of another micro-cluster, the two micro-clusters are judged to be connected. In mathematical terms, micro-clusters  $C_1$  and  $C_2$  are connected if and only if their distance  $d < \max(C_1.R + C_2.r, C_1.r + C_2.R)$ . The interconnected micro-clusters  $\{C_1, C_2, C_3\}$  in Fig. 1(b) constitutes the macro-cluster  $M_1$ .  $\{C_4, C_5, C_6\}$  constitutes the macro-cluster  $M_2$ .  $\{C_7, C_8\}$  constitutes the macro-cluster  $M_3$ .  $C_8$  and  $C_5, C_6$  only intersect in the shell region, but not in the core region, so they are not connected.  $P_1, P_2$  and  $P_3$  are potential micro-clusters. Because  $P_2$  and  $P_3$  connect to the core micro-cluster, they are in the cluster result. And  $P_1$  is considered as noise as  $S_1$  is.  $\{R_1, R_2, R_3\}$  is the final clustering result.

Micro-cluster is represented by a set of vectors  $MC(\text{center}, \text{density}, t, r, EL, M_t)$ . The parameter *center* is the position of the micro-cluster centroid, which also indicates the position of this micro-cluster in the data space, and it is calculated from the positions of all data points in the micro-cluster. The parameter *density* is the density of the micro-cluster at time  $t$ , calculated from the number of data points in the micro-cluster. The parameter  $t$  is the latest update time of this micro-cluster. The parameter  $r$  is the radius of the micro-clusters, the radius of the core region is  $r$  and the radius of the shell region is  $2r$ . The parameter  $EL$  is the set of micro-clusters connected to this micro-cluster, which facilitates the construction of macro-cluster graphs. The parameter  $M_t$  denotes the macro-cluster to which this micro-cluster belongs at moment  $t$ .

Whenever new data arrives, it may belong to an existing micro-cluster or generate a new micro-cluster. AFOCEDS defines three micro-clusters: core  $MC(MC_c)$ , potential  $MC(MC_p)$ , and sparse  $MC(MC_s)$ . They are distinguished by the density.  $MC_c$  and  $MC_p$  are used in the construction of macro-clusters. The  $MC_s$  is considered as noise and not involved in the construction of macro-clusters.

### 3.2.2 Description of AFOCEDS.

The clustering phase of AFOCEDS requires the input of the following 5 global parameters:

*decay*: The decay of the data density. It will be used in clustering. It indicates that the previous data has decreased in importance and the new input data has more weight. The value of *decay* is between 0 and 1.

$D_l, D_u$ : Boundary of three types of micro-clusters.  $D_l$  is the criterion that isolates  $MC_s$  from the rest and is used to separate

noise from normal data.  $D_u$  is the boundary between  $MC_c$  and  $MC_p$ .

$R_{min}, R_{max}$ : The minimum and maximum radius of the micro-clusters.  $R_{min}$  is used to initialize the new micro-clusters when new data form new micro-clusters.  $R_{max}$  is used in micro-clusters containing a large amount of data to limit the unrestricted expansion of the micro-clusters.

Algorithm 1 shows the processing of the data stream at time  $t$ . The algorithm puts all of the data into the appropriate micro-clusters, and the cluster graph will be updated after that. At next time  $t'$ , AFOCEDS will run algorithm 1 again.

---

#### Algorithm 1: AFOCEDS online clustering phase

---

**Input:** Data stream  $X$  at time  $t$ , Decay rate *decay*, Encoder *encoder*, Density boundary  $D_l, D_u$ , Minimum and maximum radius  $R_{min}, R_{max}$ .

**Output:** The set of clustering results  $\{R_1, \dots, R_k\}$ .

```

1 for  $x \in X$  do
2    $x' = \text{encoder}(x)$ ;
3    $MC_i = \text{Find the nearest MC} \in \{MC_c, MC_p, MC_s\}$  that
     satisfies  $d(MC, x') < 2MC.r$ ;
4   if  $MC_i$  exists then
5     Update the parameters of  $MC_i$  and the set
        $\{MC_c, MC_p, MC_s\}$  it belongs to;
6   else
7     New MC in  $MC_s$ ;
8   end
9 end
10 Update cluster graph;
11 return  $\{R_1, \dots, R_k\}$ 

```

---

The global parameters only need to be set once. After the global parameters are set, the algorithm starts reading the data. When new data comes in, it is firstly encoded by an encoder to generate a low-dimensional representation  $x'$ , and then finds the micro-clusters in the data space which that data belongs to. If the distance of the data to any of the micro-clusters is greater than the shell radius  $2r$  of the micro-clusters, it means that the data does not belong to any of the micro-clusters. Then the algorithm creates a new micro-cluster  $MC$  and initializes the properties of the  $MC$ .  $MC.\text{center}$  is set to that data point.  $MC.\text{density}$  is set to 1.  $MC.t$  is set to the timestamp entered for that data point. And the  $MC.r$  is set to  $R_{min}$ . Because this micro-cluster's density is only 1 (generally less than the lowest density  $D_l$ ), it will be classified as  $MC_s$  and will not participate in clustering, so both  $MC.EL$  and  $MC.M_t$  are empty.

If the micro-cluster  $MC_i(\text{center}, \text{density}, t, r, EL, M_t)$  with distance less than  $2r$  at time  $t_{now}$  exists, the new data  $x_j$  needs to be mapped to  $MC_i$  and the attributes of  $MC_i$  need to be updated.

$$\text{density} = \text{density} \times \text{decay}^{t_{now}-t} + 1 \quad (1)$$

$$\text{center} = \frac{x_j + \text{center} \times (\text{density} - 1)}{\text{density}} \quad (2)$$

$$t = t_{now} \quad (3)$$

$$r = \min(R_{max}, (r + \frac{1}{\text{density}}) \times \text{decay}^{t_{now}-t}) \quad (4)$$

The density of  $MC_i$  are updated as Eq. (1). The algorithm updates the centroid positions as Eq. (2) to take in new data points. The algorithm updates timestamp and radius according to Eq. (3)(4). The radius of  $MC_i$  depends mainly on the density. The higher the density is, the larger the radius is. The radius will not be larger than  $R_{max}$ , which limits the unrestricted expansion of micro-cluster. When  $MC_i$  is updated, the micro-cluster is first removed from the set it originally belonged to. Then compare  $MC_i.density$  with  $D_l$ ,  $D_u$ . When  $MC_i.density < D_l$ ,  $MC_i$  is added to  $MC_s$ . When  $D_l \leq MC_i.density < D_u$ ,  $MC_i$  is added to  $MC_p$ . When  $MC_i.density \geq D_u$ ,  $MC_i$  is added to  $MC_c$ . If  $MC_i$  is  $MC_s$ , it will clear  $MC_i.EL$  and  $MC_i.M_t$ , and also remove it in the  $EL$  of the micro-clusters connected to him. If  $MC_i$  is  $MC_p$  or  $MC_c$ ,  $MC_i.EL$  and  $MC_i.M_t$  should be updated, but this update process is complicated and time consuming. So we do not update it immediately. When the cluster graph is updated, the  $MC_c$  and  $MC_p$  are updated uniformly.

---

**Algorithm 2:** Update cluster graph

---

**Input:** core micro-cluster set  $MC_c$ , potential micro-cluster set  $MC_p$ , spares micro-cluster set  $MC_s$ .  
**Output:** The set of clustering results  $\{R_1, \dots, R_k\}$ .

```

1 for  $MC_i \in \{MC_s, MC_p, MC_c\}$  do
2   | Update the parameters of  $MC_i$ ;
3   | Classify  $MC_i$  into  $MC_s$ ,  $MC_p$  or  $MC_c$ ;
4   | Delete  $MC_i$  if  $MC_i.density < 1$ ;
5 end
6 for  $MC_i \in MC_c$  do
7   | Send  $MC_i$  to thread p;
8 end
9 Set all  $MC \in MC_c$  unvisited;
10 for  $MC_i \in MC_c$  do
11   | Dfs( $MC_i$ ) according to  $MC_i.EL$  and cluster the searched
      |  $MC$  as a macro-cluster ;
12 end
13 for  $MC_i \in MC_p$  do
14   | Send  $MC_i$  to thread q;
15 end
16 return  $\{R_1, \dots, R_k\}$ 
17 The thread p performs:
18   Find all core micro-cluster  $MC_j$  that satisfies
       $d(MC_i, MC_j) < \max(r_i + R_j, r_j + R_i)$ ;
19   Push them into  $MC_i.EL$ ;
20 The thread q performs:
21    $MC_j = \text{find the nearest core micro-cluster ;}$ 
22   Attach  $MC_i$  to the macro-cluster which the  $MC_j$  belongs
      to if  $d(MC_i, MC_j) < \max(r_i + R_j, r_j + R_i)$ ;
```

---

Once the data stream at this moment is all read and assigned to the corresponding micro-cluster completely, it is time to update the cluster graph. As shown in Algorithm 2, firstly, update all of the micro-clusters at time  $t$ . Because there are some of the micro-clusters without data input. We have to update the parameters of them. The parameter  $density = density \times decay^{t_{now}-t}$ . The parameter  $r = \max(R_{min}, r \times decay^{t_{now}-t})$ . The parameter  $t = t_{now}$ .

The micro-clusters with too low density in  $MC_s$  are removed, which not only eliminates obsolete micro-clusters as well as noise, but also prevents too many micro-clusters from affecting the efficiency.

When constructing macro-clusters, the connections are only made between core micro-clusters. The algorithm uses the potential micro-clusters to determine the boundaries of the macro-clusters because they are less dense. AFOCEDs uses a multi-threaded approach to calculate the distance between core micro-clusters to determine connectivity, and then stores the connectivity information of each micro-cluster in its own  $EL$ . Later there is no need to repeatedly compute connectivity when connecting micro-clusters to construct macro-clusters. The algorithm directly connects the core micro-clusters based on their  $EL$ . The clustering problem is then converted to the problem of finding the connected components of an undirected graph. This can be done with a simple traversal, which can greatly reduce the time needed for clustering. Connected core micro-clusters constitute a macro-cluster, and disconnected core micro-clusters constitute different macro-clusters. After the macro-clusters are constructed, the potential micro-clusters are connected to the nearest and intersecting core micro-clusters and thus assigned to the corresponding macro-clusters. The algorithm is also accelerated here in a multi-threaded manner. Unassigned potential micro-clusters are treated as noise. The macro-cluster result returned by the algorithm is the final clustering result.

## 4 EXPERIMENTS

To test the performance of AFOCEDs, we conducted experiments on four real-world datasets. Table 1 summarizes the basic information of the four datasets.

**Table 1: Description of the datasets**

Dataset	Instances	Attributes	Clusters
HAR	10299	561	6
PAMAP2	22590	53	12
Forest Covertype	581012	54	7
KDDCUP99	4,898,431	39	5

The HAR[3] (The human activity recognition) is based on the records of 30 volunteers aged 19 - 48. Each person performed six activities (walking, walking up, walking down, sitting, standing and lying down). The dataset contains 561 features generated by accelerometer and gyroscope sensors. The PAMAP2[15] (The Physical Activity Monitoring) contains data related to 12 different physical activities (watching TV, driving, playing soccer...). The dataset contains 9 subsets. We randomly select one of them called "Subject 101" and a subset of 22,590 records. The Forest Covertype[11] contains 7 categories of tree species with 54 attributes, the first 10 attributes are quantitative variables, the next 4 attributes are binary variables indicating wilderness areas and 40 binary variables indicating land types. The KDDCUP99 is 9 weeks of network connectivity data collected on the US LAN and contains 5 classes (one normal and four attack classes). All datasets are normalized.

The offline process of AFOCEDs has to pre-train the SDAE module before SDAE can be applied in the online process. The structure of SDAE is closely related to the number of attributes of the dataset.

**Table 2: ARI results of algorithms**

	AFOCEDS	AFOCEDS without encoder	Deepstream	CEDAS	DBstream	Denstream
HAR	0.875	0.637	0.867	0.532	0.560	0.460
PAMAP2	0.659	0.326	0.420	0.353	0.199	0.221
Forest Coverttype	0.634	0.421	0.489	0.468	0.403	0.396
KDDCUP99	0.826	0.780	0.814	0.771	0.734	0.715

**Table 3: Time consumption of algorithms (unit: s)**

	AFOCEDS	AFOCEDS without encoder	Deepstream	CEDAS	DBstream	Denstream
HAR	21.51	21.15	62.88	34.37	25.99	20.90
PAMAP2	23.64	23.27	43.25	25.43	33.51	19.14
Forest Coverttype	52.23	51.81	1586.51	153.60	209.43	162.45
KDDCUP99	213.46	212.13	6889.31	734.25	1205.24	401.56

In practice, for the HAR, the number of neurons per layer for SDAE is [561-220-100-30-5]. For the PAMAP2, the number of neurons per layer for SDAE is [53-34-18-5]. For the Forest Coverttype, the number of neurons per layer for SDAE is [54-34-18-5]. For the KDDCUP99, the number of neurons per layer for SDAE is [39-34-18-5]. Each layer uses relu as the activation function, and the optimizer is set to Adam with a learning rate of 0.1. The reconstruction loss is set to mean square error ( $MSE = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2$ ).

We compare AFOCEDS with four advanced data stream clustering algorithms Denstream[5], DBstream[7], CEDAS[9], Deepstream[8]. Algorithms are implemented in python. All experiments are conducted on a win10 laptop (CPU: 11th Gen Intel(R) Core(TM) i5-11400H, 6 cores/chip, 2 thread/core, 16GB; GPU: NVIDIA GeForce RTX 3050Ti, 4GB).

In the experiments, we use the Adjusted Rand Index (ARI) as the evaluation metric. The ARI is the most commonly used clustering metric, and it takes values from -1 to 1. The closer its value is to 1, the better its effect is. The label of the dataset itself is used as the ground truth.

The experimental results are shown in Table 2, and the best ARI indices for these algorithms were obtained by using the grid search method for the parameters of the algorithms. Obviously, AFOCEDS provides better results than several other algorithms. For real-world high-dimensional datasets, clustering algorithms generally have the problem of poor clustering results. However, both AFOCEDS and Deepstream add a neural network dimensionality reduction process, which allows the clustering results to be significantly improved. And because AFOCEDS uses a more optimized neural network structure than Deepstream, it makes AFOCEDS perform better.

AFOCEDS, CEDAS, DBstream and Denstream are all two-stage micro-cluster-based algorithms. As shown in Table 3, although AFOCEDS adds an additional dimensionality reduction stage, the time consumed in the encoder process is very little. The time for AFOCEDS to construct macro-clusters is greatly reduced because of the reasonable classification of micro-clusters and the multi-threaded process during re-clustering. The multi-threaded process

makes it possible to save more time in the case of more micro-clusters, and it is the fundamental reason why AFOCEDS can run fast.

## 5 CONCLUSION

In this paper, we have proposed a fast data stream clustering algorithm AFOCEDS. It uses a multi-threaded approach for re-clustering of micro-clusters, which successfully reduces its running time and improves the response speed of the algorithm. The algorithm uses a method with decaying and automatic update on parameters that can detect and react to evolution of the data stream. Facing the high-dimensional realistic data, the existing data stream clustering algorithms have the problem of poor clustering results, which is difficult to meet the needs of realistic practice. And AFOCEDS has the ability of nonlinear dimensionality reduction by using deep neural network, which can compress the high-dimensional data more effectively, so that it has better performance in the face of realistic data. More importantly, the SDAE module is an unsupervised training mechanism that does not require labels. Experiments show that AFOCEDS has good performance in terms of effectiveness and speed.

In the future work, the algorithm can incorporate mechanisms for dynamic updates to the encoder module to become a truly intelligent machine learning framework.

## REFERENCES

- [1] Charu C Aggarwal, S Yu Philip, Jiawei Han, and Jianyong Wang. 2003. A framework for clustering evolving data streams. In *Proceedings 2003 VLDB conference*. Elsevier, 81–92.
- [2] Amineh Amini, Teh Ying Wah, and Hadi Saboohi. 2014. On density-based data streams clustering algorithms: A survey. *Journal of Computer Science and Technology* 29, 1 (2014), 116–141.
- [3] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra Perez, and Jorge Luis Reyes Ortiz. 2013. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*. 437–442.
- [4] Jean Paul Barddal, Heitor Murilo Gomes, and Fabricio Enembreck. 2015. SNC-Stream: a social network-based data stream clustering algorithm. In *Proceedings of the 30th annual ACM symposium on applied computing*. 935–940.
- [5] Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou. 2006. Density-based clustering over an evolving data stream with noise. In *Proceedings of the 2006*

- SIAM international conference on data mining*. SIAM, 328–339.
- [6] Wei Fan and Albert Bifet. 2013. Mining big data: current status, and forecast to the future. *ACM SIGKDD explorations newsletter* 14, 2 (2013), 1–5.
  - [7] Michael Hahsler and Matthew Bolaños. 2016. Clustering data streams based on shared density between micro-clusters. *IEEE Transactions on Knowledge and Data Engineering* 28, 6 (2016), 1449–1461.
  - [8] Shimon Harush, Yair Meidan, and Asaf Shabtai. 2021. DeepStream: Autoencoder-based stream temporal clustering and anomaly detection. *Computers & Security* 106 (2021), 102276.
  - [9] Richard Hyde, Plamen Angelov, and Angus Robert MacKenzie. 2017. Fully online clustering of evolving data streams into arbitrarily shaped clusters. *Information Sciences* 382 (2017), 96–114.
  - [10] Md Kamrul Islam, Md Manjur Ahmed, and Kamal Z Zamli. 2019. A buffer-based online clustering for evolving data stream. *Information Sciences* 489 (2019), 113–135.
  - [11] Brian Johnson, Ryutaro Tateishi, and Zhixiao Xie. 2012. Using geographically weighted variables for image classification. *Remote sensing letters* 3, 6 (2012), 491–499.
  - [12] Hanyang Liu, Junwei Han, Feiping Nie, and Xuelong Li. 2017. Balanced clustering with least square regression. In *Thirty-first AAAI conference on artificial intelligence*.
  - [13] Edwin Lughofer and Moamar Sayed-Mouchaweh. 2015. Autonomous data stream clustering implementing split-and-merge concepts-towards a plug-and-play approach. *Information Sciences* 304 (2015), 54–79.
  - [14] Xi Peng, Shijie Xiao, Jiashi Feng, Wei-Yun Yau, and Zhang Yi. 2016. Deep subspace clustering with sparsity prior.. In *IJCAI*. 1925–1931.
  - [15] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*. IEEE, 108–109.
  - [16] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. 1096–1103.
  - [17] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. PMLR, 478–487.
  - [18] Jianwei Yang, Devi Parikh, and Dhruv Batra. 2016. Joint unsupervised learning of deep representations and image clusters. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5147–5156.

# Estimation of Distribution Algorithm with Discrete Hopfield Neural Network for GRAN3SAT Analysis

Yuan, Y.G, Gao

School of Medical Information Engineering, Chengdu  
University of Traditional Chinese Medicine, Chengdu  
611137, China; School of Mathematical Sciences, Universiti  
Sains Malaysia, Penang 11800, Malaysia  
gaoyuan@student.usm.my

Ju, J.C, Chen\*

School of Medical Information Engineering, Chengdu  
University of Traditional Chinese Medicine, Chengdu  
611137, China; School of Mathematical Sciences, Universiti  
Sains Malaysia, Penang 11800, Malaysia  
chenju@cdutcm.edu.cn

Chengfeng, C.Z, Zheng

School of Mathematical Sciences, Universiti Sains  
Malaysia, Penang 11800, Malaysia  
1002953832@qq.com

Yueling, Y.G, Guo

School of Mathematical Sciences, Universiti Sains  
Malaysia, Penang 11800, Malaysia  
guoyueling1982@163.com

## ABSTRACT

The Discrete Hopfield Neural Network introduces a G-Type Random 3 Satisfiability logic structure, which can improve the flexibility of the logic structure and meet the requirements of all combinatorial problems. Usually, Exhaustive Search (ES) is regarded as the basic learning algorithm to search the fitness of neurons. To improve the efficiency of the learning algorithm. In this paper, we introduce the Estimation of Distribution Algorithm (EDA) as a learning algorithm for the model. To study the learning mechanism of EDA to improve search efficiency, this study focuses on the impact of EDA on the model under different proportions of literals and evaluates the performance of the model at different phases through evaluation indicators. Analyze the effect of EDA on the synaptic weights and the global solution. From the discussion, it can be found that compared with ES, EDA has a larger search space at the same efficiency, which makes the probability of obtaining satisfactory weights higher, and the proportion of global solutions obtained is higher. Higher proportions of positive literals help to improve the model performance.

## CCS CONCEPTS

• **CCS CONCEPT Computing methodologies** → Artificial intelligence.

## KEYWORDS

Hopfield Neural Network, Exhaustive Search, Estimation of Distribution Algorithm, Meta-heuristic

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590021>

## ACM Reference Format:

Yuan, Y.G, Gao, Chengfeng, C.Z, Zheng, Ju, J.C, Chen, and Yueling, Y.G, Guo. 2023. Estimation of Distribution Algorithm with Discrete Hopfield Neural Network for GRAN3SAT Analysis. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590021>

## 1 INTRODUCTION

The principle of Hopfield neural network (HNN) simulating human memory, proposed by J. J. Hopfield [1] in 1982, can solve pattern recognition problems and combinatorial optimization problems. HNN guarantees convergence to a local minimum, but it is possible to converge to a non-global minimum. Discrete Hopfield neural network (DHNN) is the earliest proposed binary neural network. The input and output of the neuron only take  $\{0, 1\}$  or  $\{-1, 1\}$ . In 1990, W. McCulloch and W. Pitts [2] proposed that the relationship between neurons can be handled by propositional logic. In 1992, Wan Abdullah [3] proposed the model of introducing the satisfiability problem into DHNN. Since then, the DHNN based on Satisfiability (SAT) logic programming has undergone a series of developments, including 3SAT [4], MAX3SAT [5], RAN3SAT [6], YRAN2SAT [7], r2SAT [8] combination with DHNN. In 2022, Gao Yuan proposed G-Type Random 3 Satisfiability (GRAN3SAT) [6] so that all combination problems can be satisfied. As the number of clauses increases, the model's task of finding consistent explanations becomes more complex. The exhaustive search (ES) is used as the learning algorithm in the learning phase [9]. ES is less efficient when the search space is large. Estimation of Distribution Algorithm (EDA), also known as the genetic algorithm based on the probability distribution model, was first proposed in 1996 [10]. In 2010, Peralta [11] evaluated methods for evolving neural network architectures using EDA. In 2013, Donate [12] used EDA to introduce the automatic design of artificial neural networks for forecasting time series to improve the final forecast accuracy. According to these related works, we can deduce that the EDA can be combined with the neural network to obtain faster convergence. Based on this, EDA can be considered a promising algorithm to

facilitate the learning phase. However, no attempt has been made to exploit the EDA algorithm as an optimal learning method in Discrete Hopfield Neural Networks (DHNN), especially in optimizing GRAN3SAT logic representation and analysis.

In this paper, the GRAN3SAT logic structure is applied in DHNN. To efficiently find a consistent solution to the satisfiability problem, EDA is used as a learning algorithm to search the fitness of neurons. EDA achieves the search and convergence of satisfactory solutions through continuous updating and testing. This study focuses on different proportions of positive and negative literals, analyzes the performance improvement of EDA as a learning algorithm for GRAN3SAT compared with ES, and evaluates the behavior changes of the model at different phases through evaluation indicators.

## 2 EDA IN THE DHNN BASED ON GRAN3SAT

### 2.1 Logic Rules of GRAN3SAT.

GRAN3SAT is a novel non-systematic SAT logic structure represented in conjunctive normal form. Logic consists of a set of different literals and clauses. GRAN3SAT mainly consists of third-order, second-order, and first-order logic clauses randomly. Each literal value is of the form  $\{-1, 1\}$ . The general formula of GRAN3SAT is  $P_G$ , detailed as follows.

- A set of  $NN$  literals:  $A_1, A_2, A_3, \dots, A_{NN}$ . Randomly generated for each literal state.
- The number of clauses  $\{x_i, y_i, z_i\}$  is randomly generated.  $x_i$  is the number of third-order logic clauses,  $y_i$  is the number of second-order logic clauses, and  $z_i$  is the number of first-order logic clauses.
- The representation of a clause:

Third-order logic clause:  $C_1^{(3)}, C_2^{(3)}, \dots, C_{x_j}^{(3)}$ , whereby  $C_{m_j}^{(3)} = (A_m \vee A_n \vee A_k), m, n, k \in N^*$ .

Second-order logic clause:  $C_1^{(2)}, C_2^{(2)}, \dots, C_{y_j}^{(2)}$ , whereby  $C_{n_j}^{(2)} = (A_m \vee A_n), m, n \in N^*$ .

First-order logic clause:  $C_1^{(1)}, C_2^{(1)}, \dots, C_{z_j}^{(1)}$ , whereby  $C_{k_j}^{(1)} = A_m, m \in N^*$ .

$$P_G = \bigwedge_{i=1}^{x_j} C_i^{(3)} \wedge \bigwedge_{i=1}^{y_j} C_i^{(2)} \wedge \bigwedge_{i=1}^{z_j} C_i^{(1)} \quad (1)$$

### 2.2 DHNN

The bipolar neurons of DHNN used in this study are represented by  $\{-1, 1\}$ , and the update equation of neuron state of DHNN is as follows:

$$S_i = \begin{cases} 1, & \sum_{jk} W_{ijk} S_j S_k \geq \theta_i \\ -1, & \sum_{jk} W_{ijk} S_j S_k < \theta_i \end{cases} \quad (2)$$

whereby  $S_i$  is the neuron state,  $W_{ij}$  is the synaptic weight in the two neurons.  $\theta_i$  is the threshold. The cost function of GRAN3SAT-DHNN is  $\text{Cost}_{P_G}$ , and the formula is as follows:

$$\text{Cost}_{P_G} = \frac{1}{2^3} \sum_{j=1}^{x_i} \delta_{j_1}^{(3)} \delta_{j_2}^{(3)} \delta_{j_3}^{(3)} + \frac{1}{2^2} \sum_{j=1}^{y_i} \delta_{j_1}^{(2)} \delta_{j_2}^{(2)} + \frac{1}{2} \sum_{j=1}^{z_i} \delta_{j_1}^{(1)} \quad (3)$$

$$\delta_{j_x}^{(k)} = \begin{cases} 1 + S_{A_{j_x}}, & \text{when } A_{j_x} \\ 1 - S_{A_{j_x}}, & \text{when } \neg A_{j_x} \end{cases} \quad (4)$$

Whereby  $k = 1, 2, 3$ . The value of  $\eta_{P_G}$  describes the logical consistency of  $P_G$ . When the logical inconsistency reaches the minimum value, the weight can be obtained through the Wan Abdullah method [5]. The probability formula for consistent interpretation is as follows:

$$\lambda(\eta_{P_G} = 0) = \left(1 - \frac{1}{2^3}\right)^{x_i} \left(1 - \frac{1}{2^2}\right)^{y_i} \left(1 - \frac{1}{2}\right)^{z_i} \quad (5)$$

In the testing phase, the Formula (6)(7) represent the local field formula and the update neuron state. The activation function is the Hyperbolic Tangent Activation Function (HTAF) [7].

$$h_i = \sum_{k \neq i, j} \sum_{j \neq i} W_{ijk} S_j S_k + \sum_{j \neq i} W_{ij} S_j + W_i \quad (6)$$

$$S_{u_i} = \begin{cases} 1, & \sum_{k \neq i, j} \sum_{j \neq i} W_{ijk} S_j S_k + \sum_{j \neq i} W_{ij} S_j + W_i \geq 0 \\ -1, & \sum_{k \neq i, j} \sum_{j \neq i} W_{ijk} S_j S_k + \sum_{j \neq i} W_{ij} S_j + W_i < 0 \end{cases} \quad (7)$$

$S_i$  and  $S_{u_i}$  represent the initial state and the update state.  $W_{ijk}, W_{ij}$ , and  $W_i$  represent the weights of the third, second and first order of DHNN. The Lyapunov energy function  $En_{P_G}$  is obtained by formula(8), and the minimum energy  $En_{P_G}^{min}$  is obtained by formula(9).

$$En_{P_G} = -\frac{1}{3} \sum_i \sum_{j \neq i} \sum_{k \neq i, j} W_{ijk} S_i S_j S_k - \frac{1}{2} \sum_i \sum_{j \neq i} W_{ij} S_i S_j - \sum_i W_i S_i \quad (8)$$

$$En_{P_G}^{min} = -\left(\frac{x_i}{2^3} + \frac{y_i}{2^2} + \frac{z_i}{2}\right) \quad (9)$$

The current DHNN convergence formula (10) is as follows,  $tv$  means tolerance value.

$$\left| En_{P_G} - En_{P_G}^{min} \right| \leq tv \quad (10)$$

Introduce GRAN3SAT into DHNN to become  $GRAN3SAT_{DHNN}$ .

### 2.3 EDA

The EDA is a population evolution algorithm based on statistical theory. By establishing a probability formula to describe the distribution information of satisfiable solutions in the search range, a new population is generated by random sampling. The evolution of the population is achieved through repeated iterations. The EDA flow is as follows:

Step 1: Initialization.

Initialize the number of populations  $N_p$ . The dimension of each individual is  $N_n$ . The initial population is  $X_i = \{x_{ij} | i = 1, 2, \dots, N_p; j = 1, 2, \dots, N_n\}$ , where  $x_{ij} \in \{1, -1\}$ .

Step 2: Calculate the fitness function.

The neuron fitness of the above  $X_i$  is calculated using the following formula:

$$f(X_i) = NC - \left( \sum_m C_i^{(3)} + \sum_n C_i^{(2)} + \sum_k C_i^{(1)} \right) \quad (11)$$

$$C_i^{(x)} = \begin{cases} 1, & \text{if clause is satisfied} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

where  $NC$  is the number of clauses. The larger the  $f(X_i)$ , the greater the number of unsatisfied clauses.

Step 3: probability model.

Select dominant populations based on  $f(X_i)$ . Construct a probability model [13] based on  $N$  dominant population, where  $N <$

**Table 1: Main parameters of  $GRAN3SAT_{DHNN}$ .**

Parameter	Value
Different proportions of negative literals ( $P_N$ )	0.1, 0.3, 0.5, 0.7, 0.9
Number of neurons ( $NN$ )	$6 \leq NN \leq 100$
Activation function	HTAF [7]
Number of neuron combination	100
Relaxation rate	2 [7]
Number of learn and test trial	100
Tolerance value	0.001 [7]

**Table 2: List of main parameters in the indicator.**

Parameter	Explanation.
$f_{NC}$	Maximum fitness achieved
$f_i$	Current fitness achieved
$W_{WAN}$	Satisfactory synaptic weights
$W_i$	Number of local minimum solution
$v_w$	Current number of weights
$v_{wc}$	$v_{wc} = v_w \cdot v_{combmax}$
$\varepsilon_{min}$	Minimum energy value
$\varepsilon_f$	Final energy function value
$v_G$	Number of global solutions
$v_t$	Number of testing trials
$v_{tc}$	$v_{tc} = v_t \cdot v_{combmax}$

$N_p$ , the formula is as follows.

$$P(x_j) = \frac{1}{N} \sum_{i=1}^N \chi_{ij} \quad (13)$$

$$\chi_{ij} = \begin{cases} 1, & x_{ij} = 1 \\ 0, & x_{ij} = -1 \end{cases} \quad (14)$$

Step 4: Update data.

Randomly generate a new population with a size of  $N_p$  according to the probability model.

Step 5: Judgment.

Judging whether the conditions for ending the loop are met, if not, jump to Step 2.

Introduce EDA into  $GRAN3SAT_{DHNN}$  to become  $GRAN3SAT_{DHNN}^{EDA}$ .

### 3 EXPERIMENTAL SETTINGS

This section introduces the experimental parameters and evaluation metrics of  $GRAN3SAT_{DHNN}$ . In the system, the main experimental parameters involved are defined in Table 1. This experiment mainly uses MATLAB 2021a for the simulation.

This paper uses four performance indicators to evaluate the effectiveness of the  $GRAN3SAT_{DHNN}$ . These metrics are evaluated by Mean Absolute Error (MAE) for learning error analysis, weight analysis, energy analysis, and global solution analysis. Table 2 describes the parameters used for evaluation in the testing and learning phases. The formula for the learning phase and testing

phase is as follows.:

$$MAE_{\text{learn}} = \sum_{i=1}^{v_l} \frac{|f_{NC} - f_i|}{v_l} \quad (15)$$

$$MAE_{\text{weight}} = \frac{\sum_{i=1}^{v_{wc}} |W_{WAN} - W_i|}{v_{wc}} \quad (16)$$

$$MAE_{\text{energy}} = \frac{\sum_{i=1}^{v_{tc}} |\varepsilon_{min} - \varepsilon_f|}{v_{tc}} \quad (17)$$

$$ZM_{\text{test}} = \frac{v_G}{v_{tc}} \quad (18)$$

### 4 RESULT AND DISCUSSION

The purpose of this work is to analyze the impact of EDA as a learning algorithm on the overall behavior of  $GRAN3SAT_{DHNN}$  in the learning phase, testing phase. In  $GRAN3SAT_{DHNN}$ , the MAE index is used for evaluation. Compared with ES, EDA improves neuron fitness through update iterations and narrows down the search space. This paper discusses the advantages of EDA in the learning phase and testing phase.

Figure 1 and 2 show the performance changes of EDA and ES when the neuron states have different proportions under indicators  $RMSE_{\text{learn}}$  and  $RMSE_{\text{weight}}$ .  $RMSE_{\text{learn}}$  and  $RMSE_{\text{weight}}$  respectively quantify the fitness and weight error of neurons. It can be seen that the error of each index of EDA is better than that of ES. There is no obvious difference between  $GRAN3SAT_{DHNN}^{EDA}$  and  $GRAN3SAT_{DHNN}^{ES}$  under different  $P_N$ , different proportions of literals do not affect the fitness of neuron states of different

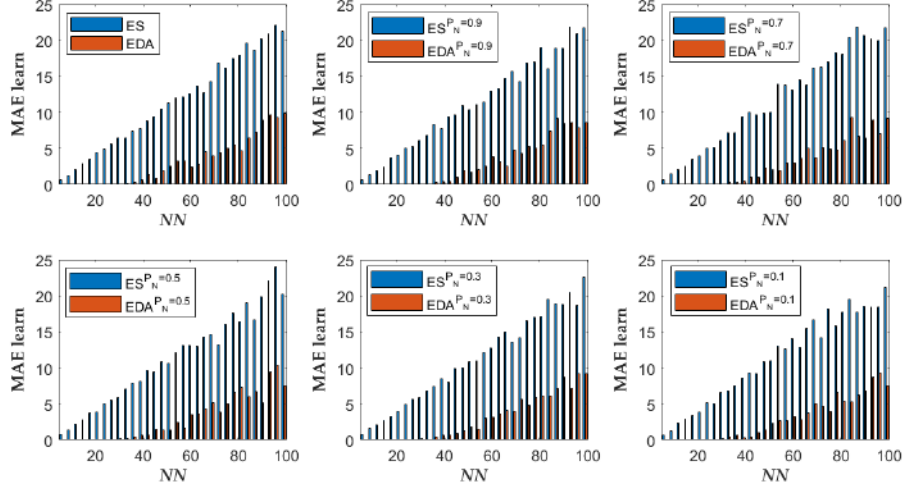


Figure 1: Changes in  $MAE_{learn}$  of EDA and ES under different  $P_N$ .

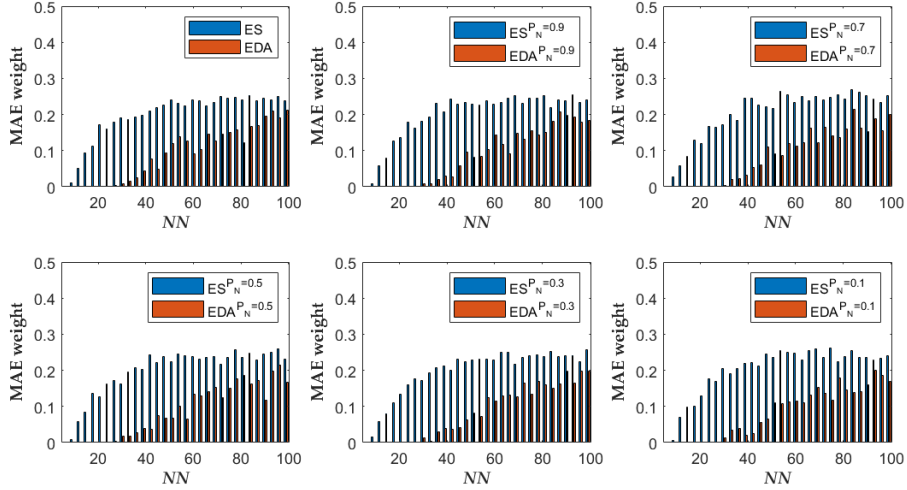


Figure 2: Changes in  $MAE_{weight}$  of EDA and ES under different  $P_N$ .

models,  $RMSE_{weight}^{ES}$  generally shows a steady state after rising, and  $RMSE_{weight}^{EDA}$  shows a linear rise. This is because as the NN increases, the solution search space of EDA expands, the fitness of neurons decreases, and satisfactory weights can no longer be obtained. To sum up, it is easier to find the optimal weight value for  $GRAN3SAT_{DHNN}^{EDA}$  than ES in the learning phase. When adjusting the state ratio of neurons, different proportions of neuron states have no significant impact on  $GRAN3SAT_{DHNN}^{EDA}$ .

Figures 3 and 4 show the  $GRAN3SAT_{DHNN}^{EDA}$  energy error distribution  $MAE_{energy}$  and the global solution proportion  $ZM_{test}$  under different  $P_N$  during the test phase. It can be obtained that the average energy distribution of  $GRAN3SAT_{DHNN}^{ES}$  with different

$P_N$  is between 4.6 and 5.0. With the decrease of  $P_N$  under EDA, the energy distribution decreases gradually from 4.5 to 2.8. The energy distribution of ES is generally higher than that of EDA. It can be seen that as the positive state of neurons increases, it is easier to obtain a smaller energy error and  $GRAN3SAT_{DHNN}^{EDA}$  larger global solution ratio than  $GRAN3SAT_{DHNN}^{ES}$ . It is mainly due to two factors, the first is that it is easier to obtain satisfactory synaptic weights in the learning phase, and the second is that as  $P_N$  becomes larger, the generated clauses are easier to increase the number of global solutions.

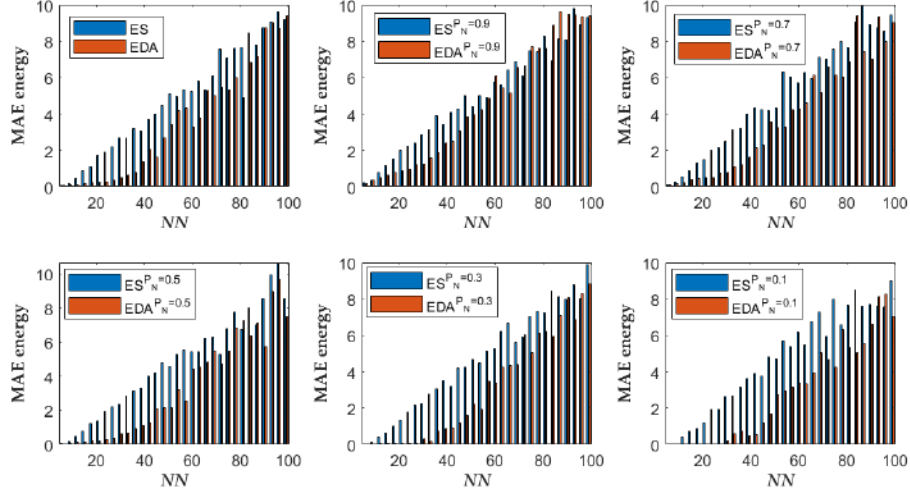


Figure 3: Changes in  $MAE_{energy}$  of EDA and ES under different  $P_N$ .

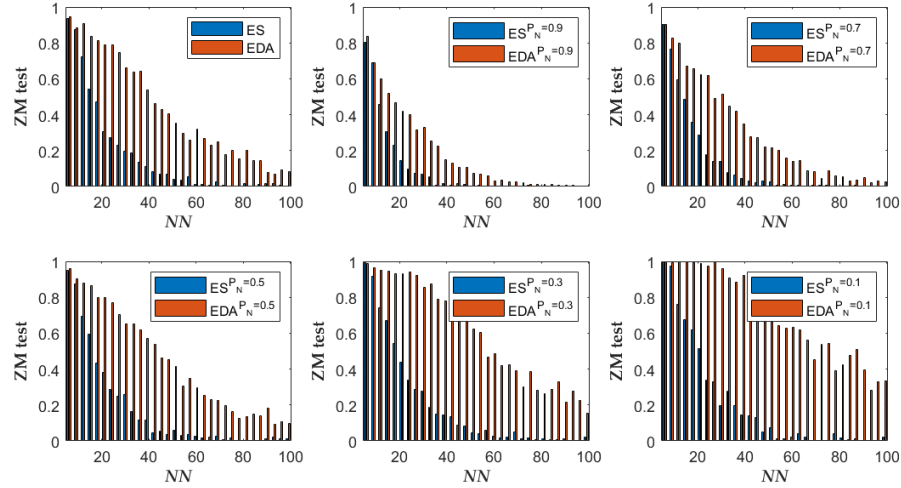


Figure 4: Changes in  $ZM_{test}$  of EDA and ES under different  $P_N$ .

## 5 CONCLUSIONS

The learning mechanism of  $GRAN3SAT_{DHNN}$  based on the EDA algorithm has the following conclusions: Compared with ES, it has a larger search space at the same efficiency, so the probability of obtaining satisfactory weights in the learning phase is higher, and the proportion of global solutions obtained in the testing phase is higher. As the  $NN$  increases, the advantages of EDA become more obvious. Different proportions of negative literals  $P_N$  has no significant impact on neuron fitness and weight error in the learning phase, and a smaller  $P_N$  in the testing phase helps to reduce the energy distribution and increase the proportion of the global solution, thereby improving  $GRAN3SAT_{DHNN}^{EDA}$  performance.

## ACKNOWLEDGMENTS

Chengdu University of Traditional Chinese Medicine Xinglin Scholar Project.ZRQN2018013,QNXZ2018042.

## REFERENCES

- [1] Atul Adya,Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the national academy of sciences, 79(8), 2554-2558. <https://doi.org/10.1073/pnas.79.8.2554>
- [2] McCulloch W S, Pitts W. 1990. A logical calculus of the ideas immanent in nervous activity. Bulletin of mathematical biology, 52(1), 99-115.W. A. T. W. Abdullah, International journal of intelligent systems. 7, 513-519 (1992). <https://doi.org/10.1007/BF02459570>
- [3] Abdullah, W. A. T. W. 1992. Logic programming on a neural network. International journal of intelligent systems, 7(6), 513-519. <https://doi.org/10.1002/int.4550070604>

- [4] Mansor M A, Sathasivam S. 2016. Accelerating activation function for 3-satisfiability logic programming. *International Journal of Intelligent Systems and Applications*, 8(10), 44. <https://doi.org/10.5815/ijisa.2016.10.05>
- [5] Kasihmuddin M S M, Mansor M A, Sathasivam S. 2018. Discrete Hopfield neural network in restricted maximum k-satisfiability logic programming[J]. *Sains Malaysiana*, 47(6): 1327-1335. <http://dx.doi.org/10.17576/jsm-2018-4706-30>
- [6] Karim, S. A., Zamri, N. E., Alway, A., Kasihmuddin, M. S. M., Ismail, A. I. M., Mansor, M. A., Hassan, N. F. A. 2021. Random satisfiability: A higher-order logical approach in discrete Hopfield Neural Network. *IEEE Access*, 9, 50831-50845. <https://doi.org/10.1109/ACCESS.2021.3068998>
- [7] Guo, Y., Kasihmuddin, M. S. M., Gao, Y., Mansor, M. A., Wahab, H. A., Zamri, N. E., & Chen, J. 2022. YRAN2SAT: A novel flexible random satisfiability logical rule in discrete hopfield neural network. *Advances in Engineering Software*, 171, 103169. <https://doi.org/10.1016/j.advengsoft.2022.103169>
- [8] Zamri, N. E., Azhar, S. A., Mansor, M. A., Alway, A., & Kasihmuddin, M. S. M. (2022). Weighted Random k Satisfiability for k= 1, 2 (r2SAT) in Discrete Hopfield Neural Network. *Applied Soft Computing*, 109312. <https://doi.org/10.1016/j.asoc.2022.109312>
- [9] Gao, Y., Guo, Y., Romli, N. A., Kasihmuddin, M. S. M., Chen, W., Mansor, M. A., Chen, J. 2022. GRAN3SAT: Creating Flexible Higher-Order Logic Satisfiability in the Discrete Hopfield Neural Network. *Mathematics*, 10(11), 1899. <https://doi.org/10.3390/math10111899>
- [10] Mühlenbein, H., Paass, G. 1996. From recombination of genes to the estimation of distributions I. Binary parameters. In *International conference on parallel problem solving from nature*. 178-187. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/3-540-61723-X\\_982](https://doi.org/10.1007/3-540-61723-X_982)
- [11] Peralta, J., Gutierrez, G., Sanchis, A. 2010. Time series forecasting by evolving artificial neural networks using genetic algorithms and estimation of distribution algorithms. In *The 2010 international joint conference on neural networks (IJCNN)* (pp. 1-8). IEEE. <https://doi.org/10.1109/IJCNN.2010.5596892>
- [12] Donate, J. P., Li, X., Sánchez, G. G., de Miguel, A. S. 2013. Time series forecasting by evolving artificial neural networks with genetic algorithms, differential evolution and estimation of distribution algorithm. *Neural Computing and Applications*, 22(1), 11-20. <https://doi.org/10.1007/s00521-011-0741-0>
- [13] Mühlenbein, H. 1997. The equation for response to selection and its use for prediction. *Evolutionary computation*, 5(3), 303-346. <https://doi.org/10.1162/evco.1997.5.3.303>

# Face Anti-spoofing Method Based on Deep Supervision

Hongxia Wang  
School of Computer and Artificial  
Intelligence, Wuhan University of  
Technology, Wuhan, China  
whx\_green@whut.edu.cn

Li Liu  
School of Computer and Artificial  
Intelligence, Wuhan University of  
Technology, Wuhan, China  
ll\_qhx@whut.edu.cn

Ailing Jia  
School of Computer and Artificial  
Intelligence, Wuhan University of  
Technology, Wuhan, China  
jal@whut.edu.cn

## ABSTRACT

Although face recognition technology is extensively used, it is vulnerable to various face spoofing attacks, such as photo and video attacks. Face anti-spoofing is a crucial step in the face recognition process and is particularly important for the security of identity verification. However, most of today's face anti-spoofing algorithms regard this task as an image binary classification problem, which is easy to over-fit. Therefore, this paper builds the basic deep supervised network as the baseline model and designs the central gradient convolution to extract the pixel difference information within the local region. To reduce the redundancy of gradient features, the central gradient convolution is decoupled to replace the vanilla convolution in the baseline model to form two cross-central gradient networks. A cross-feature interaction module is then built to effectively fuse the networks. And a depth uncertainty module is built for the problem that most face datasets are noisy and it is difficult for the model to extract fuzzy region features. Compared with existing methods, the proposed method performs well on the OULU-NPU, CASIA-FASD, and Replay-Attack datasets.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision;

## KEYWORDS

Face anti-spoofing, Deep supervision, Central gradient convolution, Depth uncertainty learning

### ACM Reference Format:

Hongxia Wang, Li Liu, and Ailing Jia. 2023. Face Anti-spoofing Method Based on Deep Supervision. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590023>

## 1 INTRODUCTION

Face anti-spoofing is determining whether a face captured by the camera is a live face or a non-live face. In practical application scenarios, face anti-spoofing is an important prerequisite for face recognition tasks. Deep learning has made some progress in the

field of face anti-spoofing in recent years. However, as printed photos and videos become increasingly high-definition, the distinguishing features between real and fake faces are becoming fewer and fewer, and classical binary classification models are easy to over-fit. And vanilla convolution is very sensitive to illumination changes, resulting in models that are more affected by illumination changes.

The depth information has characteristics such as illumination invariance, which makes face anti-spoofing robust. And the depth maps of live faces have the contour features of a 3D face and are significantly different from the depth maps of photo faces and video faces. Therefore, the depth feature is very suitable as a distinguishing feature between live faces and non-live faces. However, due to the high cost of depth camera equipment, most algorithms directly use DepthNet [1] to extract features when designing a depth-supervision network to predict depth features through pixel-level supervision, resulting in certain limitations in the learned depth features. Therefore, this paper designs a depth-supervised network that allows the network to learn depth features from RGB images and then discriminate between real and fake faces based on depth information.

The main contributions of this paper are as follows. A basic depth-supervised network DepthNet is built as the baseline model, and CGC is designed to replace the vanilla convolution. And CGC is decoupled to replace the vanilla convolution in the baseline model to form two cross-central gradient networks, C-CGN(HV) and C-CGN(DG). CFIM is built to effectively utilize the relationship between the two networks. Finally a lightweight DUM [2] is introduced to mitigate the adverse effects of partial distortion in the generated images.

## 2 RELATED WORKS

Face anti-spoofing has been considered as a binary classification task early on, treating live faces as 1 and non-live faces as 0. As a result, many end-to-end deep learning approaches use binary cross-entropy loss for supervision. Yang et al. [3] first proposed to use of an 8-layer shallow convolutional neural network for face anti-spoofing and trained a classification network. However, the network structure of this algorithm is simple and does not extract face features sufficiently.

Despite the robustness of features extracted by deep learning methods, the performance of deep learning-based methods for face anti-spoofing has not been comparable to traditional methods until 2018 due to the generally large size of these network models, the large amount of data required to train a better model, and the extreme complexity of real scenes. The proposed pixel-level supervision method has greatly improved the performance. Pixel-level supervision can provide more fine-grained and context-relevant

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590023>

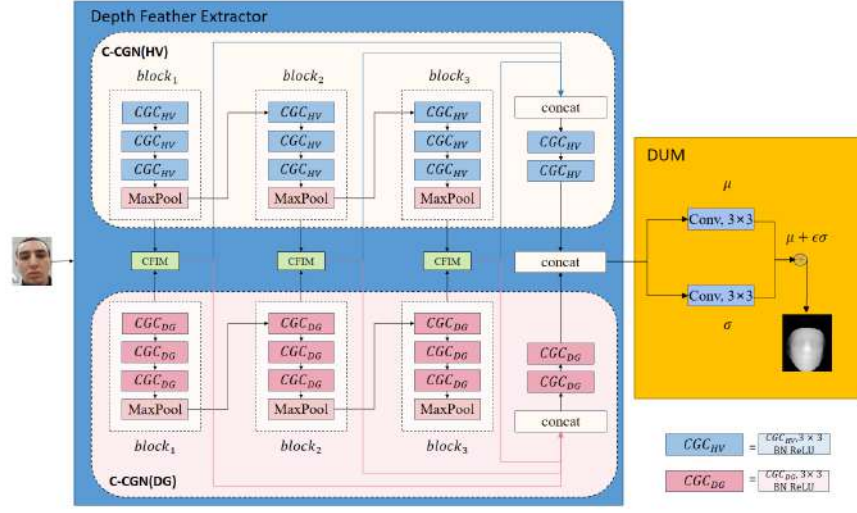


Figure 1: Overall structure of our proposed network.

cues for better intrinsic feature learning. One of the most commonly used auxiliary supervised signals is the pseudo-depth label.

Atoum et al. [4] first used the pseudo-depth label to guide multi-scale FCN, and used RGB images to generate depth maps as labels for supervision. This approach better exploits the fine-grained information between faces. George and Marcel et al. [5] introduced deep pixel-level binary supervision as auxiliary supervision to train face anti-spoofing networks, further illustrating the importance of pixel-level supervision. Liu’s team [1] used the pseudo-depth labels for supervision, i.e. live faces were labeled using face 3D reconstruction techniques to generate labels, and non-live faces were labeled using all-0 representation. And they proposed a CNN-RNN network that combines CNN and RNN, using the CNN part to extract the depth texture features of faces and the RNN part to extract the rPPG signal, combining the two signals and using them to uncover features of faces. This method is the more effective of the depth-supervised methods, but the RNN part of the network is more redundant, making the whole detection process unable to achieve real-time feedback and suffering from problems such as insufficient fine-grained extraction capability. Yu et al. [6] proposed a new convolutional operator, called central differential convolution, to replace the vanilla convolution in DepthNet, and designed a central differential convolutional network, using the depth map to guide the model that improves the model performance.

### 3 PEOPOSED METHOD

#### 3.1 Overview

In this paper, the mainstream DepthNet network is used as the backbone network of the depth feature extractor, and the convolution is optimized using the CGC. Further, to reduce the redundancy of gradient features, CGC is decoupled into two cross directions (horizontal/vertical and diagonal), forming C-CGN(HV) and C-CGN(DG). These two networks fuse features through CFIM adaptively. The extracted deep feature maps are fed into the DUM to mitigate the adverse effects of partial distortion in the generated images and

to enhance the network’s ability to extract features. The overall structure is shown in Figure 1.

#### 3.2 Depth feature extractor

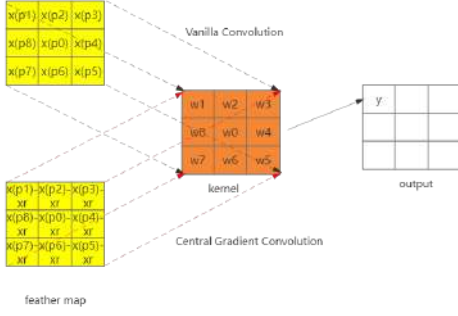
**3.2.1 Convolutional optimization.** The task of the deep-supervised network is to find the difference between live faces and non-live faces by learning its corresponding depth map from an RGB image. Vanilla convolution has a limited ability to extract such fine-grained features and is sensitive to illumination changes. CGC focuses more on local gradient relationships at the pixel level compared to vanilla convolution and enhances the network’s robustness to illumination changes.

To facilitate the understanding of CGC, this paper takes a single channel image as an example and uses a  $3 \times 3$  convolution kernel for CGC calculation, as shown in Figure 2. The process of vanilla convolution can be expressed as Eq. 1), where the input feature map  $x$ , each pixel point of the corresponding convolution kernel is multiplied with the weight of the convolution kernel, and then the weighted 9-pixel points are summed up as the new feature value  $y_1$ . CGC also consists of two steps: weighted multiplication and summation. The summation step is the same as the vanilla convolution, but the weighted multiplication process is different. CGC multiplies the difference between each pixel point of the corresponding convolution kernel and the average value of the pixel points in the local area with the weight of the convolution kernel to learn the gradient information between pixels in the local area, and the output feature map  $y_2$  can be expressed as Eq. 2) and Eq. 3).

$$y_1(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \quad (1)$$

$$y_2(p_0) = \sum_{p_n \in R} w(p_n) \cdot (x(p_0 + p_n) - x_r) \quad (2)$$

$$x_r = \frac{\sum_{p_n \in R} x(p_0 + p_n)}{n} \quad (3)$$



**Figure 2: Schematic diagram of the calculation process of vanilla convolution and CGC.**

Where  $x$  denotes the input feature map,  $w$  denotes the convolution kernel,  $y_1$  denotes the feature map computed by vanilla convolution,  $y_2$  denotes the feature map computed by CGC,  $x_r$  denotes the pixel average of the local receptive field  $R$ ,  $p_0$  denotes the current position on the input and output feature maps, and  $p_n$  denotes all positions on  $R$ .

Vanilla convolution focuses only on semantic information, while in the face anti-spoofing task, both semantic information and gradient-level detail information are crucial to distinguish live faces from spoofed faces. This suggests that combining vanilla convolution with CGC is a feasible approach to extract both intensity-level semantic information and gradient-level detail information, enhancing the robustness of the features. Therefore, the combination of vanilla convolution and CGC in a certain ratio, denoted as CGC (generalized), is shown in Eq. 4).

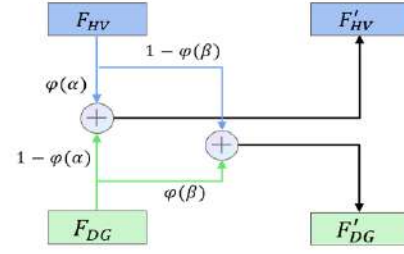
$$y(p_0) = \theta \cdot y_2(p_0) + (1 - \theta) \cdot y_1(p_0) \quad (4)$$

Where, the hyper-parameter  $\theta \in [0, 1]$  measures the contribution between semantic information and gradient information. The larger  $\theta$  is, the more important the gradient information is.

CGC calculates the entire local neighboring region  $R$  of the central gradient, resulting in redundancy. Therefore, the convolution is further sub-divided on the basis of the central gradient convolution and decoupled into two directions to obtain a convolution that extracts the horizontal and vertical cross-adjacent gradient features  $CGC_{HV}$  and a convolution extracting the diagonal cross-adjacent gradient features  $CGC_{DG}$ . Use  $CGC_{HV}$  and  $CGC_{DG}$  replacing the vanilla convolution, so the feature extraction network has two Cross-Central Gradient Networks C-CGN(HV) and C-CGN(DG).

**3.2.2 CFIM.** If two independent networks C-CGN(HV) and C-CGN(DG) are used for feature extraction of the input image and simply concatenating the results of the two networks together, will lack the message passing from the previous network stage. The performance improvement of CGC will be limited and the features extracted throughout the convolution process will not be maximized.

To effectively mine the relationship between the two networks and enhance the local detail representation, CFIM is used to adaptively fuse the multi-level features of the two networks. Specifically, the extracted horizontal-vertical cross-adjacent domain features



**Figure 3: Adaptive fusion process.**

$F_{HV}$  and diagonal cross-adjacent domain features  $F_{DG}$  are adaptively merged to generate two new fused improved features. The corresponding fusion equation is shown in Eq. 6) and Eq. 7), and the process is shown in Figure 3.

$$\varphi(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

$$F'_{HV} = \varphi(\alpha) \cdot F_{HV} + (1 - \varphi(\alpha)) \cdot F_{DG} \quad (6)$$

$$F'_{DG} = \varphi(\beta) \cdot F_{DG} + (1 - \varphi(\beta)) \cdot F_{HV} \quad (7)$$

Where  $\varphi(\cdot)$  denotes the Sigmoid function.  $\alpha$  and  $\beta$  are the weights of  $F_{HV}$  and  $F_{DG}$  respectively.

### 3.3 DUM

In face anti-spoofing algorithms, the quality of the collected face images and the accuracy of the generated 3D face depth map have a significant impact on the accuracy of the algorithm. However, the process of collecting face images and pre-processing the data always generates unavoidable errors and noise, making it difficult for the model to predict accurate depth values and therefore hindering depth-supervised optimization, resulting in some bias in the predicted depth values. To address this problem, a lightweight DUM is introduced to mitigate the adverse effects of noisy samples.

For each patch  $x_{i,j}$  of the input feature map  $x$ , instead of predicting a fixed depth value  $d_{i,j}$  in the training phase, the standard normal distribution  $z_{i,j}$  is learned in the latent space, as shown in Eq. 8). The final depth values are randomly sampled from the depth distribution  $z_{i,j}$ .

$$p(z_{i,j}|x_{i,j}) = N(z_{i,j}; \mu_{i,j}, \sigma_{i,j}^2) \quad (8)$$

Where  $\mu_{i,j}$  is the mean of the expected depth values of the image patches and  $\sigma_{i,j}$  is the standard deviation of the uncertainty in the prediction of  $\mu_{i,j}$ .

The structure of the DUM is illustrated in Figure 1 and it includes two separate convolutional layers behind the depth feature extractor. One is used to predict the mean  $\mu_{i,j}$  and the other for predicting the standard deviation  $\sigma_{i,j}$ .

However, the sampling operation is not differentiate during model training, thus hindering gradient back-propagation. To make the process learnable, a random noise  $\epsilon$  is sampled from a standard normal distribution  $N(0, I)$  independent of the model parameters as the uncertainty weight. The depth value shown in Eq. 9).

$$d_{i,j} = \mu_{i,j} + \epsilon \sigma_{i,j}, \quad \epsilon \sim N(0, I) \quad (9)$$

### 3.4 Loss function design

In this paper, the deep supervision method is used for face anti-spoofing, which requires pixel-level supervision of the depth map to estimate the difference between the real and predicted depth values. Therefore, as in [4], the MSE mean square loss function  $L_{MSE}$  is used for pixel-level supervision to help the model learn more discriminative distinguishing features, as shown in Eq. 10).

$$L_{MSE} = \sum_{i \in H, j \in W} (y_{i,j} - d_{i,j})^2 \quad (10)$$

Where  $y_{i,j}$  denotes the model predicted output depth value, and  $d_{i,j}$  denotes the depth map label corresponding to the RGB map.

However, [7] pointed out that the MSE mean square loss function only helped the network to learn the absolute distance between the face and the camera, ignoring the depth difference between adjacent pixels, yet the relative distance relationship between adjacent pixels is also important for the learning of depth maps. Therefore, they proposed the Contrastive Depth Loss function  $L_{CDL}$ , which performs adjacent pixel difference-level supervision.

DUM is introduced in our method for learning, therefore it needs to be supervised. As the representation of each image patch in the feature space is defined as a standard normal distribution, the learned feature distribution needs to be made to approximate the standard normal distribution of the face image labels as closely as possible. Therefore, the KL-Divergence loss function [8]  $L_{kl}$  is used.

The loss functions are combined to form a joint loss function, as shown in Eq. 11).

$$L = L_{MSE} + L_{CDL} + \lambda_{kl} L_{kl} \quad (11)$$

Where  $\lambda_{kl}$  is used to control the regularization term.

## 4 EXPERIMENTS

### 4.1 Datasets and Evaluation Metrics

This paper uses the commonly used datasets OULU-NPU dataset [9], CASIA-FASD dataset [10], and Replay-Attack dataset [11] in the face anti-spoofing algorithm. The results of the model are evaluated using the original evaluation metrics [12] of each dataset. The OULU-NPU dataset is evaluated using three commonly used evaluation metrics APCER, BPCER, and ACER. APCER is the Attack Presentation Classification Error Rate, which represents the proportion of prediction errors for spoofed face samples; BPCER is the Bona Fide Presentation Classification Error Rate, which represents the proportion of live face samples that are incorrectly classified; and ACER is the Average Classification Error Rate, which represents the combined error rate of the model. The half-error rate HTER is the mean of the false acceptance rate (FAR) and false rejection rate (FRR) and is used to evaluate cross-dataset experiments on the CASIA-FASD dataset and the Replay-Attack dataset.

### 4.2 Implementation Details

In this experiment, frames are sampled from the video datasets, and the ratio of the number of samples of live faces and spoofed faces is set to 1:1. The depth map labels of live faces are generated using PRnet, while the depth map of spoofed faces is set to 0. The model in this paper is implemented based on the PyTorch deep learning framework, and the Adam optimizer is used in the training phase to

Table 1: Results of ablation experiments

Model name	ACER(%)
DepthNet	3.8
DepthNet+DUM	3.6
CGN	0.9
C-CGN	0.7
Ours	<b>0.5</b>

set the initial learning rate is  $1e-4$  and the weight decay is  $5e-5$ . The hyper-parameter  $\theta$  of the CGC is optimally valued at 0.7 according to the experiments. In the test phase, the depth map is averaged as the final prediction value. If this value is greater than the threshold then the face is judged to be a live face, otherwise, it is judged to be a spoofed face.

### 4.3 Ablation Study

In order to verify the effectiveness of the model in this paper, ablation experiments were designed for DepthNet, DepthNet+DUM, CGN, C-CGN and our model. Among them, the DepthNet model is the baseline, DepthNet+DUM is a new DUM added to the baseline, CGN is the replacement of vanilla convolution with CGC, and C-CGN is the depth feature extractor of the improved model of this paper. The ablation experiments were conducted on protocol 1 of the OULU-NPU dataset, and the experimental results are shown in Table 1.

The results in the table 1 show that the ACER value of the model decreases by 0.1% after adding DUM to DepthNet, indicating that DUM improves the effectiveness of the model. The ACER value of C-CGN decreases by 0.9% compared to the original DepthNet network, and the performance is significantly improved. It shows that the CGC proposed in this paper has a stronger detail gradient feature extraction ability than vanilla convolution. The lowest ACER value of our model illustrates its effectiveness and improves the accuracy.

### 4.4 Intra Testing

To further validate the effectiveness of our model, experiments were conducted on protocol 1 and protocol 2 of the OULU-NPU dataset, and the final results were compared with other mainstream models.

As can be seen from table 2, the ACER value of our model on the OULU-NPU dataset protocol 1 and protocol 2 are 0.6 and 1.3, respectively, which is better than the mainstream methods. The experimental results for protocol 1 demonstrate that the model has better robustness under changing environmental conditions, that is, the effect of the model is less affected by changes in illumination and background scenes; the experimental results for protocol 2 demonstrate that the model has a lower error rate under new attacks, indicating that the model is less affected by new types of attacks and more generalizable.

### 4.5 Inter Testing

To further validate the generalization ability of the model, cross-dataset experiments are designed. As the CASIA-FASD dataset is relatively simple compared to the Replay-Attack dataset, its generalization ability can be better demonstrated by training on the

**Table 2: Results of experiments within the OULU-NPU dataset**

protocol	Model	APCER(%)	BPCER(%)	ACER(%)
protocol 1	GRADIANT [14]	1.3	12.5	6.9
	STASN [15]	1.2	2.5	1.9
	Auxiliary [1]	1.6	1.6	1.6
	FaceDs [16]	1.2	1.7	1.5
	FAS-TD [17]	2.5	0.0	1.3
	CDCN [6]	0.4	1.7	1.0
	Ours	0.3	0.7	<b>0.5</b>
protocol 2	GRADIANT [14]	3.1	1.9	2.5
	STASN [15]	4.2	0.3	2.2
	Auxiliary [1]	2.7	2.7	2.7
	FaceDs [16]	4.2	4.4	4.3
	FAS-TD [17]	1.7	2.0	1.9
	CDCN [6]	1.5	1.4	1.5
	Ours	0.7	1.9	<b>1.3</b>

**Table 3: Results of cross-dataset experiments**

Model	HTER(%)
FaceDs [16]	28.5
STASN [15]	31.5
Auxiliary [1]	27.6
FAS-TD [17]	17.5
CDCN [6]	15.5
Ours	<b>8.9</b>

CASIA-FASD dataset and testing on the Replay-Attack dataset. Therefore, only this experiment was done to compare other mainstream models.

From table 3, it can be seen that our model has a lower HTER relative to other mainstream models under the condition of cross-dataset. This indicates that the model has better generalization ability in different environments and can have better results in the case of environmental changes, and can effectively distinguish between live and fake faces in unknown environments.

## 5 CONCLUSION

In this paper, we propose a depth-supervised algorithm for face anti-spoofing based on the design of the central gradient convolution to extract pixel difference information within a local region, decouple the central gradient convolution to build two cross-central gradient networks, and fuse the networks using a cross-feature interaction module. Finally, a depth uncertainty module is built after the depth feature extractor. The algorithm of this paper is verified experimentally with ablation experiments, intra-dataset experiments, and cross-dataset experiments designed to verify the effectiveness and good generalization of the proposed model from multiple perspectives.

## REFERENCES

- [1] Liu Y, Jourabloo A, Liu X. Learning deep models for face anti-spoofing: Binary or auxiliary supervision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 389-398.
- [2] Wu H, Zeng D, Hu Y, *et al.* Dual Spoof Disentanglement Generation for Face Anti-spoofing with Depth Uncertainty Learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2021.
- [3] Yang J, Lei Z, Li S Z. Learn convolutional neural network for face anti-spoofing[J]. arXiv preprint arXiv:1408.5601, 2014.
- [4] Atoum Y, Liu Y, Jourabloo A, *et al.* Face anti-spoofing using patch and depth-based CNNs[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 319-328.
- [5] George A, Marcel S. Deep pixel-wise binary supervision for face presentation attack detection[C]//2019 International Conference on Biometrics (ICB). IEEE, 2019: 1-8.
- [6] Yu Z, Zhao C, Wang Z, *et al.* Searching central difference convolutional networks for face anti-spoofing[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5295-5305.
- [7] Wang Z, Yu Z, Zhao C, *et al.* Deep Spatial Gradient and Temporal Depth Learning for Face Anti-Spoofing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 5042-5051.
- [8] Chang J, Lan Z, Cheng C, *et al.* Data uncertainty learning in face recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 5710-5719.
- [9] Boulkenafet Z, Komulainen J, Li L, *et al.* OULU-NPU: A mobile face presentation attack database with real-world variations[C]//2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, 2017: 612-618.
- [10] Zhang Z, Yan J, Liu S, *et al.* A face anti-spoofing database with diverse attacks[C]//2012 5th IAPR international conference on Biometrics (ICB). IEEE, 2012: 26-31.
- [11] Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing[C]//2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). IEEE, 2012: 1-7.
- [12] Goicoechea-Telleria I, Fernandez-Saavedra B, Sanchez-Reillo R. An evaluation of presentation attack detection of fingerprint biometric systems applying ISO/IEC 30107-3[C]//International Biometric Performance Testing Conference. 2016.
- [13] Feng Y, Wu F, Shao X, *et al.* Joint 3d face reconstruction and dense alignment with position map regression network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 534-551.
- [14] Boulkenafet Z, Komulainen J, Akhtar Z, *et al.* A competition on generalized software-based face presentation attack detection in mobile scenarios[C]//2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2017: 688-696.
- [15] Yang X, Luo W, Bao L, *et al.* Face anti-spoofing: Model matters, so does data[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3507-3516.
- [16] Jourabloo A, Liu Y, Liu X. Face de-spoofing: Anti-spoofing via noise modeling[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 290-306.
- [17] Wang Z, Zhao C, Qin Y, *et al.* Exploiting temporal and depth information for multi-frame face anti-spoofing[J]. arXiv preprint arXiv:1811.05118, 2018.

# Genetic algorithm in hopfield neural network with probabilistic 2 satisfiability

Ju, J.C, Chen

School of Medical Information Engineering, Chengdu  
University of Traditional Chinese Medicine, Chengdu  
611137, China; School of Mathematical Sciences, Universiti  
Sains Malaysia, Penang 11800, Malaysia  
chenju@student.usm.my

Yuan, Y.G, Gao\*

School of Medical Information Engineering, Chengdu  
University of Traditional Chinese Medicine, Chengdu  
611137, China; School of Mathematical Sciences, Universiti  
Sains Malaysia, Penang 11800, Malaysia  
gaoyuan@student.usm.my

Chengfeng, C.Z, Zheng

School of Mathematical Sciences, Universiti Sains  
Malaysia, Penang 11800, Malaysia  
1002953832@qq.com

Yueling, Y.G, Guo

School of Mathematical Sciences, Universiti Sains  
Malaysia, Penang 11800, Malaysia  
guoyueling1982@163.com

## ABSTRACT

Genetic Algorithm (GA) is to convert the problem-solving process into a process similar to the chromosomal changes in biological evolution using the mathematical method and computer simulation operation. This meta-heuristic algorithm has been successfully applied to system logic and non-system logic programming. In this study, we will explore the role of the Bipolar Genetic Algorithm (GA) in enhancing the learning process of the Hopfield neural network based on the previous study of PRO2SAT, and generate global solutions of the Probabilistic 2 Satisfiability model. The main purpose of the learning phase of the PRO2SAT model is to obtain consistent interpretations and calculate the optimal prominence weights, and the GA algorithm is introduced to improve the ability of PRO2SAT to obtain consistent interpretation using its selection, crossover, and mutation operators, and thus to improve the ability of the logic programming model to get a global solution. In the experimental phase, simulation data are used for result testing, and three performance metrics are used to test the consistency interpretation and global solution acquisition ability of the proposed model, including mean absolute error, logic formula satisfaction ratio, and global minimum ratio. Experimental results show that GA, as a meta-heuristic algorithm, has better searching ability for optimal solution and can effectively assist logic programming.

## CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence.

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590024>

## KEYWORDS

Probabilistic 2 satisfiability, Machine learning, Systematic logic, Genetic algorithm, Explanation neural network

## ACM Reference Format:

Ju, J.C, Chen, Chengfeng, C.Z, Zheng, Yuan, Y.G, Gao, and Yueling, Y.G, Guo. 2023. Genetic algorithm in hopfield neural network with probabilistic 2 satisfiability. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590024>

## 1 INTRODUCTION

Artificial intelligence is an important topic of science and technology in the 21st century. It has great potential and value in many fields and is also a vane for the future development of an industry [1] [2]. Discrete hopfield neural network is one of the most typical feedback neural networks. The traditional DHNN network uses the asynchronous method to adjust the network state, and the synaptic weight matrix uses the symmetric matrix. This method can ensure that the network will eventually converge to a stable state for any initial state. However, for fully connected network DHNN, with the increase of neurons, the repeated calculation of DHNN synaptic weight will increase rapidly until the optimal synaptic weight is obtained. This greatly increases the network computing overhead. Therefore, it is necessary to implement symbolic rules to control DHNN modeling to ensure that DHNN can quickly determine the optimal synaptic weight and converge to the optimal solution.

To solve this problem, Wan Abdullah [3] proposed the method of logical programming on the Hopfield neural network in 1992, which is the first time that the SAT structure is embedded into DHNN by minimizing the cost function and finding a consistent interpretation. Sathasivam [4] used a new relaxation method to upgrade the method of Wan and presented a new concept of SAT named Horn Satisfiability (Horn SAT). Since then, a large number of researchers have started the study of logic programming. They can be divided into system logic and nonsystem logic. They can be divided into system logic and nonsystem logic. For system logic, Kasihmuddin. [5-7] first proposed the systematic logic by

embedding 2 Satisfiability (2SAT) in DHNN, which required that each clause contained only two literals, the literals were connected by disjunctions, and the clauses were connected by conjunctions. Mansor [8] proposed 3SAT in DHNN. The number of characters in the extended clause was 3, which innovatively expanded the cost function and energy function and made them applicable to the 3SAT logical structure. The accuracy of these two models can reach more than 90%, with a good global optimal solution and acceptable stability. Chen et al. [9] proposed PRO2SAT in DHNN. This paper introduces the control of positive literal proportion and positive literal distribution in system logic 2SAT, This is the first attempt to control the distribution and quantity of positive literals in system logic. For nonsystem logic, The main recent results are RAN3SAT [10, 11], RAN2SAT [12, 13], r2SAT [14], Maj2SAT [15].

Metaheuristic algorithms are methods for finding optimal or satisfactory solutions to complex optimization problems based on computational intelligence mechanisms, Sometimes referred to as an intelligent optimization algorithm, it is popular with researchers because it is easy to implement and provides feasible solutions to problems at a reasonable cost (i.e., computational time and space). Currently, more multivariate heuristic algorithms have been applied to logic programming. For example, Genetic Algorithm (GA) [5, 16], Election Algorithm (EA) [17], and Artificial Bee colony Algorithm (ABC) [18], etc. All these studies have proved the auxiliary optimization function of the meta-heuristic algorithm for logic programming. Therefore, in this paper, we will also embed the meta-heuristic algorithm on the previous work done by the author to optimize the PRO2SAT model.

## 2 PROBABILISTIC 2 SATISFIABILITY REPRESENTATIONM

The probability 2 Satisfiability is known as a systematic logic rule represented in the Conjunctive Normal Form (CNF). This was introduced as the new System SAT curriculum. Given the expected number of body words, each clause of a logical expression must contain 2SAT. It controls the distribution of positive literals by the preset probability of positive literals. Meanwhile, it tries to make the two literals in each clause not to be positive at the same time. The PRO2SAT logical formula consists of the following contents:

- Variable:  $x_1, x_2, x_3, \dots, x_i, \dots, x_n$ , which have values of 1 or -1.
- Literal: For any variable  $x_i$ , both  $x_i$  and  $\neg x_i$  are literals, Where  $x_i$  is positive literal and  $\neg x_i$  is negative literal.
- Clauses:  $F_1, F_2, F_3, \dots, F_i, \dots, F_n$ , Where  $n$  is the number of clauses, and each clause has only two characters.  $F_i = x_{2i-1} \vee x_{2i}$ .
- control probability:  $p_1, p_2, p_3, \dots, p_i, \dots, p_n$ , each variable has a control probability. This probability represents the probability that the variable will be positive. It can be expressed as:  $p_i x_i$ , if  $p_i = 0.3$ ,  $x_i \in \{\neg x_i\}$  has a probability of 0.3, otherwise,  $x_i \in \{x_i\}$  has a probability of 0.7.

Hence the general formula for PRO2SAT is presented as follows:

$$P_{PRO2SAT} = \wedge_{i=1}^m F_i \quad (1)$$

$$F_i = \vee_{l=2i-1}^{2i} p_l x_l \quad (2)$$

$$p_l = \begin{cases} \xi & , \text{ if } (\eta_{l-1} \geq \eta) \text{ or } (l = 2i \text{ and } x_{l-1}) \\ \max(\eta, 1 - \eta) & , \text{ otherwise} \end{cases} \quad (3)$$

$$\eta_i = \frac{1}{i} \sum_{j=1}^i N_j \quad (4)$$

$$N_j = \begin{cases} 1, & \text{if } x_j \\ 0, & \text{if } \neg x_j \end{cases} \quad (5)$$

Among them, I indicates the relationship of the male role in the previous I verb. The value range is [0.1, 0.9]. The value of  $\xi$  is 0.01, which is used to ensure that the variable is negative literal.

## 3 LOGIC PROGRAMMING IN DISCRETE HOPFIELD NEURAL NETWORK

Discrete Hopfield Neural Network (DHNN) has good dynamic characteristics and associative memory function. In the absence of learning samples, it can replace a large number of learning samples through memory learning mode to improve the accuracy of network output. DHNN once revealed new research paths for the development of the artificial neural network and was the pioneer in seeking to resolve problems in linear programming. The update equation of DHNN is as follows:

$$S_i = \begin{cases} 1, & \text{if } \sum_j W_{ij} S_j \geq \theta \\ -1, & \text{otherwise} \end{cases} \quad (6)$$

Where  $S_i$  is the state of the  $i$ -th neuron, and  $S_j$  is the state of the  $j$ -th neuron connected to  $S_i$ .  $W_{ij}$  is the Synaptic weight between neuron  $i$  and neuron  $j$ , and there is no self-loop in the Hopfield neural network, so  $W_{ii} = W_{jj} = 0$ ,  $W_{ij} = W_{ji}$ .  $\theta$  is the pre-determined threshold value. When  $\sum_j W_{ij} S_j \geq \theta$ , the neuron is in the activated state which can be represented by 1. Otherwise, the neuron is in the inhibitory state which can be represented by 0. The updating rule of neuron state is shown in Equation (7):

$$S_i(t) = \begin{cases} 1, & \text{if } \tanh[h_i] \geq 0 \\ -1, & \text{otherwise} \end{cases} \quad (7)$$

Where  $S_i(t)$  represents neuron state at  $t$ , and  $\tanh$  represents the activation function, which is hyperbolic activation function (HTAF).  $h_i$  represents the local field of the  $i$ -th neuron, as shown in Equation (8):

$$h_i = \sum_{j=1, i \neq j}^{2n} W_{ij}^{(2)} S_j + W_i^{(1)} \quad (8)$$

The energy function is used to test the validity of the propositional logic embedding HNN, and the final energy value of the system is calculated by the Lyapunov Energy Function, as shown in Equation (9):

$$E_{PRO2SAT} = -\frac{1}{2} \sum_{i=1, i \neq j}^{2n} \sum_{j=1, i \neq j}^{2n} W_{ij}^{(2)} S_i S_j - \sum_{i=1}^{2n} W_i^{(1)} S_i \quad (9)$$

When the energy value reaches the global minimum energy, the system reaches a stable state.

**Table 1: List of parameters used for PRO2SAT-GA**

Parameter	Parameter Value
Neuron combination ( $\alpha$ )	100
Number of learning ( $\beta$ )	100
Tolerance value ( $Tol$ )	0.001
Proportion of positive literal ( $\eta$ )	$\{0.1, 0.2, \dots, 0.9\}$
Control probability ( $p_l$ )	$\{\xi, \max(\eta, 1 - \eta)\}$
Number of generations	100
Number of chromosomes	100
Crossover rate	0.9
Mutation rate	0.01

**Table 2: List of parameters used for PRO2SAT-ES**

Parameter	Parameter Value
Neuron combination ( $\alpha$ )	100
Number of learning ( $\beta$ )	100
Tolerance value ( $Tol$ )	0.001
Proportion of positive literal ( $\eta$ )	$\{0.1, 0.2, \dots, 0.9\}$
Control probability ( $p_l$ )	$\{\xi, \max(\eta, 1 - \eta)\}$

#### 4 GENETIC ALGORITHM IN PRO2SAT

In this paper, we combine the GA algorithm with the training phase of PRO2SAT and seek a consistent explanation. The purpose of PRO2AT genetic algorithm is to find the maximum fitness, and the objective function equation is as follows:

$$Max [f_{PRO2SATGA}] \quad (10)$$

Where,  $f_{PRO2SATGA}$  is the number of satisfied clauses, which is also fitness. The GA algorithm helps PRO2AT find consistent interpretation during the training phase. Here is the fitness equation:

$$f_{PRO2SATGA} = \sum_{i=1}^m C_i \quad (11)$$

Where,  $C_i$  is the second-order logical clause. It is defined as follows:

$$C_i = \begin{cases} 1, & \text{The clause is satisfied} \\ 0, & \text{Otherwise} \end{cases} \quad (12)$$

The maximum fitness value of each chromosome is equal to the number of clauses, that is,  $m$ .  $C_i = 1$  if the states of variable in PRO2SAT clause has consistent interpretations. Otherwise,  $C_i = 0$ .

#### 5 EXPERIMENTAL SETUP

Three performance metrics,  $MAE_{learn}$ ,  $LRTR$  and  $ZM$ , were used to compare the behavior of DHNNPRO2SAT-GA and DHNNPRO2SAT-ES. The operating environment of the experiment was MACOS edition with Apple M1 PRO chip and 16GB of RAM. The experimental model was coded using the Python programming language. Table 1, 2 listed the parameters used in PRO2SAT for GA, ES.

Mean absolute error ( $MAE_{learn}$ ): This metric is able to show evenly distributed errors to evaluate good error estimates. A good

logic will attain low values of  $MAE_{learn}$ .

$$MAE_{learn} = \frac{1}{\alpha} \sum_{j=1}^{\alpha} \sum_{i=1}^n \frac{|f_{\max} - f_i|}{n} \quad (13)$$

Where  $f_i$  represent the number of satisfied clauses,  $f_{\max}$  represent the total number of clauses.  $n$  is the number of iterations before  $f_{\max} = f_i$ .

Logic formula satisfaction ratio ( $LRTR$ ): This metric can show a uniform distribution of errors to provide a good estimate of the error,  $LRTR$  formulation is shown as follows:

$$LRTR = \frac{1}{\alpha} \sum_{i=1}^{\alpha} D_{satisfaction-learn} \quad (14)$$

When  $P_{PRO2SAT} = 1$ ,  $D_{satisfaction-learn} = 1$ , indicating that the logic is satisfied.

Global Minimum Ratio ( $ZM$ ): The global minimum ratio is the generalized metric for evaluating the efficiency of the solutions.  $ZM$  formulation is shown as follows:

$$ZM = \frac{1}{\alpha\beta} \sum_{i=1}^{\alpha\beta} G_p \quad (15)$$

Where  $G_p = 1$  when the solution of the model is global, otherwise  $G_p = 0$ .

#### 6 RESULT AND DISCUSSION

$MAE_{learn}$  and  $LRTR$  are used to analyze the ability of logic rules to find consistent interpretation in the initial state. But they have different units of computing,  $MAE_{learn}$  computes clause satisfiability,  $LRTR$  computes the satisfiability of logical rule ( $P_{PRO2SAT}$ ).  $ZM$  is a key metric to measure model performance. The larger the  $ZM$ , the more global solutions.

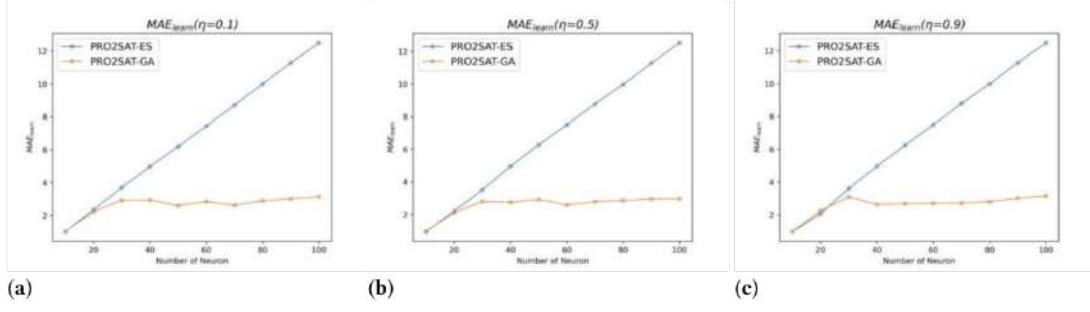
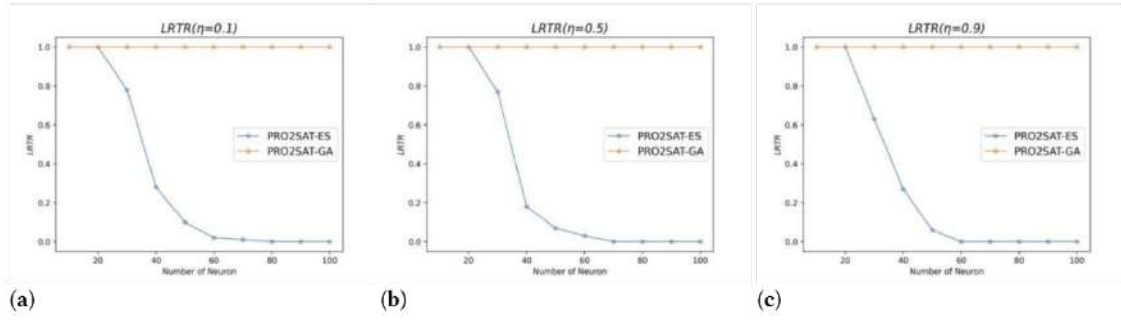
Figure 1:  $MAE_{learn}$  for PRO2SAT-GA and PRO2SAT-ES.

Figure 2: LRTR for PRO2SAT-GA and PRO2SAT-ES.

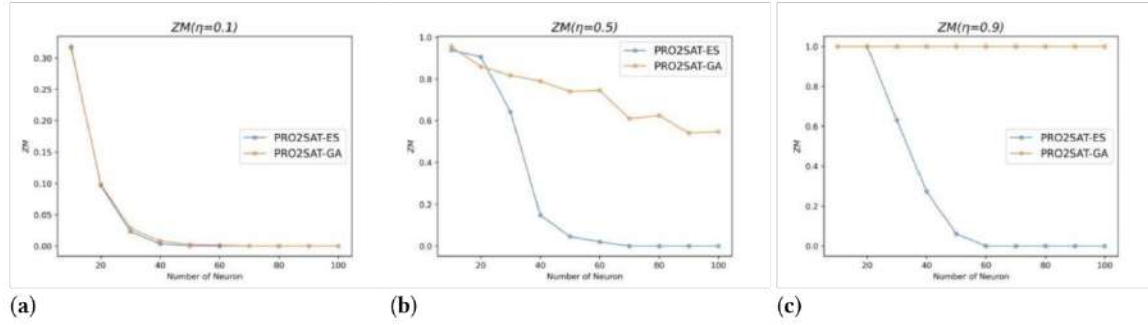


Figure 3: ZM for PRO2SAT-GA and PRO2SAT-ES.

Simulation data were used in our experiments to compare the experimental results. PRO2SAT-ES used the exhaustive search (ES) method to find consistent interpretations, and it was used as a baseline model to compare with the proposed PRO2SAT-GA model. PRO2SAT-GA used the genetic algorithm (GA) for consistent interpretation. Figure 1 shows a metric comparison trend between the two models with several neurons ranging from 10 to 100. The complexity of all models increases as the number of neurons increases. As can be seen from Figure 1, the ability of the PRO2SAT model to find consistent interpretation is significantly improved by adding GA, especially when the number of neurons is 100, it still has a good learning effect. In Figure 2, the LRTR value of the

PRO2SAT-ES model presents a curve downward trend with the growth of neurons. On the contrary, PRO2SAT-GA with the help of the GA algorithm has an excellent consistent interpretation-finding ability. Figure 3 shows the comparison trend of ZM metrics for the two models. When the proportion of positive literals is greater than 0.3, the PRO2SAT-GA model can converge to a more accurate final neuron state and obtain more global minimum solutions. When the positive literal proportion is less than or equal to 0.3, there are many clauses that make the model trap into local minimum solutions.

PRO2SAT-GA has more obvious performance advantages. The reason is that the GA algorithm introduces the learning phase of logical transformation, which improves the ability of PRO2SAT

to find consistent interpretations. If the model can find consistent interpretations, it can obtain the optimal prominence weight, and it has a higher probability to have the global minimum solution under the guidance of the optimal weight.

## 7 CONCLUSION

In this paper, based on the PRO2SAT-ES model, which is the previous research of the author, the GA meta-heuristic algorithm is introduced to generate a new Hopfield neural network logic programming model (PRO2SAT-GA). And it can be used as the symbolic instruction of HNN. We compare the results before and after model optimization, and the finding shows that PRO2SAT-GA outperforms PRO2SAT-ES, due to its effective learning mechanism. GA can effectively aid the computation of the learning phase of the model by obtaining lower errors and improving the quality of the solution in DHNN. The three performance metrics introduced in this paper can demonstrate the performance of the proposed model in the learning and testing phases. The next step in the research of this model will be to use real-world data for logic mining to solve optimization problems that exist in life.

## ACKNOWLEDGMENTS

Chengdu University of Traditional Chinese Medicine Xinglin Scholar Project. QNXZ2018042,ZRQN2018013

## REFERENCES

- [1] Beam, A. L.; Kohane, I. S. Big Data and Machine Learning in Health Care. *JAMA* 2018, 319, 1317-1318. <https://doi.org/10.1001/jama.2017.18391>
- [2] Vigilante, K.; Escaravage, S.; Mc Connell, M. Big Data and the Intelligence Community-Lessons for Health Care. *N Engl J Med* 2019, 380, 1888-1890. <https://doi.org/10.1056/NEJMp1815418>
- [3] Abdullah, W. A. T. W. 1992. Logic programming on a neural network. *International journal of intelligent systems*, 7(6), 513-519. <https://doi.org/10.1002/int.4550070604>
- [4] Sathasivam, S. "Upgrading logic programming in Hopfield network," *Sains Malaysiana* 2010, 39, 115–118.
- [5] Kasihmuddin, M. S. M., Mansor, M. A., & Sathasivam, S. (2017). Hybrid Genetic Algorithm in the Hopfield Network for Logic Satisfiability Problem. *Pertanika Journal of Science & Technology*, 25(1).
- [6] Mohd Kasihmuddin, Mohd Shareduwan, Mohd. Asyraf Mansor, Md Faisal Md Basir, and Saratha Sathasivam. 2019. "Discrete Mutation Hopfield Neural Network in Propositional Satisfiability" *Mathematics* 7, no. 11: 1133. <https://doi.org/10.3390/math7111133>
- [7] Kasihmuddin, Mohd Shareduwan Mohd, Siti Zulaikha Mohd Jamaludin, Mohd. Asyraf Mansor, Habibah A. Wahab, and Siti Maisharah Sheikh Ghadzi. 2022. "Supervised Learning Perspective in Logic Mining" *Mathematics* 10, no. 6: 915. <https://doi.org/10.3390/math10060915>
- [8] Mansor, M. A.; Sathasivam, S. Accelerating activation function for 3-satisfiability logic programming. *International Journal of Intelligent Systems and Applications* 2016, 8, 44-50. <https://doi.org/10.5815/ijisa.2016.10.05>
- [9] Chen, Ju, et al. "PRO2SAT: Systematic Probabilistic Satisfiability logic in Discrete Hopfield Neural Network." *Advances in Engineering Software* 175 (2023): 103355. <https://doi.org/10.1016/j.advengsoft.2022.103355>
- [10] Gao, Yuan, Yueling Guo, Nurul Atiqah Romli, Mohd Shareduwan Mohd Kasihmuddin, Weixiang Chen, Mohd. Asyraf Mansor, and Ju Chen. 2022. "GRAN3SAT: Creating Flexible Higher-Order Logic Satisfiability in the Discrete Hopfield Neural Network" *Mathematics* 10, no. 11: 1899. <https://doi.org/10.3390/math10111899>
- [11] Karim, S. A., Zamri, N. E., Alway, A., Kasihmuddin, M. S. M., Ismail, A. I. M., Mansor, M. A., Hassan, N. F. A. 2021. Random satisfiability: A higher-order logical approach in discrete Hopfield Neural Network. *IEEE Access*, 9, 50831-50845. <https://doi.org/10.1109/ACCESS.2021.3068998>
- [12] Guo, Y., Kasihmuddin, M. S. M., Gao, Y., Mansor, M. A., Wahab, H. A., Zamri, N. E., & Chen, J. (2022). YRAN2SAT: A novel flexible random satisfiability logical rule in discrete hopfield neural network. *Advances in Engineering Software*, 171, 103169. <https://doi.org/10.1016/j.advengsoft.2022.103169>
- [13] Sathasivam, S., Mansor, M. A., Ismail, A. I. M., Jamaludin, S. Z. M., Kasihmuddin, M. S. M., & Mamat, M. (2020). Novel Random k Satisfiability for  $k \leq 2$  in Hopfield Neural Network. *Sains Malays*, 49, 2847-2857. <http://dx.doi.org/10.17576/jism-2020-4911-23>
- [14] Zamri, N. E., Azhar, S. A., Mansor, M. A., Alway, A., & Kasihmuddin, M. S. M. (2022). Weighted Random k Satisfiability for  $k = 1, 2$  (r2SAT) in Discrete Hopfield Neural Network. *Applied Soft Computing*, 109312. <https://doi.org/10.1016/j.asoc.2022.109312>
- [15] Alway, A., Zamri, N. E., Karim, S. A., Mansor, M. A., Mohd Kasihmuddin, M. S., & Mohammed Bazuhair, M. (2022). Major 2 satisfiability logic in discrete Hopfield neural network. *International Journal of Computer Mathematics*, 99(5), 924-948. <https://doi.org/10.1080/00207160.2021.1939870>
- [16] Zamri, Nur Ezlin, et al. "Multi-discrete genetic algorithm in hopfield neural network with weighted random k satisfiability." *Neural Computing and Applications* 34.21 (2022): 19283-19311. <https://doi.org/10.1007/s00521-022-07541-6>
- [17] Karim, Syed Anayet, Mohd Shareduwan Mohd Kasihmuddin, Saratha Sathasivam, Mohd. Asyraf Mansor, Siti Zulaikha Mohd Jamaludin, and Md Rabiul Amin. 2022. "A Novel Multi-Objective Hybrid Election Algorithm for Higher-Order Random Satisfiability in Discrete Hopfield Neural Network" *Mathematics* 10, no. 12: 1963. <https://doi.org/10.3390/math10121963>
- [18] Muhammad Sidik, Siti Syatirah, Nur Ezlin Zamri, Mohd Shareduwan Mohd Kasihmuddin, Habibah A. Wahab, Yueling Guo, and Mohd. Asyraf Mansor. 2022. "Non-Systematic Weighted Satisfiability in Discrete Hopfield Neural Network Using Binary Artificial Bee Colony Optimization" *Mathematics* 10, no. 7: 1129. <https://doi.org/10.3390/math10071129>

# Construction of Scene Library System for Commercial Vehicle Products Based on Multidimensional Terminal

Zheng Yunshuang, Z, Zheng\*  
China Auto Information Technology  
(Tianjin) Co., Ltd  
zhengyunshuang@catarc.ac.cn

Hu Shilan, H, Hu  
China Auto Information Technology  
(Tianjin) Co., Ltd  
hushilan@catarc.ac.cn

Xue Nannan, X, Xue  
China Auto Information Technology  
(Tianjin) Co., Ltd  
xuenannan@catarc.ac.cn

## ABSTRACT

Scene-based product design is an effective way to improve user experience, and static scene library is an important basis for commercial vehicle product planners to carry out their design and planning. This paper explores the characteristics of commercial vehicle's dynamic scene and static scene, as well as the current situation and problems of commercial vehicle's static scene library. Furthermore, using the research methods of combinatorial design and task analysis, a digital solution for the construction of commercial vehicle's static scene library is proposed in this paper with reference to the construction process of vehicle's dynamic scene library. This paper innovatively proposes the combination of scene information collection and scene creation scale by means of structured questionnaire, which provides a reference for the construction of commercial vehicle's static scene library. This research helps automobile enterprises systematically manage the survey results, expand the application scope of the survey data, extend the application time of the survey results, and reduce the enterprise research costs.

## CCS CONCEPTS

• Information systems; • Information systems applications;

## KEYWORDS

Digital, Static Scene, Commercial Vehicle, Construction Mode

### ACM Reference Format:

Zheng Yunshuang, Z, Zheng\*, Hu Shilan, H, Hu, and Xue Nannan, X. 2023. Construction of Scene Library System for Commercial Vehicle Products Based on Multidimensional Terminal. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590025>

## 1 INTRODUCTION

Scene-based product design is an effective way to improve user experience. More and more car companies are aware of the importance of the scene, and the product research around the scene

has become an important basis for commercial vehicle product researchers and product planners to design and plan products. OEMs spend a lot of money on scene research every year. The essential difference between static scene and dynamic scene is whether the data source is car networking data. The dynamic scene library based on car networking data can directly reflect users' usage behavior, but it is deficient in users' subjective attitude. The car static scene library is different from the dynamic scene library constructed by car networking. It is mainly constructed by user research. Through in-depth mining of user perception, high-value application scenes are found, typical features and elements are extracted, and static scenes of automobiles are constructed. The research of static scenes has irreplaceable value compared with that of dynamic ones. Building the static scene library of commercial vehicles systematically and efficiently by digital means can provide important support for automobile product planning. Taking the field of commercial vehicles as an example, this paper gives an analysis on the digital construction method of static scene library, and provides reference for the digital construction of automobile static scene.

## 2 PRESENT RESEARCH ON THE THEORY OF SCENE

The "scene" first refers to the space scenery of artistic works such as theatrical performances [1]. At present, the commonly mentioned scenes come from the concept of user experience put forward by the Internet industry [2]. Robert's "Five Forces of Scenery" provides a new idea for the study of scenes [3]. Since then, the study of scenes has been deepened. Scenes are the bearing containers that connect users with their behaviors. The relationship of scenes-people-behaviors proposed by Jiang Haiyang makes it clear the necessity of scene research [4]. Scenes describe the background information about users, product use environment, users' purpose or goal, and a series of activities and events of users [5]. Scene is a form of story-telling transformation of the contact process between users and products [6].

Compared with other product fields, the research on scenes in automobile has richer dimensions. First of all, from the perspective of content, influenced by the development of advanced technologies such as 5G technology, car networking technology, unmanned driving technology, etc., it extends a wealth of scene research categories and research directions such as automatic driving scene, driving scene, accident analysis scene, car test scene, road simulation scene, etc [7-8]. The main scene construction methods are as follows: 1) extracting static scenes from user research; 2) recording the user's driving data from the vehicle-mounted terminal equipment and vehicle networking technology, and building a dynamic scene library; 3) extracting special scenes from monitoring image

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590025>

analysis of special scenes, such as the construction of accident analysis scene library [9]. Among them, it has become a hot research topic this year to extract dynamic scenes from the information of driving behavior recorded by the vehicle-mounted terminal equipment. In this process, the elements are directly extracted through the existing algorithm logic, and the scene library construction and application are basically digitized. However, it is mainly used for risk prediction, behavior warning, and vehicle state monitoring during driving [10-11]. There are few references to successful cases of digital schemes and products in static scene extraction and construction based on user research.

### 3 ANALYSIS OF THE CONSTRUCTION PROCESS OF COMMERCIAL VEHICLE STATIC SCENE LIBRARY

#### 3.1 Problems in the construction of static scene library

According to the research results published by various enterprises and research institutions, it is found that all OEMs unanimously agree on the importance of scene research, and scenes once became the high-frequency vocabulary of international forums and conferences, concepts like "scenes define cars" frequently put forward. However, after careful study, it was found that the scene research of each enterprise is not systematic, and the research methods and ideas are not unified. During the study, by contacting commodity planners of commercial vehicle enterprises, a survey on the current situation of scene research was carried out, and it was found that there are still many problems in the scene research process, mainly reflected in the following aspects.

1) The definition of scenes is difficult to be clear: different OEMs have different definition standards for scenes, and even the definition dimension of products in the same enterprise has not formed a standard, which leads to the lack of comparability among various scenes obtained in research. 2) Scene division is difficult to form a system: the main data extracted from static scenes comes from users. The unpredictable characteristics of users make it difficult to exhaust all kinds of scenes and find a clear basis of division for those scenes. 3) It is difficult to set the weight of the scene: even if the scene is divided and extracted based on various dimensions, and the index system for constructing the scene is formulated, the weight system of each index is still difficult to establish. 4) Scene extraction is difficult to be scaled. Scene research is based on user research, and different batches of research are separated from each other. As a consequence, the research results are difficult to integrate, and the extracted static scenes are difficult to be scaled, which cannot meet the demand of a scene library.

#### 3.2 Analysis of static scene library construction process

There are differences between static scene and dynamic scene in the data source of scene extraction, the construction target of scene library and the application object of scene library. However, from the perspective of the construction process of scene library, the construction method of dynamic scene library can be a reference for the construction of digital static scene library. The construction



Figure 1: Core steps of dynamic scene library construction

process of dynamic scene library is shown in the figure. The construction process of dynamic scene library is the process of dynamic data flow being processed based on the behavior data related to car networking. The essence of the construction process of static scene library is the process of collecting, extracting and storing related data of static library.

Aiming at the construction method and process of commercial vehicle static library, combined with the digital method of dynamic library construction, this paper explores the construction of commercial vehicle static scene library by digital method. As shown in Table 1, the construction process of the scene library of various enterprises is sorted out, and it is found that the process of commercial vehicle scene research mainly includes the following steps.

During the construction of static scene library in commercial vehicle field, it is found that the data types related to static scene construction are more diverse than those of dynamic scene, and the research of static scene involves various data types such as text, pictures, videos, etc. The biggest difficulty in the construction of static library is to realize the structured storage and application of the data related to static scene. Compared with the construction method of dynamic scene library, the direct data source of static scene is obtained through user investigation.

Combining with the four main problems faced by scene research, this paper explores the following problems to be solved in the digitalization of commercial vehicle static scene library construction: 1) Defining the concept of scenes and constructing the scene index system; 2) Integrating the data resources of each stage of scene research and standardizing management; 3) Accumulating data in the process of user research, and realizing standardized and large-scale management.

### 4 DESIGN OF COMMERCIAL VEHICLE STATIC SCENE LIBRARY SYSTEM

#### 4.1 Task extraction and analysis in static scene library construction process

The static scene library includes scene information, user information and vehicle information, as shown in Figure 2. The static scene library of commercial vehicles will integrate various database information such as vehicle database, questionnaire database and user portrait. The combination design method is used to design the system of commercial vehicle static scene library [12].

Table 1: Research Steps of Commercial Vehicle Field Scenes

Step	Content	Way
one	Establish scene index elements.	Literature research method
two	Scene listing based on vehicle type.	brainstorming
three	Disassemble elements and divide index weights for each scene.	Expert evaluation /SWOT analysis, etc.
four	Select specific scenes to conduct user research.	Questionnaire/interview and other research methods
five	Store scene research results in reported form.	without



Figure 2: Related elements of static scene library

Table 2: Task List of Scene Research in Commercial Vehicle Field

Main task	subtask
Create a scene	Setting scene indicators Define a new scene Index weight setting
Scene information collection	Research content design Scene information collection Scene information warehousing
Scene information application	Scene information retrieval Scene information supplement/update

The construction process of the digital commercial vehicle static scene library is shown in Figure 3. Firstly, the digital solutions of each step are designed based on the scene research process in the commercial field, and then the digital solutions of each module are integrated according to the task order. Finally, verify product effectiveness through end-user testing.

According to the steps of commercial vehicle scene research shown in Table 1, task extraction is carried out. The construction process of scene library can be summarized into three main tasks: scene creation, scene information collection and scene information application. Sub-tasks are continuously divided according to the main task, and the results are shown in Table 2.

Create a management module whose scene task is static scene library, and realize the multidimensional management of the scene. The field scene creation task of commercial vehicles needs to realize multi-level scene management, such as dividing according to the number of indicators. Scenes at the same level need to realize the classification of scene values, such as important scenes and common scenes. Scenes at each level have the same index elements, which is convenient for comparison and analysis of scenes at the same level in the application process of scenes.

The process of information collection mainly involves scene research, and the application of questionnaire research tools in this process can realize the digital storage of information. Information collection includes structured index information collection and personalized supplementary information collection of each scene. After the collection is completed, it can be directly linked to the warehouse.

Easy-to-use scene information application products should enable users to conduct multidimensional analysis. In system design, content tagging is an effective way to improve the flexibility in the use of data content.

4.2 Digital scheme of static scene library construction

Based on the task process analysis of the construction of commercial vehicle scene library, it can be concluded that the whole process of realizing the digitalization of static scene library is to integrate the scattered digital tools in all links. The last chapter mentioned four major problems in the process of scene construction. From the perspective of digital tools, the problems of creating and dividing scene definitions and indicators can be solved through standardized and structured storage of digital products in the process of scene library construction, but the definition methods and rules still depend on the input of concepts by researchers. It is difficult to set the index weight automatically by digital tools, and it is difficult to add it automatically based on the survey data with relatively small sample size. However, in terms of scene scale, it is found that commercial vehicle enterprises spend a lot of money on user research every year, and the scene library can be integrated and accumulated on a large scale with the help of the research system. The design scheme of commercial vehicle static scene library is shown in Figure 3.

Using big data thinking, scene thinking and research thinking to redesign the research system, the collected information is tagged, structured and stored in the smallest meaningful data unit, and the tag matching and retrieval are realized in the data analysis and application process. Combined with mobile tools such as applet

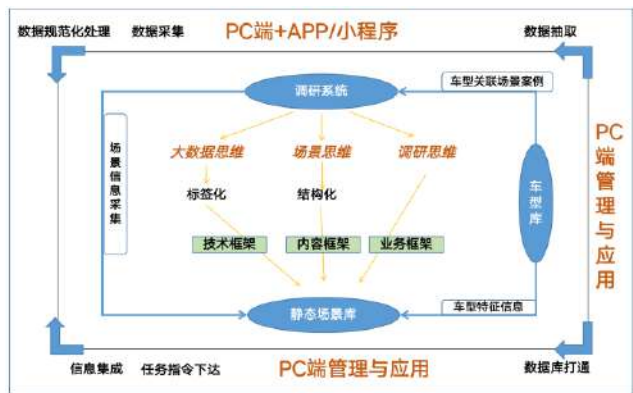


Figure 3: Solution of commercial vehicle static scene library)



Figure 5: scene information collection in applet



Figure 4: Screenshot of Tagged Design Page of Research System

/App, the informatization of the acquisition process is realized, and the directional data transmission process design of the research and scene module is realized by the way of task assignment and execution. The whole scene library creation process totally realizes digital management. In addition, the association between the enterprise scene library and the existing vehicle database is realized by opening the database interface. Figure 4 shows the data tagging management and application page of the research system after optimization based on the design scheme. Figure 5 shows the main page of information collection in applet.

4.3 Influence

The solution realizes the standardized collection and storage of index data in scene database through the combination of mobile phone and computer. The database can carry out long-term accumulation and storage of scene index data, which can provide support for the continuous research of scene database, reduce the repeated investment of data acquisition in scene research, improve the user experience in scene data collection and application, and reduce the time and labor cost of scene database construction.

5 SUMMARY

Scene-based product design is an effective way to improve user experience, and static scene library is an important basis for commercial vehicle product planners to carry out their design and planning. This paper mainly explores the characteristics of commercial vehicle dynamic scene and static scene, the present situation and problems of the construction of commercial vehicle static scene library, and puts forward a digital solution for the construction of commercial vehicle static scene library by applying research methods such as combination design and task analysis with reference to the construction process of commercial vehicle dynamic scene library, providing a reference for the construction of commercial vehicle static scene library. However, this paper only gives an analysis from the perspective of static scene library construction, and the combination with dynamic scene library is ignored. Furthermore, the long-term thinking of scene research still has certain limitations. A mature scene research system should be a systematic work that integrates users' static perception scenes and users' dynamic behavior scenes. The future design research process should be deeply analyzed from the perspective of integration of dynamic and static scenes.

REFERENCES

[1] Yuan Yuqing. Analysis of social media scene application in the era of mobile internet [J]. China Media Science and Technology, 2018, 000(003):111-114

[2] Peng Lan. Scene: a new element of media in the mobile era [J]. Journalist, 2015

[3] Robert Scober. The Coming Scene Era [M]. Beijing: Beijing United Publishing Company, 2014

[4] Jiang Haiyang, Mei Yun, Gu Xiansong. Analysis and Research of Scene Interaction Design Theory [J]. Packaging Engineering, 2019, v.40; No.408(18):281-287

[5] GO K, CARROLL J M, IMAMIYA A. Surveying Scene Based Approaches in System Design[J]. IPSJ SIG Notes, 2000: 43–48

[6] Zhao Wanru. On the application of scene stories in user experience design [J]. Design, 2014(9): 174–175.

[7] Chen Tao, Cai Bo, Hui Chun. Research on scene construction of intelligent networked car based on scene elements [J]. Highway and Automobile Transportation, 2019(6).

[8] Shu Hong, Yuan Kang, Xiu Hailin, et al. Research on the construction of basic test scene of self-driving car [J]. china journal of highway and transport, 2019, 32(11).

[9] Su Jiangping, Chen Junyi, Wang Hongyan, et al. Extraction and analysis of typical scenes of pedestrian traffic conflicts based on dangerous conditions in China [J]. Traffic and Transportation (Academic Edition), 2017, 01:216-221.

- [10] Ren Hongxia. Research on networked remote monitoring of engineering vehicles based on embedded processing [J]. Computer Measurement and Control, 2013, 21(011):2972-2974,2978.
- [11] Yang Le. Vehicle networking data monitoring system based on real-time streaming data platform [D]. University of Electronic Science and Technology of China, 2016.
- [12] Yang Bo, Huang Kezheng, Shine Wong, *et al.* Task allocation and planning for modular collaborative product design [J]. Modular Machine Tool and Automatic Processing Technology, 2004(07):22-24

# Research on Epidemic Big Data Monitoring and Application of Ship Berthing Based on Knowledge Graph-Community Detection

Shang,Dongfang

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China  
527226058@qq.com

Li Yuesong\*

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China  
lys25cn@126.com

Xu Jiashuai

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China,  
5220243@sohu.com

Bao Kexin

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China,  
363709838@qq.com

Wang Ruixi

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China,  
394540999@qq.com

Qin Liu

Tianjin Research Institute of Water  
Transport Engineering, Ministry of  
Transport, Tianjin 300451, China,  
qinliu0204@163.com

## ABSTRACT

The COVID-19 epidemic has been raging overseas for more than three years, and inbound goods and people have become the main risk points of the domestic epidemic. As the main window for China to exchange materials and personnel with foreign countries, under the dual pressure of the global economic downturn and the China-US economic confrontation, ports' pressure and responsibility to ensure material transportation and foreign trade are particularly heavy. However, the risk screening of ship and crew epidemic information based on manual methods is extremely time-consuming and labor-intensive, and it is difficult to take into account the efficiency and accuracy requirements of the port's own business and disease control and traceability. To this end, this study proposes an epidemic risk screening method based on knowledge graphs. This method is based on shipping big data and community discovery algorithms, analyzes the geospatial similarity of ship information, crew information and real-time epidemic policy information, and quickly establishes a structure. Map data, quickly screen high-risk ships and crew members, and access the business system to arrange nucleic acid testing tasks. When the time cost is only one thousandth of that of manual labor, the detection accuracy rate approaches and exceeds the accuracy level of manual screening, with an average precision advantage of 8.18% and an average time advantage of 1423 times. It is further found that it is more capable of performing heavy screening tasks than humans, and its AUC decline rate with the increase of the amount of measured data is only 34% of that of the manual method. The research results have been initially applied in Ningbo Port, which has greatly improved

the informatization level and screening efficiency of Ningbo Port's risk screening during COVID-19 epidemic.

## CCS CONCEPTS

• **Social and professional topics** → Computing / technology policy; Medical information policy; Health information exchanges.

## KEYWORDS

port, epidemic, risk screening, big data, knowledge graph, community detection

### ACM Reference Format:

Shang,Dongfang, Li Yuesong, Xu Jiashuai, Bao Kexin, Wang Ruixi, and Qin Liu. 2023. Research on Epidemic Big Data Monitoring and Application of Ship Berthing Based on Knowledge Graph-Community Detection. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590026>

## 1 INTRODUCTION

China's water transport industry has developed rapidly and is currently a world-class powerhouse in this sector. China's ports are the largest in the world. The national port's cargo throughput surpassed 14,55 billion tons (2020). In addition, the port has the highest container and cargo throughput in the world. Nonetheless, with the global outbreak of the new crown pneumonia in early 2020, the port, as the primary gateway for China's opening to the outside world, has become a frontline battleground to prevent the epidemic's spread abroad. How to realize the monitoring and early warning of massive epidemic information of ships and crew arriving in Hong Kong under the premise of ensuring smooth cargo transportation and orderly personnel shift duty are related to the timely and effective emergency handling of disease control in Hong Kong [1]. Consequently, the task of improving the use of epidemic data gathering and risk monitoring in port settings is particularly challenging [2].

Information technology has evolved into a potential tool for coordinating "anti-epidemic" efforts. Numerous academics have offered diverse information solutions for the complex duties of epidemic monitoring in logistics [3], ports [4], cruise ships [5] and so on [6]. Many industries have implemented information systems

\*Li Yuesong(1977-)male, Han nationality, from Langfang, Hebei, senior engineer. Research interests: hydraulic structure detection, diagnosis and reinforcement technology.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590026>

for monitoring epidemic risk. Huang [7] believed that information technology should be utilized to strengthen the tracking and early warning of the trajectory of epidemic areas and ships where the epidemic occurs, and to further reduce unnecessary contact during the epidemic through "paperless" and "contactless" management in order to increase the online order processing rate [8]. The processing and analysis of data and information are fundamental to the informatization process. Chatterjee [9] proposed a paradigm for research and evaluation of pandemic risk based on information aggregation. Xue [10] believed that it is of the utmost importance to save and evaluate pandemic information and data in a variety of sectors and industries, as well as to grasp the personnel information of key regions, key groups, and important industries. In addition, we should optimize the data management process, fully discover data information, and increase monitoring and identification links by examining the to-be-developed mechanism for assessing epidemic risk [11].

Intelligent monitoring based on epidemic data is also one of the most important research directions. Zhang [12] implemented the spatial analysis and knowledge graphs for maritime dangerous goods; Jiang [13] incorporated the control intelligence of the active distribution network during the epidemic through the similarity matching of knowledge graphs, confirming the efficiency improvement of knowledge graph algorithms under complex business logic and judgment requirements.

Numerous advances have been made in the field of industrialized epidemic risk information monitoring by a large number of scholars [14], which has strengthened the general stability and order of national epidemic prevention and control efforts. In terms of application depth, however, there is still potential for improvement. The most significant issues can be summed up in two points:

- (1) Due to the volume of epidemic data information monitoring and screening, it is challenging to meet the efficiency requirements of emergency management in emergency scenarios [15]. Most existing information platforms in the engineering business connect information such as production, manufacture, sales, warehousing, and logistics, but they lack deep integration with epidemic information and lack consistent information collaboration [16]. If you rely solely on manual control of continuously updated risk list information, it is time-consuming and laborious to quickly screen risk information, making it challenging to provide rapid and effective business support for the epidemic management of disease control departments during emergencies.
- (2) There is space for improvement in the accuracy of information [17]. From Alpha through Omicron BA.5, the infectivity and concealment of new forms of the new coronavirus have been regularly updated, putting a strain on manual screening [18] and posing a higher challenge to the discovery of risk information.

This paper provides a set of shipping big data platforms based on a knowledge graph community discovery algorithm that can be used for rapid epidemic risk screening in light of the aforementioned issues. Under the premise of ensuring the accuracy of screening for epidemic risk, it is possible to expand the screening scope and greatly enhance screening efficiency.

## 2 DATA STORAGE AND FUSION BASED ON SHIPPING BIG DATA PLATFORM

This study's algorithm-supporting platform is a scientific big data sharing service platform for water transport engineering (hereinafter referred to as the platform). This study collected current and historical data on coastal ports and waterways, inland waterways, water depth, topography, geology, hydrology, meteorology, waves, sediment, environmental and hydraulic buildings, infrastructure, and other current and historical data on coastal ports and inland waterways. Several examples of topical big data applications are now functioning on this platform. The "epidemic risk monitoring" application scenario of this study is a sub-module of the "Ningbo Port Line Cargo Safety Supervision Information System" platform module. The system has a B/S (browser/server) design, which allows users to simply log in, view, and control system content using a desktop computer, tablet, or mobile device. In addition, the architecture of decoupling data, functions, and display terminals is particularly advantageous for future zero-activity adjustments and dynamic additions of data and functions.

This platform's big data on water transport engineering has multi-source heterogeneous properties. Using HDFS as the underlying storage, distributed application coordination service Zookeeper, data warehouse Hive, and integrated distributed database HBase's highly available, elastic, and scalable storage architecture, the platform provides a foundation for intelligent data storage, processing, and modeling. In addition, the platform combines parallel file systems and highly redundant unstructured databases to achieve the dynamic expansion and material-understanding connection of compute units and storage units. This allows for the elastic expansion of data and hardware resources to fulfill the system's high concurrent access requirements.

In this research, multi-source heterogeneous data are sampled, screened, cleaned, formatted, and reconstructed, and integrated on logical and physical devices, so as to decrease the impact of the soft and hard environment on the data itself, in order to meet the requirements of data security. Finally, heterogeneous hazards are compiled into a set of abstract data models for the analysis and decision-making of various business topics. This study then encapsulates the data to create a specialized interface necessary for business applications. In this study, the business application comprises the "epidemic risk monitoring" application, which can readily collect the necessary data via the data engine and data access interface. In addition, this study develops a platform for the storage and fusion of large amounts of data that can accommodate the requirements of numerous shipping business applications, heterogeneous data sources, and multi-level distributed storage.

Therefore, this research can not only easily access and evaluate the vast topic data of the big data platform, but also conduct a more thorough and extended thematic data analysis and mining. This ensures the construction of a comprehensive knowledge map pertaining to ship pandemic hazards and improves the precision of similarity analysis.

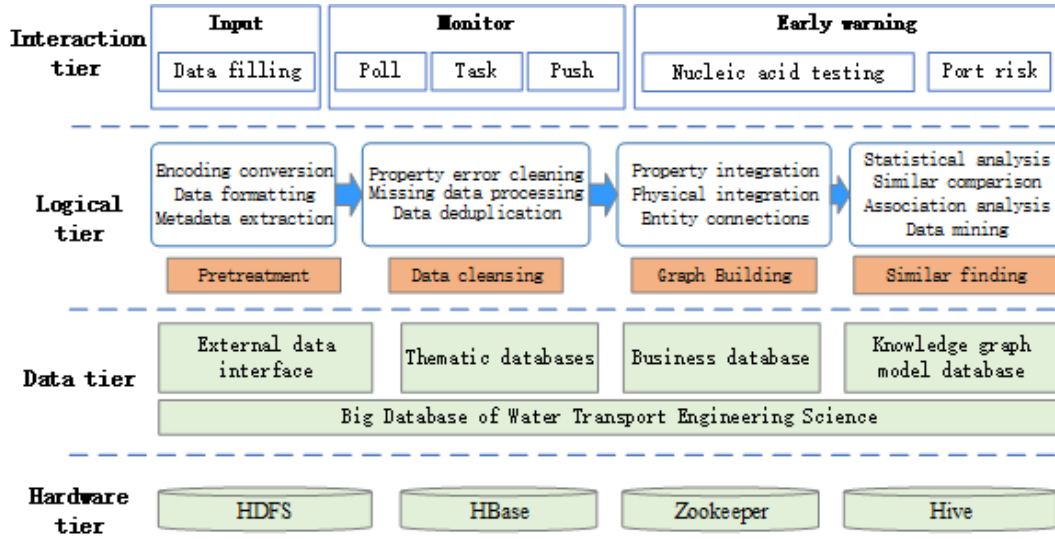


Figure 1: Schematic illustration of the risk control system for ship epidemic prevention and control in Ningbo port

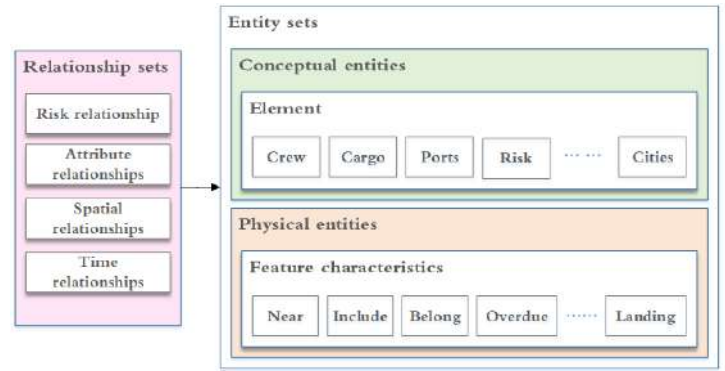


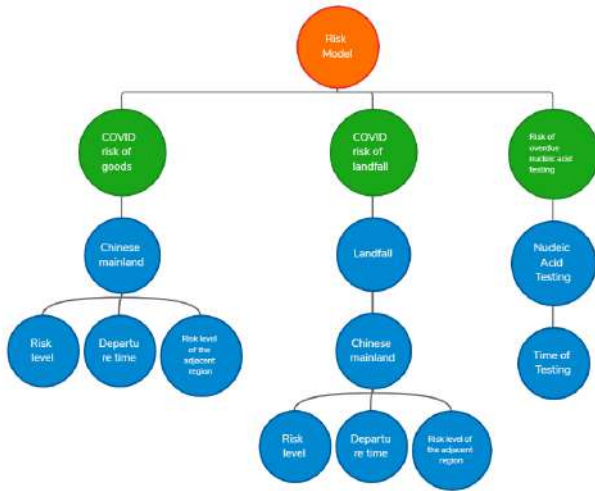
Figure 2: Schematic diagram of the construction idea of knowledge graph

### 3 METHOD

#### 3.1 Knowledge graph construction ideas

Developing a knowledge graph of ship pandemic prevention and control serves as the foundation for comparing the existing risk model with the characteristics of ships entering port. The key components of the knowledge graph structure are the conceptual ontology set and the relationship ontology set, which correspond to the entity and relationship sets in the ship epidemic prevention and control knowledge graph (Figure 2). The ship epidemic prevention and control model's core idea ontology is abstracted separately from the port, crew, ship, and cargo. Ship Features and Feature Features comprise the major tier, whereas the second tier is successively populated as a supplement or annotation to the primary tier. In consideration of the stringent requirements for risk areas and travel time in China's new crown pneumonia prevention and control policies, the relationship ontology set includes not only the belonging relationship between the basic sets, but also the spatial and temporal relationships between ships, crews, and port elements.

In the process of constructing the knowledge map, the relational ontology sets are categorized into four groups: risk relational sets, attribute relational sets, geographical relational sets, and temporal relational sets. The risk relationship refers to the relationship between the epidemic level and each port, the source of goods, and the ship's dock; the attribute relationship refers to the affiliation relationship between entities, such as a crew member belonging to a certain ship, etc.; and the spatial relationship refers to the relationship in spatial orientation, such as Port A is adjacent to Port B, and Province C includes Ports A, B, and C, etc. In the past, the geographical relationship consisted of an exact search for "if" a particular city is present. In the past two years, the COVID-19 epidemic has frequently been marked by its abruptness and regional spread. Therefore, there are still hidden dangers in the conventional placing of route points that are accurate with respect to cities. For instance, "28-day route cities" and "near cities to dangerous cities" have been added to Ningbo Port's screening standards for epidemic hazards. This study employs the neighborhood analysis function of GIS technologies to compile data on all prefecture-level cities in



**Figure 3: Examples of knowledge graphs for three epidemic risk models**

the nation and describes the map spatial relationship development process. The temporal relationship influences crew risk mostly after leaving an area at risk for an epidemic. Consequently, the information structure of ship epidemic risk is typically hierarchical, and a knowledge graph system with hierarchical and strong semantic linkage can be developed.

The entity type is primarily a concept describing the ship, crew, and cargo epidemic risk's formation, evolution, nature, and characteristics. This item primarily describes this type of risk object for the risk model. Different tags are used to divide and classify representative descriptions. This research must establish several labels with practical meanings for words with fuzzy part-of-speech boundaries in order to eliminate the ambiguity of Chinese part-of-speech boundaries. This can not only increase the diversity of entity categories, but also significantly reduce map redundancy. To determine the port, crew, and ship's passage location, the relationship type is mostly based on the national epidemic risk level issued by the National Health and Health Commission, and the risk level is updated in real time to meet national criteria. This research must add specific qualities to entities in order to avoid problems such as unclear entity matching, disordered relationship links, and low node placement precision caused by the huge number of entities. The attribute type is primarily responsible for elaborating on entities and relationships and complementing specific data.

### 3.2 Knowledge graph construction methods

The knowledge map of ship epidemic prevention and control is primarily generated by a combination of humans and machines, encompassing both manual and semi-automatic processes. The artificial construction component involves formatting unstructured material, transforming it into structured or semi-structured data, and then sorting the produced data to produce a knowledge graph. Using data processing scripts, the semi-automatic component converts unstructured fundamental information data such as ships,

crews, and cargo into semi-structured data. In addition, this research extracts entities, properties, and relationships from ship pandemic data using regular expressions and then turns them into knowledge graphs according to the expression rules of the data. Knowledge graph construction typically involves knowledge acquisition, knowledge representation, knowledge modeling and fusion, and knowledge visualization. Upon completion of the construction, this study also consists of two sections: knowledge query and knowledge reasoning.

#### (1) Knowledge acquisition

The extraction of knowledge (entities, relationships, and attributes) from disparate data sources and structures is the foundation and precondition for building a knowledge graph. The majority of the knowledge acquisition objects for this project are derived from the above-mentioned shipping database, and include crew basic information/health information, port call/landing information, crew change information, and activity information. This study must clean the corpus in order to collect useful information. This involves word segmentation, stop words, keyword extraction, and other measures to increase the accuracy of corpus information. In this study, following data cleaning, just a portion of the descriptive corpus of epidemic information remained, and the corpus data lost their normal semantic structure.

#### (2) Knowledge representation

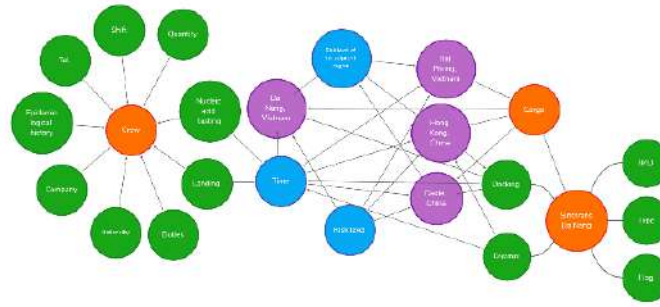
In order to produce a consistent data schema for knowledge, this study manually creates a knowledge map of ship epidemic risk using the National Health and Construction Commission's announced epidemic location as an example of a fake ship docking and crew logging into a dangerous region. And the acquired knowledge is stored in accordance with a standardized data structure to establish a knowledge base. This project's knowledge graph is centered on describing the properties of information on epidemics. The primary layer is comprised of a geographic feature and a time feature, whereas the second layer is sequentially populated as a supplement or annotation to the primary layer.

#### (3) Modelling and integration of knowledge

To achieve the fusion of heterogeneous data, knowledge from various sources must be standardized, verified, disambiguated, and processed inside a unified system and framework due to the various sources of system data. Data fusion is not only a crucial method for updating the knowledge graph, but also a prerequisite for community discovery and difference extraction from the standpoint of the early warning model. The knowledge map of ship epidemic prevention and control focuses primarily on two factors: epidemic risk zone and risk time. This study is data-driven, based on the risk prediction logic of each ship's epidemic, and further constrains the risk rules through the knowledge drive of the current new crown epidemic prevention and control policy, resulting in the formation of a knowledge data model. In addition, this study identifies, categorizes, and matches essential knowledge so that data and knowledge-driven map information are strongly integrated.

#### (4) Visualisation of knowledge

In order to generate the related knowledge graph, knowledge must be input in batches into the program used to create the knowledge graph. In this study, the knowledge graph is displayed using Neo4j "entity-relationship-entity" triples. Entities correspond to independent conceptual ontologies as nodes in the knowledge graph.



Each object contains zero to  $n$  attributes, which exist in "key-value" format. The directionality of linkage and constraint of knowledge is particularly stringent. "From" represents the direction of the node, while "to" represents the direction of the node. In this study, the data is organized in the format of entity "from" relationship "to" entity, with appropriate labels added to the entity based on the characteristics of hierarchical elements, and "label" added to emphasize the hierarchical nature of the dataset and facilitate retrieval and querying.

In network science, a "community" is a network with more tightly interconnected child nodes than other surrounding networks. There are few connections between communities, but the connections between nodes within the same community are robust. There are many "communities" in the network of ship pandemic big data's knowledge graph, similar to the numerous tiny groups that exist in real life due to shared interests. These communities frequently share characteristics with unusual knowledge graphs. Therefore, it makes sense to seek additional abstraction and migration of the knowledge network across domains via these nodes/communities. In general, given map  $G = G(V, E)$ , community discovery refers to the identification of  $nc (\geq 1)$  communities in figure G:

The collection of individual community vertices can overrule V. A distinction must be established between communities that overlap and communities that do not overlap. We refer to communities where the intersection of two vertex sets is empty as disjoint; otherwise, we refer to them as overlapping. By comparing community findings to a random graph, the Modularity value can be utilized to assess their impact. If distinct communities are found using various approaches in the same atlas, it is prudent to select the method with the highest Modularity value (Equation 2):

$A_{vw}$  represents the weight of the edge between nodes  $V$  and  $W$ ;  $k_v$  represents the sum of all weights linked to node  $v$ ;  $c_v$  represents the community to which the current node  $V$  belongs; When  $v$  is equal to  $w$ , the value of the function  $\delta(c_v, c_w)$  is 1, otherwise it is 0, which can be used to determine whether two nodes belong to

In equation 3),  $\Sigma_{in}$  represents the number of connections within a community.  $\Sigma_{tot}$  represents the sum of degrees of all nodes in a community. As can be seen from formulas (2) and (3), the  $\delta(c_v, c_w)$  function used to determine whether two nodes belong to the same community consumes a lot of computing resources, simplifying it and saving a lot of calculations.

In this study, all stages of dividing and obtaining  $\Delta Q$  are repeated until the structure of all node communities is stable and does not change. Then, this study examines the various communities produced in the preceding phase in order to construct a new network using those communities as nodes. The weight of an edge in the new network is the total of the weights of all edges between every pair of nodes in the two communities. This research then repeats the steps of dividing and determining  $\Delta Q$  until the maximum Modularity value is determined. Not only are the communities in the network’s knowledge graph detected, but also whether the communities overlap. Shared nodes/communities not only indicate shared entities/features at discrete scales, but also reflect inter-entity relationships and attribute characteristics.

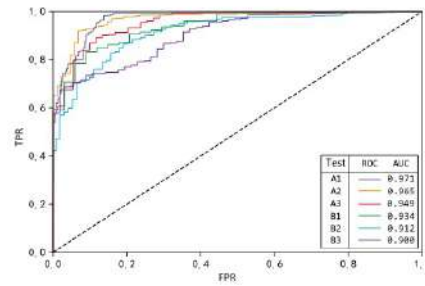
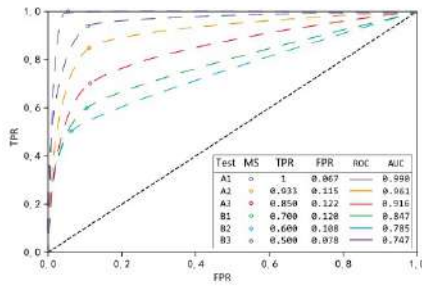
Simulation data were utilized to compare the accuracy of this approach to that of manual screening. In this study, the community

**Table 1: Test data statistics**

Test	Number of ships	Numbers of people	Risk model	Number of positives
A1	3	70	Risky city landing	10
A2	3	160	Risky city landing	15
A3	3	270	Risky city landing	20
B1	3	70	Risk proximity city landing	10
B2	3	160	Risk proximity city landing	15
B3	3	270	Risk proximity city landing	20

**Table 2: Test results of manual screening for epidemic risk**

Test	TP	FP	FN	TN	TPR	FPR
A1	10	4	0	56	1.00	0.067
A2	14	15	1	115	0.93	0.115
A3	17	28	3	202	0.85	0.122
B1	7	6	3	44	0.70	0.120
B2	9	14	6	116	0.60	0.108
B3	10	18	10	212	0.50	0.078

**Figure 5: ROC and AUC of the epidemic risk test, Image 1:based on traditional manual methods, Image 2: based on knowledge graph-community detection algorithm**

discovery algorithm and traditional manual screening methods were used to assess the risk level of this data, respectively. The risk is based on the nationwide list of high-risk regions released by the National Health and Construction Commission on 11 June 2022. This study evaluated a total of 6 rounds. The main tested data and virtual positives are shown in the following table:

In addition to these primary parameters, this study's test data also simulated ship name, flag, IMO number, ship type, last port, crew nationality, crew position, telephone, ID card, home address, activity trajectory within 28 days, data filler, shipping agency, contact information, time and location of the most recent nucleic acid test, and other data for testing by randomly combining real data. The source of these data is the shipping big data platform.

In this study, ROC curves were used to evaluate the risk detection results of the knowledge graph method. The manual screening method cannot draw the native ROC curve, but it can draw (TPR, FPR) points based on its test results. The Bézier curve passing through (0,0), (1,1), and (TPR,FPR) points in the same coordinate system can be used as the manual screening method's approximate

ROC curve for horizontal comparison. Among these, the manual screening produced the following results:

In this study, the points of manual screening methods (TPR, FPR) and their approximate ROC were drawn, and the ROC of the knowledge graph intelligent screening algorithm was drawn. From the results, the manual screening method can obtain a relatively good true positive ratio (TPR) in the test with less data. But when the amount of data increases, its true positive rate begins to decline. When the demand for geographic analysis such as spatial proximity is increased, its true positive ratio decreases rapidly. However, the ROC curve of the knowledge graph algorithm has little to do with the amount of measured data, and its AUC does not decrease much after the introduction of nearby risk cities, and the overall level can still be maintained at a high level.

To quantify the relationship between the AUC and the measured data volume, the measured AUC was plotted as the ordinate and the measured data volume as the abscissa, and a linear fit was conducted. The slope of the fitted curve displays the rate of AUC decline for different risk screening approaches as the amount of

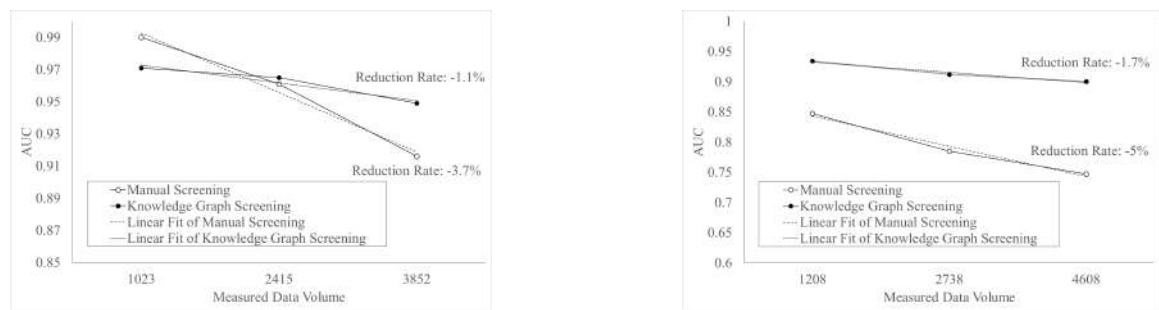


Figure 6: The relationship between the increase in the amount of test data and the decrease in AUC, Image 1:the two epidemic risk screening methods, Image 2: the risk of neighboring cities and the two epidemic risk screening methods

Table 3: The total amount of test data and test time of the two screening methods

Test	Number of ships	Numbers of people	Number of risks	Total data volume	Manual time (min)	Time taken for this method (s)	Multiple of dominance
A1	3	70	10	1023	51	2.19	1397
A2	3	160	15	2415	121	5.30	1370
A3	3	270	20	3852	194	8.40	1386
B1	3	70	10	1208	61	2.47	1482
B2	3	160	15	2738	138	5.76	1438
B3	3	270	20	4608	232	9.50	1465

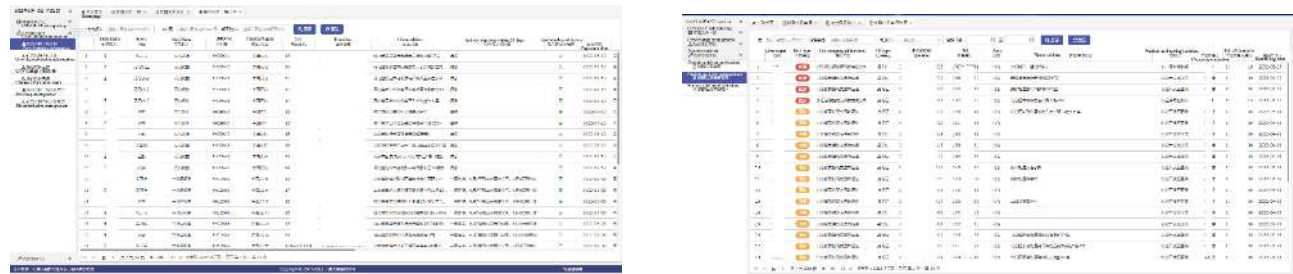


Figure 7: Basic interface of Image 1: incoming ships and crew members, Image 2: Crew nucleic acid detection imminent reminder based on knowledge graph-community detection algorithm (all identity information has been desensitized)

collected data rises. As the amount of data increases, the detection accuracy drops for both screening methods, as shown in Figure 7 the decrease rate for manual screening is -3.7%(Figure 6-1), while the decline rate for knowledge graph screening is -1.1%. (Figure 6-1). After the introduction of risk in neighboring cities, when the quantity of data increased, the accuracy of both screening methods decreased to a greater extent, with the reduction rate of manual screening reaching -5% (Figure 6-2) and the reduction rate of the knowledge graph method reaching -1.7%. (Figure 6-2). Nevertheless, the difference in time and expense between manual screening and knowledge graph screening is more than 1,000 times bigger. Specifically, because the B1, B2, and B3 tests bring the necessity for screening in risk areas—neighboring cities, which exacerbates the difficulties of manual screening and extends the screening time. The results of the test times are displayed in the table below:

5 APPLICATION

Knowledge graph-community discovery algorithms can achieve similar screening accuracy under varying degrees of labor pressure, which is crucial for the prevention and control of epidemics. This study is incorporated into the dangerous goods safety supervision information system of Ningbo Port, as well as real-time monitoring and pushing the epidemic risk information of port ships and vaccination advent information via an epidemic risk statistical summary and reminder, owing to the efficiency and precision advantages of the algorithm.

6 DISCUSSION

This study demonstrates that the knowledge graph-community discovery algorithm can approach or even surpass the accuracy of manual epidemic screening. However, this study is restricted by the lack

of clarity on responsible ownership and the rigidity of emergency response methods in the knowledge-driven screening mode of the knowledge graph. Therefore, the knowledge graph-community discovery approach cannot totally replace manual screening methods in a short amount of time. However, at the application level, this study indicates that it be utilized as a supplement to the pre- and post-screening of manual data, which can significantly reduce the screening process' labor and time costs. This study offers a more modern and data-driven method for ensuring the quality and quantity of port business duties during an epidemic, constructing a more robust barrier for epidemic protection, and ensuring the orderly functioning of the national society.

## 7 CONCLUSION

Based on the multi-source heterogeneous shipping big data system of Ningbo Port, this study creates a series of methodologies for identifying pandemic risk using the knowledge graph community discovery algorithm. On the basis of this method, the ship epidemic risk knowledge graph is developed and applied, and the following findings are drawn:

- (1) In this study, a set of epidemic risk screening logic based on a knowledge graph-community discovery method is constructed using the big data of the shipping epidemic in Ningbo Port as the research objective. This effectively identifies risk data, which is typically more accurate than manual screening. In the six tests, the AUC advantages of the knowledge graph compared with manual screening were -1.92%, 0.42%, 3.6%, 10.3%, 16.2%, and 20.5%, and the average advantage was 8.18%.
- (2) Based on the knowledge graph-community discovery algorithm, this study can greatly improve the efficiency of ship epidemic risk screening. In the 6 experiments, the efficiency of obtaining results similar to or exceeding manual screening in this study was 1397 times, 1370 times, 1386 times, 1482 times, 1438 times, and 1465 times that of manual screening, respectively, with an average advantage of 1423 times.
- (3) This study demonstrates that the greater the difficulty of the epidemic information screening task, the more the knowledge graph-community discovery algorithm surpasses manual methods. Knowledge graph-community found that the AUC decrease rate of the algorithm with the increase of the measured data was -1.1% and -1.7%. The AUC reduction rate for manual screening was -3.7% and -5%.

## ACKNOWLEDGMENTS

Grant sponsor: Key R&D Project of Guangxi Science and Technology Department, Grant no:2021AB07045.

## REFERENCES

- [1] Liu Nan, Wang Zengyun, Liao Nvnan. 2022. The Dilemma of Protecting the Rights and Interests of Port Employees under Public Health Emergencies and Its Solution. *Medicine & Jurisprudence*, 14, 2, 60-63.
- [2] Michail, Nektarios A., Kostis D. Melas. 2020. Shipping markets in turmoil: An analysis of the Covid-19 outbreak and its implications. *Transportation Research Interdisciplinary Perspectives*, 20, 7, 100178.
- [3] Wang Hailing, Yang Yaning, Kang Ziyi, *et al.* 2022. Epidemic prevention risk assessment of logistics system using combination algorithm. *Journal of Safety and Environment*, 7 (July 2022), 1-12. <https://doi.org/10.13637/j.issn.1009-6094.2022.1080>.
- [4] Zhang, Yan, Zhikuan Sun. 2021. The coevolutionary process of maritime management of shipping industry in the context of the COVID-19 pandemic. *Journal of Marine Science and Engineering*, 9, 11, 1293.
- [5] Zhang Ruizhen. 2022. China's cruise ship health and safety supervision from the perspective of COVID-19 epidemic prevention and control. *Shipping Management*, 44, 2, 28-31.
- [6] Gilmore, Brynne, *et al.* 2020. Community engagement for COVID-19 prevention and control: a rapid evidence synthesis. *BMJ global health*, 5, 10, e003188.
- [7] Huang Bo. 2022. Thoughts and suggestions on port production under the normalization of epidemic prevention and control. *China Water Transport*, 22, 5, 13-14.
- [8] Yamagishi T, Kamiya H, Kakimoto K, *et al.* 2020. Descriptive study of COVID-19 outbreak among passengers and crew on Diamond Princess cruise ship, Yokohama Port, Japan, 20 January to 9 February 2020. *Eurosurveillance*, 25, 23, 2000272.
- [9] Chatterjee R, Bajwa S, Dwivedi D, *et al.* 2020. COVID-19 Risk Assessment Tool: Dual application of risk communication and risk governance. *Progress in Disaster Science*, 20, 7, 100109.
- [10] Xue Fengping. 2021. Construction of information sharing platform in epidemic prevention and control system. *Journal of the Party School of C.P.C. Qingdao Municipal Committee and Qingdao Administrative Institute*, 21, 5, 53-60.
- [11] Chen Dan, Shan Shuangshuang. 2022. Inspection and reconstruction of the risk communication system for epidemic prevention and control. *Journal of Nanjing Medical University (Social Sciences)*, 22, 3, 260-266.
- [12] Zhang Q, Wen Y, Zhou C, *et al.* 2019. Construction of knowledge graphs for maritime dangerous goods. *Sustainability*, 11, 10, 2849.
- [13] Jiang Lei, Zhan Wenhua, Zhang Guoyan, *et al.* 2021. Active distribution network voltage control strategy based on knowledge graph. *Science Technology and Engineering*, 21, 30, 12982-12989.
- [14] Shen Hejiang, Zhen Huigang, Li Xujiao. 2022. Construction of Rural Community Tourism Destination Environment Monitoring System Under COVID-19 Epidemic Geographical Diffusion Effect. *Journal of Hebei Normal University(Natural Science)*, 46, 4, 410-416.
- [15] Wang Shukun, Zhao Shiwen, Fu Xiaoqing, *et al.* 2021. Detection, monitoring and early warning of infectious disease outbreaks or epidemics. *Chinese Journal of Epidemiology*, 42, 5, 941-947.
- [16] Zhou Licheng, Wen Zhe, Luo Weiquan, *et al.* 2021. Research on the establishment of 3m-spr system of global infectious disease surveillance, early warning and response to prevent and deal with biosafety risks at poes in china. *Port Health Control*, 26, 5, 27-30+34.
- [17] Yan Jiaqi, Song Jinbei, Da Jingwei, *et al.* 2021. A Blockchain-Based Early Warning System for Infectious Diseases: Risk Measurement Combined with Complex Network. *Journal of Information Resources Management*, 11, 4, 90-99.
- [18] Erkhembayar R, Dickinson E, Badarch D, *et al.* 2020. Early policy actions and emergency response to the COVID-19 pandemic in Mongolia: experiences and challenges. *The Lancet Global Health*, 8, 9, 1234-1241.

# Early warning of corporate financial crisis based on sentiment analysis and AutoML

Wei CHENG\*

College of Information and Electrical  
Engineering, China Agricultural  
University, Beijing  
2799619818@qq.com

Shiyu CHEN†

School of Safety Science and  
Emergency Management, Wuhan  
University of Technology,  
Wuhan, China  
chensy@whut.edu.cn

Xi Liu

College of Humanities and  
Development Studies, China  
Agricultural University, Beijing  
2019312100112@cau.edu.cn

Jiali Kang

School of Safety Science and  
Emergency Management, Wuhan  
University of Technology,  
Wuhan, China  
Kangjiali@whut.edu.cn

Jiahao Duan

School of Safety Science and  
Emergency Management, Wuhan  
University of Technology,  
Wuhan, China  
duanjiahao@whut.edu.cn

Shixuan LI ‡

School of Safety Science and  
Emergency Management, Wuhan  
University of Technology,  
Wuhan, China  
shixuanli@whut.edu.cn

## ABSTRACT

Establishing an early warning model for corporate financial crises is important for managing risks and ensuring the continued stability of the capital market. A financial crisis early warning indicator system for listed companies was constructed, which includes financial indicators, management indicators and annual report text tone features. Using techniques such as web crawlers and text sentiment analysis, we collected data related to 820 listed companies in mainland China from 2017 to 2021. Six models were then constructed and their results were compared. The results of the comparative analysis showed that: there is room for AutoML to be applied and explored in this area; the model performance and inference speed of integrated learning CatBoost are substantially improved compared with traditional methods; feature importance rankings help to understand the formation of corporate financial distress. Thus, textual information such as corporate annual reports can help predict financial crises.

## CCS CONCEPTS

• Machine learning; • Learning paradigms; • Supervised learning; • Supervised learning by classification;

## KEYWORDS

Financial Distress Prediction, Sentiment Analysis, AutoML, CatBoost

\*The authors contribute equally

†The authors contribute equally

‡The authors contribute equally

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590027>

## ACM Reference Format:

Wei CHENG, Shiyu CHEN, Xi Liu, Jiali Kang, Jiahao Duan, and Shixuan LI. 2023. Early warning of corporate financial crisis based on sentiment analysis and AutoML. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590027>

## 1 INTRODUCTION

The global financial crisis has had a tremendous impact on businesses, sparking widespread concern. The development of an effective tool to predict company bankruptcy is important for accurate decision-making [1]. The essence of enterprise bankruptcy prediction is to use statistical and intelligent models to predict whether a given enterprise will face a financial crisis in the future based on the various information resources available at present [2].

Building on the pioneering research of Beaver [2] and Altman [3] on bankruptcy prediction, research in this field has seen remarkable progress. However, there are two shortcomings in the existing studies. On one hand, the existing indicator systems are unsuitable, often relying on quantitative financial indicators due to technical and information constraints. On the other hand, most of the existing studies lack mechanism explanations (black-box models) and are therefore less useful for domain knowledge mining.

The true reliability of financial data needs to be considered, especially because Chinese enterprises have accounting fraud problems. Therefore, the indicator system of previous studies still needs to be improved. For another, models in previous research are mostly black boxes with limited interpretability, making it difficult for relevant stakeholders to understand the factors that cause financial crises based on a single prediction result. Meanwhile, most of the current studies are industry-specific and do not focus on gaining knowledge from large samples. The prediction accuracy of current models is generally 80%-85%. To improve the accuracy of prediction, training on larger samples and gaining domain knowledge in multi-model comparisons should be on the agenda.

To address these issues, we expanded the sample, further refined the early warning indicators, and introduced an automatic modeling

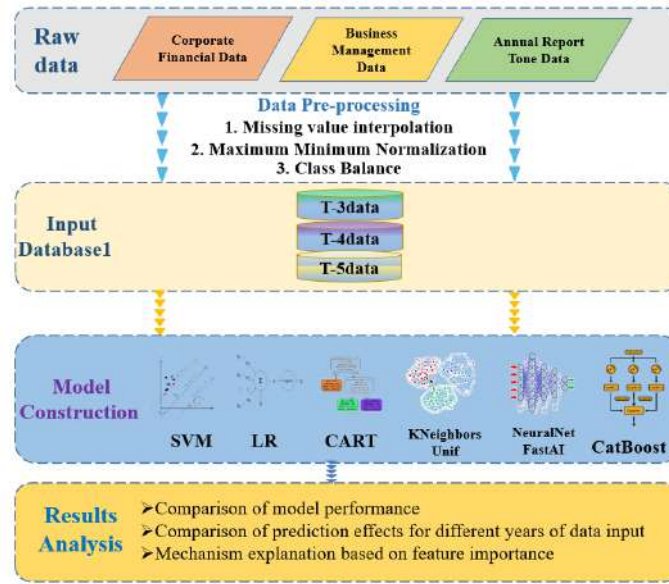


Figure 1: Structure of the paper

approach. The structure of this paper is shown in Figure 1 The primary contributions of this paper are:

- 1) Introduction of textual metrics and validation of their validity.
- 2) The first use of automatic modeling methods, significantly improved model performance and generalization ability.
- 3) Acquired additional domain knowledge on financial crisis formation using feature importance ranking.

## 2 LITERATURE REVIEW

### 2.1 Index System

The choice of indicators directly affects the effect of prediction. Zhou [4] pointed out that early studies mainly used quantitative financial information from financial statements to build models. They stated that economic inefficiency due to poor management is the main reason why companies are in financial distress.

However, some recent studies argued that the predictive capacity of financial features is limited due to many inherent flaws [5]. Therefore, to enhance the effectiveness of the model, non-financial features can be used as a complement to financial indicators [6]. Jones stated that governance style proxies may have an association with corporate distress [7]. In 2018, Jiang et al. [8] tried to use more information that can reflect the operating conditions of a company, such as board characteristics, executive shareholding ratio, and internal governance structure.

In addition, the qualitative textual information contained in the annual reports of listed companies can also describe the operating conditions and development tendencies of companies to a certain extent [9]. In 2011, Loughran and McDonald [10] extracted positive and negative tones from the annual reports. After that, making full use of non-financial information as assistant information to financial indicators for financial distress prediction has become an important development direction.

In this paper, the above three types of indicators are all included in the index system to achieve more efficient financial distress prediction.

### 2.2 Model Building

Since Beaver pioneered this field of research in 1966, the academic community has achieved fruitful results. The current methods can be summarized into three main categories. The first category is statistical discriminant methods, typified by the discriminant analysis model containing five factors established by Altman in 1968. However, traditional statistical methods have strict requirements on the data, such as its linearity, normality and independence [11]. The second category is machine learning methods, and the models widely used today are support vector machines [12], logistic regression [13], and decision trees. The last one is the emerging deep learning methods, such as Odom and Sharda who pioneered the application of NNs for early warning of financial crises in 1990. However, their interpretability is currently very limited and difficult to apply.

In general, all the above methods have certain problems. First, the fixed weights of each variable allow most variables to substitute each other, which in turn leads to the overestimation of the evaluated object [14]. Secondly, they tend to ignore the interactions between the characteristic variables. Finally, the vast majority of models focus on correlation and prediction, but the mechanisms of crisis formation are not analyzed [15].

In 2018, Dorogush et al. proposed CatBoost [16], a new open-source gradient boosting library that successfully handles classification features. In 2020, Mangalathu et al [17] evaluated eight machine learning models. In these models, CatBoost modified the computation of gradients to avoid prediction bias and thus improve

the accuracy of the models. The combination of AutoML and CatBoost offers many advantages for data science research. The first is the ability to improve the efficiency and accuracy of model creation. AutoML automates the process of data preprocessing, feature engineering, model selection, and hyperparameter optimization, while CatBoost provides a powerful and accurate library of gradient boosts. By using AutoML and CatBoost together, researchers can create powerful models quickly and accurately, improving interpretability and reducing training time. So far, AutoML has been successfully applied to many important problems, such as automatic model selection [18] and automatic feature engineering [19].

Therefore, in this paper, we use an automatic modeling framework to accelerate the modeling and deployment process while ensuring model performance.

### 3 MATERIALS AND METHODS

#### 3.1 Sample and Data

The China Securities Regulatory Commission requires the stock exchange to apply the special treatment (ST) to trading in the shares of listed companies when their financial or other conditions are abnormal. The above abnormal conditions mainly include two main aspects: net profits of the listed company are negative in both accounting years after the audit; the net asset per share of the listed company is lower than the face value of the stock in the latest accounting year after the audit. Therefore, we use listed companies that received the label 'ST' as a negative sample.

Based on this principle, We select 205 Chinese companies that received 'ST' from 2017 to 2021 from the Shanghai Stock Exchange and the Shenzhen Stock Exchange. To upgrade the applicability of the model, we adopt the paired sampling method [20] to select 615 listed companies randomly in the corresponding industries with the most similar asset size as healthy companies. In addition, since it is less practical to utilize data from the year prior to the ST judgment and the previous two years for early warning, data from the first three, four and five years prior to the ST judgment are collected for follow-up work. Due to the special nature of the financial industry, none of the above companies are financial enterprises.

#### 3.2 Indicator System

**3.2.1 Financial indicators.** Financial indicators can provide crucial insight into a company's likelihood of bankruptcy, which are often seen as the best way to capture and assess the performance of the company. Considering that Kordestani stated that the statements of cash flows are more difficult to be distorted than other financial ones [21], we integrate the financial indicators obtained from experts in the same field with those obtained from the literature (cash flow, etc.). Therefore, we collect financial indicators from two major databases, CSMAR and CNRDS, for the sample companies in seven areas, including 9 cash flow indicators, 11 earnings capacity indicators, 8 solvency indicators, 9 ratio structure indicators, 14 operating capacity indicators, 10 development capacity indicators, and 16 per share indicators. Due to the limitation of space, the specific calculation formula is not listed in this article, please contact the author if you need it. The same applies to the following article.

**3.2.2 Management indicators.** Jones [22] demonstrated that governance variables, such as institutional ownership and insider ownership, could potentially affect a company's operations. As a result, we collected 33 indicators from four aspects of corporate management: board structure, shareholding structure, internal control information, and auditor's opinion. The data was sourced from the same source as the financial indicators.

**3.2.3 Annual report text indicators.** The texts of annual reports of listed companies, especially the analysis and discussion sections of the management, contain a lot of information about companies' operating conditions and development tendencies. Based on the above study, we first obtained the texts of annual reports from the real disclosures of the Juchao Information Website (<http://www.cninfo.com.cn/new/index>), Shenzhen Stock Exchange and Shanghai Stock Exchange. Considering the specificity of the domain, we adopted the Chinese sentiment dictionary in the financial domain constructed by Yao et al. based on dictionary restructuring and deep learning. Their constructed annual report tone index can effectively predict market factors such as returns and trading volume of listed companies and outperforms the indexes constructed based on other widely used sentiment dictionaries [23]. Seven textual sentiment indicators including the positive sentiment index and sentiment consistency index are finally obtained.

#### 3.3 Selection of model

Traditional classification algorithms such as SVM, LR and CART have been used to predict corporate financial distress. However, with the advent of AutoML, it becomes streamlined and more efficient to design and optimize machine learning algorithms. AutoML can help to better uncover the underlying mechanism of corporate financial distress, while also offering improved generalization and prediction accuracy. To demonstrate its efficacy, we have selected AutoGluon-Tabular [24], a highly accurate open-source tool, as our AutoML framework. Using AutoML, we have optimized the hyperparameters of 14 machine learning algorithms to accurately predict corporate financial distress. Three of the best-performing algorithms, KNN, NeuralNetFastAI and CatBoost, have been selected based on their AUC, training and validation runtimes.

The architecture was implemented using PyTorch (version 1.12.1) on Python (version 3.9.7). Autogluon version 0.3.1 was deployed for financial distress prediction.

#### 3.4 Model evaluation indicators

**3.4.1 AUC.** The AUC is a widely used metric to evaluate the performance of a binary classification model. It is well-recognized for its ability to measure a model's discriminatory power [25]. Moreover, policymakers are often interested in more than just binary bankruptcy predictions. They may use the probability of insolvency to generate credit portfolios or to determine loan rates [26]. Even in the case of unbalanced sample data, the classifier can still produce an accurate evaluation thanks to AUC, which considers the model's ability to classify both positive and negative samples.

**3.4.2 Err-2(Type II error rate).** Classification accuracy is an inadequate measure of model performance when it comes to predicting bankruptcy, as this metric assumes that Type I and Type II errors are

**Table 1: Average performance of the model on three annual data sets**

Model	AUC	Acc	Precision	Recall	Err-1	Err-2
SVM	0.8452	0.7456	0.7377	0.7938	0.2062	0.2213
LR	0.8574	0.7604	0.7228	0.8004	0.1996	0.2538
CART	0.8045	0.7765	0.7431	0.7981	0.2019	0.2436
KNeighborsUnif	0.8994	0.8415	0.8911	0.8901	0.1099	0.0533
NeuralNetFastAI	0.9211	0.8873	0.8861	0.8977	0.1023	0.0434
CatBoost	0.9376	0.9208	0.9233	0.9144	0.0856	0.0237

equally costly. However, the cost of false negatives is usually much more severe than the cost of false positives. While it is possible to assign increased costs to false negatives, the cost structure remains specific to each context. Therefore, accuracy scores should not be used as the sole measure of model performance when predicting bankruptcy.

Most investors, whose primary goal is to preserve value, have varying sensitivities to different errors. In general, the assumption of the decision maker is that the firm is operating as usual (which is usually the case). However, the Type II error occurs when the model incorrectly identifies a sample with an impending financial crisis as normal. This error can lead to significant losses, such as when investors believe the company is functioning properly, invest more money, or fail to divest in a timely manner. Therefore, we evaluate the Type II error rate, which has received little attention in previous studies. Type II error rate is defined as follow:

$$Err_2 = \frac{FN}{FN + TN}$$

where  $FN$ (False Negtive) represents the number of negative samples predicted by the classifier and the actual positive samples, i.e., the number of positive samples missed..  $TN$ (True Negtive) represents the number of negative samples predicted by the classifier that are actually negative samples, i.e., the number of negative samples correctly identified.

## 4 EXPERIMENT RESULTS

The data obtained from the sampling pairs are unbalanced data and therefore need to be upsampled. After balancing the data using SMOTE, the data sample was expanded to 1230. Subsequently, we split the training and test sets in a 4:1 ratio. In the training phase, the data collected in each of the three years were used to train six models. In the testing phase, we evaluated the model using 10-fold cross-validation. Due to space limitation, we do not include the complete experimental results in this paper. The reader is invited to contact the authors if needed. Next, we focus on the performance of the model and explain the corporate financial crisis based on the importance of the features.

### 4.1 Prediction accuracy analysis

Table 1 and Figure 2 respectively illustrate the average performance of the six models, as well as the prediction effects across the three years of data. To ensure accuracy, we calculate the average performance of the assessment metrics as the average of the model's assessment metrics over the three annual datasets.

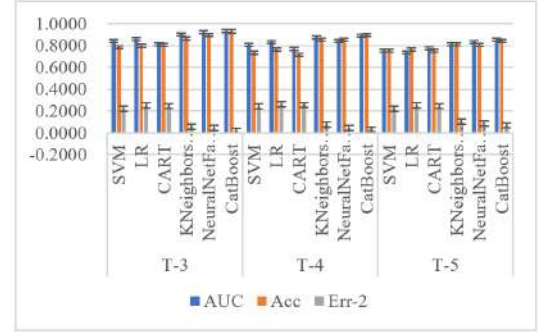
**Figure 2: Prediction effect of different data**

Table 1 shows that AutoML-based models, such as CatBoost, KNN, and NeuralNetFastAI, have higher prediction accuracy than traditional algorithms like SVM, LR, and CART. The average AUC of CatBoost models was 10.93% higher (averaged across years) compared to SVM, and 12.53%, 9.40% and 10.86% higher in each year, respectively. Notably, the three models constructed based on AutoML showed a significant improvement in the Err-2 compared to the traditional model (p-value of Wilcoxon test = 0.04953).

The potential cause of such an outcome may be attributed to two factors. Firstly, integrated learning stabilizes and increases the hypothesis space, thus allowing for more accurate approximations. While deep learning models have been shown to perform well with a large quantity of labeled data. The more data samples, the higher the model prediction accuracy. What's more, AutoML is a scalable, efficient, and easily deployable solution that automates the time-consuming, iterative tasks of machine learning model selection and development. This leads to a substantial improvement in the prediction results.

The automatically constructed CatBoost model exhibits superior prediction results compared to the traditional model. This is because CatBoost automatically turns category-based features into numerical features, enhancing feature dimensionality, while AutoML configures it with ideal hyperparameters and simplifies the iterative process.

In addition, we find that data collected closer to the time of the financial crisis had better prediction results. As shown in Figure 2, the accuracy of most models increases as the crisis approaches and their percentage of type II errors decreases. This is consistent with previous findings [27] that as a firm approaches a financial crisis, its risk will be manifested in the firm's operational management, annual reports, etc.

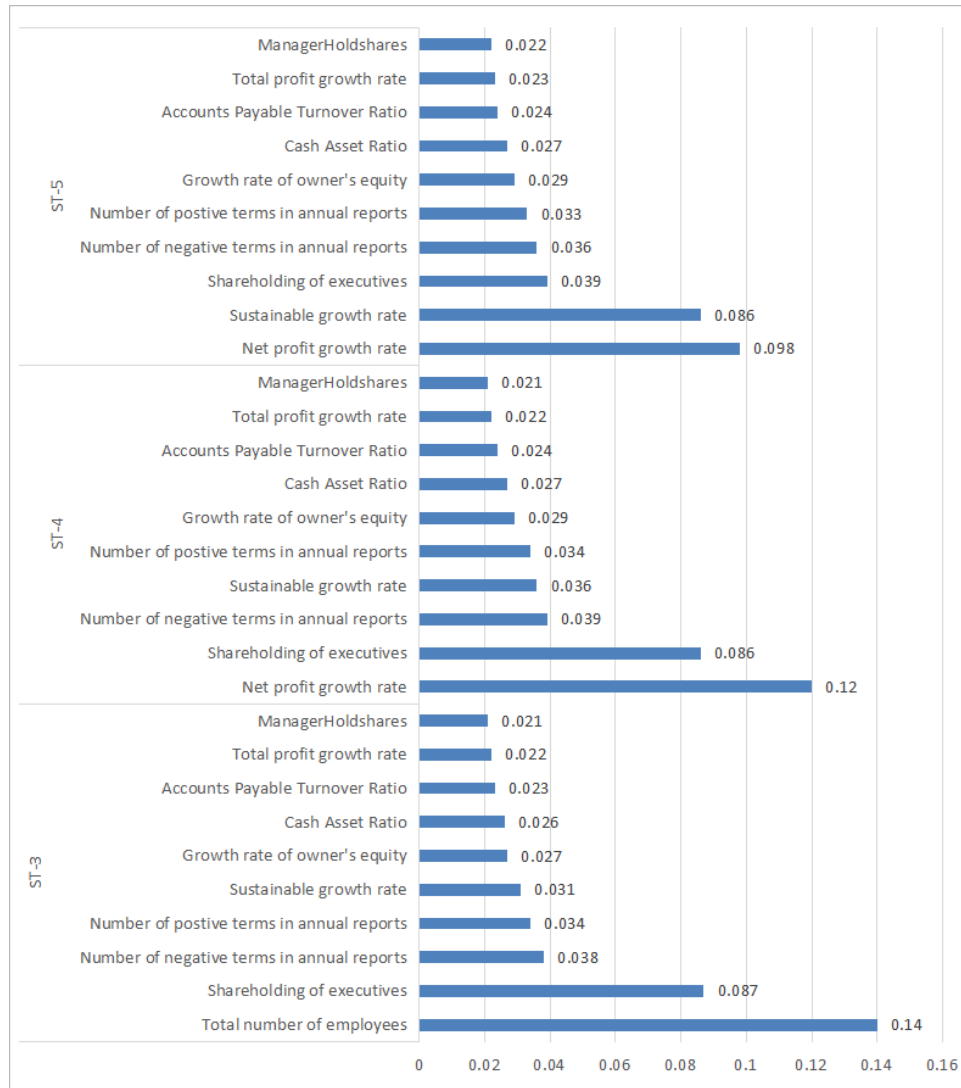


Figure 3: Ranking the importance of features by year

## 4.2 Feature Importance Analysis

We noted a lack of mechanistic explanations for the lack of attention in previous studies. Based on integrated learning, we ranked feature importance using SHapley Additive exPlanations (SHAP). SHAP has many advantages, such as its ability to visualize the contribution of features to understand feature importance [28]. At the same time, this method has a good explanation for the tree integration model. Figure 3 shows the ranking of feature importance for different years of data.

The above figure shows that financial and management indicators are always important factors in the success or failure of a company, and as the crisis approaches, the equity structure and management capabilities become more and more important (the importance of related indicators, such as the ranking of the percentage of executive shareholding, keeps rising). Wu Jing's study in 2020 proves a similar point [29]. It is worth noting that in any given year,

the text sentiment indicator is in the top 5 in terms of importance and gradually increases. In other words, the textual information in the annual report can be a good aid for early warning of financial crises.

## 5 CONCLUSIONS AND DISCUSSION

In this paper, we used a financial sentiment dictionary to extract text features from annual reports of listed companies to build a more complete indicator system, and subsequently we constructed six classification models. The results of the comparative analysis show that the CatBoost model constructed under the AutoML framework has the best performance, with an accuracy rate of 93.76% and a significantly lower class II error rate; textual information, such as corporate annual reports, can assist in financial crisis prediction; and feature importance ranking of the causes of financial crises

can not only further understand the formation process of financial crises, but also promote its practical application.

Future research can consider introducing more textual information, such as financial news and online web commentaries, while the influence of uncertainties such as national policies, markets and unforeseen events should be taken into account in the textual analysis process. In addition, as the research progresses, fine-grained exploration, such as delineating more crisis categories, should be on the agenda.

## ACKNOWLEDGMENTS

We thank Jie Cheng, Bin Ye and Xiang Zhang for helpful discussions on topics related to this work. Financially supported by National Innovation and Entrepreneurship Training Program for College Students (S202210497002). Finally, we would like to express our heartfelt gratitude to the associate editor and the reviewers for their useful feedback that improved this paper.

## REFERENCES

- [1] Sun, J., Li, H., Huang, Q.-H., & He, K.-Y. (2014). Predicting financial distress and corporate failure: A review from the state-of-the-art definitions, modeling, sampling, and featuring approaches. *Knowledge-Based Systems*, 57, 41–56. <https://doi.org/10.1016/j.knsys.2013.12.006>
- [2] Beaver, W. H. (1966). Financial Ratios As Predictors of Failure. *Journal of Accounting Research*, 4, 71–111. <https://doi.org/10.2307/2490171>
- [3] Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, 23(4), 589–609. <https://doi.org/10.1111/j.1540-6261.1968.tb00843.x>
- [4] Zhou, L., Tam, K. P., & Fujita, H. (2016). Predicting the listing status of Chinese listed companies with multi-class Classification Models. *Information Sciences*, 328, 222–236. <https://doi.org/10.1016/j.ins.2015.08.036>
- [5] Wang, G., Chen, G., & Chu, Y. (2018). A new random subspace method incorporating sentiment and textual information for financial distress prediction. *Electronic Commerce Research and Applications*, 29, 30–49. <https://doi.org/10.1016/j.eleap.2018.03.004>
- [6] Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, 50(1), 164–175. <https://doi.org/10.1016/j.dss.2010.07.012>
- [7] Jones, S., Johnstone, D., & Wilson, R. (2016). Predicting corporate bankruptcy: An Evaluation of Alternative Statistical Frameworks. *Journal of Business Finance & Accounting*, 44(1-2), 3–34. <https://doi.org/10.1111/jbfa.12218>
- [8] Jiang, Y., & Jones, S. (2018). Corporate distress prediction in China: A machine learning approach. *Accounting & Finance*, 58(4), 1063–1109. <https://doi.org/10.1111/acfi.12432>
- [9] Balakrishnan, R., Qiu, X. Y., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3), 789–801. <https://doi.org/10.1016/j.ejor.2009.06.023>
- [10] LOUGHRAN, T. I. M., & MCDONALD, B. I. L. L. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-KS. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- [11] Deakin, E. B. (1972). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 10(1), 167. <https://doi.org/10.2307/2490225>
- [12] HHrdle, W. K., Moro, R. A., & Schhfer, D. (2005). Predicting bankruptcy with support Vector Machines. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2894426>
- [13] Martin, D. (1977). Early warning of bank failure. *Journal of Banking & Finance*, 1(3), 249–276. [https://doi.org/10.1016/0378-4266\(77\)90022-x](https://doi.org/10.1016/0378-4266(77)90022-x)
- [14] Li, H., Wen, Z., & Li, Z. (2020). Financial crisis early warning research: a literature review. *Communication of Finance and Accounting*, 24(860), 12–15. <https://doi.org/10.16144/j.cnki.issn1002-8072.2020.24.002>
- [15] Li, S., Chen, Y., Shi, W., & Yang, D. (2021). Research on early warning of financial crisis of listed companies based on knowledge reasoning. *Journal of Wuhan University of Technology(Information & Management Engineering)*, 4(43), 322–329.
- [16] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *ArXiv*. <https://doi.org/10.48550/arXiv.1810.11363>
- [17] Mangalathu, S., Jang, H., Hwang, S.-H., & Jeon, J.-S. (2020). Data-driven machine-learning-based seismic failure mode identification of reinforced concrete shear walls. *Engineering Structures*, 208, 110331. <https://doi.org/10.1016/j.engstruct.2020.110331>
- [18] Feurer, M., Klein, A., Eggenberger, K., Springenberg, J.T., Blum, M., & Hutter, F. (2015). Efficient and Robust Automated Machine Learning. *NIPS*.
- [19] G. Katz, E. C. R. Shin and D. Song, "ExploreKit: Automatic Feature Generation and Selection," 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016, pp. 979–984, doi: 10.1109/ICDM.2016.0123.
- [20] Agarwal, S., Kumar, S., & Goel, U. (2019). Stock market response to information diffusion through internet sources: A literature review. *International Journal of Information Management*, 45, 118–131. <https://doi.org/10.1016/j.ijinfomgt.2018.11.002>
- [21] Ibrahim, M., & Alagidede, P. (2018). Nonlinearities in financial development-economic growth nexus: Evidence from sub-saharan africa. *Research in International Business and Finance*, 46, 95–104. <https://doi.org/10.1016/j.ribaf.2017.11.001>
- [22] Jones, S. (2017). Corporate bankruptcy prediction: A high dimensional analysis. *Review of Accounting Studies*, 22(3), 1366–1422. <https://doi.org/10.1007/s11142-017-9407-1>
- [23] Yao, G., Feng, X., Wang, Z., Ji, R., & Zhang, W. (2021). Tone, sentiment and market impact: based on financial sentiment dictionary. *Journal of Management Sciences in China*, 24(5), 26–46. <https://doi.org/10.19920/j.cnki.jmsc.2021.05.002>
- [24] Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., & Smola, A. (2020). AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *ArXiv*. <https://doi.org/10.48550/arXiv.2003.06505>
- [25] Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159. [https://doi.org/10.1016/s0031-3203\(96\)00142-2](https://doi.org/10.1016/s0031-3203(96)00142-2)
- [26] Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, 9(1), 5–34. <https://doi.org/10.1023/b:rast.0000013627.90884.b7>
- [27] Liang, D., Tsai, C.-F., & Wu, H.-T. (2015). The effect of feature selection on financial distress prediction. *Knowledge-Based Systems*, 73, 289–297. <https://doi.org/10.1016/j.knsys.2014.10.010>
- [28] Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2019). Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods. *ArXiv*. <https://doi.org/10.48550/arXiv.1911.02508>
- [29] Wu, J., & Yuan, F. (2020). Equity pledges, corporate governance and financial crisis warning. *Communication of Finance and Accounting*, 848(12), 45–49. <https://doi.org/10.16144/j.cnki.issn1002-8072.2020.12.009>

# Graph representation learning and software homology matching based A study of JAVA code vulnerability detection techniques

Yibin Yang

Department of Information, Beijing  
City University, Beijing 100191  
1045181534@qq.com

Xin Bo

Department of Information, Beijing  
City University, Beijing 100191  
wakle262@outlook.com

Zitong Wang

Department of Information, Beijing  
City University, Beijing 100191  
2199406944@qq.com

Xinrui Shao

Department of Information, Beijing  
City University, Beijing 100191  
1374930852@qq.com

Xinjie Xie\*

Department of Information, Beijing  
City University, Beijing 100191  
3053325409@qq.com

## ABSTRACT

In nowadays using different tools and apps is a basic need of people's behavior in life, but the security issues arising from the existence of source code plagiarism among tools and apps are likely to bring huge losses to companies and even countries, so detecting the existence of vulnerabilities or malicious code in software becomes an important part of protecting information and detecting software security. This project is based on the aspect of JAVA code vulnerability detection based on graph representation learning and software homology comparison to carry out research. This project will be based on the content of deep learning, using a large number of vulnerable source code, extracting its features, and transforming it into a graph so that it can be tested source code for comparison and report the vulnerability content.

The main work and results of this project are as follows:

1.By extracting the example dataset and generating json files to save the feature information of relevant java code; by generating vector files, bytecode files and dot files, and batch extracting nodes and edges based on the contents of the dot files for subsequent machine learning use, the before and after steps and operations form a logical self-consistency to ensure the integrity of the project.

2.Through the study of graph neural networks and graph convolutional neural networks, relevant models are selected for machine learning using predecessor files and manual model tuning is performed to ensure good learning results and feedback for the machine learning part of the project.

3.This project training dataset negative samples for sard above the shared dataset, which contains 46636 java vulnerability source code, and dataset support environment, test dataset negative samples dataset also from sard, positive samples dataset are generated from the relevant person in charge.

\*Supported by Beijing City University in 2022 "the innovation and entrepreneurship training program for college students"

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590028>

4.Based on Graph Neural Network (GNN) and Graph Convolutional Neural Network (GCN), this project will design and implement a whole set of automated vulnerability detection system for java code.

5. All the related contents of this project, after the human extensive search of domestic and foreign related papers and materials, there are not all projects or contents similar to this project, the same papers and materials appear, all the problems involved in this project and related ideas are for the project this group of people thinking, looking for solutions.

## CCS CONCEPTS

• **Security and privacy** → Systems security; Vulnerability management; Vulnerability scanning.

### ACM Reference Format:

Yibin Yang, Xin Bo, Zitong Wang, Xinrui Shao, and Xinjie Xie. 2023. Graph representation learning and software homology matching based A study of JAVA code vulnerability detection techniques. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3590003.3590028>

## 1 INTRODUCTION

### 1.1 Background and significance of the study

With the continuous progress of scientific research, the technical level of the network and its applications are also developing rapidly, and computer technology has penetrated into all aspects of people's daily life. Software systems, software applications, and other areas of security issues. Nowadays, the booming development of the Internet has inevitably led to an increase in the attack surface, resulting in the probability of security threats to the systems and applications on the market nowadays is also increasing. Although developers have done their best to program securely, the use of open source projects and source code has become the trend, which leads to many vulnerabilities are often created inadvertently, and it is difficult to have effective means to prevent and protect against such vulnerabilities. Nowadays, many software or application modules are very similar, which further leads to the increasing rate of reuse of open source code, and the combination of different open source code is likely to generate some unpredictable and unpredictable vulnerabilities.

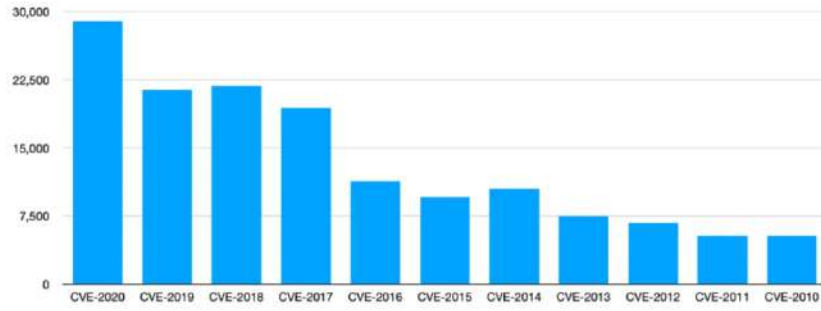


Figure 1: CVE Vulnerability



Figure 2: 2021 Number of new vulnerability information per month

The number of vulnerabilities has continued to increase each year in recent years and has shown an exponential increase. Figure 1 and Figure 2 show the number of CVE (Common Vulnerabilities & Exposures) vulnerabilities published from 2010 to December 2021. The need to accelerate the speed of vulnerability discovery and the ability to fix vulnerabilities.

Along with the continuous development of machine learning and deep learning, a large number of automated vulnerability mining techniques based on machine learning have been bred, covering the typical scenarios of static analysis, dynamic analysis and combined dynamic and static analysis in the past, because the false alarm rate and leakage rate of dynamic analysis is relatively too high compared to static analysis and combined analysis of both, so most of the automated vulnerability mining based on machine learning is now based on static or combined static and dynamic analysis to progress.

The research of this project has the following implications:

Solve the problem of relatively low efficiency of manual mining and discovery of source code vulnerabilities. Nowadays, with the gradual development of network technology, the number of vulnerabilities is increasing, but the ability and speed of manual auditing is relatively limited, which leads to the use of manual vulnerability mining and vulnerability analysis of large projects often takes a long time, the pressure of manual auditing is too much, many relatively complex code, manual detection and audit is difficult to

quickly understand the meaning, which leads to the inability to quickly find vulnerabilities and fix them quickly. The approach proposed in this project effectively avoids inefficient vulnerability detection, which reduces labor costs and allows vulnerabilities to be found and fixed as quickly as possible.

This project effectively reduces the difficulty of manually verifying the existence of vulnerabilities in the source code. Even if the vulnerabilities identified using the code vulnerability detection system generally require the use of manual verification of the existence of vulnerabilities in this step, even though the use of the code vulnerability detection system reduces the workload of inspection, it still requires manual review of the last step, which consumes a lot of manpower and time. This project uses graph neural network audit to directly display the vulnerability content through graphs and reports, and uses a large amount of data learning to ensure the correct rate of audit, effectively reducing the difficulty of auditing source code vulnerabilities and saving time costs.

## 1.2 Current status of related research at home and abroad

In the literature and materials searched and read by the authors, the domestic and foreign materials involved are only similar to the technical content of this project, and there are no papers and materials with all the same technology as this project. The machine learning of java source code is limited to the use of LSTM (Long

Short Term Memory Neural Network) and BLSTM (Bidirectional Long Short Term Memory Neural Network) which are two methods, and even the code related to other languages with the graph neural network (GNN) to be used in this project for machine learning and batch vulnerability mining are very rare.

At present, static vulnerability scanning of project code before the project goes live at home and abroad has become a priority for many enterprises, and in practice each enterprise will use self-developed or third-party vulnerability detection systems for testing to ensure the security of the source code. Now the technology of static vulnerability scanning on line products is mainly lexical analysis, data flow analysis and other technologies. Lexical analysis technology is mainly through the source code and sensitive functions, statements, etc. to match, once the match may be reported as a vulnerability, but because the detection is limited to a single statement and function, without fully taking into account the logic of the context, which leads to a high probability of false positives, and even inefficiencies. The main role of the data flow analysis method is now to use its extension of the taint analysis method for source code vulnerability detection, because when there are some situations that should not appear in the data flow, through the data tracking is to identify these characteristics, because this method covers a wide range of vulnerabilities, so now companies favor to use this method for source code vulnerability detection.

With the continuous development and advancement of machine learning, in order to solve the problem of high leakage and high false positives of traditional scanners, people are constantly exploring how to use machine learning related neural networks for automated vulnerability scanning and mining. Initially, a semi-automated solution was proposed, but it was quickly abandoned because it required manual definition of features and other important characteristics, which resulted in a solution that was similar to manual code vulnerability detection.

Later, a method to identify vulnerabilities by the range of similarity was proposed. The main idea of this method is that there is a source code to be tested<sup>1</sup> and an existing source code fragment<sup>2</sup>, and there is a high degree of similarity between them in some vulnerability features, and the similarity reaches a determinable range, if the source code fragment<sup>2</sup> is now known to be the vulnerability code, then we can assume that the source code<sup>1</sup> is likely to also be vulnerable source code. By using the existing vulnerability source code dataset for similarity training, and finally by using machine learning to get the relationship network related to the vulnerability code, when there is a new source code to be tested, we can quickly compare and identify whether there is a vulnerability in the source code to be tested. This method is convenient, but it requires a large amount of data inside the data set for training to ensure the accuracy of the final results, and once the training data is relatively small, the accuracy of the detection results will be greatly reduced.

Zhou[1] and others in their paper published in 2020 are based on publicly available function datasets, using convolutional neural network (CNN) model, recurrent neural network (RNN) model, convolutional neural network combined with bidirectional recurrent neural network model (CNN - BIGRU), recurrent neural network combined with bidirectional recurrent neural network model (RNN - BIGRU) to extract features from static codes and analyze them after find potential vulnerabilities. in Zhou's paper, he first uses

word2vec method to convert text data into an effective vector representation, which improves the density and structure of the data. by choosing the appropriate network structure and parameters, an optimized vulnerability detection model is formed, and relevant experiments are conducted on the test dataset. Zhou mainly studies the methods of source code vulnerability mining and feature extraction, in different network models, selecting features with different dimensions, and then testing them by model training to compare the performance of different models.

In a research paper published in 2020, Xu [2] and others implemented a static code scanning system based on taint analysis, program slicing, and BLSTM, which is oriented to the Java language and provides accurate vulnerability scanning reports using the Jar package obtained by developers after compilation as input. The paper first analyzes the principles and advantages and disadvantages of the currently used program security analysis techniques, and selects static taint analysis as the basic analysis method of the system, then chooses Joana-based backward program slicing and BLSTM as the false alarm prediction method according to the cutting-edge research results in academia, and finally analyzes the user requirements to design the system architecture, dividing the system into taint analysis module, program slicing module, data The system is divided into taint analysis module, program slicing module, data preprocessing module and false alarm prediction module, and the system provides services to users with C/S architecture.

The paper published by Gu [3] and others in 2021 mentions the project they have done gives a general working framework for deep learning based software security vulnerability mining models which includes 3 phases of data collection, learning and detection. There are many more directions to choose about the future research of the project, there are mentioned that there is maximum abstraction of deeper information about vulnerability features, building new ways of code characterization will help to improve the performance of existing vulnerability mining models, using migration learning and attention mechanisms for further analysis of problems such as cross-project detection and vulnerability location, overcoming the limitations of existing research methods, and many other aspects elaborated .

### 1.3 Research objectives and related results and contents

The key point of automated detection of source code based on machine learning is to extract the important features, and the features should contain comprehensive and detailed vulnerability information, Java source code is basically long, and the code logic is relatively complex, so it is impossible to extract the relevant information directly, even if it is taken directly for comparison, the rate of missing and false positives is also very high, which is basically equivalent to doing useless work, so it is necessary to compile the java file for compilation, compile the byte code file, after generating the dot file, so you will get the relevant node and edge feature information, through the corpus generated by the vectorization file, through the existing, the existence of machine learning neural network selection, design and implementation of a set of java source code based on graph convolutional neural network automated vulnerability mining system. Finally, the paper will be published and

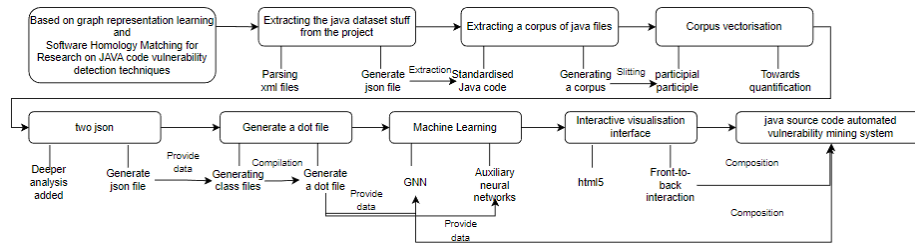


Figure 3: General flow chart of the project

the project results will be open source for all those who need it. The general flow chart of the project is shown in Figure 3.

## 2 PROJECT-RELATED THEORETICAL KNOWLEDGE AND RELATED TECHNICAL DESCRIPTION

### 2.1 Java code vulnerability concept understanding and elaboration

Nowadays, there are many different forms of vulnerabilities in Java code, and any kind of vulnerability that is not fixed accordingly will have relatively serious consequences. Java code has arbitrary file manipulation class vulnerabilities, command execution vulnerabilities, deserialization vulnerabilities, middleware vulnerabilities, business logic vulnerabilities, XSS vulnerabilities, SSRF vulnerabilities and XXE vulnerabilities, etc. In machine learning, each kind of vulnerability source code will be involved, learned to ensure that the final identification of the source code to be tested when the accuracy rate can be maintained at the normal level.

Java code has its own unique syntactic features compared to code in other languages, so it is important to choose a machine learning approach to ensure the relative completeness and relative accuracy of the final automated vulnerability detection system.

### 2.2 Introduction to Java code vulnerability mining technology

The use of manual Java code vulnerability mining techniques is mainly through code auditing techniques to Java source code vulnerability mining, but this requires a high level of technical literacy for the code auditor, experience in code auditing and a keen Java process thinking to be able to find potential vulnerability threats through path retrieval.

Java code vulnerability mining using scanning tools is mainly done by static analysis, dynamic analysis, and hybrid analysis methods to perform vulnerability scanning. Code vulnerability scanning using static analysis method depends mainly on the availability of rules with specific vulnerabilities, because static analysis method involves underlying logic and rules, so trying to create rules for all these vulnerabilities is basically an impossible operation. The accuracy of the static analysis tool is determined by calculating the ratio of false positives and false negatives to measure, in the ideal case, there is no false positive, but in practice the results are impossible to achieve, the only thing that can be achieved is to

continuously improve the vulnerability rules and reduce the false positive rate. Using the tools of dynamic analysis method is mainly to monitor the code when it is running, and to overall, systematic analysis of the entire Java source code, execution and tracking operations. Because dynamic analysis does not cover all paths, dynamic analysis is likely to miss many critical vulnerability code, and because it is dynamic analysis, it requires a lot of resources and time for real-time computing, so dynamic analysis can not be used for very large Java vulnerability analysis projects. The hybrid analysis method mainly combines the advantages and disadvantages of static analysis and dynamic analysis to make up for each other's shortcomings, and the above two methods used alone in the steps are not different, only in the choice of more room.

### 2.3 Machine Learning and Deep Learning

Machine learning can be simply defined as enabling a computer to use already existing facts as well as experience to speculate or compare certain events that are similar to the learned example in the present by continuously learning iteratively. The difference between machine learning and traditional programs is that traditional programs are coded operations to solve a specific task and are an object-oriented kind of operation, but machine learning requires a large amount of data to train and later to know how to accomplish the task.

In the broad concept of machine learning there are four types of learning: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning, each with its own characteristics and advantages and disadvantages, and different questions are answered using different types of learning.

Supervised learning, which literally means learning with criteria, is a process in which a model is trained and each set of training data has a clear label that is used to supervise this learning. The whole process of learning is constantly compared with the results of the process and the criteria to make adjustments, and the learning stops when the results reach expectations. The common types of application problems for supervised learning are classification problems and regression problems.

Unsupervised learning, as opposed to supervised learning, is a concept in which the data set is not clearly labeled during the training of the model, and because unsupervised learning does not have completely clear criteria for learning, the structure of the code can be inferred from the analysis of the data. Unsupervised learning

is often applied to clustering problems and association rule related problems.

Semi-supervised learning is both supervised and unsupervised learning. The whole learning process refers to the training process of the model, where part of the dataset is clearly labeled and part of the dataset is not clearly labeled, and the model first learns and then relabels the unlabeled dataset with a custom label based on the learned content. Semi-supervised learning is often used to solve classification problems and regression problems.

Supervised learning checks whether the model is correct by inputting the labels labeled by the data in the dataset and checking the model, but reinforcement learning directly inputs the data so that the model feeds back the relevant content, and then the model adjusts the parameters based on the feedback.

Deep learning is a special kind of machine learning. The concept of deep learning originates from artificial neural networks (ANNs), where a standard complete neural network model is constructed using neurons and adjusting the values of the weights so that the final neural network performs as expected. In 2006, a new training method, layer-by-layer greedy learning, was proposed, marking the birth of deep learning as a technology. Common deep learning architectures, models such as Restricted Boltzmann Machines (RBM), Deep Belief Networks (DBN), Convolutional Neural Networks (CNN), etc. are available.

## 2.4 GNN Graph Neural Network

The core features of CNN are: local connectivity, weight sharing and multi-layer superposition, which are also very applicable in graph neural networks (GNN). GNN is a connectivity model that obtains the dependencies in a graph by means of information transfer between nodes in the network, and GNN updates the state of a node by neighbors of any depth from that node, and this state is able to represent state information.

GNN has three advantages over CNN: nodes, edges and inference.

The GNN model is limited in that it is not efficient to use iterations to update the model state for nodes that are essentially immobile. The very first GNN uses the same parameters in the iterations, while other more well-known models use different parameters in different network layers for hierarchical feature extraction, allowing the model to learn deeper feature representations, and, the update of the hidden layer of nodes is a sequential process that can be further optimized using other neural networks, but some edges may have certain informative features that cannot be efficiently considered into it.

## 2.5 GCN Graph Convolutional Neural Network

GCN (Graph Convolutional Neural Network) is similar to CNN (Convolutional Neural Network), except that CNN is used for two-dimensional data structures and GCN is used for graph data structures. Even if the node does not have any category criteria or other criteria, the machine learning can be performed using GCN and the final result is also usable.

This method is different from the global pooling method in the SAGpool method. Global pooling is suitable for cases where there are fewer nodes and more information can be extracted, but this project has many nodes, reaching an amount of about 380,000, so

the hierarchical pooling method is chosen so that a larger range of node information can be obtained, which is suitable for the environment of this project.

## 2.6 Summary

This chapter explains what is the java source code vulnerability type, how should be mined and detected, machine learning and deep learning to exploit the learning method and GNN, GCN related knowledge and related considerations.

## 3 RESEARCH AND OVERVIEW OF MACHINE LEARNING RELATED CONTENT

### 3.1 Machine learning related problems and the main ideas of operation

Nowadays, many machine learning automated vulnerability mining techniques are based on CNN models, for example, the example of Zhou and others from Beijing Jiaotong University is a machine learning based source code vulnerability mining project [1], CNN models are fundamental to the whole project, in his project demonstrates the use of machine learning to detect software vulnerabilities directly from the source code, using a data labeled by CWE A unique C/C++ lexical analyzer was created to create a simple generalized representation of function source code for machine learning training, but the problems involved are obvious. The second problem is that the preprocessing of data is not completely done.

This project is based on graph neural network, graph convolutional neural network for machine learning, so to learn from the above lessons, the data set should try to be sufficient in number, its data set integrity, comprehensiveness are up to standard, in doing java source code pre-processing to improve, try to make each data can be perfect, in order to serve the later machine learning.

### 3.2 The advantages of graph convolutional neural network for project selection

This project is about Java source code automated vulnerability mining, because the Java language itself is special, can be compiled to finally extract the code related nodes and edges, and graph neural network can also use nodes and edges as code labels to machine learning, so this project selected graph neural network (GNN) extended graph convolutional neural network (GCN) as the GCN itself is a very good model because it does not need very complicated layers or clear nodes and edges to have good training results, and when the characteristics of nodes and edges are clear, the GCN can be used to obtain better experimental results.

### 3.3 Machine learning in graph convolutional neural network for the operation of the project and ideas

First of all, the data should be pre-processed, the data will be processed to get the normalized results, the nodes and other information will be unified and holistic, using the layered pooling method (poolh) in the SAGpool method to define the model, creating layers of GCN, merging a layer of graph pooling layers. Each layer of GCN layer, graph pooling layer as a group, erect three groups, the processed node data into the model, after the data over a group out

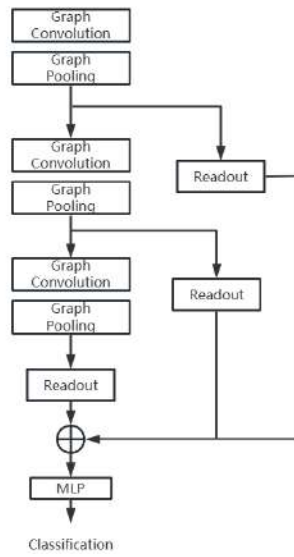


Figure 4: SAGpool hierarchical pooling method

of a group of relevant results passed into the multilayer perceptron (MLP) below for processing, as shown in Figure 4. After the above work is done, the model can be trained iteratively according to the neural network operation process. Through continuous iterative optimization, the changes of the loss function during training, the training dataset will be recorded to verify the accuracy of the dataset, and then the effect of the model will be tested by the test dataset after the training is completed to record the accuracy. Also the tested code will be tested for vulnerabilities by outputting the results with or without the words containing vulnerabilities, saved in a file containing the given statements.

### 3.4 Summary

This section provides an overview of machine learning related content, discusses the strengths and weaknesses of other projects and the areas that need attention in this project. The reasons for selecting graph neural networks and graph convolutional neural networks for this project are explained, and the operational ideas for the machine learning part of this project are described as well as the operational ideas.

## 4 A STUDY OF MACHINE LEARNING PRE-STEP DOCUMENTATION

### 4.1 Problems related to the previous steps and the main ideas of operation

In the antecedent steps of machine learning, there are six steps, the first six steps are detailed in Figure 5, the first step is to first extract the information related to Java source code from the positive and negative samples of the training data set and the positive and negative samples of the test data set and put them in the json file, there are 92113 related file information in total, the second step is because the path of Java source code has been extracted for saving, after that The third step is based on the corpus, the corpus

is vectorized, the fourth step is because the features of the json file are not clear enough, several features are re-added to generate the json file again, the fifth step is related to the compilation of Java source code to generate bytecode files (.class), and finally through the script The sixth step is to extract the nodes and edges for data processing.

The project was stuck and difficult in the fifth step, because a lot of Java source code needs to have testcasesupport package support, so you need to go to the relevant support environment to find out, and finally went through countless error reports finally solved the problem.

### 4.2 Summary

The whole chapter introduces the work of machine learning pre-steps, each step is important and indispensable. Each step of the whole project is interlinked, and the operations related to the pre-steps lay an important foundation for the subsequent machine learning steps.

## 5 JAVA CODE AUTOMATION VULNERABILITY DETECTION SYSTEM

### 5.1 The purpose of designing an automated vulnerability detection system for Java code

Solving the problem of relatively inefficient manual source code vulnerability mining. The method proposed in this project effectively avoids inefficient vulnerability detection, which reduces labor costs and allows vulnerabilities to be discovered and fixed as soon as possible.

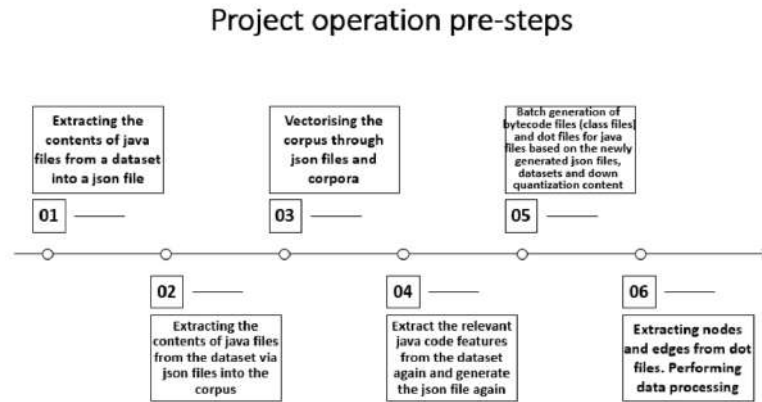
The detection system effectively reduces the difficulty of manually verifying the existence of vulnerabilities in the source code. The graph neural network audit directly displays the vulnerability content through graphs and reports, and uses large volume of data learning to ensure the correct audit rate, effectively reducing the difficulty and saving time costs of auditing source code vulnerabilities.

Solves problems such as security risks due to software homology. Because most of the java source code is open source, many of the tools use the same set or sets of open source code and use the same third-party libraries, so it will involve the problem of extremely similar software source code, and such behavior will lead to software homology and some of the same security problems, etc. This automated vulnerability detection system basically solves the above security problems caused by code homology.

### 5.2 Java code automation vulnerability detection system framework design ideas, problem solving and implementation

The front-end uses Vue for writing the front-end framework and element UI for framework layout and interface rendering. element is a set of desktop component library based on Vue 2.0 that provides various components for writing the Vue framework, such as forms, tables and navigation, etc. Using these components can meet our needs for interface design implementation.

The core design of the front-end page of the whole project is: put in the test code; output the test results; save the test results.



**Figure 5: Project operation pre-step**

Concrete process: when the entire page after the first request, was taken over by Vue. Vue and then according to the user operation to obtain data, rendered as html, Vue and then listen for events, waiting for events to occur, and finally processing events. In order to deal with responsive data, by intercepting operations, modify a data while collecting its dependencies, after modifying the data while updating the DOM, so that the modification of a JavaScript object while the view layer is also modified to complete, greatly improving the efficiency of building the framework. However, as the complexity of building the framework increases, the number of listeners in Vue also increases, but too many listeners can also lead to performance crashes, in order to solve this problem introduced components, components as the granularity, the use of responsive notifications at the component level, the internal components use DOM diff to calculate, such an approach to solve the problem of performance crashes.

The advantage of using Vue is that although the syntax of templates is relatively limited, the syntax is relatively convenient, so you can make more predictions at the pre-compilation level and improve the efficiency of writing.

Once the front-end is written it will be interfaced with the final saved optimal model of machine learning for the purpose of putting in test Java code and outputting test reports.

There are many problems associated with the operation of the project predecessor environment.

In the first step, because of the huge dataset to be involved, the variety of java source code in the dataset and the complexity of the logic, so it adds some difficulties to write the code, in identifying the absolute path of the java source code, there are many wrong paths, but these wrong paths are mixed with the correct path, which adds a lot of trouble to the subsequent steps, so without solving this problem Before we did not solve this problem, we kept trying in the code, and finally got the correct path result after 5 or 6 experimental modifications, so that the content generated in the first step can be called in the second step. The related step idea is shown in Figure 6.

In the second step, because it is to put each java source code into one line, it is time consuming to think about how to write the code,

and because it will involve the problem of separating lines and code between lines, I choose to separate each line with a space at the end, and because the dataset is huge, the corpus has to be generated by calling the virtual memory in order to generate it completely. The related step-by-step idea is shown in Figure 7.

In the third step, because we want to quantize the corpus, we need to use this three-party library, but this three-party library is not easy to download directly in the programming tool, it will report an error, so we tried to experiment by putting the environment directly through the NLTK package, the result shows that this is not stable support, so we choose to download this three-party library in the programming software again, when installing it, it will report no This is because there is no underlying environment for C. So, we went to the official website of Mvs to download some C-related underlying environment to solve the problem, and finally the vectorization file was generated smoothly. The related steps are shown in Figure 8.

In the fourth step, because the features of the java source code in the previously generated json file were not detailed enough, a secondary optimization was carried out on the basis of the original one, and the features were artificially optimized and added to improve their accuracy for the subsequent steps of identification. The relevant step ideas are shown in Figure 9.

In the fifth step, the files generated in the previous steps will be used to batch compile Java source code and generate dot files, but because all Java source code requires support, some support is up, some three-party library support is not up, but the support environment exists, the relevant Java source code is not called, the error message is that the environment does not exist, so it is Therefore, we are looking for the original support for the relevant dataset, and looking for a compatible environment so that we can compile and generate the relevant dot file in batch. At present, we have found 10 relevant supports, and after compiling them into a jar package named testcasesupport.jar for project support, we found that the project can be run, so the problem is solved. The related step-by-step idea is shown in Figure 10.

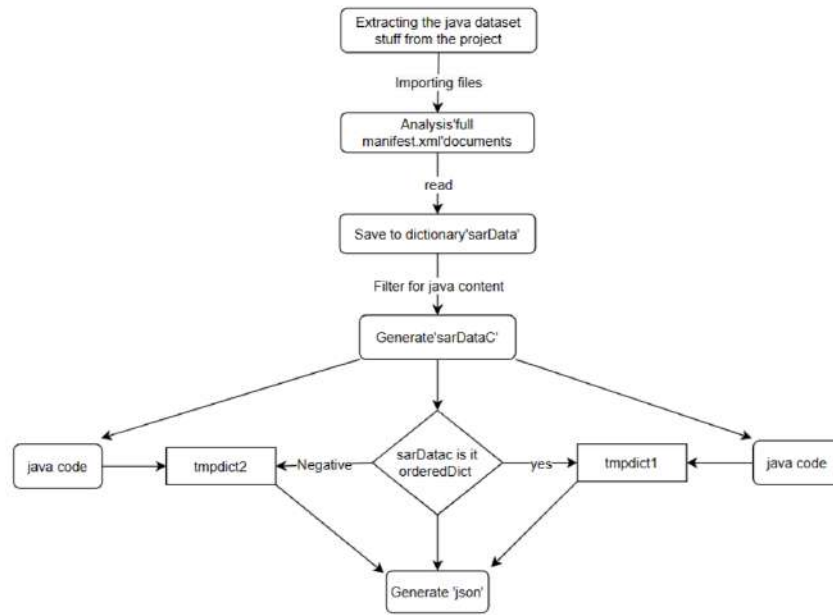


Figure 6: Machine learning pre-steps step 1

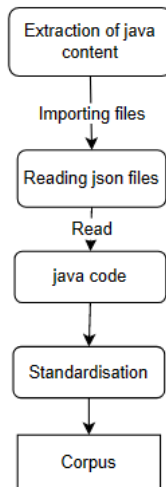


Figure 7: Machine learning pre-steps step 2

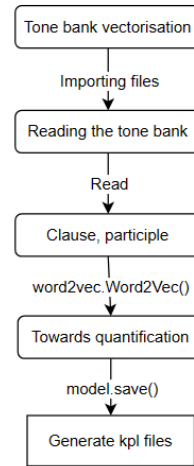


Figure 8: Machine learning pre-steps step 3

The sixth step is to take out the nodes and edges in the dot file in bulk and process the data, thus laying the foundation for subsequent machine learning of the associated graph neural network.

The dot file is a kind of generated file after secondary compilation of bytecode files compiled by Java code through soot environment. The dot file will have the information of nodes and edges corresponding to the relevant code, which can be used for the subsequent GNN and GCN related contents. the contents of the dot file are shown in Figure 11.

The dot file can be compiled to generate a png file by doing the command again, which is shown in Figure 12, 13, and 5-9.

There are also many problems associated with the operation of machine learning in the project

When the data processing script `data_pre_dispose.py` was run, it would report an error and the path was not recognized, after searching the problem, I found that it might be caused by the presence of Chinese, and changed the Chinese to English, but the problem was still not solved. After changing the file naming, the problem was solved. In the case that part of the dataset can be run and part of

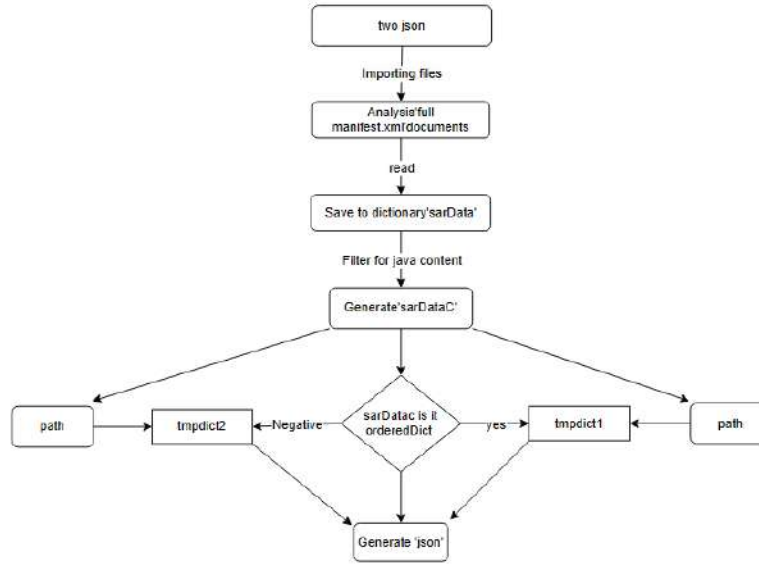


Figure 9: Machine learning pre-steps step 4

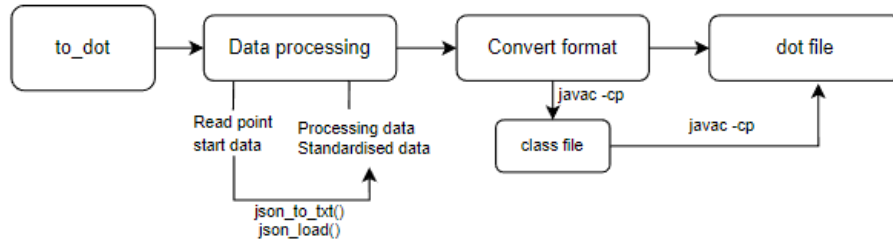


Figure 10: Machine learning pre-steps step 5

the dataset is not recognized, changing the relevant path can solve this kind of problem.

Sometimes there are problems related to main.py not being able to run, and after reviewing related materials and searching for related problems, I identified that the reason may be because the version of PyTorch is low, and changed from pytorch1.12.0+cu116 to pytorch1.13.0+cu117, which is the same as the native CUDA version, and then except for torch-geometric everything else was replaced with torch-scatter2.1.0+pt113cu117 and torch-sparse0.6.15+pt113cu117 which removed the two related supports. The solution to this problem is mainly to find the corresponding version, which corresponds to the local machine.

When the project iteration is completed, the problem also occurs when drawing images, is not normal after running the model, and then retrieve the relevant articles, the problem may be the cause of the problem of the three-party library NumPy, delete and replay the problem is solved, or import os and os.environ["KMP\_DUPLICATE\_LIB\_OK"] = "TRUE" these two statements at the top of the code, the problem can also be solved, can be normal out of the image.

### 5.3 Summary

This section lists what aspects of the Java code automation vulnerability detection system produced by this project can solve, and explains in detail how the Java code automation vulnerability detection system is implemented, the problems encountered by the Java code automation vulnerability detection system and related solutions, and how it works.

## 6 JAVA CODE AUTOMATED VULNERABILITY DETECTION SYSTEM TESTING AND ANALYSIS RESULTS

### 6.1 Training project design

The machine learning part of this project is designed to use 6 py scripts in the project, two of them put some common network modules, such as SAGPool network and backbone network, such as the model directly used in this project. One script is a data preprocessing script, and after running this script, 5 files are generated for subsequent use, as shown in Figure 14. A script is the main

```

23":>"24";
25" [label="label08: if i5 == 20 goto label09"];
26":>"25";
27" [label="$i20 = java.lang.System.out"];
28":>"26";
29" [label="label09: i6 = 100"];
30":>"31";
31" [label="$i19 = new java.lang.StringBuilder"];
32":>"27";
33" [label="specialinvoke $r19.<init>()"];
34":>"28";
35" [label="$i21 = $r19.append('value of x: \t')"];
36":>"29";
37" [label="$i22 = $r21.append(5)"];
38":>"30";
39" [label="$i23 = $r22.toString()"];
40":>"31";
41" [label="$r20.print('$i23')"];
42":>"32";
43" [label="i5 = i5 + 1"];
44":>"33";
45" [label="$i24 = java.lang.System.out"];
46":>"34";
47" [label="$i24.print('\n\t')"];
48":>"35";
49" [label="goto label08"];
50":>"36";
51" [label="$i6 = 100"];
52":>"41";
53" [label="label10: if i6 == 1000 goto label12"];
54":>"38";
55" [label="i0 = i6 % 10"];
56":>"39";
57" [label="label12: if i5 == 20 goto label13"];
58":>"63";
59" [label="$i1 = i6 / 10"];
60":>"40";
61" [label="i2 = $i1 % 10"];
62":>"41";
63" [label="i3 = i6 / 100"];
64":>"42";
65" [label="$d0 = (double) i3"];
66":>"43";
67" [label="$d1 = java.lang.Math.pow($d0, 3.0)"];
68":>"44";
69" [label="$d2 = (double) i2"];
70":>"45";
71" [label="$d3 = java.lang.Math.pow($d2, 3.0)"];
72":>"46";
73" [label="$d5 = $d1 + $d3"];
74":>"47";
75" [label="$d4 = (double) i0"];
76":>"48";
77" [label="$d6 = java.lang.Math.pow($d4, 3.0)"];
78":>"49";
79" [label="$d8 = $d5 + $d6"];
80":>"50";
81" [label="$d7 = (double) i6"];
82":>"51";
83" [label="$b4 = $d8 cmpl $d7"];
84":>"52";

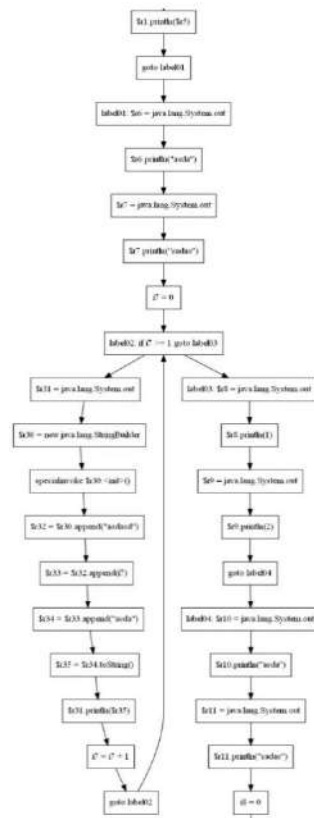
```

**Figure 11: dot file content**

program execution function script, which is mainly responsible for iterations and so on. A script has a class defined inside it that will be called in the main program execution function script, and four of the five files generated by the data preprocessing will be called in the script with the class defined. The last script is a script that verifies whether there are vulnerabilities in the code content of the test dataset. The role is mainly to verify the existence of vulnerabilities and save the results in the specified files, as shown in Figures 15 and 16. The project will produce images and save the optimal model after the completion of the iteration.

## 6.2 Test result display and discussion

The current machine learning part uses a training dataset of 56,636 positive and negative sample codes and a test dataset of 35,477 positive and negative sample codes to generate the relevant files, and after data processing, there are 379,922 nodes, 379,922 node attribute information and 129,524 edge information of the relevant graphs, using the SAGpool method, iterated 54 times, and the current model accuracy is approximately 98%. The related machine learning results are shown in Figure 17.

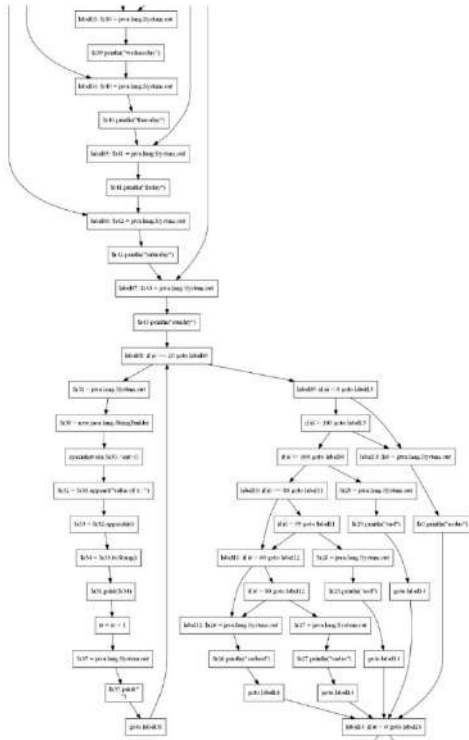


**Figure 12: Related code diagram indicates the conten**

### 6.3 Comparison testing and analysis with similar systems already in existence

In 2019 Yu [4] and others proposed a sample collection scheme based on code block sequences, designed a Block2Vec-based machine learning model, input a basic block of code as a whole into the Doc2Vec model, using the collection of assembly code blocks of the collected programs as a training data set, not limited to the previously collected assembly code blocks, and tested by constructing The experiments of Yu et al. apply to assembly, which is also a vulnerability prediction model like this project, but the differences are based on different language directions and the neural networks utilized in machine learning are also very different.

Liu [5] and others used bidirectional LSTM neural network for static java code vulnerability mining, while in the experiments and two more similar work VulDeePecker [6] and AE-KNN [7] for comparison tests, Liu and others used static analysis to extract semantic features of the source code and generate intermediate representations, after which the generated intermediate representations were mapped to vectors, while for After that, the generated intermediate representations are mapped into vectors, and at the same time labeled as safe or unsafe, and finally after machine learning, the overall idea is not similar to this project except that the tested language is the same. This project is directly based on the nodes in the dot file, and the data processing is performed and then machine



**Figure 13: Related code diagram indicates the content**

```
100% |██████████| 84220/84220 [24:35<00:00, 57.07it/s]
100% |██████████| 20002/20002 [02:48<00:00, 118.50it/s]
train_size=21283
100% |██████████| 144133/144133 [31:42<00:00, 75.76it/s]
100% |██████████| 20000/20000 [03:09<00:00, 105.59it/s]
train_size and test_size is 52167
total_size=52167
not_fount_file_count=0
```

**Figure 14: Data pre-processing script results demonstration**

[illegible]

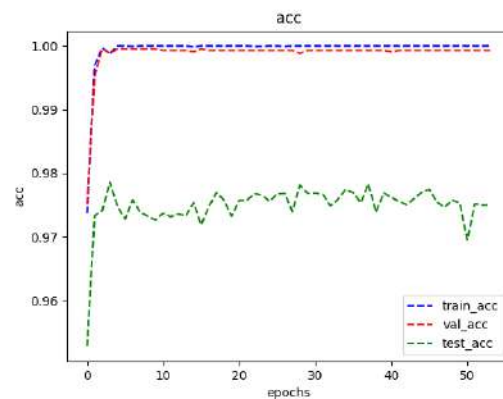
**Figure 15: Verify the test dataset code content for vulnerability script results show**

```

75749 test10021.class,0,0,ftest10021.class,无属性
75750 test10020.class,0,0,ftest10020.class,无属性
75751 test10019.class,0,0,ftest10019.class,无属性
75752 test10018.class,0,0,ftest10018.class,无属性
75753 test10017.class,0,0,ftest10017.class,无属性
75754 test10016.class,0,0,ftest10016.class,无属性
75755 test10015.class,0,0,ftest10015.class,无属性
75756 test10014.class,0,0,ftest10014.class,无属性
75757 test10013.class,0,0,ftest10013.class,无属性
75758 test10012.class,0,0,ftest10012.class,无属性
75759 test10011.class,0,0,ftest10011.class,无属性
75760 test10010.class,0,0,ftest10010.class,无属性
75761 test10009.class,0,0,ftest10009.class,无属性
75762 test10008.class,0,0,ftest10008.class,无属性
75763 test10007.class,0,0,ftest10007.class,无属性
75764 test10006.class,0,0,ftest10006.class,无属性

```

**Figure 16: Verify the test dataset code content for vulnerability script results show**



**Figure 17: Machine learning results showcase**

learning is performed using the methods in GCN, which is different from what Liu and others have done.

Zhuang [8] and others construct graph neural networks to perform vulnerability mining in C. The graph containing vertex attributes and edge attributes (control dependencies and data dependencies) generated by the source code is used as input to the graph network, and the attributes of the graph are updated in the graph network, and the attributes obtained through learning are updated as features of the source code, and the source code is classified into two categories containing vulnerabilities and not containing vulnerabilities using these features. The idea and logic of Zhuang and others are similar to the logic of this project, but the languages they detect are different, and although they both use GNN, this project uses GCN, which is an extension of GNN, for machine learning, so it does not have exactly the same points as this project.

## 7 SUMMARY

## 7.1 Project Summary and Outlook

This project is based on the generation of documents in the front environment, the subsequent machine learning steps and the visual interactive interface to form a back-and-forth linkage, and the overall logic of the project is self-consistent. Although many problems

were encountered during the completion of the project, they were solved perfectly through continuous efforts and exploration.

This project is currently only the initial machine, the future will have more optimization into the project, and after the project open source will also have more like-minded people to continue to improve the project. I hope that the content of this project can be more Java-related projects in the future, the program escort.

## ACKNOWLEDGMENTS

I would like to express my sincere thanks to my classmates and tutors for their help and support for this project!

## REFERENCES

- [1] Zhou Keqiang. Machine learning based source code vulnerability mining [D]. Beijing Jiaotong University, 2020. DOI:10.26944/d.cnki.gbfju.2020.001006
- [2] Xu Wenyuan. Design and implementation of a Java static vulnerability scanning system based on machine learning [D]. Nanjing University, 2020. doi:10.27235/d.cnki.gniju.2020.001404.
- [3] Gu Mianxue, Sun Hongyu, Han Dan, Yang Su, Cao Wanying, Guo Zhen, Cao Chunjie, Wang Wenjie, Zhang Yuqing. Deep learning-based software security vulnerability mining[J]. Computer Research and Development, 2021, 58(10): 2140-2162.
- [4] Yu Ting. Research on binary vulnerability mining technology based on machine learning[D]. Xi'an University of Electronic Science and Technology, 2019. doi:10.27389/d.cnki.gxadu.2019.002805.
- [5] Liu Jiahua, Wan Ming, Zhou Chenxi, Zhang Pan. Bi-directional LSTM-based vulnerability detection for Java open source software[J]. Computer Applications and Software, 2020, 37(12): 322-327.
- [6] Li Z, Zou D Q, Xu S H, *et al.* Vul Dee Pecker: A deep learning-based system for vulnerability detection [C]//The Net-work and Distributed System Security Symposium, 2018.
- [7] Li Yuancheng, Huang Rong, Lai Fenggang, *et al.* A deep clustering-based vulnerability detection method for open source software [J]. Computer Application Research, 2020, 37(4): 1107-1110, 1114
- [8] Zhuang, R.F.. Research on key technologies of vulnerability mining based on graph networks [D]. Harbin Institute of Technology, 2020. doi:10.27061/d.cnki.ghgdu.2020.003472.

# PhyGNNet: Solving spatiotemporal PDEs with Physics-informed Graph Neural Network

Longxiang Jiang

Liyuan Wang

Xinkun Chu

Yonghao Xiao

Hao Zhang\*

jianglx@whu.edu.cn

wangly\_xjtu@163.com

neocosmos@163.com

xiao\_yonghao@caep.com

linusec@163.com

Institute of Computer Application, China Academy of Engineering Physics  
Mianyang, Sichuan, China

## ABSTRACT

Partial differential equations (PDEs) are a common means of describing physical processes. Solving PDEs can obtain simulated results of physical evolution. Currently, the mainstream neural network method is to minimize the loss of PDEs thus constraining neural networks to fit the solution mappings. By the implementation of differentiation, the methods can be divided into PINN methods based on automatic differentiation and other methods based on discrete differentiation. PINN methods rely on automatic backpropagation, and the computation step is time-consuming, for iterative training, the complexity of the neural network and the number of collocation points are limited to a small condition, thus abating accuracy. The discrete differentiation is more efficient in computation, following the regular computational domain assumption. However, in practice, the assumption does not necessarily hold. In this paper, we propose a PhyGNNet method to solve PDEs based on graph neural network and discrete differentiation on irregular domain. Meanwhile, to verify the validity of the method, we solve Burgers equation and conduct a numerical comparison with PINN. The results show that the proposed method performs better both in fit ability and time extrapolation than PINN. Code is available at <https://github.com/echowve/phygnet>.

## CCS CONCEPTS

• **Computing methodologies** → *Causal reasoning and diagnostics*.

\*The corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590029>

## KEYWORDS

Physics-informed neural networks, Partial differential equation, Graph neural networks, Surrogate modeling

### ACM Reference Format:

Longxiang Jiang, Liyuan Wang, Xinkun Chu, Yonghao Xiao, and Hao Zhang. 2023. PhyGNNet: Solving spatiotemporal PDEs with Physics-informed Graph Neural Network. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590029>

## 1 INTRODUCTION

With the development of physics, biology, chemistry, and other fields, a large number of partial differential equations [3] have been accumulated in the related fields, the evolution process and results of specific problems can be obtained by solving the PDEs. However, solving PDEs is a challenging task, except for a few equations which exist analytical solutions, recent years, most of the equations are solved numerically [5]. Along with the rapid progress and popularization of deep learning, benefit from the excellent fit ability of neural networks, a series of novel methods have emerged nowadays.

The current representative method is PINN [9], which mostly adopts the fully connected network as the solver, and constructs a multi-objective loss function with the loss of PDE through automatic differentiation to optimize the network parameters. The solver takes time-domain coordinates tuple as input and predicts the corresponding solution, the differential between prediction and input is calculated through back-propagation. The method is widespread in solving fluid flow [1], heat conduction [13], and other problems. However, the current PINN method has two limitations. On the one hand, automatic differentiation is time-consuming, and the consumption is significantly increased when the amount of parameters of the fully connected network is large, or the differentiation is high-order. On the other hand, PINN approximates the real solution by piecewise fitting, to ensure the correctness of the solution, a large number of collocation points are required for iterative training, which further increases the time consumption of solving.

In addition to the above PINN method, some methods refer to numerical calculation and adopt discrete differences to approximate differentiation. Different from PINN, these methods take the solution at a certain time step as neural network input, the neural network outputs the predicted solution on a time interval behind the input, by minimizing the loss of PDE constructed by discrete difference, and the prediction tends to the real solution, which is an iterative method on time step. Geo et al [6] proposed a method to calculate spatial discrete differential using convolution kernel with fixed weights and adopt convolution neural network (CNN) as the solver. Similar to [6], [10] introduced Conv-LSTM neural network to solve spatiotemporal partial differential equations. Furthermore, with FV discretization scheme and two-point flux approximation [4], [11] applies CNN to solve transient Darcy flows.

In this paper, we present a method for solving PDEs in irregular domain based on GNN and discrete difference. This method divides the computational domain into meshes and treats the meshes as an undirected graph. The solution of PDE at the vertices of the mesh is predicted by message passing mechanism of the graph neural network. In addition, this paper propose a discrete difference method for calculating laplace and gradient values on irregular grids based on Taylor expansion and least squares regression, which is used to construct PDE loss to constraint graph neural network to predict solution. What's more, in this paper, we solve burgers equation using the method presented and conduct comparative numerical experiments with PINN. The results show that this method has better solution accuracy and time extrapolation ability.

## 2 METHOD

The general form of PDEs can be expressed as:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathcal{N}[\mathbf{u}, \dots, \nabla \mathbf{u}, \Delta \mathbf{u}, \nabla \mathbf{u} \cdot \mathbf{u}, \dots; \lambda] = 0 \quad (1)$$

where,  $\mathbf{u} \in \mathbb{R}^{d \times 1}$  denotes solutions which have  $d$  components in domain  $\Omega$ ,  $\frac{\partial \mathbf{u}}{\partial t}$  is time derivative,  $\nabla \mathbf{u}$  represents the gradient in space.  $\Delta \mathbf{u}$  is Laplace item, which is equal to  $\nabla^2 \mathbf{u}$ .  $\lambda$  is parameters of the PDE. In addition, the Initial Condition (IC) and Boundary Condition (BC) have the following definitions:

$$\mathcal{I}(\mathbf{u}, \nabla \mathbf{u}, \Delta \mathbf{u}, \dots; t = 0) = 0 \quad (2)$$

$$\mathcal{B}(\mathbf{u}, \nabla \mathbf{u}, \Delta \mathbf{u}, \dots; \mathbf{x} \in \partial\Omega) = 0 \quad (3)$$

where,  $\partial\Omega$  denotes the boundary area of  $\Omega$ .

Given the above PDE equation and Initial/Boundary Conditions, a solution can be found with several approaches. However, In this paper, we aim to solve the problem with GNN. Similar to PINN, we yearn to devise a method that is unsupervised to solve PDEs.

The framework of our approach to solving the above equation is illustrated in Fig.1. We first divide the domain into an irregular mesh and express the mesh as an undirected graph, then assign the solution in a time step to the nodes of the graph, note that at the very first time step, the solution is the IC. The network takes the graph and encodes the node attributions and edges into features, updates features with message passing, and then decodes the features into solution for next time which behind the input time  $\Delta t$  interval, the procedure is detailed in Sec.2.1. When training, the BC is assigned to the predicted solutions to compute PDE loss, the solution is then detached from the calculation graph and fed into

the network to predict solutions of all  $T$  time steps by repeating the above procedure, and the losses of  $T$  steps are cumulated to update the parameters of the network at once. The above process is repeated many times until the specified number of times is reached or the network converges to get final solutions.

### 2.1 Network

We regard the mesh as an undirected graph to train the network. That is, we express the grid points as graph nodes, and assign edges to nodes that are nearest neighbors to each other. Similar to [8], the features of nodes consist of the current solution and 2-dimensional one-hot code of node type, indicating whether a specific node is located on  $\partial\Omega$ . The features of edges contain euclidean distance and the coordinate difference between sender and receiver nodes.

The framework of our network is shown in Fig.1. The network put the solution at time  $t$  to obtain the solution at the next time step  $t + \Delta t$ . It mainly has three parts. The encoder transforms node and edges features mentioned above with MLP, the processor predicts latent feature variation of nodes via Graph Network [2] (GN) and the decoder decodes node features with MLP as correction of the input  $\mathbf{u}_t$  to create final predicts. The dotted lines in the figure represent residual connection.

Specifically, the calculation process of GN in the framework contains edges update and nodes update steps. To illustrate the procedure, here, we define the edge  $\mathbf{e}_{i,j}$  connected to two nodes with features  $\mathbf{v}_i$  and  $\mathbf{v}_j$  respectively. The GN conducts edges update step at first, which can be described as:

$$\tilde{\mathbf{e}}_{i,j} = f_e(\mathbf{e}_{i,j} || \mathbf{v}_i || \mathbf{v}_j) \quad (4)$$

where  $||$  denotes concatenating operator and  $f_e$  denotes MLP for edges. With the updated edge features, GN then update node features as:

$$\tilde{\mathbf{v}}_i = f_n(\mathbf{v}_i || \sum_{j \in \mathcal{N}(i)} \tilde{\mathbf{e}}_{i,j}) \quad (5)$$

and,  $f_n$  denotes MLP for nodes.

### 2.2 Discrete format

The above PDE equation consists of three derivative operators, namely  $\mathbf{u}_t$ ,  $\nabla \mathbf{u}$ , and  $\Delta \mathbf{u}$ . The three operators can be approximated by the following discrete format.

For operator  $\mathbf{u}_t$ , it can be approximated with backward difference on time  $t$ , denoted as:

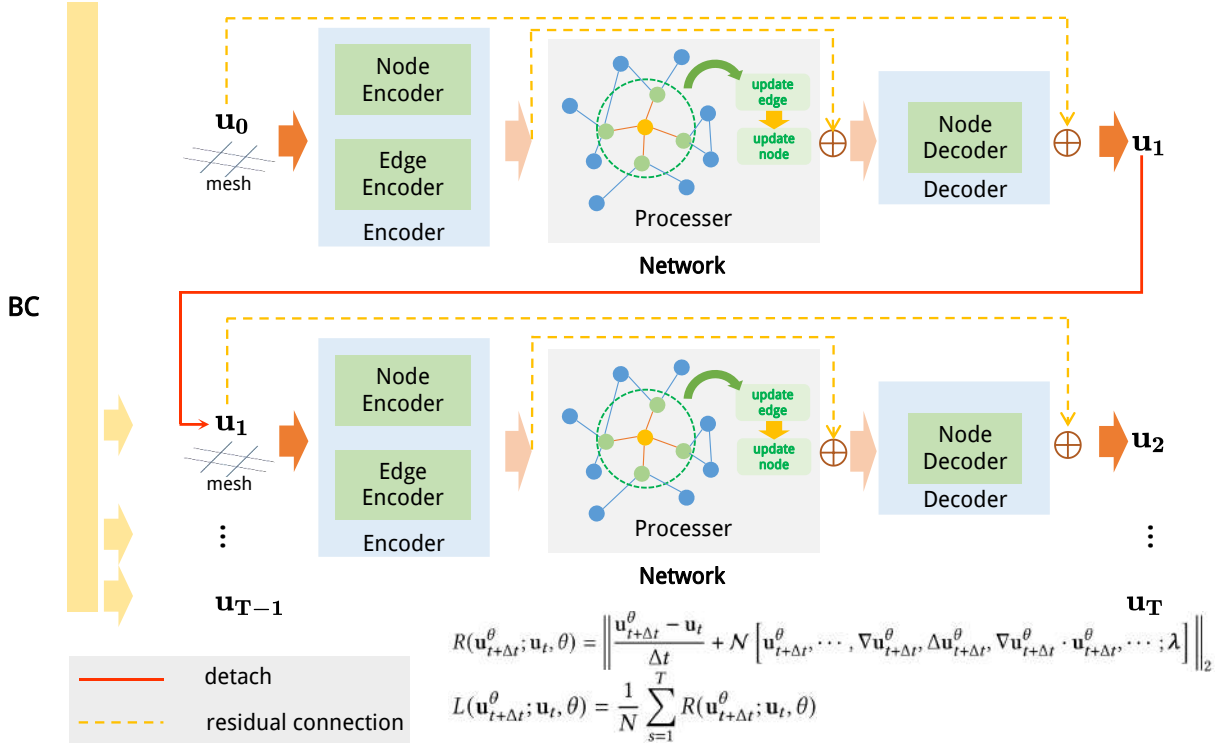
$$\frac{\partial \mathbf{u}}{\partial t} = \frac{\mathbf{u}_{t+\Delta t} - \mathbf{u}_t}{\Delta t} \quad (6)$$

where  $\Delta t$  denotes the single time interval in the time evolution process, is a superparameter.

Operators  $\nabla \mathbf{u}$  and  $\Delta \mathbf{u}$  are differential on spatial area, when the area  $\Omega$  is divided into irregular grids, the operators can be defined on the grids as well. For simplicity, we only take two-dimension as an example to illustrate the solution method of the operator, for others, the same.

According to first-order Taylor expansion the value at  $\mathbf{u}_{\mathbf{x}+\Delta \mathbf{x}}$  can be approximated with the value at  $\mathbf{u}_{\mathbf{x}}$  as:

$$\mathbf{u}_{\mathbf{x}+\Delta \mathbf{x}} = \mathbf{u}_{\mathbf{x}} + \nabla \mathbf{u}_{\mathbf{x}} \Delta \mathbf{x} \quad (7)$$



**Figure 1: The framework of our proposed method. Given initial and boundary conditions, we predict the solutions of multi time steps with the network consisting of an encoder, processor, and decoder blocks, where the processor is a graph neural network block.**

Then, given  $\mathbf{u}_{x+\Delta x}$  and  $\mathbf{u}_x$ , the gradients can be solved with least squares regression as:

$$\nabla \mathbf{u}_x = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{U} \quad (8)$$

where,  $\mathbf{A} \in \mathbb{R}^{m \times 2}$  is coordinate differences matrix with  $m$  nearest neighbors, denoted as:

$$\begin{bmatrix} \Delta x_1, \Delta y_1 \\ \Delta x_2, \Delta y_2 \\ \vdots \\ \Delta x_{m-1}, \Delta y_{m-1} \\ \Delta x_m, \Delta y_m \end{bmatrix} \quad (9)$$

and  $\mathbf{U} \in \mathbb{R}^{m \times d}$  is value differences matrix with  $m$  nearest neighbors, denoted as:

$$\mathbf{U} = [\mathbf{u}_{x+\Delta x_1} - \mathbf{u}_x; \mathbf{u}_{x+\Delta x_2} - \mathbf{u}_x; \dots; \mathbf{u}_{x+\Delta x_m} - \mathbf{u}_x] \quad (10)$$

And, the Laplace value follow the form as:

$$\Delta \mathbf{u}_i = \sum_{j \in \mathcal{N}(\mathbf{u}_i)} w_{ij} (\mathbf{u}_j - \mathbf{u}_i) \quad (11)$$

which indicates that the laplacian value is a weighted difference of values of nearest neighbors, and the weights  $w$  are only related to the nearest neighbor structure of the graph, therefore, here, we determine the weights based on a group of test function as following:

We expect to select representative functions to solve the weights, due to the linearity of the laplacian operator, here, we choose the basic function of the second-order Taylor expansion as  $\{x, y, xy, x^2, y^2\}$ . Given the coordinates of a point and its neighbors, the function values and laplacian values of the group can be numerically obtained. It is worth noting that to ensure numeric stability, all the coordinates are subtracted from the point coordinates. With the above test function values and corresponding Laplacian values,  $w$  can also be solved with least squares regression as:

$$\mathbf{w}_i = (\tilde{\mathbf{A}}^T \tilde{\mathbf{A}})^{-1} \tilde{\mathbf{A}}^T \tilde{\mathbf{u}} \quad (12)$$

where  $\tilde{\mathbf{A}} \in \mathbb{R}^{5 \times m}$  is test function value difference matrix, and  $\tilde{\mathbf{u}} \in \mathbb{R}^{5 \times 1}$  is a vector filled with corresponding laplacian values.

### 2.3 PDE loss construction

The objective of our approach is to solve PDE on an irregular domain with GNN when the PDE equation, IC, and BC are explicitly presented. In PINN, the PDE equation, IC, and BC are softly satisfied when minimizing a multi-objective loss function. However, the loss function requires tuning the weight parameters of multiple losses carefully to avoid falling into local minima. In this paper, we construct a loss function with the same methodology as [10], given  $\mathbf{u}_t$ , the network predicts  $\mathbf{u}_{t+\Delta t}^\theta$ , the loss value of grid points has the

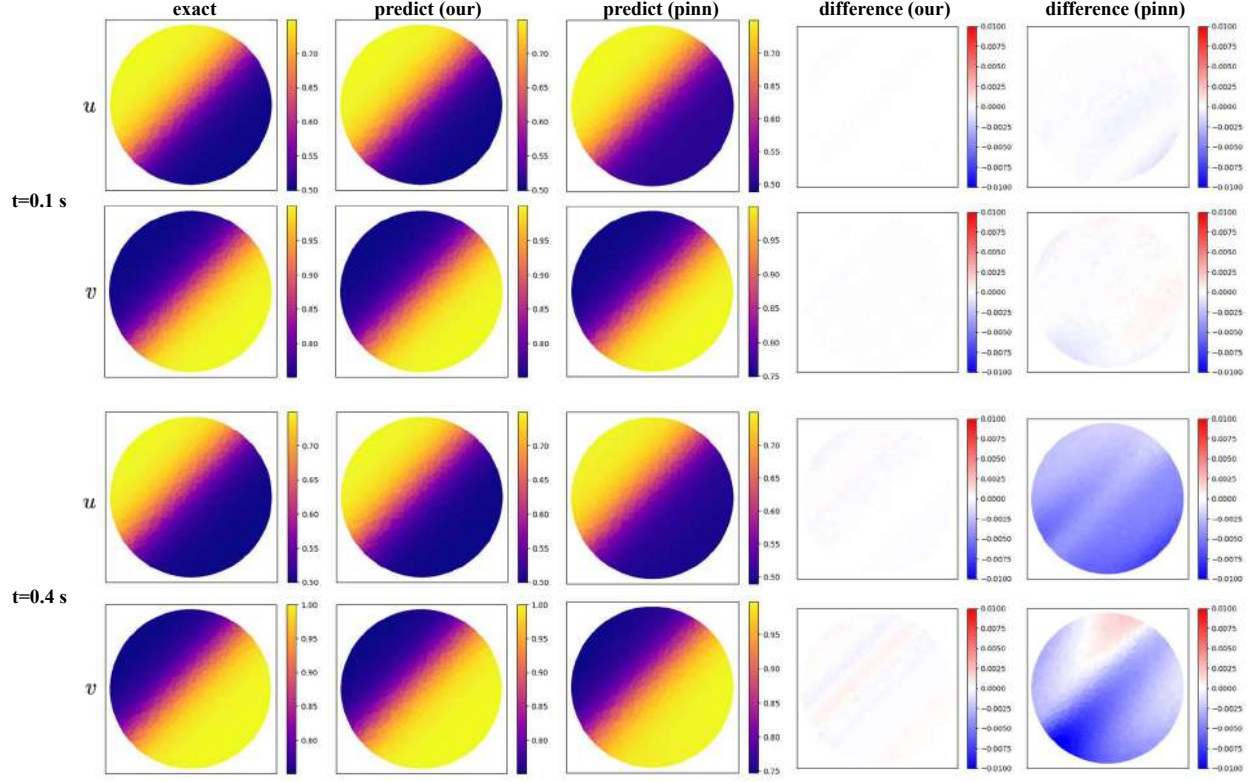


Figure 2: The results of burgers equation with  $u, v$  components at different time steps. The predicted results are compared with the exact analytical solutions and the difference is also presented.

following form:

$$R(\mathbf{u}_{t+\Delta t}^\theta; \mathbf{u}_t, \theta) = \left\| \frac{\mathbf{u}_{t+\Delta t}^\theta - \mathbf{u}_t}{\Delta t} + \mathcal{N}[\mathbf{u}_{t+\Delta t}^\theta, \dots, \nabla \mathbf{u}_{t+\Delta t}^\theta, \Delta \mathbf{u}_{t+\Delta t}^\theta, \nabla \mathbf{u}_{t+\Delta t}^\theta \cdot \mathbf{u}_{t+\Delta t}^\theta, \dots; \lambda] \right\|_2 \quad (13)$$

and the loss function to optimize is:

$$L(\mathbf{u}_{t+\Delta t}^\theta; \mathbf{u}_t, \theta) = \frac{1}{N} \sum_{s=1}^T R(\mathbf{u}_{t+\Delta t}^\theta; \mathbf{u}_t, \theta) \quad (14)$$

Notice that when training, we update the network parameters by cumulating gradients of multi time-steps. For each time step, the input is the network prediction of the last time step, and the IC is as input at the very first step of the network. Also, similar to [10], the BC is hard assigned when organizing loss. In detail, the values of boundary nodes in prediction are assigned according to the BC before PDE loss construction.

### 3 EXPERIMENTS

In this section, we conduct numerical experiments on burgers equation [12] to solve the propagation and reflection of waves to evaluate our proposed method.

#### 3.1 Setup

As aforementioned, there are MLPs in the encoder, processor, and decoder of the neural network. In our experiments, the MLPs are with two hidden layers, each with 128 neurons, and the ReLU activation function is applied to transform the output of the input layer and hidden layer. For the variable parameters in discrete format, the time interval  $\Delta t$  is set to 0.001. The computational domain is set to a disk with 0.5 radius and (0.5, 0.5) as center. The learning rate of our method is set to  $1 \times 10^{-4}$ .

#### 3.2 Burgers Equation

Here, we consider the two-dimensional burgers equation as example, which has the following form:

$$\begin{aligned} u_t + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} - \frac{1}{R} \Delta u &= 0 \\ v_t + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} - \frac{1}{R} \Delta v &= 0 \end{aligned} \quad (15)$$

where the parameter  $R$  is the Reynolds number, which controls the wave dynamics. Specifically, To verify the accuracy of our method, we choose to solve the equation with the exact solution same as

**Table 1: The aRMSE at different time steps of Burgers Equation.**

Step	1	100	200	300	400
PINN	2.06e-04	2.01e-04	3.73e-04	8.60e-04	1.69e-03
OUR	1.09e-06	3.09e-05	5.67e-05	8.00e-05	1.02e-04

[12], that is, the solution is shown below:

$$\begin{aligned} u(x, y, t) &= \frac{3}{4} - \frac{1}{4(1 + e^{(R(-t-4x+4y))/32})} \\ v(x, y, t) &= \frac{3}{4} + \frac{1}{4(1 + e^{(R(-t-4x+4y))/32})} \end{aligned} \quad (16)$$

The IC is the exact solution at time  $t = 0$  and the boundary value changes along with time  $t$  based on the solution. The parameter  $R$  is set to 80. We train the network to solve the PDE at the first 10 time steps and expect the model to have the ability to infer the subsequent solutions. In detail, we set  $T = 10$  in this situation and repeat the training process 10000 epochs, taking the model with minimal PDE loss to evaluate performance.

In addition, based on DeepXDE [7] framework, we construct a PINN baseline as a comparison of our method, that is, we build an MLP model with 4 hidden layers, each layer containing 20 neurons and we choose the Tanh as activation function. When training, for a fair comparison, the computation area, and train time steps are the same as the items in our approach, we randomly sample  $1.2 \times 10^5$  collocation points and train the model with two stages, in the first stage, we train the model  $1 \times 10^4$  epochs by Adam optimizer with  $1 \times 10^{-4}$  learning rate and then, in the second stage, the L-BFGS optimizer is adopted to further minimize the loss, which repeats  $1 \times 10^5$  epochs.

The results are demonstrated in Tab.1, which contains the aRMSE (accumulated Root Mean Square Error) [8, 10] at different time steps. Note that the model is built at the very first 10 steps, thus the results are an extrapolation of the model. As is shown in the Table, our method maintains a low level of error over time, demonstrating the better fitting ability and long-time generalization performance of the model compared with the PINN approach. Besides, we also visualize our results in Fig.2. In the Figure, the first two rows in the figure are the results at 0.1s and the remaining indicates the results at 0.4s. As we can observe, the predicted patterns are close to the exact solutions, and the differences are near zero.

## 4 CONCLUSION

In this paper, we proposed a method to solve PDEs with a graph neural network on an irregular computational domain. That is, specifying the status of a time step, we supervised the network to predict the solution of the next time step by minimizing a PDE loss. Compared with the typical PINN methods, our approach doesn't rely on automatic differentiation, the differential term is constructed with spatial and time differences. In particular, to obtain spatial differences on an irregular computational domain, we propose an approach that approximates gradients and laplacian values with least squares regression. The experiments conducted on the burgers

equation show that our method performs better in fit ability and time extrapolation in contrast to the PINN method.

## REFERENCES

- [1] Muhammad M Almajid and Moataz O Abu-Al-Saud. 2022. Prediction of porous media fluid flow using physics informed neural networks. *Journal of Petroleum Science and Engineering* 208 (2022), 109205.
- [2] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261* (2018).
- [3] Haim Brezis and Haim Brézis. 2011. *Functional analysis, Sobolev spaces and partial differential equations*. Vol. 2. Springer.
- [4] Zhangxin Chen, Guanren Huan, and Yuanle Ma. 2006. *Computational methods for multiphase flows in porous media*. SIAM.
- [5] John R Dormand. 2018. *Numerical methods for differential equations: a computational approach*. CRC press.
- [6] Han Gao, Luning Sun, and Jian-Xun Wang. 2021. PhyGeoNet: physics-informed geometry-adaptive convolutional neural networks for solving parameterized steady-state PDEs on irregular domain. *J. Comput. Phys.* 428 (2021), 110079.
- [7] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. 2021. DeepXDE: A deep learning library for solving differential equations. *SIAM Rev.* 63, 1 (2021), 208–228. <https://doi.org/10.1137/19M1274067>
- [8] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. 2020. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409* (2020).
- [9] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378 (2019), 686–707.
- [10] Pu Ren, Chengping Rao, Yang Liu, Jian-Xun Wang, and Hao Sun. 2022. PhyCRNet: Physics-informed convolutional-recurrent network for solving spatiotemporal PDEs. *Computer Methods in Applied Mechanics and Engineering* 389 (2022), 114399.
- [11] Zhao Zhang. 2022. A physics-informed deep convolutional neural network for simulating and predicting transient Darcy flows in heterogeneous reservoirs without labeled data. *Journal of Petroleum Science and Engineering* (2022), 110179.
- [12] Hongqing Zhu, Huazhong Shu, and Meiyu Ding. 2010. Numerical solutions of two-dimensional Burgers' equations by discrete Adomian decomposition method. *Computers & Mathematics with Applications* 60, 3 (2010), 840–848.
- [13] Navid Zobeiry and Keith D Humfeld. 2021. A physics-informed machine learning approach for solving heat transfer equation in advanced manufacturing and engineering applications. *Engineering Applications of Artificial Intelligence* 101 (2021), 104232.

# Multi-strategy Improved Multi-objective Harris Hawk Optimization Algorithm with Elite Opposition-based Learning

Fulin Tian<sup>1\*</sup>, Jiayang Wang<sup>1</sup>, Fei Chu<sup>1</sup>, Lin Zhou<sup>1</sup>

School of Computer Science, Central South University, Changsha 410006, China. e-mail: 709128915@qq.com

## ABSTRACT

**Abstract:** To make up for the deficiencies of the Harris hawk optimization algorithm (HHO) in solving multi-objective optimization problems with low algorithm accuracy, slow rate of convergence, and easily fall into the trap of local optima, a multi-strategy improved multi-objective Harris hawk optimization algorithm with elite opposition-based learning (MO-EMHHO) is proposed. First, the population is initialized by Sobol sequences to increase population diversity. Second, incorporate the elite backward learning strategy to improve population diversity and quality. Further, an external profile maintenance method based on an adaptive grid strategy is proposed to make the solution better contracted to the real Pareto frontier. Subsequently, optimize the update strategy of the original algorithm in a non-linear energy update way to improve the exploration and development of the algorithm. Finally, improving the diversity of the algorithm and the uniformity of the solution set using an adaptive variation strategy based on Gaussian random wandering. Experimental comparison of the multi-objective particle swarm algorithm (MOPSO), multi-objective gray wolf algorithm (MOGWO), and multi-objective Harris Hawk algorithm (MOHHO) on the commonly used benchmark functions shows that the MO-EMHHO outperforms the other compared algorithms in terms of optimization seeking accuracy, convergence speed and stability, and provides a new solution to the multi-objective optimization problem.

## CCS CONCEPTS

• Computing methodologies;

## KEYWORDS

Multi-objective optimization, Harris hawk algorithm, Elite opposition-based learning, Adaptive grid, Gauss walk learning

### ACM Reference Format:

Fulin Tian<sup>1\*</sup>, Jiayang Wang<sup>1</sup>, Fei Chu<sup>1</sup>, Lin Zhou<sup>1</sup>. 2023. Multi-strategy Improved Multi-objective Harris Hawk Optimization Algorithm with Elite Opposition-based Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590030>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590030>

## 1 INTRODUCTION

Multi-objective optimization problems (MOP) are a class of problems in which there are multiple optimization objectives in the optimization problem, and multiple objectives must be optimized simultaneously to obtain the best solution. With the development of computers and artificial intelligence, many novel swarm intelligence algorithm have been proposed and used to solve multi-objective optimization problems. Deb applied genetic algorithms to multi-objective optimization and proposed NSGA-II<sup>[1]</sup> and NSGA-III<sup>[2]</sup>. Xue et al. proposed multi-objective differential evolution (MODE)<sup>[3]</sup>. Coello et al. proposed multi-objective particle swarm optimization, (MOPSO)<sup>[4]</sup>, which applied the particle swarm algorithm, which can only be applied to a single objective, to multiple objectives. The algorithm has attracted many scholars to study it. Since then, numerous multi-objective versions of swarm intelligence optimization algorithms have emerged, such as the nondominated neighbor immune algorithm (NNIA)<sup>[5]</sup>, multi-objective ant colony optimization (MOACO)<sup>[6]</sup>, multi-objective cuckoo search algorithm (MOCS)<sup>[7]</sup>, multi-objective grey wolf optimizer (MOGWO)<sup>[7]</sup>, etc.

The Harris Hawks Optimization (HHO) algorithm<sup>[8]</sup> is a novel swarm intelligence algorithm inspired by Herdari et al.. The current research on multiobjective HHO is in its initial stage, and Yüzge U et al.<sup>[9]</sup> proposed Multi-Objective Harris Hawks Optimizer (MOHHO) by combining external profile strategy and roulette rules with the HHO algorithm, but there are problems of convergence accuracy and Pareto front distribution Uniformity is difficult to balance, easy to fall into local optimum at the later stage of search and slow convergence speed. Selim et al.<sup>[10]</sup> used gray correlation analysis to obtain the best compromise solution in the non-dominated Pareto scheme to increase the uniformity of the solution set. Devarapalli et al.<sup>[11]</sup> proposed an escape energy control formulation to improve the convergence speed by introducing an exponential function and balancing the relationship between exploration and exploitation to improve the convergence accuracy. In this paper, we address the defects in the original Harris Hawks algorithm and propose the Multi-objective multi-strategy improved Harris hawks optimization with elite opposition-based learning (MO-EMHHO).

## 2 HARRIS HAWKS OPTIMIZATION (HHO)

The basic Harris hawk optimization algorithm is described as follows.

### 2.1 Exploration Phase

The global search phase is mainly determined by the location information of the Harris hawk population, and its update strategy is as

**Table 1: Location update strategy in exploitation phase**

Strategy	Location update	Conditions
Soft besiege	$X(t+1) = \Delta X(t) - E JX_{rabbit}(t) - X(t) $ (4) $\Delta X(t) = X_{rabbit}(t) - X(t)$ (5)	$ E  \geq 0.5$ and $r \geq 0.5$
Hard besiege	$X(t+1) = X_{rabbit}(t) - E \Delta X(t) $ (6)	$ E  < 0.5$ and $r \geq 0.5$
Soft besiege with progressive rapid dives	$Y = X_{rabbit}(t) - E JX_{rabbit}(t) - X(t) $ (7) $Z = Y + S \times LF(D)$ (8) $X(t+1) = \begin{cases} Y & \text{if } F(Y) < F(X(t)) \\ Z & \text{if } F(Y) < F(X(t)) \end{cases}$ (9)	$ E  \geq 0.5$ and $r < 0.5$
Hard besiege with progressive rapid dives	$X(t+1) = \begin{cases} Y & \text{if } F(Y) < F(X(t)) \\ Z & \text{if } F(Y) < F(X(t)) \end{cases}$ (10) $Y = X_{rabbit}(t) - E JX_{rabbit}(t) - X_m(t) $ (11) $Z = Y + S \times LF(D)$ (12)	$ E  < 0.5$ and $r < 0.5$

follows:

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 |X_{rand}(t) - 2r_2 X(t)| & q \geq 0.5 \\ (X_{rab}(t) - X_m(t)) - r_3 (LB + r_4 (UB - LB)) & q < 0.5 \end{cases} \quad (1)$$

where  $X(t+1)$  is the position vector of the hawks in the next iteration,  $X_{rab}(t)$  is the position of the prey,  $X(t)$  is the current position vector of the hawks,  $r_1, r_2, r_3, r_4$ , and  $q$  are random numbers inside  $(0,1)$ ,  $UB$  and  $LB$  are the upper and lower bounds of the variables,  $X_{rand}(t)$  is the random individual, and  $X_m(t)$  is the average position of the hawks in the current population. The average position of the hawks is obtained from equation Eq. 2):

$$X_m(t) = \frac{1}{n} \sum_{k=1}^n X_k(t) \quad (2)$$

where  $X_k(t)$  denotes the position of hawk  $k$  in the iteration  $t$ , and  $n$  denotes the number of hawks.

## 2.2 Transition from Exploration to Exploitation

The energy equation controlling the escape of prey is as follows:

$$E = 2E_0(1 - t/T) \quad (3)$$

where  $t$  is the current number of iterations,  $T$  denotes the maximum number of iterations, and the value of  $E_0$  is a random number within  $(-1,1)$  that indicates the initial state of its energy.

## 2.3 Exploitation Phase

where  $\Delta X(t)$  is the distance between the prey and the Harris hawks in iteration  $t$ ,  $r_5$  is a random number inside  $(0,1)$ , and  $J = 2(1 - r_5)$  denotes the distance of random jumps when the prey escapes,  $D$  is the dimension of the problem, and  $LF$  represents the escape behavior of the prey simulated using Levy function.

## 3 MO-EMHHO

### 3.1 Sobol Sequence Initialization Populations

In the basic HHO algorithm, the initialized population is generated by randomization. However, the individuals generated in this way are not uniformly distributed throughout the search space, which in turn affects the speed of convergence and precision of the algorithm. The Sobol sequence [12] is a deterministic low-difference sequence that has the feature of distributing the points in the space

as uniformly as possible compared to the random sequence. The expression for the original population generated by the Sobol sequence can be expressed as:

$$X_i = lb + S_n \times (ub - lb) \quad (4)$$

where  $lb$  and  $ub$  are the lower and upper bounds of the search space, and  $S_n$  is the random number generated by the Sobol sequence,  $S_n \in [0, 1]$ .

### 3.2 Elite Opposition-Based Learning

Opposition-Based Learning (OBL) [13] is a new method of intelligent computing proposed by Tizhoosh in 2005. In recent years, this strategy has been applied to the improvement of various algorithms and has achieved good optimization results. Assume that a feasible solution on a  $d$ -dimensional search space is  $X = (x_1, x_2, \dots, x_d)$  ( $x_j \in [a_j, b_j]$ ), then its opposition-based solution is defined as  $\bar{X} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_d)$ , where  $\bar{x}_j = r(a_j + b_j) - x_j$ ,  $r$  is the coefficient of uniform distribution inside  $[0, 1]$ .

The inverse solution generated by the opposition-based learning strategy is not necessarily search for the global optimal solution more easily than the current search space. To address this problem, elite opposition-based learning (EOBL) is proposed. Assuming that the extreme-point of the current population on the search space as the elite individual  $X_e = (x_1^e, x_2^e, \dots, x_d^e)$ , its inverse solution  $\bar{X}_e = (\bar{x}_1^e, \bar{x}_2^e, \dots, \bar{x}_d^e)$  can be defined as follows:

$$\bar{x}_j^e = k \cdot (a_j + b_j) - x_j^e \quad (5)$$

where  $x_j^e \in [a_j, b_j]$ ,  $k$  is a random value inside  $[0, 1]$ ,  $b_j$  and  $a_j$  are the upper and lower bounds of the dynamic boundary,  $a_j = \min(x_j^e)$ ,  $b_j = \max(x_j^e)$ . Replacing the fixed boundary with a dynamic boundary is beneficial to make the generated inverse solution gradually reduce the search space and speed up the convergence of the algorithm. Since the elite inverse solution may jump out of the boundary and lose its feasibility, the following approach is taken to reset the value.

$$\bar{x}_i^e = rand(a_j, b_j) \quad (6)$$

### 3.3 Escape Energy Update Optimization

In the basic HHO, Harris hawk relies on the energy factor  $E$  to manage the transition of the algorithm from the global search to

the local search phase. However, as shown in Eq. 3), its energy factor  $E$  is reduced from 2 to 1 using a linear update, which tends to trap into a local optimum in the second half of the iteration. To overcome the deficiency of only local search when the algorithm proceeds to the later phase,  $E_1$  as a new update of the energy factor is used:

$$E = 2 \times (1 - (t/T)^{1/3})^{1/3} \quad (7)$$

$$E_1 = E \times (2 \times r - 1) \quad (8)$$

where  $t$  is the current number of iterations,  $T$  is the maximum number of iterations, and  $r$  is the random number inside  $[0, 1]$ .

### 3.4 External archive and its maintenance

In multi-objective optimization algorithms, since the Pareto-optimal solution of the current population is not necessarily non-dominated by the solutions generated before the algorithm, so it is necessary to store the previously obtained non-dominated solutions in a collection of archives set up outside the population. In this paper, a grid partitioning method is used to maintain the external archives to ensure the population diversity. In the grid mechanism, the target space is divided into multiple grid regions, and the distribution density is determined by calculating the number of particles in each grid, and the probability of each grid being chosen is inversely proportional to the distribution density, and grid selection is performed by roulette selection.

**3.4.1 Adaptive grid strategy.** For the multi-objective optimization problem containing  $n$  objectives, a grid region containing  $2n$  boundaries needs to be determined, with the minimum boundary denoted as  $Lb_i$  and the maximum boundary denoted as  $Ub_i$ . To ensure that the particles on the boundary are also within the grid, the boundary needs to be expanded, and the modal representation of the grid is defined as follows:

$$\begin{cases} W_i = \max f_i - \min f_i \\ Lb_i = \min f_i - \alpha \cdot W_i \\ Ub_i = \max f_i + \alpha \cdot W_i \\ M_i = \frac{Ub_i - Lb_i}{D} \end{cases} \quad (9)$$

where  $f_i$  donates the objective function value in dimension  $i$ ,  $\alpha$  indicates the grid expansion factor,  $Ub_i$  and  $Lb_i$  are the upper and lower bounds of the grid space,  $i = 1, 2, \dots, n$ , and  $D$  is the number of divisions of the grid in each dimension.

**3.4.2 Grid selection strategy based on roulette selection.** After dividing the target space into a grid, the grid density is determined by calculating the number of particles in the grid. The smaller the density of the grid, the sparser the particles are, and the greater the probability of individuals being selected, thus ensuring a uniform distribution of particles. For the particles in the external file, the position of the grid in which they are located can be expressed as:

$$\left( \left\lceil \frac{f_1 - Lb_1}{M_1} \right\rceil, \left\lceil \frac{f_2 - Lb_2}{M_2} \right\rceil, \dots, \left\lceil \frac{f_n - Lb_n}{M_n} \right\rceil \right) \quad (10)$$

The particle number determines the position of the grid in which it is located, and thus the probability of each grid being selected, which is calculated and then normalized so that the probability of being selected for all grids sums to 1:

$$p_j = 1/C_j \quad (11)$$

$$P_j = \frac{p_j}{\sum_k p_k} \quad (12)$$

Where  $C_j$  indicates the number of particles in the grid  $j$ ,  $k$  donates the number of grids,  $P_j$  is the normalized probability.

### 3.5 Adaptive mutation strategy

To avoid the algorithm converging too quickly near the local optimum and obtaining the wrong Pareto frontier solution, an appropriate variation strategy can achieve the further exploitation capability of the algorithm, thus improving the algorithm's optimality finding ability and convergence accuracy. Gauss walk learning (GWL) is a classical stochastic walk strategy with strong exploitation capability<sup>[14]</sup>, so this paper uses this strategy to mutate the population individuals to improve the diversity of the population while helping it to leap out of the local optimum trap. In each iteration of the algorithm, there is a probability to select some population individuals for mutation operation, and the mutation rate decreases with the increasing of the number of iterations, as shown in Eq. 13).

$$Mu = \left( 1 - \frac{t-1}{T-1} \right)^{\frac{1}{\beta}} \quad (13)$$

where  $\beta$  is the variation rate factor,  $t$  is the number of current iterations, and  $T$  is the maximum number of iterations.

The Gauss walk learning model is shown in Eq. 15):

$$X(t+1) = \text{Gaussian}(X(t), \tau) \quad (14)$$

$$\tau = \cos(\pi/2 \times (t/T)^2) \times (X(t) - X_r(t)) \quad (15)$$

where  $X(t)$  indicates the individual in the generation population  $t$ ,  $\text{Gaussian}(X(t), \tau)$  is the Gaussian distribution with  $X(t)$  as the expectation and  $\tau$  as the standard deviation, and  $X_r(t)$  is the location of the random individual in the generation population  $t$ .

### 3.6 Algorithm structure of MO-EMHHO

Combining the above improvement strategies, the steps of the MO-EMHHO are as follows.

## 4 EXPERIMENT AND RESULT

### 4.1 Benchmark Functions and Numerical Experiment

To inspect and verify the capability of the MO-EMHHO algorithm proposed in this paper, MOHHO<sup>[9]</sup>, MOPSO<sup>[4]</sup>, and MOGWO<sup>[7]</sup> are selected for comparison experiments, and the same experimental environment, platform and basic parameters are chosen for the experiments. Let the population size  $N$  be 100, the maximum number of iterations  $T$  be 300, and the external archive capacity *Archive* be 100.

The ZDT series<sup>[15]</sup> functions ZDT1 to ZDT4 are selected in this paper to verify the effectiveness of the algorithm, these test functions have different characteristics and the true Pareto front interpretation of the functions is known, so the convergence of different optimization algorithms can be tested. The information of the functions are shown in Table 2.

Inverted generational distance (IGD)<sup>[16]</sup> is a comprehensive performance evaluation metric that evaluates the performance of the

**Table 2: The information of ZDT1-ZDT4**

Function	Information	range
ZDT1	$f_1(x) = x_1, f_2(x) = g(x)(1 - \sqrt{x_1/g(x)})$ $\{g(x) = 1 + 9 \sum_{i=2}^n xi/(n-1)$	[0,1]
ZDT2	$f_1(x) = x_1, f_2(x) = g(x)(1 - (x_1/g(x))^2)$ $\{g(x) = 1 + 9 \sum_{i=2}^n xi/(n-1)$	[0,1]
ZDT3	$f_1(x) = x_1, f_2(x) = g(x)[1 - \sqrt{\frac{x_1}{g(x)}} - \frac{x_1}{g(x)} \sin(10\pi x_1)]$ $\{g(x) = 1 + 9 \sum_{i=2}^n xi/(n-1)$	[0,1]
ZDT4	$f_1(x) = x_1, f_2(x) = g(x)[1 - \sqrt{\frac{f_1(x)}{g(x)}}]$ $\{g(x) = 1 + 10(n-1) + \sum_{i=2}^n (x_i^2 - 10 \cos(4\pi x_i))$	$x_1 \in [0, 1]$ $x_i \in [-5, 5]$

**Table 3: Test results of IGD**

Function	Item	MO-EMHHO	MOPSO	MOHHO	MOGWO
ZDT1	Ave	<b>4.83E-04</b>	1.98E-03	8.32E-04	2.82E-03
	Best	2.88E-04	2.98E-04	2.90E-04	<b>2.35E-04</b>
	Std	<b>8.77E-05</b>	2.61E-03	1.76E-04	1.25E-03
ZDT2	Ave	<b>3.46E-04</b>	4.32E-02	3.68E-03	1.92E-02
	Best	<b>2.91E-04</b>	3.01E-04	2.94E-04	2.26E-04
	Std	<b>7.50E-05</b>	2.66E-02	7.75E-03	7.97E-03
ZDT3	Ave	<b>1.37E-03</b>	2.41E-02	1.84E-03	1.17E-02
	Best	<b>3.01E-04</b>	7.66E-04	5.05E-04	3.87E-04
	Std	<b>1.76E-03</b>	3.49E-02	3.71E-03	2.96E-02
ZDT4	Ave	<b>4.56E-04</b>	9.81E-01	1.63E-03	7.49E-02
	Best	2.75E-04	3.14E-01	2.80E-04	<b>2.38E-04</b>
	Std	<b>9.40E-05</b>	7.61E-01	5.07E-03	1.91E-01

convergence and distribution of the algorithm by calculating the minimum sum of distances between the obtained Pareto front solution and the true Pareto front, The smaller the calculation result, the better the overall performance of the algorithm. The calculation formula is defined as follows.

$$IGD = \sum_{i=1}^N d_i / N \quad (16)$$

$$d_i = \sqrt{\sum_{j=1}^m (f_j^i - f_j^{true})^2} \quad (17)$$

where  $N$  indicates the number of Pareto optimal solutions obtained,  $m$  is the number of objectives in the test problem,  $f_j^i$  indicates the  $j$ th objective value of the  $i$ th solution obtained, and  $f_j^{true}$  is the nearest true Pareto frontier solution to  $f_j^i$ .

## 4.2 Results and Analysis

The dimensionality of the test function set in the experiment is 30, and the algorithms were run 30 times independently on each function to prevent chance from bringing bias to the experimental results, and the mean, best value, and standard deviation of

the results of each algorithm run were displayed, and the IGD results of each algorithm on benchmark functions were obtained by calculation as shown in Table 2.

As can be seen from Table 2, the MO-EMHHO is slightly behind MOGWO in terms of optimal values on ZDT1 and ZDT4 but achieves the optimal IGD mean values on the test functions ZDT1 to ZDT4. On ZDT2 and ZDT4, the MO-EMHHO algorithm has an order of magnitude improvement compared to the other algorithms, indicating that its convergence is significantly better than the comparison algorithms. In terms of stability, the standard deviation of the MO-EMHHO algorithm is also significantly smaller than that of the other comparison algorithms, indicating that the algorithm has excellent stability.

To show the uniformity and convergence of the algorithms more intuitively, Figure 1 shows the iteration curves of IGD on the test functions for each algorithm. Figure 2~Figure 5 shows the optimal solution set of each algorithm with the true Pareto front distribution. From the convergence curves shown in Figure 1, it can be seen that MOPSO and MOGWO converge slowly on the ZDT1 function and significantly lag behind MOHHO and MO-EMHHO in terms of accuracy. The MOHHO algorithm converges to the optimum after 50 iterations but shows large fluctuations in the later stages, while

**Algorithm 1** MO-EMHHO

---

Inputs: Population size  $N$ , the maximum number of iterations  $T$ , decision vector dimension  $dim$ , upper and lower bounds  $Ub$  and  $Lb$ , external file capacity  $R$ , number of grid divisions  $G$

According to Eq.(16) initialize the population

Determine the non-dominated solution and initialize *Archive*

While( $t < T$ )

Generate the reverse population using the elite opposition-based learning mechanism and calculate the fitness of the original population and its reverse population individuals

for  $i=1:N$

According to Eq.(18) update the escape energy  $E$

if  $|E| \geq 1$

According to Eq.(1) update the location

end if

else if  $|E| < 1$

If  $|E| \geq 0.5$  and  $r \geq 0.5$

According to Eq.(4) update the location

end if

else if  $|E| < 0.5$  and  $r \geq 0.5$

According to Eq.(6) update the location

end if

else if  $|E| \geq 0.5$  and  $r < 0.5$

According to Eq.(10) update the location

end if

Else if  $|E| < 0.5$  and  $r < 0.5$

According to Eq.(11) update the location

end if

end if

If mutation occurs

According to Eq.(23) update the location

end if

end for

Determine non-dominated solutions and update *Archive*

if *Archive* is full

Use roulette selection to eliminate redundant solutions from the archive grid

end if

Detects boundaries to prevent individuals from being located beyond the scope of the solution

end While

Return *Archive*

---

MO-EMHHO gradually smoothes out and has the best accuracy after 10 iterations. The convergence curves as shown in ZDT2, ZDT3, and ZDT4 show more intuitively that the MO-EMHHO algorithm has obvious advantages in convergence accuracy and speed. From Figure 2 to Figure 5, it can be seen that the MOPSO algorithm is easily trap into the local optimum in each test function, the MOHHO algorithm can converge to the true Pareto frontier but its solution set is not uniformly distributed, the distribution of the MOGWO algorithm is limited to a part of the true frontier, the MO-EMHHO algorithm shows the best Pareto frontier in the benchmark function, and the rest of the compared algorithms have a large gap with the improved algorithm proposed in this paper.

**5 CONCLUSION**

A multi-strategy improved multi-objective Harris hawk optimization algorithm (MO-EMHHO) with elite opposition-based learning is proposed to address the shortcomings of low algorithm accuracy, slow rate of convergence, and easily fall into local optima in solving multi-objective optimization problems. The experiments are compared with several other commonly used multi-objective optimization algorithms on standard test functions by three aspects: distribution of optimal solution sets, the convergence speed of the algorithm, and evaluation index values, experimental results show that the performance of MO-EMHHO proposed in this paper surpasses that of other comparative algorithms and provides a new solution to the multi-objective optimization problem and contributes a new scheme for solving multi-objective optimization problems.

**ACKNOWLEDGMENTS**

The authors would like to acknowledge the National Natural Science Foundation of China(61772031), and the Natural Science Foundation of Hunan Province(2020JJ4753).

**REFERENCES**

- [1] Deb K, Agrawal S, Pratap A, et al. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II[J]. Lecture notes in computer science, 2000: 849-858.
- [2] Deb K, Jain H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints[J]. IEEE Transactions on Evolutionary Computation, 2014, 18(4):577-601.
- [3] Xue F, Sanderson A C, Graves R J. Pareto-based multi-objective differential evolution[C]//Congress on Evolutionary Computation, 2003. CEC'03. IEEE, 2003, 2: 862-869.
- [4] Coello C, Lechuga M S. MOPSO: A proposal for multiple objective particle swarm optimization[C]// Congress on Evolutionary Computation. IEEE Service Center, 2002.
- [5] Gong M, Jiao L, Du H, et al. Multiobjective immune algorithm with nondominated neighbor-based selection[J]. Evolutionary computation, 2008, 16(2): 225-255.
- [6] Chaharsooghi S K. An intelligent multi-colony multi-objective ant colony optimization for the 0-1 knapsack problem[J]. Proc. of IEEE/CEC, 2008, 2008.
- [7] He X S, Li N, Yang X S, et al. Multi-objective Cuckoo Search Algorithm[J]. Journal of System Simulation, 2015,27(04):731-737.
- [8] Heidari A A, Mirjalili S, Faris H, et al. Harris hawks optimization: Algorithm and applications[J]. Future generation computer systems, 2019, 97: 849-872.
- [9] Yüzge U, Kusoglu M. Multi-objective harris hawks optimizer for multiobjective optimization problems[J]. BSEU Journal of Engineering Research and Technology, 2020, 1(1): 31-41.
- [10] Selim A, Kamel S, Alghamdi A S, et al. Optimal placement of DGs in distribution system using an improved harris hawks optimizer based on single- and multi-objective approaches[J]. IEEE Access, 2020, 8: 52815-52829.
- [11] Devarapalli R, Bhattacharyya B. Optimal parameter tuning of power oscillation damper by MHHO algorithm[C]//2019 20th International conference on intelligent system application to power systems (ISAP). IEEE, 2019: 1-7.
- [12] Bratley P, Fox B L. Implementing sobols quasirandom sequence generator (algorithm 659)[J]. ACM Transactions on Mathematical Software, 2003, 29(1): 49-57.
- [13] Tizhoosh, H. R. Opposition-Based Learning: A New Scheme for Machine Intelligence[C]// International Conference on International Conference on Computational Intelligence for Modelling, Control & Automation. IEEE, 2005:695-701.
- [14] Peng H, Zeng Z, Deng C, et al. Multi-strategy serial cuckoo search algorithm for global optimization[J]. Knowledge-Based Systems, 2021, 214: 106729.
- [15] Zhang Q, Zhou A, Zhao S, et al. Multiobjective optimization test instances for the CEC 2009 special session and competition[J]. University of Essex, Colchester, UK and Nanyang technological University, Singapore, special session on performance assessment of multi-objective optimization algorithms, technical report, 2008, 264: 1-30.
- [16] Yanan Sun and Gary G. Yen and Zhang Yi 0001. IGD Indicator-Based Evolutionary Algorithm for Many-Objective Optimization Problems[J]. IEEE Trans. Evolutionary Computation, 2019, 23(2): 173-187.

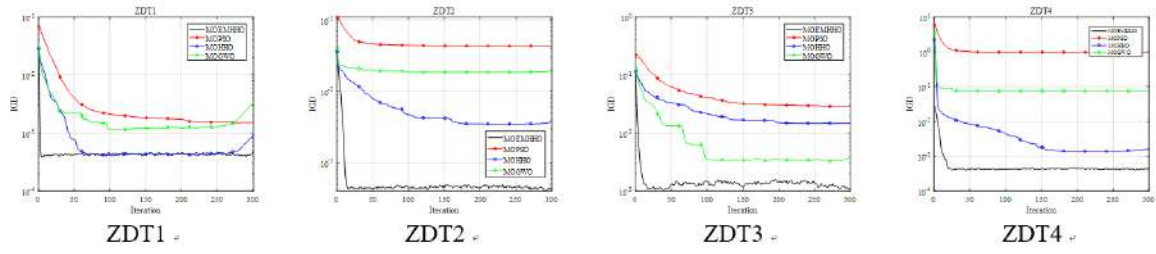


Figure 1: IGD convergence curves of each algorithm

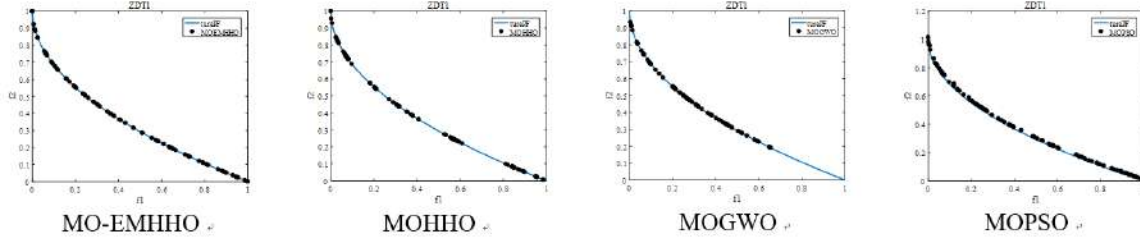


Figure 2: The optimal set of solutions for each algorithm of the ZDT1 function and the real PF frontier distribution

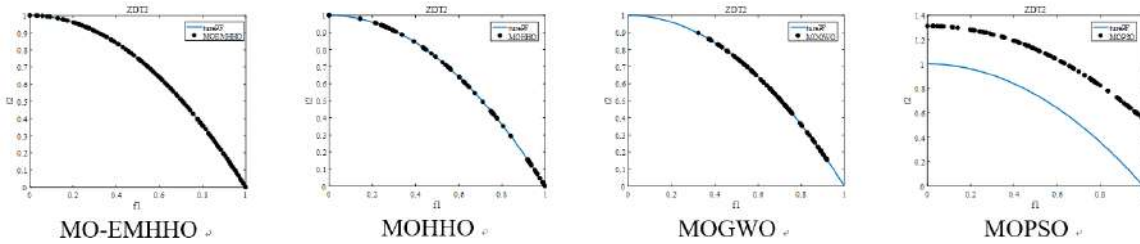


Figure 3: The optimal set of solutions for each algorithm of the ZDT2 function and the real PF frontier distribution

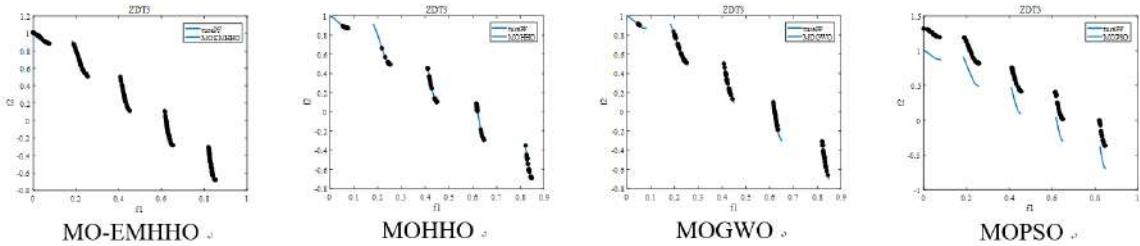


Figure 4: The optimal set of solutions for each algorithm of the ZDT3 function and the real PF frontier distribution

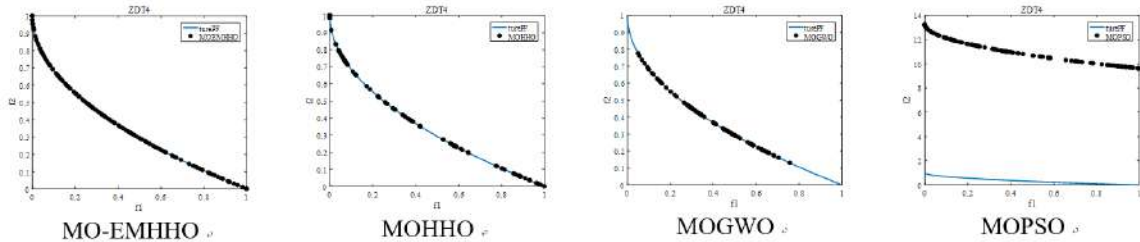


Figure 5: The optimal set of solutions for each algorithm of the ZDT4 function and the real PF frontier distribution

# Elastic Detection Mechanism Aimed at Hybrid DDoS Attack

Yubo Wang

wyb@mail.ustc.edu.cn

University of Science and Technology of China

Hefei, China

Anhui Province Key Laboratory of Cyberspace Security

Situation Awareness and Evaluation

Hefei, China

Jinyu Wang

wangjinyubu@bupt.edu.cn

Beijing University of Posts and Telecommunications

Beijing, China

## ABSTRACT

In Distributed Denial of Service(DDoS) attack, the attacker uses a remotely controlled botnet to attack the target server at the same time to prevent legitimate users from obtaining information services. Previous studies focused on the detection of DDoS attacks on offline datasets, but ignored the detection of specific DDoS types, and some reports showed that the number of DDoS hybrid attacks was increasing significantly. In this paper, we propose an elastic detection mechanism(EDM), which can economize the server's idle computing power. The framework integrates a variety of pre-trained lightweight CNN detect models, which are suitable for on-line rapid detection of DDoS hybrid attacks. We focus on evaluating the response accuracy and the detection speed of the EDM. The experimental results show that the model can achieve excellent hybrid attack detection performance, and meet the actual requirements of real-time detection.

## CCS CONCEPTS

• **Networks** → *Network monitoring*; • **Computing methodologies** → *Artificial intelligence*; • **Security and privacy** → **Denial-of-service attacks**.

## KEYWORDS

DDoS hybrid attack, DDoS detection, malicious traffic analysis

### ACM Reference Format:

Yubo Wang and Jinyu Wang. 2023. Elastic Detection Mechanism Aimed at Hybrid DDoS Attack. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590031>

## 1 INTRODUCTION

DDoS attack has a long history and has become one of the most threatening network attack means on the Internet, with complex and changeable attack methods, easy to achieve attack threshold, low attack cost and difficult to trace. In recent years, the number of DDoS attacks has shown a rapid growth trend, and the attack

methods are also evolving<sup>[2, 9]</sup>, and many new types of DDoS attacks have emerged. DDoS attacks are divided into many types<sup>[8]</sup>. They often aim at various defects of different network protocols or application services, and then saturate some key resources of the target, such as memory space, computing power, bandwidth, network connectivity, etc.

Reflective DDoS attacks, which have become more popular recently(shown in figure 1), are reportedly<sup>[2]</sup> becoming steadily preferred by attackers due to their advantages of successfully hiding the attackers' locations and having a noticeable magnification impact. However, SYN flooding, UDP flooding and ACK flooding are still the significant means. There are also many open source related tools and datasets on the network.

The security community has already done numerous research on DDoS detection. Numbers of precious traffic datasets for enterprise network, cloud server, Internet of Things and other scenarios were collected. These studies had achieved satisfactory detection accuracy with the help of machine learning. However, it is usually achieved on offline datasets to determine whether there is a DDoS attack, making it challenging to identify the precise attack type and subsequently implement tailored protection.

This study focuses on the low-density DDoS attack scenarios that small-sized Web servers often confront. We concentrate on the detection of five common attack types, such as SYN flooding, UDP flooding, ACK flooding, ICMP flooding, and HTTP Get flooding.

The rest of paper is organized as follows: Section2 reviews and discusses the relevant research work in this field, Section3 describes the elastic detection mechanism and the detect model for DDoS hybrid attacks in detail. Section4 introduces the relevant environment contents, then gives conclusions. Section5 explores some constraints faced by this work. Section6 summarizes the whole work.

## 2 RELATED WORKS

To cope with the challenge of DDoS attacks on network services, many related work of DDoS detection technology has been proposed. This section briefly reviews these efforts.

### 2.1 Statistical Analysis Based DDoS Detection

Traditional methods such as statistical analysis(SA) and information theory(IT) used to be the mainstream for the DDoS detection.

In [4], the authors suggested MULTOPS, a tree structure that records traffic rate characteristics, with the goal of detecting bandwidth attacks. MULTOPS identifies subnet with aberrant rate of data packets of incoming and outgoing victims as attack sources.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590031>

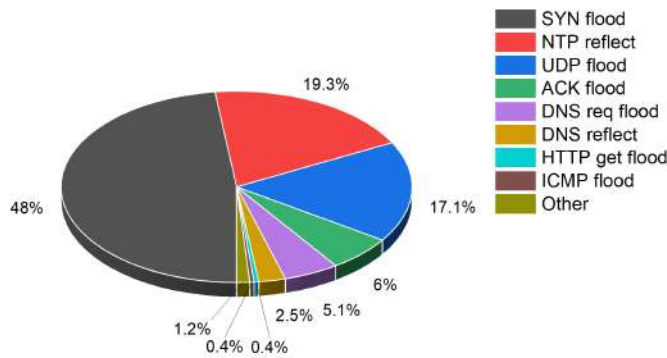


Figure 1: Frequency distribution of DDoS attacks in 2021.

However, it is incapable of stopping DDoS attacks with random source IP addresses. In [6], the authors developed a dynamic entropy model (DEM) after analyzing the network's activity state using the request flow and response flow. Relevant experiments show that the DEM has a better detection rate on the datasets, which contains several types of DDoS attacks, when compared to the conventional static entropy model.

The difficulty in using these conventional statistical analysis-based methodologies comes from the necessity to precisely characterize the profile of legitimate user activity. When an attacker has controlled over a sizable botnet, simulating the statistical behavior of legal users is effortless.

## 2.2 Machine Learning Based DDoS Detection

Machine learning (ML) technology has also injected new vitality into the abnormal traffic detection. The malicious traffic detection can be reduced to a classical dichotomous task. Many DDoS detection methods based on support vector machine (SVM), k-Means, decision tree (DT), etc., have achieved decent results in some scenarios.

In [1], a semi-supervised ML approach was proposed to detect DDoS attacks. At the phase of labelling, the data is clustered by two different algorithms and a voting method to decide the final label of traffic flows. This core idea is beneficial to solve the problem of high false positives caused by the single clustering model. At the phase of classification, three classic ML models (i.e. kNN, SVM and Random Forest) are applied on labeled data to classify DDoS attacks. Relevant results show that the RF model can get a highest accuracy score in a dataset which obtained with the network traffic generated in OPNET Modeler simulator. The study [7] proposed a feature selection based DDoS attacks detection framework which combines the information gain (IG) and correlation (CR) techniques with threshold mechanism. The IG and CR technique are based on entropy and Pearson correlation coefficient (PCC) respectively, which can obtain scores for each feature of datasets. The higher score value associated with the feature implies more valueable. The system is trained with J48 classifier using the reduced features from the original feature set. The results show that the proposed system outperforms at performance in terms of accuracy, detection rate and FAR (false alarm rate) compared to the relevant existing feature selection methods.

However, complex and time-consuming feature engineering often causes confusion to users, and ML model has limited capacity to capture spatial and temporal features from massive regular traffic.

## 2.3 Deep Learning Based DDoS Detection

In the past decade, deep learning (DL) technology has developed rapidly. Its features include great fitting ability and high detection accuracy. It has been extensively used in various disciplines, including computer vision, natural language processing, and recommendation systems. Naturally, DL based DDoS detection technology is the most cutting-edge trend.

In [5], the author combined auto encoder (AE) and logistic regression classifier for intrusion detection. AE was used to achieve feature extraction and compression. In both binary and multi-class classification tasks, this approach performs well. In [11], the author proposed a model, named DeepDefense, and designed a training sample generation method. Extracted 20 features from the ISCX2012 datasets for training. The experimental results showed obvious advantages in multiple indicators compared with other ML models such as random forest (RF). In [3], the author proposed a lightweight convolution neural network (CNN) model, named LUCID. At the phase of preprocessing, a new data structure is adopted to ensure the model architecture's simplicity. The adjustment of the model parameters uses grid search technology to ensure the good adaptability of the model. Compared with LSTM and other models, the LUCID has better detection performance and faster detection speed in CICIDS2017, CSECIC2018 and other datasets.

Compared with other methods, various detection schemes based on DL have the best detection accuracy and are favored by the research community. Unfortunately, the distribution of attack behaviors in the current DDoS datasets is unbalanced, which inhibits the detect model's portability. A comprehensive detect model can be trained on the datasets containing multiple types of DDoS attacks, but it cannot detect new attack type and must be retrained, which weakens the practicability. Therefore, ultimate detection accuracy is not the primary focus of this paper. The original intention of this study is to redesign DDoS attack detection schemes and improve their portability and practicability.

## 2.4 Multiple Types of DDoS Detection

DDoS hybrid attack refers to a single attack event that contains multiple types of attack, such as large-capacity flooding, application-layer attacks, and connection exhaustion attacks. The attacker can flexibly combine the attack events according to the specific conditions of the target system and launch multiple attack types at the same time, exploiting the defects of the protocol and system to obtain maximum attack benefits at the lowest attack cost.

In [12], the author designed a packet threshold algorithm (PTA) to detect SYN flood, UDP flood, ICMP flood and Smurf DDoS attacks in sequence. Then, DDoS attack types are detected based on different packet rate thresholds. However, setting an appropriate threshold is challenging. In [10], the author trained a DT model and divided system traffic into four categories: normal, SYN flooding, UDP flooding and ICMP flooding.

In brief, a CNN model is adopted in our work to ensure the detection accuracy of malicious traffic. With the main objectives

**Table 1: Comparison of some DDoS detection works.**

Methods	Work	Main techniques	Advantage
SA	[4]	statistical analysis	scheme deployment is simple
	[6]	dynamic entropy analysis	possess effective diagnostic power in detecting anomalies
ML	[1]	ensemble learning	unlabeled traffic can be used for detection
	[7]	feature reduction, J48 model	the worthless dataset's features are trimmed
DL	[5]	auto encoder model	the compression of features increases efficiency and accuracy
	[11]	LSTM model's variants	high accuracy
	[3]	CNN model	fast, high accuracy

of saving server idle computing overhead, real-time and hybrid attacks detection. Related content is detailed in Section 3.

### 3 PROPOSED SYSTEM

#### 3.1 Concept of Network Flow

The interaction of network terminal nodes generates traffic. Multiple network services run on server's ports, and the port reuse is often not permitted. Diverse network protocols are utilized based on the real service demands, enhancing the user experience. Limited by transmission bandwidth and other factors, flow is divided into many packets with a certain sequence. Formally, the flow is described in terms of a quintuples:

$$\langle Srcip, Srcport, Dstip, Dstport, Protocol \rangle \quad (1)$$

$$\langle Dstip, Dstport, Srcip, Srcport, Protocol \rangle \quad (2)$$

All packets matching a quintuples belong to the same flow. Flows are bidirectional, it is generally assumed that the foregoing two quintuples belong to the same flow. In the classic malicious traffic detection task, all flows are assigned label '0' or '1'. Generally, label '0' represents benign flow and label '1' represents malicious flow.

#### 3.2 EDM's Architecture and Workflow

The elastic detection mechanism aimed for DDoS hybrid attacks proposed by us is shown in figure 2, which specifically includes the system state anomaly percept module and real-time traffic monitor module, respectively abbreviated as **percept module** and **monitor module** later. The monitor module contains a pre-trained models library for various attack types.

The workflow of EDM in the server mainly includes:

- (1) The percept module periodically checks network bandwidth and server performance related parameters. If these parameters are normal, the module will do nothing and continue running.
- (2) If the percept module finds that the parameters are abnormal, then several possible DDoS attack types set (marked as S) will be inferred and transmitted to a new process, which runs the monitor module.
- (3) Every time the monitor module is called by the new process, it will run for a period of T. It invokes the relevant pre-trained models (marked as M) in S. Every model in M detects malicious network flows in real-time.

- (4) If any detect model predicts the occurrence of DDoS attack, the system will log the attack information and warn rapidly. If no DDoS attacks are detected across the period, the monitor module will be turned off.

Obviously, the server deployed EDM can theoretically realize the efficient utilization of computing and storage resources, without the need to continuously run complex detection algorithms. The pre-training model library in the monitor module can be expanded according to the demand. Each module's implementation details are described here.

#### 3.3 Percept Module

Eight monitored objects are chosen after analyzing the characteristics of the five DDoS attack types: network bandwidth, CPU usage, memory usage, UDP rate, ICMP rate, TCP rate with ack-flag, TCP rate with syn-flag, and HTTP rate. Many common shell commands provide access to these system state information. The range of system state values is strongly related to the different types of DDoS attacks. We train a Gaussian mixture model (GMM) to establish implicit mapping relation. Assuming that all network flows are normal except for 5 types of DDoS attacks, thus the number of mixed models  $K = 6$  is used in this work. The GMM's mathematical formula is as follows:

$$P(\vec{X}) = \sum_{k=1}^K \alpha_k * N(\mu_k, \sigma_k), \quad s.t. \sum_{k=1}^K \alpha_k = 1 \quad (3)$$

In equation 3, there are  $K$  normal distribution models with different expectation and variance. An 8-dimensional vector  $\vec{X}$  of the system state is taken as input and a 6-dimensional vector  $\vec{Y}$  is output, representing the probability distribution of prediction for 6 network behaviors. If GMM's prediction category for the new sample is not 'Normal', then three attack types with the largest probability value in vector  $\vec{Y}$  will be selected as outputs, representing the inferred set of potential attack types S.

$$S = \{ 'attack\_type1', 'attack\_type2', 'attack\_type3' \} \quad (4)$$

#### 3.4 Monitor Module

The monitor module activates several processes and loads the pre-trained detect model corresponding to the attack type in S. It monitors network ports in real-time to provide more accurate flow-level analysis. In this paper, all pre-training models uniformly adopt a relatively mature CNN architecture [3], whose main advantage is

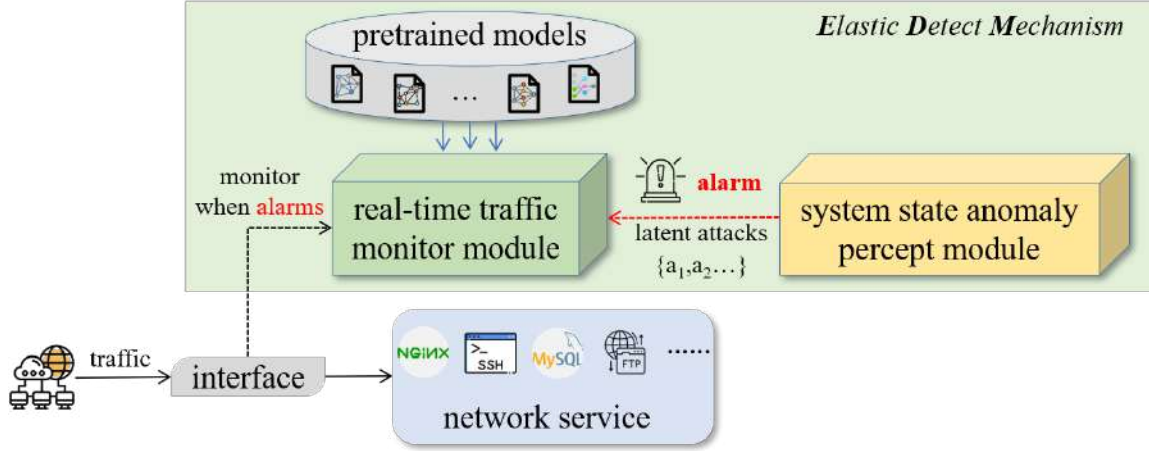


Figure 2: EDM's architecture.

its simple structure, which allows it to predict the real-time traffic behavior quickly.

**3.4.1 Traffic preprocessing.** Real-time traffic at network ports must be stored as samples with a regular data structure in order to be converted into usable data for model training or prediction (as shown in figure 3).

Each flow is divided into multiple samples based on the adjacent time slice  $T$ , with a maximum of  $N$  packets in each sample. If the length of a sample exceeds  $N$ , subsequent packets will be discarded. Otherwise, the vacancy is padded with 0. This data structure ensures that time and space dimension information is recorded in an orderly manner.

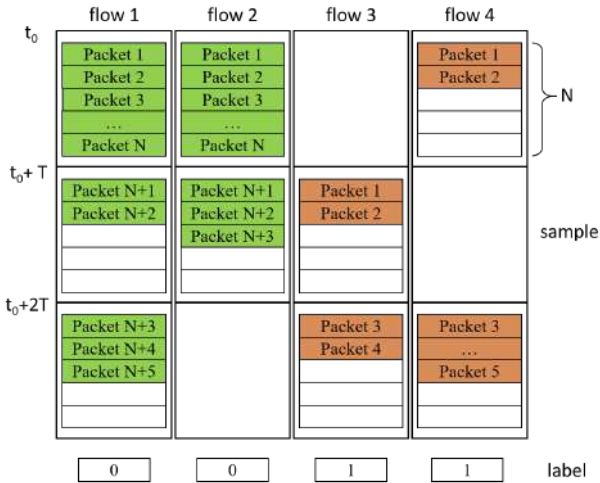


Figure 3: Data structure of flow and sample.

11 packet attributes are extracted, including the timestamp, packet length, highest protocol, IP flags, TCP flags, TCP windows size, TCP length, UDP length, and ICMP type etc. All features are normalized, and meaningless features should be set to 0. All samples generated

by the flow share the same label, which can be obtained from the dataset. The pseudo-code for the preprocessing process is shown by algorithm 1.

---

**Algorithm 1:** Traffic preprocessing algorithm

---

**Input:** Traffic dataset( $\mathcal{D}$ ), flow-level labels( $\mathcal{L}$ ), time slice length( $T$ ), max packets per sample( $N$ )

**Output:** Set of labelled samples( $S$ )

```

1 Function TrafficPreprocess( $\mathcal{D}, \mathcal{L}, T, N$ ):
2    $S \leftarrow \emptyset$ 
3    $\tau \leftarrow -1$ 
4   foreach  $pkt \in \mathcal{D}$  do
5      $id \leftarrow pkt.id$ 
6     if  $\tau == -1$  or  $pkt.time > \tau + T$  then
7        $\tau \leftarrow pkt.time$ 
8     end
9     if  $S[\tau, id].length < N$  then
10       $S[\tau, id].append(pkt.features)$ 
11    end
12  end
13  foreach  $s \in S$  do
14     $s.label = \mathcal{L}[s.id]$ 
15  end
16   $S \leftarrow \text{NormalizeAndPadding}(S, N)$ 
17  return  $S$ 
18 end

```

---

**3.4.2 CNN-based detect model.** The labelled sample set is used to be trained or predicted by a CNN-based detect model (shown in figure 4). Each sample sequentially undergoes 1D convolution layer, relu activation, maximum pooling layer, full connection layer, and sigmoid activation to obtain the final prediction result.

$S_i$  represents a random sample with label, whose shape is  $[N, m]$ .  $m = 11$  represents feature dimensions of packets.  $k$  represents the number of CNN kernels.  $kernel_i$ 's shape is  $[h, m]$ .

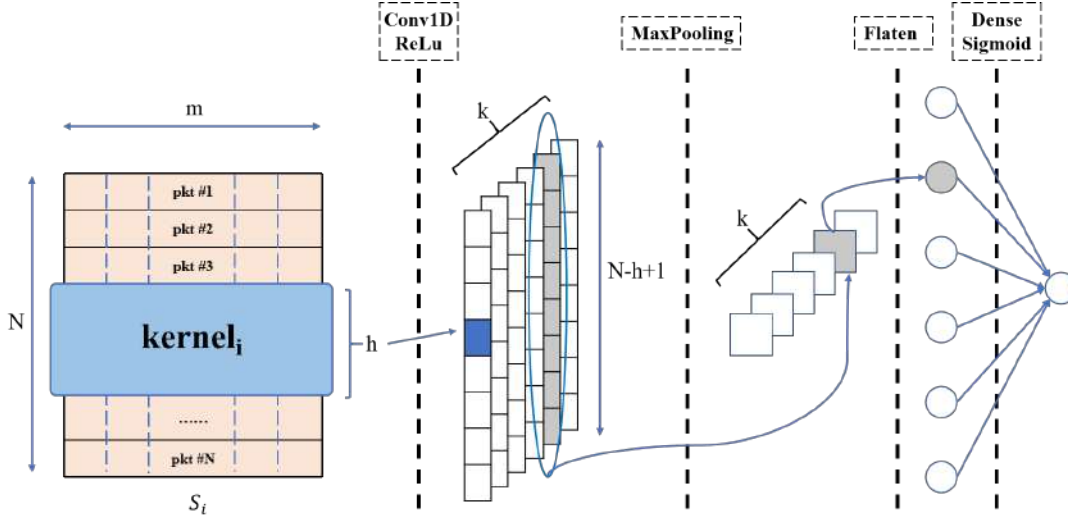


Figure 4: Structure of CNN-based detect model.

The detection model's final layer produces a scalar with values ranging from 0 to 1. If the value is less than or equal to 0.5, the sample is considered normal; otherwise, the sample is considered malicious. Because the traffic characteristics generated by various DDoS attacks differ greatly, different models are trained for different malicious behavior data sets. In addition to the CNN model discussed in this paper, the model library can also expand more pre-trained detection algorithm.

## 4 EXPERIMENTAL EVALUATION

### 4.1 Dataset

Open source datasets contain raw traffic files in PCAP format, providing all the characteristics for user training detection models. We train SYN flooding, UDP flooding, and HTTP Get flooding attack detection models using several datasets that have become popular in recent years. Some practical tools are used to generate datasets for two other types of attacks in the LAN, called ACK2022 and ICMP2022 respectively. All datasets are divided into the training set, verification set and test set in a ratio of 6:2:2, relevant information is shown in table 2.

### 4.2 Environments

We developed the EDM and trained the model on a Ubuntu20-based server, all work realized by the python3. In addition, cloud server open certain Web services to support remote user access to generate more realistic traffic behaviors. Configuration information is shown in table 3.

In the experiment, we integrate a DDoS hybrid attack script to allow for the simultaneous generation of traffic from multiple attack types. It can not only change the intensity of the attack, but also forge packets with any IP source address(besides the HTTP flood attack). We install the script on three hosts to generate hybrid attack flows and use another five hosts to access the web server normally to provide normal traffic behaviors. In this work, all of

the hyper-parameters' values are:

$$K = 8, T = 10, N = 15, m = 11, h = 3, kernelstep = 1, k = 64$$

### 4.3 Results

We test the warning rate of EDM's percept module against multiple DDoS hybrid attacks at different attack rates(AR, the unit is packets/seconds). The warning rate is defined as the GMM model successfully sensing and alarming of every attack types. Related results are shown in table 4. For convenience and illustrative purposes, 'I' denotes ICMP flooding attack, 'U' denotes UDP flooding attack, 'S' denotes SYN flooding attack, 'A' denotes ACK flooding attack, 'H' denotes HTTP flooding attack, '+' denotes hybrid attack.

We fix the attack rate value at 50, the PTA-SVM model<sup>[12]</sup> was compared to measure the detection response time of EDM against multiple DDoS hybrid attacks. The average value of multiple tests was taken, and the results were shown in figure 5.

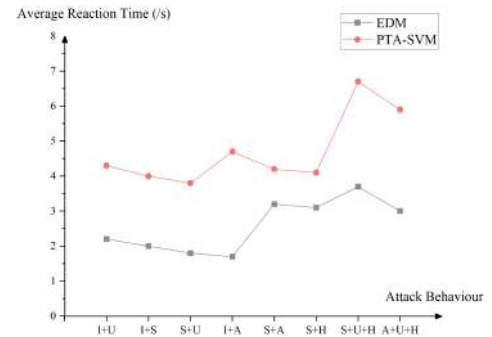


Figure 5: Average reaction time in different methods.

The above results intuitively reveal that the increase of attack rate has a significant positive impact on the successful alarm of

**Table 2: Details of the datasets used.**

Datasets	Traffic Slice Used	Attack Type	Extra Tool
CICDDoS2017	Jul. 7th 15:56-16:16	HTTP flood	editcap
CIC2018	Feb. 20th 10:12-11:17, 13:13-13:32	HTTP flood, UDP flood	editcap
CIC2019	Mar. 11th 10:53-11:03, 11:28-12:00	UDP flood, SYN flood	editcap
ACK2022	-	ACK flood	tcpdump, hping3
ICMP2022	-	ICMP flood	http_load

**Table 3: Related configuration information.**

Hardware	Service Architecture	Web Service	Layer7 Protocol
intel i7 1060	apache	phpcms	HTTP
32G RAM	mysql	wordpress	SSH
RTX 2060	php7		FTP

**Table 4: EDM’s warning rates with different AR.**

Attack Behavior	Warning Rate	
	AR=50	AR=100
I+U	90.4%	98.4%
S+I	96.0%	99.0%
S+U	91.5%	98.3%
A+I	95.7%	99.1%
S+A	96.4%	99.6%
S+H	87.7%	97.8%
S+U+H	87.0%	99.0%
A+U+H	89.0%	98.5%

the precept module. In addition, the TCP three-way handshake is closely related to any HTTP access. Therefore, in the case of massive HTTP requests, SYN flooding and ACK flooding targeting TCP connection resources bring great challenges to the alarm rate and system response time.

## 5 DISCUSSION

There are some limitations to our current work. Firstly, only five types of DDoS attack are focused in EDM framework at present, so the system can be further improved. Secondly, this work is only applicable to the detection and verification of small traffic DDoS attack scenarios. Thirdly, in the percept module, system states need to be selected elaborately when new DDoS attack type are extended, which relies on specific expert experience. Finally, the EDM can only be deployed in linux-based operating system, because the percept module runs on some bash command.

## 6 CONCLUSION

DDoS attacks make a long-term, hidden impact on Internet services all over the world. In this work, an elastic detection mechanism is proposed to cope with the challenge of DDoS hybrid attacks. The percept module checks the status of the server with a low cost, and the monitor module integrates the pre-training model library

to accurately classify real-time traffic behaviors. The feasibility of the scheme is verified by relevant experiments, and it has decent warning rate and response speed.

## ACKNOWLEDGMENTS

This work is supported by the Open Fund of Anhui Province Key Laboratory of Cyberspace Security Situation Awareness and Evaluation, under grant CSSAE-2021-003.

## REFERENCES

- [1] Muhammad Aamir and Syed Mustafa Ali Zaidi. 2021. Clustering based semi-supervised machine learning for DDoS attack classification. *Journal of King Saud University-Computer and Information Sciences* 33, 4 (2021), 436–446.
- [2] Alibaba Cloud. 2021. DDoS offense-defense situation observation in 2020-2021.
- [3] Roberto Doriguzzi-Corin, Stuart Millar, Sandra Scott-Hayward, Jesus Martinez-del Rincon, and Domenico Siracusa. 2020. LUCID: A practical, lightweight deep learning solution for DDoS attack detection. *IEEE Transactions on Network and Service Management* 17, 2 (2020), 876–889.
- [4] Thomer M Gil and Massimiliano Poletto. 2001. MULTOPS: A Data-Structure for Bandwidth Attack Detection.. In *USENIX security symposium*. 23–38.
- [5] Ahmad Javaid, Quamar Niyaz, Weiqing Sun, and Mansoor Alam. 2016. A deep learning approach for network intrusion detection system. In *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*. 21–26.
- [6] Zhu Jian-Qi, Fu Feng, Yin Ke-Xin, and Liu Yan-Heng. 2013. Dynamic entropy based DoS attack detection method. *Computers & Electrical Engineering* 39, 7 (2013), 2243–2251.
- [7] Deepak Kshirsagar and Sandeep Kumar. 2022. A feature reduction based reflected and exploited DDoS attacks detection system. *Journal of Ambient Intelligence*

- and Humanized Computing* (2022), 1–13.
- [8] Jelena Mirkovic and Peter Reiher. 2004. A taxonomy of DDoS attack and DDoS defense mechanisms. *ACM SIGCOMM Computer Communication Review* 2 (2004), 39–53.
- [9] Tencent NSFOCUS. 2022. Global DDoS threat report in 2021.
- [10] Yi-Chi Wu, Huei-Ru Tseng, Wu Yang, and Rong-Hong Jan. 2011. DDoS detection and traceback with decision tree and grey relational analysis. *International Journal of Ad Hoc and Ubiquitous Computing* 7, 2 (2011), 121–136.
- [11] Xiaoyong Yuan, Chuanhuang Li, and Xiaolin Li. 2017. DeepDefense: identifying DDoS attack via deep learning. In *2017 IEEE international conference on smart computing (SMARTCOMP)*. IEEE, 1–8.
- [12] Mohd Azahari Mohd Yusof, Fakariah Hani Mohd Ali, and Mohamad Yusof Darus. 2018. Detection and defense algorithms of different types of DDoS attacks using machine learning. In *Computational Science and Technology: 4th ICCST 2017, Kuala Lumpur, Malaysia, 29–30 November, 2017*. Springer, 370–379.

# Detecting Arbitrary-oriented Objects in Remote Sensing Imagery with Segmentation-Aware Mask

Jiali Wei  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
wei\_jiali0821@163.com

Bo Hua  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
264530474@qq.com

Fei Gao  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
18911990452@163.com

Huan Zhang  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
18110079659@163.com

Jiangwei Fan  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
1836710221@qq.com

Shuran Zhang  
CSSC Systems Engineering Research  
Institute  
Beijing, China  
zsr199210@163.com

## ABSTRACT

Arbitrary-Oriented object detection in remote sensing images is a hot topic in recent years. Currently, most arbitrary-oriented object detectors adopt the oriented bounding box (OBB) to represent targets in remote sensing imagery. However, OBB representation suffers from suboptimal regression problems caused by the ambiguity of the angle definition. In this paper, we propose a novel framework to Learning Segmentation-aware Mask for arbitrary-oriented object Detection (LSM-Det) in remote sensing imagery. LSM-Det predicts the mask of the object, and then converts the mask prediction into a minimum external OBB to achieve arbitrary-oriented object detection. Moreover, we designed a segmentation-aware branch to select high-quality predictions via the output matching score. Our method achieves superior performance on multiple remote sensing datasets. Code and models are available to facilitate related research.

## CCS CONCEPTS

• **Computing methodologies** → **Object detection; Shape representations; Object recognition.**

## KEYWORDS

oriented object detection, remote sensing image; segmentation mask

### ACM Reference Format:

Jiali Wei, Bo Hua, Fei Gao, Huan Zhang, Jiangwei Fan, and Shuran Zhang. 2023. Detecting Arbitrary-oriented Objects in Remote Sensing Imagery with Segmentation-Aware Mask. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590032>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590032>

## 1 INTRODUCTION

With the explosive growth of available remote sensing data, efficient remote sensing image interpretation technology plays an increasingly important role. Object detection is a basic task of remote sensing image interpretation, which aims to identify the targets of interest in the remote sensing images. In recent years, the rapid development of deep learning has made major breakthroughs in the field of generic object detection. A large number of frameworks based on convolutional neural networks (CNNs) have been proposed to achieve efficient and accurate object detection [2, 9, 14–16].

Generic object detection tasks often use horizontal bounding boxes (HBB) to represent object in the images, which is not applicable in objects in remote sensing imagery. Objects in the remote sensing images are often in arbitrary-oriented. The HBB annotations contain a lot of backgrounds, which makes the localization results not intuitive and hard to recognize. Compared with HBB representation, OBB introduces the extra angle dimension to denote the orientation of the rectangular bounding boxes. OBB contains less background information and can clearly distinguish objects and backgrounds in the image. Therefore, the most mainstream representation for oriented object detection is OBB [5, 10, 11, 22, 25].

However, detectors that directly predict OBB suffer the following problems.

- First of all, the definition of OBB leads to representation ambiguity, which makes it difficult for the network to converge. Multiple equivalent OBB representations caused by representation ambiguity would lead to a suboptimal regression process of the network and the oscillation of the loss function.
- Secondly, the slight angle deviation of the two rotated boxes will cause a sharp drop in the intersection-over-union (IoU). Therefore, it's hard to weight the loss contribution of different parameters.
- Finally, the mainstream  $L_n$ -norm loss cannot accurately measure the real deviation between predictions and ground-truth (GT) boxes, which leads to inconsistency between evaluation metric and regression loss.

Based on the above considerations, we suggest that the more effective representation for oriented objects is required for better performance. We observe that the objects in remote sensing images are all from a bird's-eye view, and most of them have simple contours. Most part inside the rotated bounding box is often the foreground with little background information. Hence the OBB annotation of the oriented object can be regarded as the pixel-level instance segmentation mask inside the OBB area. In this paper, we proposed a novel segmentation framework called LSM-Det to learning segmentation-aware mask to achieve accurate oriented object detection in remote sensing imagery. LSM-Det predicts a series of masks which may contains the objects. For each mask, the bbox IoU score, the mask IoU score, and the classification confidence are generated from the segmentation-aware branch. A matching score that combines these factors to determine the confidence of the predicted mask. Then, non-maximum suppression (NMS) is conducted to filter out minimum oriented bounding boxes corresponding to the redundant masks at inference.

Compared with other OBB-based rotation detectors, LSM-Det uses the mask representation to eliminate potential representation ambiguity caused by angle definition boundary. Moreover, the matching score bridges the inconsistency between classification, regression, and segmentation subtasks in LSM-Det by combining the classification score, bbox IoU, and mask IoU. This metric helps to conduct effective mask selection during inference to ensure high-quality detection performance. Extensive experiments on publicly available dataset prove the effectiveness of our model. The contribution of our method can be summarized as follows:

- We use segmentation-aware masks to represent oriented objects in remote sensing images, thereby converting the oriented object detection task into the instance segmentation task. In this case, the representation ambiguity and hard convergence issues brought by the angle representation will be eliminated, and better performance can be achieved.
- A segmentation-aware branch is designed to achieve high-quality prediction mask selection, which combines classification confidence, mask IoU, and bounding box IoU to determine with the prediction confidence.

## 2 METHODOLOGY

### 2.1 Combination with Instance Segmentation for Oriented Object Detection

OBB annotations contain relatively little background information for objects in the bird's-eye view of remote sensing images. It is intuitively feasible to convert OBB annotations into approximate mask annotations. In this case, instance segmentation can be applied to achieve better performance. In addition, the pixel-level mask prediction eliminates the representation ambiguity caused by the angle in the OBB, so the network is easier to converge.

The framework of our method is shown in the Figure. 1. The proposed LSM-Det is generally a two-stage detection framework and is combined with segmentation branches. The flow of representation method in LSM-Det is shown in the Figure 3. The OBB annotations of GT are converted into corresponding mask annotations for the supervision of subsequent segmentation branches. Specifically, we

fill the inner region of the OBB are with a pixel-level mask. First, the region proposal network (RPN) generates a series of candidate OBB regions based on the densely preset anchors on multi-scale feature maps. Then, the area-aligned features are extracted via the RoIAlign operation [3] from the candidate regions. Finally, classification and mask prediction are performed on the extracted features respectively. At inference, the minimum enclosing rectangular boxes of the predicted masks will be output as the final detections. In the training phase, RPN is trained under the supervision of GT OBB, and the mask prediction branch adjusts the parameters under the supervision of GT mask.

Integrating semantic segmentation into CNNs has been explored in previous work. For example, MaskOBB [18] uses Mask R-CNN [3] to detect and segment objects in remote sensing imagery. But these methods have the following shortcomings: 1). Semantic segmentation cannot distinguish different instances well, especially in a scene where the objects are densely arranged. 2). It additionally predicts the HBB just like the classic Mask R-CNN, which brings limited performance gains while leads to extra inference overhead. 3). They directly conduct NMS on the output OBBs without considering the credibility of the confidence. Compared with these methods, our LSM-Det uses the instance segmentation pipeline to achieve accurate mask prediction. To further eliminate the representation ambiguity brought by angle prediction, we discard the OBB prediction branch for better convergence. Finally, we have designed a mask matching score to evaluate the confidence of output masks, which will be introduced in the following sections.

### 2.2 Segmentation-Aware Confidence for Mask Scoring

In the instance segmentation task, the classification branch evaluates the predicted mask and selects the high-score masks as the final predictions. However, the classification score is not equivalent to the mask quality. There are many scenarios where high classification scores are endowed to low-quality mask predictions. This issue stems from the decoupling of classification task and segmentation branch. Mask scoring R-CNN [4] achieved credible evaluation of mask accuracy by predicting mask IoU, thereby improving segmentation performance. Although this strategy works well in the segmentation task, there are still more factors that need to be considered in the detection task.

A major problem is the inconsistency of the evaluation criteria when introducing instance segmentation into object detection. The evaluation metric for the object detection task is the bounding box IoU (bbox IoU), while mask IoU is used in instance segmentation. We use instance segmentation to achieve object localization in LSM-Det, but the mask IoU does not always reflect the accuracy of the predictions. Ideally, as shown in Figure. 4(a), the mask IoU and bbox IoU are approximately the same. Then we can use the mask IoU to correct the classification score to achieve accurate prediction mask selection just like Mask scoring R-CNN [4]. However, for the assumed extreme cases as shown in Figure. 4(b), a high mask IoU is not equivalent to a high bbox IoU. The unreliable mask IoU will mislead the model to output low-quality OBB predictions in the inference stage.

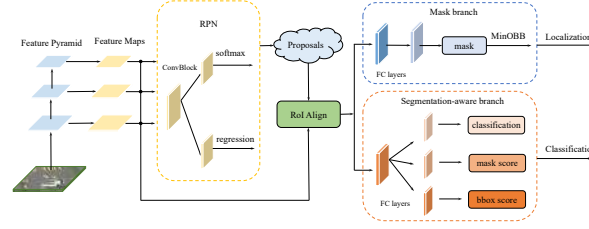


Figure 1: Overview of our method.



Figure 2: Annotations of the objects in the image in DOTA dataset.

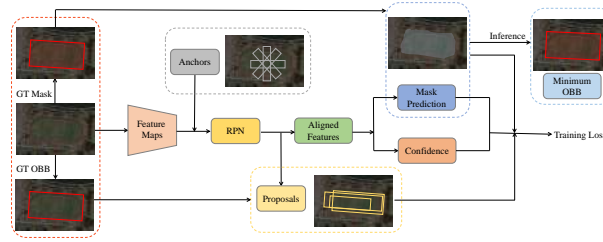


Figure 3: Representation flow in LSM-Det.

To solve the above problems, we designed the matching score to combine the mask IoU and the corresponding bbox IoU to comprehensively evaluate the output detections. For each predicted mask  $M$ , there are three scores will be attached in the designed segmentation-aware branch: classification confidence  $cc$ , mask IoU score  $ms$ , bbox IoU score  $bs$ . Then, the final matching score  $s$  is as follows:

$$s = \sqrt{cc \times (\alpha \cdot ms + \beta \cdot bs)}, \quad (1)$$

in which  $\alpha$  and  $\beta$  are the hyperparameters to adjust the contribution of different metrics. Mask IoU score  $ms$  and bbox IoU score  $bs$  are the posterior criteria, which evaluate the deviation between the predictions and GT results. The matching score  $s$  combines mask IoU score  $ms$ , bbox IoU score  $bs$ , which comprehensively considers the quality of the mask and that of the corresponding bbox, so as to achieve more accurate predictions. Specifically, this metric provides accurate evaluation of predictions in the inference stage and efficient supervision in the training stage. Take the case in Figure. 4(b) as an example. Although the mask IoU is relatively high, the low bbox IoU will further force the prediction mask to converge to the GT OBB area during training, thereby making the prediction more accurate.

However, previous work [12] has pointed out that the direct use of posterior IoU prediction makes the model hard to converge. Therefore, we integrate the prior alignment information, that is, the IoU between proposals and GT boxes, into the predictions of mask IoU and bbox IoU as follows:

$$\begin{aligned} ms' &= \gamma_1 \cdot ms + \gamma_2 \cdot p, \\ bs' &= \gamma_1 \cdot bs + \gamma_2 \cdot p, \end{aligned} \quad (2)$$

in which  $p$  is the IoU between the candidate regions and the GT box.  $\gamma_1$  and  $\gamma_2$  are the adaptive adjustment parameters. A suitable  $\gamma_2$  is expected to guarantee a smooth training process. Specifically, in the early stage of training, we pay more attention to the prior information for fast converge. When the model is relatively stable, more attention should be paid to the accuracy of predictions. Therefore, the adaptive weights  $\gamma_1$  and  $\gamma_2$  are set as follows:

$$\begin{aligned} \gamma_2 &= \cos\left(\frac{iter}{Max\_Iter} \cdot 0.5\pi\right), \\ \gamma_1 &= 1 - \gamma_2 \end{aligned} \quad (3)$$

Where  $iter$  demotes the current iterations, and  $Max\_Iter$  is the maximum iterations. At the beginning of training, we pay more attention to prior knowledge. In this way, the final matching score



(a) Consistent case



(b) Inconsistent case

**Figure 4: Visualization of the inconsistency between bbox IoU and mask IoU. The failed case in (b) shows that it is not always feasible to use the mask IoU as metric to evaluate the quality of the predicted mask.**

is defined as follows:

$$s = \sqrt{cc \times (\alpha \cdot ms' + \beta \cdot bs')}, \quad (4)$$

As the model converges well, it gradually concentrates on predicting matching scores instead of the initial bbox IoU. The strategy that is based on curriculum learning method helps the network smoothly achieve the prediction from priori bbox IoU to matching score, and avoid the case that the new metric hinder the network convergence. In our experiments, we found that the best performance is achieved when  $\alpha = 0.6$  and  $\beta = 0.4$ .

### 2.3 Loss Function

The overall loss of our method consists of the following parts: RPN loss, mask prediction loss, classification loss, bbox IoU loss, and mask IoU loss.

$$L = L_{RPN} + L_{mask} + L_{cls} + L_{bs} + L_{ms}. \quad (5)$$

$L_{RPN}$  is used for region proposal in object detection.  $L_{cls}$  denotes the general classification loss for the mask class recognition.  $L_{mask}$  is the loss to supervise the mask prediction of the object. Supposing the GT OBB of the target object is  $B$ , then the GT label of the  $L_{mask}$  is as follows:

$$m_i = \begin{cases} 1, & p_i \in B \\ 0, & otherwise \end{cases} \quad (6)$$

**Table 1: Evaluation of different supervision methods in the training loss on DOTA validation set.**

Supervision	OBB	Mask	Ours
AP <sub>50</sub>	68.7	68.2	<b>69.8</b>
AP <sub>75</sub>	32.9	35.1	<b>36.7</b>

$L_{ms}$  and  $L_{bs}$  are the posterior evaluation criteria for masks prediction and their corresponding minimum enclosing boxes, respectively. During the training stage, the GT labels of the two scores is not completely the corresponding IoU, but the weighted results as shown in Equation. 2.

## 3 EXPERIMENTS

### 3.1 Datasets and Baselines

DOTA [19] is a large public remote sensing dataset for oriented object detection in remote sensing images. It contains 2806 images with 188,282 annotated instances from different sensors and platforms. There are 15 categories in total, including plane (PL), baseball diamond (BD), bridge (BR), ground track field (GTF), small vehicle (SV), large vehicle (LV), ship (SH), tennis court (TC), basketball court (BC), storage tank (ST), soccer ball field (SBF), roundabout (RA), harbor (HA), swimming pool (SP) and helicopter (HC). Images in DOTA dataset exhibit a wide variety of scales, orientations, and shapes. Refer to the official data division strategy, we select half of the original images as the training set, 1/6 as validation set, and 1/3 as the testing set.

The baseline model in our experiments is RetinaNet[6]. A multi-feature pyramid network is constructed to detect multi-scale objects in remote sensing images. In order to achieve oriented object detection, we introduce additional angle prediction to output the oriented bounding box to locate the objects.

### 3.2 Experimental Settings

**3.2.1 Implementation Details.** Images in DOTA are too large to be fed into the model. Therefore, we crop the original images into 800×800 patches with the stride 200 for training and testing. It has been proved in previous work [8] that oriented anchors bring marginal benefits, but greatly reduce the inference speed. Therefore, we only preset horizontal anchors on feature maps to achieve oriented object detection. There are 9 anchors preset on each position of the feature maps, with an aspect ratio of {0.5, 1.0, 2.0} and a scale of  $\{2^0, 2^{\frac{1}{3}}, 2^{\frac{2}{3}}\}$ . Random flip, rotation, and multi-scale training and testing are adopted for data augmentation. We use SGD optimizer for training, and the momentum and weight decay were set to 0.9 and  $5 \times 10^{-4}$ , respectively. The learning rate is set to  $2 \times 10^{-3}$  and divided by 10 at epoch 16 and 22. We trained the model for 24 epochs on 4 RTX 2080Ti GPU with the batch size set to 8.

### 3.3 Ablation Study

We have conducted ablation experiments to verify the effectiveness of the modules in LSM-Det. For all ablation experiments, the images are resized into 416×416 for training and testing. Models are trained on the training set and evaluation is then conducted on validation set. No data augmentation is used in the ablation experiments.

**Table 2: Effects of components in the matching score.**

	Methods		
Classification score	✓	✓	✓
Mask confidence	✓	✓	×
BBox confidence	✓	×	×
AP <sub>75</sub>	<b>36.7</b>	36.1	35.6

**3.3.1 Evaluation of Different Representations.** We first compare the performance of different representation methods as shown in Table. 1. It can be seen that the AP<sub>50</sub> of the method based on mask prediction is even lower than that of OBB-based representation. However, the AP<sub>75</sub> of the model with mask prediction is 2.2% higher than that of the OBB prediction. It proves that the fine-grained prediction of the mask provides more accurate boundary information, which helps to achieve high-precision object detection. Furthermore, LSM-Det is equipped with our segmentation-aware branch. The output matching score is effective in selecting high-quality masks. With this strategy, the AP<sub>75</sub> of our method is 1.6% higher than mask representation.

**3.3.2 Evaluation of Segmentation-aware Matching Score.** Next, we discuss the role of each component in the matching score. In the instance segmentation task, the classification scores are enough to suppress redundant predictions. However, the parallel mask predictions branch and classification task are not aligned. Specifically, a high classification score cannot guarantee the accurate mask prediction. In LSM-Det, mask IoU is an accurate evaluation criteria to bridge the inconsistency between the two tasks. Therefore, it can be seen from the Table. 2 that the prediction of mask IoU increase AP<sub>75</sub> by 0.5%. Meanwhile, the misalignment also exists between mask IoU and bbox IoU, that is, the high mask IoU between the two predictions cannot guarantee their high bbox IoU. To solve the issue, our matching score in the inference stage combines the two criteria to achieve a more accurate selection of prediction results. Experimental results in Table. 2 shows that after integrating bbox IoU into the matching score, the AP<sub>75</sub> is further increased by 0.6%. Also, we tried different  $\alpha$  and  $\beta$ . Experimental results show that different combinations of  $\alpha$  and  $\beta$  can achieve stable performance gains within a reasonable range. Specifically,  $\alpha \in [0.4, 0.7]$  and  $\beta = 1 - \alpha$ . It shows that the proposed algorithm is robust. Among them, the best performance is achieved when  $\alpha = 0.6$  and  $\beta = 0.4$ .

### 3.4 Comparison with State-of-the-Art Methods

To prove the advancement and effectiveness of our algorithm, we conducted extensive experiments on multiple public remote sensing or aerial image datasets and compared our method with the existing state-of-the-art method. All compared methods show their best results reported in the corresponding papers.

DOTA is currently the largest public dataset for oriented object detection in remote sensing images. The scales and aspect ratios of objects in the images in DOTA dataset vary greatly. In addition, there are many scenes with densely arranged objects, which makes it difficult to achieve accurate detection. We train the model on the training set, then detect objects on the test set, and submit


**Figure 5: Visualization results on DOTA dataset.**

the detection results to the official test server for performance evaluation.

The evaluation results are shown in Table. 3. Our model achieves the best detection performance among the compared methods, reaching a mAP of 79.50%. MaskOBB [18] directly treats the oriented bounding box as a mask, and does not consider the misalignment between different tasks. In addition, we use a better general backbone network to extract more powerful features for better performance. In the end, the mAP of LSM-Det is 4.17% higher than MaskOBB.

The visualization of the detections on the DOTA dataset is shown in Figure. 5. The method based on pixel-level prediction will not be affected by the object aspect ratio and angle prediction of OBB. Even in scenes with densely arranged objects, our method still achieves accurate detection results.

## 4 CONCLUSIONS

In this paper, we discussed the issue in oriented object detection in remote sensing images, that is, the suboptimal regression problem caused by representation ambiguity. We suggest that the oriented rectangular bounding box approximates the mask of the ground-truth object, which contains little background. Based on this observation, we propose a novel framework LSM-Det to integrate instance segmentation branch into the detection framework to achieve better performance. A segmentation-aware branch is designed to predict bbox IoU, mask IoU, and classification scores. In the inference stage, the matching score comprehensively considers the three indicators to bridge the misalignment between different tasks and achieve more accurate predictions. Our method achieves advanced performance on public remote sensing image datasets, which further proves its effectiveness.

## REFERENCES

- [1] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. 2019. Learning RoI transformer for oriented object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2849–2858.
- [2] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.
- [3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
- [4] Zhaolin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. 2019. Mask scoring r-cnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6409–6418.
- [5] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. 2017. R2cnn: rotational region cnn for orientation robust scene text detection. *arXiv preprint arXiv:1706.09579* (2017).
- [6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

**Table 3: Comparisons with other state of the art methods on DOTA test dataset.**

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP
<i>Single-stage:</i>																	
A <sup>2</sup> -Det [20]	R-101	89.59	77.89	46.37	56.47	75.86	74.83	86.07	90.58	81.09	83.71	50.21	60.94	65.29	69.77	50.93	70.64
DAL [12]	R-101	88.61	79.69	46.27	70.37	65.89	76.10	78.53	90.84	79.98	78.41	58.71	62.02	69.23	71.32	60.65	71.78
DRN [13]	H-104	89.71	82.34	47.22	64.10	76.22	74.43	85.84	90.57	86.18	84.89	57.65	61.93	69.30	69.63	58.48	73.23
BBAVector [23]	R-101	88.35	79.96	50.69	62.18	78.43	78.98	87.94	90.85	83.58	84.35	54.13	60.24	65.22	64.28	55.70	72.32
CFC-Net [8]	R-50	89.08	80.41	52.41	70.02	76.28	78.11	87.21	90.89	84.47	85.64	60.51	61.52	67.82	68.02	50.09	73.50
SLA [10]	R-50	88.33	84.67	48.78	73.34	77.47	77.82	86.53	90.72	86.98	86.43	58.86	68.27	74.10	73.09	69.30	76.36
RIDet [11]	R-50	89.31	80.77	54.07	76.38	79.81	81.99	89.13	90.72	83.58	87.22	64.42	67.56	78.08	79.17	62.07	77.62
RDD [25]	R-101	89.15	83.92	52.51	73.06	77.81	79.00	87.08	90.62	86.72	87.15	63.96	70.29	76.98	75.79	72.15	77.75
<i>Two-stage:</i>																	
RRPN [7]	R-101	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
RoI Trans. [1]	R-101	88.64	78.52	43.44	75.92	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
CAD-Net [24]	R-101	87.80	82.40	49.40	73.50	71.10	63.50	76.70	90.90	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
Gliding Vertex [21]	R-101	89.64	85.00	52.26	77.34	73.01	73.14	86.82	90.74	79.02	86.81	59.55	70.91	72.94	70.86	57.32	75.02
MaskOBB [18]	RX-101	89.56	85.95	54.21	72.90	76.52	74.16	85.63	89.85	83.81	86.48	54.89	69.64	73.94	69.06	63.32	75.33
OPLD [17]	R-101	89.37	85.82	54.10	79.58	75.00	75.13	86.92	90.88	86.42	86.62	62.46	68.41	73.98	68.11	63.69	76.43
LSM-Det (Ours)	R-101	90.05	84.52	58.32	79.31	75.22	83.61	87.90	90.57	85.33	86.74	67.37	65.78	74.20	73.15	64.28	77.76

- [7] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. 2018. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia* 20, 11 (2018), 3111–3122.
- [8] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, and Yunpeng Dong. 2021. CFC-Net: A Critical Feature Capturing Network for Arbitrary-Oriented Object Detection in Remote-Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* (2021), 1–14. <https://doi.org/10.1109/TGRS.2021.3095186>
- [9] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, Junjie Song, Yunpeng Dong, and Xue Yang. 2023. Task interleaving and orientation estimation for high-precision oriented object detection in aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing* 196 (2023), 241–255. <https://doi.org/10.1016/j.isprsjprs.2023.01.001>
- [10] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, Junjie Song, and Xue Yang. 2021. Sparse Label Assignment for Oriented Object Detection in Aerial Images. *Remote Sensing* 13, 14 (2021), 2664.
- [11] Qi Ming, Lingjuan Miao, Zhiqiang Zhou, Xue Yang, and Yunpeng Dong. 2021. Optimization for Arbitrary-Oriented Object Detection via Representation Invariance Loss. *IEEE Geoscience and Remote Sensing Letters* (2021), 1–5. <https://doi.org/10.1109/LGRS.2021.3115110>
- [12] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. 2021. Dynamic Anchor Learning for Arbitrary-Oriented Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2355–2363.
- [13] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haojie Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. 2020. Dynamic Refinement Network for Oriented and Densely Packed Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11207–11216.
- [14] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.
- [15] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.
- [16] Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018).
- [17] Qing Song, Fan Yang, Lu Yang, Chun Liu, Mengjie Hu, and Lurui Xia. 2020. Learning Point-guided Localization for Detection in Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2020).
- [18] Jinwang Wang, Jian Ding, Haowen Guo, Wensheng Cheng, Ting Pan, and Wen Yang. 2019. Mask OBB: A semantic attention-based mask oriented bounding box representation for multi-category object detection in aerial images. *Remote Sensing* 11, 24 (2019), 2930.
- [19] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. 2018. DOTA: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3974–3983.
- [20] Zhifeng Xiao, Kai Wang, Qiao Wan, Xiaowei Tan, Chuan Xu, and Fanfan Xia. 2021. A2S-Det: Efficiency Anchor Matching in Aerial Image Oriented Object Detection. *Remote Sensing* 13, 1 (2021), 73.
- [21] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE transactions on pattern analysis and machine intelligence* (2020).
- [22] Xue Yang, Xiaojiang Yang, Jirui Yang, Qi Ming, Wentao Wang, Qi Tian, and Junchi Yan. 2021. Learning High-Precision Bounding Box for Rotated Object Detection via Kullback-Leibler Divergence. *arXiv preprint arXiv:2106.01883* (2021).
- [23] Jingru Yi, Pengxiang Wu, Bo Liu, Qiaoying Huang, Hui Qu, and Dimitris Metaxas. 2021. Oriented object detection in aerial images with box boundary-aware vectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2150–2159.
- [24] Gongjie Zhang, Shijian Lu, and Wei Zhang. 2019. CAD-Net: A context-aware detection network for objects in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* 57, 12 (2019), 10015–10024.
- [25] Bo Zhong and Kai Ao. 2020. Single-Stage Rotation-Decoupled Detector for Oriented Object. *Remote Sensing* 12, 19 (2020), 3262.

# A Review of Routing Optimization Techniques for Quality of Service Assurance in Software-Defined Networks

Guozhu Yan\*

Network information research institute Academy of  
Military Sciences, China  
1065624386@qq.com

Shuangyin Ren

Network information research institute Academy of  
Military Sciences, China  
renshuangyin@126.com

Jingchao Wang

Network information research institute Academy of  
Military Sciences, China  
wangjc.2000@tsinghua.org.cn

Chao Xue

Network information research institute Academy of  
Military Sciences, China  
xuec11@tsinghua.org.cn

## ABSTRACT

The traditional military communication network is based on IP architecture, which has the problems of rigid architecture and challenging quality of service guarantee. The rapid development of various new applications has put differentiated demands on the whole network's service quality. In recent years, software-defined network technology has been developing, which has the characteristics of decoupling the control and data planes. Its control plane has excellent global control capability and network equipment information collection capability, which has natural advantages for optimizing the quality of service. Firstly, the SDN network architecture, quality of service performance parameters, and service model are introduced; secondly, the different requirements of quality of service for standard and typical service application scenarios are analyzed, and the current research status of quality of service enhancement through routing optimization is described; finally, the outlook is summarized, and two new ideas for quality of service enhancement in SDN networks and the development trend of quality of service research in military communication networks are proposed.

## CCS CONCEPTS

• Networks; • Network architectures; • Network design principles;

## KEYWORDS

Software-defined networks, QoS optimization, Shortest path (Dijkstra) algorithm, Heuristic algorithm, Machine learning

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590033>

## ACM Reference Format:

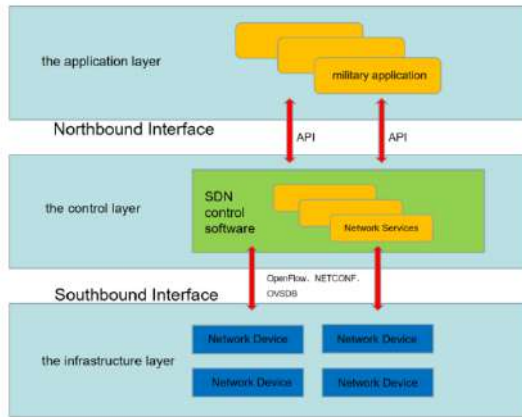
Guozhu Yan, Jingchao Wang, Shuangyin Ren, and Chao Xue. 2023. A Review of Routing Optimization Techniques for Quality of Service Assurance in Software-Defined Networks. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590033>

## 1 INTRODUCTION

To win quickly and precisely on the battlefield, military communication networks must be able to transmit information in a timely and reliable manner, support combat units, weapons platforms, and other elements in step within the battlefield space, and ensure effective command and control and weapons operations for troops. The variety of battlefield network services has been enriched in recent years, which has put forward higher requirements for the quality of service guarantee of military communication networks.

Currently, the military communication network is based on IP architecture, which has the problems of rigid architecture and complex service quality assurance. It is difficult to meet the needs of future information warfare. It faces excellent challenges regarding flexible on-demand deployment of resource services, complete determinism of information transmission services, and diverse service quality assurance. In addition, the current military communication network is a two-layer logical overlay architecture, where the upper application network and the underlying transmission network are relatively independent, and it is difficult to perceive the service quality requirements of the service during transmission, which leads to low transmission efficiency and challenging to guarantee the service quality reliably.

With the rapid development of various new applications, military communication networks must meet the differentiated quality of service requirements. In recent years, software-defined networks have been developing with the characteristics of decoupling the control and data planes. Their control planes have excellent global control and network device information collection capabilities, which have natural advantages when performing network routing optimization. Therefore, this paper focuses on the research related to SDN-based network routing optimization techniques. The SDN network architecture, performance parameters, and service model of quality of service are introduced. The necessity of applying SDN routing optimization to quality of service research is discussed by comparing the different requirements for quality of service of



**Figure 1: The architecture diagram for software-defined networks**

standard network services and typical network service scenarios. For the current research status of SDN routing optimization to improve quality of service, we summarize the shortest path algorithm, heuristic algorithm, and machine learning in 3 aspects and try to propose a new idea to improve the quality of service of the SDN networks.

## 2 RELATED WORK

### 2.1 SDN Network Architecture

The uppermost of these layers is the application layer, corresponding to the management program of the military communication network, which is mainly responsible for security protection and for dealing with software-related matters, such as firewalls and network virtualization, which are applications of this layer [1], and in the middle is the control layer corresponding to the control plane, which plays a role between the application layer and the infrastructure layer, where the controller is programmed to issue decisions, direct data devices to route, forward, and packet discard, responsible for managing the traffic of the entire network and providing network management services; at the bottom layer, the infrastructure layer is the data plane, mainly responsible for forwarding packets according to the rule policy assigned to it by the control plane, usually consisting of network devices such as switches, routers and access points; the above three structural layers are connected to each other through southbound and northbound interfaces.

The controller is the core part of SDN, the bridge connecting the upper layer application and the underlying switching equipment. On the one hand, it is responsible for forwarding data streams through the data plane. On the other hand, it has to provide programmable service interfaces to the upper application layer while unifying management, monitoring, and effectively securing routing and forwarding policies for network devices. Developers invoking the interface will be able to realize the SDN programmable capability and flexibly complete the design and management of the SDN network while writing the corresponding routing optimization algorithm in the controller to be able to get a path that meets the quality of service requirements [2].

### 2.2 QoS-related parameters and models

Quality of Service (quality of service) is a technique adopted in information and communication networks to solve the problem of link congestion and packet loss to guarantee the stable and reliable transmission of services to ensure that services are not affected by network overload [3]. This service refers explicitly to the transmission service that a packet (stream) can receive from the start through several network nodes to reach the destination end [4] finally.

**2.2.1 quality of service performance parameters.** In information and communication networks, the quality of service performance of a network is usually evaluated by performance parameters such as latency, throughput, packet loss rate, and jitter [5]:

- (1) Latency: End-to-end latency is the sum of the time a data message takes to reach its destination address from the network source. Time delay is an essential indicator of the quality of service, and some assistance with high real-time such as online games needs low time delay.
- (2) Throughput: The amount of data (measured in units of weight, bytes, etc.) that a network successfully transmits per unit of time, which means that throughput is the maximum data rate that a device can receive and forward without frame loss. Throughput can reflect the overall state of a network and can be affected by link congestion.
- (3) Packet Loss Ratio: The number of packets lost to the number of packages sent when data messages are forwarded [6]. When there is a large amount of traffic in the network, the link bandwidth is insufficient, and congestion will occur. When the queue buffer of the forwarding node cannot hold too many packets, some packets will be dropped and cause packet loss [7]. The stuttering and blurring that occurs when we watch live web videos are caused by a high packet loss rate [8].
- (4) Jitter: It refers to the jitter when packets leave the transmitter's end evenly spaced according to a regular interval. After passing through the network, this even interval is disrupted by packages experiencing different delay sizes. Jitter exists in all transmission systems, but as long as the jitter can be kept within a defined range, it will not adversely affect the quality of service. Often, too much jitter can distort multimedia applications.

**2.2.2 quality of service model.** In early communication networks, most applications, such as website browsing, file transfer, email, etc., do not need to guarantee transmission delay, bandwidth, jitter, etc. The Best-Effort service model can already meet the demand, and this service model is simple and can utilize the network resources more fully. The existing quality of service research mechanisms, according to the different requirements of the service, the QoS service model is divided into two aspects of integrated service model and differentiated service model [9-12].

#### (1) IntServ model

The Integrated Service Model (IntServ) was mainly proposed by the IntServ Working Group established by the IETF in 1994, with the primary purpose of providing end-to-end quality of service

assurance for emerging real-time multimedia applications and distributed multimedia applications. The data processing process of this model is as follows: before sending packets, the network application will calculate the network resources required to satisfy the QoS service through packet marking priority, token bucket management rate, QoS scheduling techniques for precedence queuing, etc., and submit a resource request to the network, and then allocate the reserved resources according to the Resource Reservation Protocol (RSVP) [13, 14], and secure a certain amount of network resources for the application on each relevant routing node. If the available resources of the network node are not much, thus leading to reservation failure, the data stream will be denied access to the network, and when the data transmission is completed, the corresponding resources allocated by the resource reservation protocol [15] are released. The main advantage of the integrated service model is that it can meet the quality of service requirements of each stream in this way. However, the network nodes in this model need to maintain a large number of reserved resource information of data streams, which increases the processing time and memory resources, resulting in the overloading of network equipment. The maintenance overhead will increase significantly when the network scale is further expanded. Providing quality service for each stream is more complex, requiring specific hardware and software support. The implementation mechanism is complicated and not conducive to upgrade deployment, and scalability could be better and, therefore, unsuitable for large-scale applications. Currently, it is mainly used in combination with traffic engineering [16].

#### (2) DiffServ model

In response to the problems of poor scalability and high network equipment load of the integrated service model, the IETF working group proposed a differentiated service model (DiffServ) based on the IntServ model.

The primary purpose of this model is to provide differentiated quality of service for different priorities of application services in the network. Higher priority services are provided for real-time service requirements without reserving network resources. The processing flow is as follows: the differentiated service model is mainly based on classification techniques. The aggregation and single-trunk segment behavior PHBs provide QoS for application requirements, using a coarse-grained control mechanism managed by traffic type. The differentiated service (DS) byte and type of service (ToS) byte in the IP header are used to determine the PHB, and different DSCP values are assigned according to the PHB of the packet. At the regional network boundary, the network infrastructure device examines the header marking information of data messages for grouping, distinguishes different service levels of applications by different DSCP values, and classifies messages, queue scheduling, traffic shaping, and traffic monitoring by service level [4]. This mechanism aggregates and manages other service data streams with the same quality of service requirements. At the same time, the network nodes of the differentiated service model are also able to prioritize the forwarding and scheduling of the data streams to ensure the quality of service requirements of the data streams, which is simple to implement, has good scalability, is easy to deploy and upgrade on existing networks, and can be applied to the deployment of large-scale core backbone networks. Still, due to its lack of end-to-end resource

reservation, it cannot guarantee the quality of service requirements in the case of poor network status.

## 3 BUSINESS APPLICATION SCENARIOS

### 3.1 Common business scenarios

In traditional IP networks, all information is processed indiscriminately, and the reliability, delay congestion, and other characteristics in information forwarding are not guaranteed. With the development of information networks, new applications are emerging, such as multimedia games, smart medical, smart city, live video, etc. Compared with traditional web links, email, FTP, and other applications that are not sensitive to time delays, new services have special needs for bandwidth, latency, jitter, and other transmission performance, such as video conferencing, live games, and other services need high bandwidth, low latency, and low jitter, while e-commerce, etc. Although they do not necessarily require high bandwidth, they need to ensure low latency and be able to get priority treatment in case of congestion, which puts higher requirements on network QoS. If the transmission is still performed according to the best-effort approach, the user experience will become poor.

### 3.2 Typical business scenarios

In this paper, we take three typical application scenarios of the data center, satellite network, and streaming media as examples to analyze the importance of quality of service assurance. In recent years data center has gradually become a critical project of new infrastructure all over the world, especially in China [17, 18], data center volume is enormous, but the internal traffic characteristics are apparent, manifested as elephant flow and mouse flow, different from the data center, satellite network service is diverse, traffic characteristics are other, correspondingly there is various quality of service requirements, satellite network now becomes a part of China's development of national globalization strategy [19], and streaming media is Representative of many emerging services, based on live streaming has penetrated the working life of the public, mainly in terms of high real-time requirements, the delay requirements for traffic is self-evident.

**3.2.1 Data Center.** There are mainly two kinds of traffic in data center networks, east-west and north-south [20]. With the development of big data and cloud computing, east-west traffic has become the main traffic pattern, the proportion can be more than 70% [21], and there are two main types of east-west traffic, which are elephant flow and mouse flow. In the data center, the elephant flow is mainly generated by file storage, virtual machine migration, MapReduce, cluster computing tasks, and other services. The size is generally less than 100Mbit/s, which is 20% of the total data center traffic. The data volume accounts for about 90% of the total traffic, and the duration is long, characterized by high throughput requirements, requiring high bandwidth; mouse flow is mainly generated by Web services, web search, e-commerce, and other services, and the size is generally less than 100Mbit/s. The mouse stream is primarily caused by Web services, web search, e-commerce, and other services. The size is usually less than 10Kbit/s, and the quantity is about 80% of the data center traffic. The data volume accounts for about 10% of the total traffic, and the duration will not exceed 10s

**Table 1: QoS target requirements for common services**

Business Type	Time delay	Bandwidth	Packet loss rate	Jittering
HTML Web Browsing	Allow appropriate time delay	Uncertain bandwidth requirements	Do your best to transmit	Allow for proper shaking
Email	Allowable time delay	Low bandwidth requirements	Do your best to transmit	Allowing for jitter
FTP	Time delay sensitive	Bandwidth requirements are appropriate	Sensitive to packet loss and must transmit reliably	Sensitive to jitter
E-commerce	Sensitive to time delay	Bandwidth requirements are appropriate	Sensitive to packet loss and must transmit reliably	Sensitive to jitter
Videoconferencing	Very sensitive to time delay	High bandwidth requirements	Require predictable latency and packet loss	Very sensitive to shaking
Streaming Media	Very sensitive to time delay	High bandwidth requirements	Require predictable latency and packet loss	Very sensitive to shaking

**Table 2: Comparison of elephant flow and mouse flow characteristics**

Name	Processing business	Quantity	Data volume	Duration	QoS Requirements
Elephant Flow	File storage, virtual machine migration, MapReduce, cluster computing tasks	Less, 20% of total	Larger, 90% of the total	Relatively long	High bandwidth
Mouse Stream	Web services, web search, e-commerce	More, 80% of the total	Smaller, 10% of total	Relatively short, no more than 10s at most	Low time delay

at most, characterized by delay-sensitive, need to be transmitted on time [22]. The specific characteristics of elephant flow and mouse flow are shown in Table 2.

**3.2.2 Satellite network.** Satellite communication link resources are valuable. Traditional satellite systems mostly use a distributed network architecture, the network control level, and business-level unified deployment. Satellite nodes need to complete the link maintenance, status monitoring, routing calculations, and other network control functions, consuming valuable on-board payload and interplanetary link resources [23], and satellite communication has the advantages of being less susceptible to interference, wide service area, large communication capacity; therefore The satellite network carries various services and has different demands on several indicators of quality of service such as throughput, packet loss rate, and delay, and the satellite network needs to support other QoS routing processes.

**3.2.3 Streaming Media.** Streaming Media, or streaming media, is a continuous piece of media data that is compressed and segmented into packets and transmitted over a network. The terminal can receive the package on the web and decompress the media data to present it to the user. As streaming media faces many fundamental challenges, such as network client heterogeneity, bandwidth fluctuations, and constantly changing scenarios, it will be challenging for users to get quality of service guarantee without the time and space constraints, especially for video applications such as web-casting, video conferencing, and teleconferencing, which have high real-time requirements and are sensitive to increased throughput

and jitter requirements, and require high bandwidth, low latency, and low jitter to ensure video transmission quality.

The manifestation of data flow in the network varies in the three application scenario modes. Still, all of them put higher requirements for quality of service assurance, which obviously cannot meet the QoS requirements by relying on the traditional best-effort transmission form. The centralized management mode feature of software-defined network (SDN) technology can effectively solve these problems. Its primary research idea is to rely on the SDN centralized control, mastering the global topology information, collecting network parameters to sense the current network state, combining with the feedback data cache or transmission parameters from clients in the network, and carrying out network transmission path optimization or transmission priority modification to finally achieve the optimized network transmission. The result is to maximize network transmission.

## 4 STATUS OF RESEARCH ON QOS ENHANCEMENT THROUGH ROUTING OPTIMIZATION

Routing optimization is the most direct perspective to consider in studying QoS, and there is a wealth of related research results. This paper focuses on three classical and popular schemes (shortest path algorithm, heuristic algorithm, and machine learning) in the development of SDN routing optimization and the corresponding improvement measures proposed by academia.

#### 4.1 Shortest path (Dijkstra) algorithm

The path calculation module of the SDN controller itself mainly uses the Dijkstra shortest path algorithm, which was proposed by the Dutch mathematician Dijkstra in 1959 [24]. Dijkstra's algorithm [25] divides the nodes in the network into unmarked nodes, temporarily marked nodes, and permanently marked nodes, using the source point as the initial permanently marked point, and adds more points to this set through continuous greedy selection to add more points into this set until all points are added to the collection. The SDN controller has the complete network topology and the weights on each link, based on which Dijkstra's algorithm can calculate the shortest path between a source node and other network destination nodes.

The shortest path algorithm can find the shortest path from a specified node to the rest of the nodes in the network topology graph and is a typical single-source shortest path algorithm using greedy thinking [26]. For example, Xiangbin Kong et al. [27] proposed a QoS routing algorithm based on Dijkstra's algorithm to select routes that satisfy bandwidth constraints for service flows.

The core idea of the algorithm is to select the path with the maximum available bandwidth as the transmission path in the link with the minimum guaranteed hop count, which is perfect as a routing algorithm. Still, the application in large-scale networks requires larger storage space and computation time. The selection of each stage is optimal, which is easy to fall into the optimal local solution [28], and the algorithm first selects the available path based on the hop count. It does not determine the optimal route based on the network state. This may lead to poor link utilization.

#### 4.2 Heuristic algorithm

The heuristic algorithm is mainly applied to the quality of service routing optimization calculation under multiple constraints [29], usually for the optimization of the state constraints of the full link of the network, solving a path with the minimum cost between the source and the destination node under the satisfaction of multiple regulations defined [30], the algorithm requires accurate modeling of the application scenario and relevant parameters, and in searching the path first selects the optimal node to expand, and finally calculates the feasible solution.

The ant colony algorithm is the most representative heuristic and has the most related optimization studies. Ant colony algorithm is better used in routing, and its advantage is that the performance is better when the number of network nodes is a small number of network exit branches, such as PARSAEIMR, etc. [31] based on the ant colony algorithm proposes an adaptive routing scheme, which can effectively improve the quality of the remotely received video. However, the global search capability of the ant colony algorithm is weak, the reliability of calculating the optimal path result is not high, the convergence speed is slow, and the convergence time of the ant colony algorithm increases exponentially as the complexity of the network becomes more elevated. This deficiency is also the main problem in improving numerous ant colony optimization algorithms [32]. Ming Li proposed an SDN routing strategy (DP-ACO) based on the ant colony optimization algorithm, which is significantly better than Dijkstra's algorithm and the traditional ant colony algorithm in terms of delay and throughput. [33]. Amikal et

al. [34] attempted to solve the network path congestion problem using the ant colony optimization method and applied the ant colony algorithm in two phases to optimize the network performance. Wang C et al. [35] first proposed An ant colony optimization-based link load balancing algorithm, which takes link load and delays as influencing factors and calculates the most comprehensive and shortest paths among all paths to maintain link load balancing and reduce end-to-end delay. Zhenpeng Liu et al. [36] proposed an SDN network routing strategy based on an ant colony optimization algorithm for handling large flows that occupy large bandwidth in the network, which has the advantages of high average throughput and high link utilization compared with the strategy based on the shortest path algorithm.

In addition to the ant colony algorithm, the more commonly used practical heuristic algorithms include the annealing algorithm, genetic algorithm, etc. [37]. Hedera [38] proposes the use of equivalent multipath ECMP network routing in data centers, which predicts the link bandwidth required for network data transmission before the network data is forwarded and uses a simulated annealing algorithm to obtain the path between source and destination hosts, freeing the link bundle during packet forwarding, and improves the utilization of network links. WANG J et al. [39] used a genetic algorithm to optimize the routing problem with good results.

Although heuristic algorithms can compute results quickly and relatively close to the optimal solution in practical applications, the algorithms have strict applicability scenarios and poor generalization performance, and the model parameters of these heuristic algorithms need to be adjusted when the network state changes, making it difficult for the algorithms to work stably, less flexible and requiring higher computation time costs, so they cannot compute optimal solutions in real-time, and in SDN controller to compute flow paths is not feasible [37].

#### 4.3 Machine Learning

Machine learning can calculate SDN routing policies in real-time, unlike heuristic algorithms. It is widely used in SDN routing optimization with its excellent performance in massive data processing, intelligent data classification, and high-frequency action decision-making [40]. Machine learning does not require an accurate network model but mainly uses its powerful characterization capabilities to learn different traffic patterns and understand the relationship between complex network environments and network services to manage routing and better help network decisions dynamically, and once the learning model is trained, it can calculate the near-optimal solution in a short time [41].

**4.3.1 Supervised learning-based routing optimization.** Supervised learning is a machine learning technique that makes inferences based on labeled training data [42], mainly using samples with known inputs and outputs to train models capable of obtaining features and complex structures between high-dimensional data. Supervised learning-based QoS routing optimization improves routing forwarding decisions by transforming network topology information and labels into vectorized parameters to train deep neural networks [43].

In supervised learning-based routing optimization, many studies have combined supervised learning and heuristic algorithms with

much less running time than heuristic algorithms. At the same time, the performance of delay and jitter is similar to the heuristic short-hair routing results, achieving good results. Azzouni A et al.[44] proposed a deep neural network-based dynamic routing framework for NeuRoute and incorporated heuristic algorithms into it, and the results showed that the method dramatically reduces network latency. Li Yanjun et al.[45] proposed a dynamic routing framework based on supervised learning with a heuristic algorithm input layer incorporated into the framework. Experiments show that: the method is computationally efficient and has better network performance.

Supervised learning algorithms based mainly on deep learning methods develop routing decisions by learning deep relationships between routing decisions or in-depth features between network data and forwarding decisions. Kato N et al.[46] proposed a deep neural network system based on the input of traffic features, and experimental results show that the method minimizes delay. Zirui Zhuang [47] proposed a graph-aware deep learning-based routing decision scheme that yields higher prediction accuracy in less training time than a pure neural network model. A. Azzouni et al.[48] proposed a deep neural network-based routing framework by estimating and predicting the traffic matrix (TM) and using the prediction results and real-time network state parameters as input and the best path as output to train the model.

However, supervised learning still has non-negligible drawbacks in route optimization, requiring pre-access to a large amount of training data for training in the learning process, poor scalability, and often resulting in higher computational complexity [49].

**4.3.2 Routing optimization based on reinforcement learning.** Reinforcement learning (RL) [50], one of the main branches and methods of machine learning, is used to solve the problem of an intelligent body that learns to interact with its environment to obtain an optimal action strategy to try to get the maximum action for the current environmental reward. The reinforcement learning-based quality of service routing algorithm, on the other hand, continuously interacts with the environment in the network, uses the self-learning property of reinforcement learning to perform quality of service routing exploration using a trial-and-error strategy, uses the quality of service as the reward function and trains the model intending to maximize the cumulative reward [42]. S. C. Lin et al.[51] introduced a distributed hierarchical control plane structure and a Markovian decision process to implement a reinforcement learning-based QoS-aware adaptive routing algorithm. Xianbei Che et al.[52] introduced a near-end optimal algorithm to optimize the routing policy dynamically by adjusting the reinforcement learning reward function according to different optimization objectives. The results show that the algorithm has significantly reduced the average and maximum end-to-end delay on the network compared to the traditional shortest-path routing algorithm.

The classical reinforcement algorithm Q-learning [53] is a value iteration algorithm that stores Q-values through Q-tables, which updates the stored states, actions, and rewards by learning Q-values. Kochi Liu [54] uses a Q-learning algorithm for QoS routing optimization decisions. Dynamically varying exploration rates, as well as reward functions, are designed to meet different demands. The results show that the algorithm effectively improves network

throughput and average bandwidth utilization and reduces average packet loss. JIANG J et al.[55] is an application of the Q-learning method to end-to-end adaptive HTTP streaming intelligent delivery architecture. WU LINGLING [56] proposes an  $\epsilon$ -Q-Learning traffic scheduling model based on  $\epsilon$ -Q-Learning. The results show that the method performs better in average throughput, link utilization, and bandwidth utilization.

However, the Q-learning algorithm brings great memory overhead due to the limitation of the Q-table, and Q-learning can only handle the problem that the state space and action space can be manageable and discrete. So many scholars used neural networks to fit the value or policy functions in reinforcement learning and created various deep reinforcement learning models. Xu et al.[57] proposed a deep reinforcement learning-based control framework (DRL-TE) that assigns three candidate paths to each end-to-end communication session and uses DRL to determine the proportion of each stream on these three paths, utilizing the throughput and delay of each stream as network state, the split path ratio of each flow as the action space, and the performance metrics of each flow as the reward function, and the optimal split ratio is determined by feedback. The results show that the method significantly reduces the end-to-end delay and improves the throughput simultaneously. Machine learning, intense learning techniques, has attracted much attention for its excellent performance in large-scale data processing, classification, and intelligent decision-making. Many studies are using it in SDN networks to solve problems in network operation and management [58, 59].

Deep reinforcement learning achieves intelligent network forwarding and has received much attention from researchers. The algorithm does not rely on the exact system model, which is especially suitable for complex networks with random and unpredictable behavior, and the algorithm only needs simple optimization to get the approximate optimal solution after training. In contrast, the heuristic algorithm needs several iterations, and compared with supervised learning, deep reinforcement learning simplifies the processing of data. However, the current deep reinforcement learning models are black-box optimization networks, reliability and robustness cannot be guaranteed, the research on routing optimization based on deep reinforcement learning is still in the preliminary stage, and its solutions are too large for large-scale complex networks, the model input and output dimensions are too large, the convergence speed and computational cost of the model are not ideal, and the model generalization capability is limited. Thus, the current research on reinforcement learning QoS routing optimization techniques is basically at the simulation stage and has not been applied to real SDN networks for performance evaluation [60].

## 5 SUMMARY OUTLOOK

### 5.1 Propose two new ideas for improving QoS in SDN networks

#### (1) SDN-based microservice architecture

By embedding the microservice mechanism on the SDN network, the microservice invocation process load balancing can be moved entirely up to the control plane. The load of the service provider's hosts and the network path's bandwidth can be considered comprehensively, and the service invoker can be made aware of the load

balancing policy. The SDN control plane provides finer-grained load balancing capability and enables optimization of load balancing for microservice features.

## (2) SDN and intelligence

With the control plane programmable idea of SDN, AI algorithms can be deployed to each node of computer networks more flexibly to optimize the routing computation of the web, but the machine learning-based algorithms mentioned before have considerable time and space complexity, in this case, network representation learning can be applied to network modeling by mapping network topological nodes into low-dimensional vectors with fixed dimensional size, and these vectors can retain information such as topology and node characteristics, thus reducing the complexity of machine learning algorithms.

## 5.2 Trends in Quality of Service Research for Military Communication Networks

This paper focuses on the control level of the SDN framework. It summarizes targeted QoS optimization solutions based on routing algorithm optimization based on the analysis of SDN controllers and QoS-related parameter models. With the emergence of new services and scenarios, the research thinking on QoS optimization solutions will continue to deepen. The connection problems and topology changes caused by device movement in future networks are significant challenges affecting the quality of service and need to be given great attention.

Modern warfare faces a more complex situation due to the uncertainty and complexity of the battlefield environment and the diversity of missions. To make correct and timely command decisions, we must rely on information and communication networks with a higher quality of service based on the SDN global network view, comprehensive assessment, and measurement of the whole picture of operations, using the characteristics of SDN control and transfer separation, command decisions are more flexible and efficient. They will not be due to subjective consciousness as well as the command and control will be more scientific and reasonable.

With the rapid development of 5G and cloud networks, military communication networks should also take advantage of the trend, transform and upgrade so that the network can move with the cloud, thus realizing the on-demand allocation of network resources, dynamic scheduling, and elastic adjustment, and quality of service research will mainly focus on edge computing, IoT, orbiting satellites. Other mobile quality of service research will mainly focus on edge computing, IoT, orbiting satellites and other mobile networks, and command and decision special networks. Quality of service optimization problems related to frequent switching of network access nodes and high-quality transmission of video traffic in mobile devices will be hot spots for research.

## REFERENCES

- [1] Rawat D B, Reddy S R. Software Defined Networking Architecture, Security and Energy Efficiency: A Survey [J]. IEEE Communications Surveys & Tutorials, 2017, 19(1): 325-346.
- [2] Guozhu YAN, Qiongyu Wu, Rongbing Chen, Linfeng Du, Shuangyin Ren. "A Literature Review of Resiliency Technologies in Military Software Defined Networks", 2022 5th International Conference on Data Science and Information Technology (DSIT), 2022
- [3] Zhao, Regina. Research on Service-Oriented End-to-End QoS Routing Policy [D]. Beijing University of Posts and Telecommunications, 2021. DOI:10.26969/d.cnki.gbydu.2021.000775.
- [4] Song Zhikun. Research on multi-constraint routing algorithm for software-defined networks based on nonlinear annealing [D]. Xi'an University of Electronic Science and Technology, 2015.
- [5] Yuan Qijie. Research on QoS routing technology in software-defined networks (SDN) [D]. Beijing: Beijing University of Posts and Telecommunications, 2019.
- [6] Dai, Z., Cheng, G., Zhou, Y., Yang. Research on measurement methods for software-defined networks [J]. Journal of Software, 2019, 30(06): 1853-1874.
- [7] Yin H, Li F. Research on the development of Internet performance measurement technology [J]. Computer Research and Development, 2016, 53(01): 3-14.
- [8] Lu Y. Research on intelligent QoS routing optimization based on software-defined networks [D]. Hunan Normal University, 2021. doi:10.27137/d.cnki.ghusu.2021.002694.
- [9] Ponnappan A, Yang L, Pillai R, et al. Policy Based QoS Management System for the IntServ/DiffServ Based Internet [C]// International Workshop on Policies for Distributed Systems & Networks. IEEE, 2002.
- [10] Lin B., Shan Z. G., Sheng L. J., Wu J. P. Research on Internet Distinguished Service and its several hot issues [J]. Beijing: Journal of Computer Science, 2000.
- [11] Venkatesh K, Srinivas L N B, Krishnan M B M, et al. QoS improvisation of delay sensitive communication using SDN based multipath routing for medical applications [J]. Future Generation Computer Systems, 2019, 93: 256-265.
- [12] Cao X, Popescu I, Chen G, et al. Optimal and dynamic virtual datacenter provisioning over metro-embedded datacenters with holistic SDN orchestration [J]. Optical Switching and Networking, 2017, 24: 1-11.
- [13] Zhang Y, Wang HJ, Wei G. Design and implementation of distributed resource reservation protocol in UWB networks [J]. Communication Technology, 2008, 41(12): 146-148.
- [14] Hong D, Kim J, Jeong J P. A congestion contribution based traffic engineering scheme using software-defined Networking [C]// 2018 International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2018: 263-267.
- [15] Li Shenhao. Research and implementation of traffic classification-based routing optimization technology in SDN [D]. Beijing University of Posts and Telecommunications, 2021. doi:10.26969/d.cnki.gbydu.2021.000431.
- [16] Li Y, Han L. MPLS traffic engineering and its path management mechanism implementation [J]. China Management Informationization, 2020, 23(24): 117-118.
- [17] White paper on "New Infrastructure" policy (above) [N]. Institute of Industrial Policy, Institute of Policy and Regulation, Sadie Intelligence. China Computer News. 2020-09-07 (008).
- [18] White Paper on "New Infrastructure" Policy (below) [N]. Institute of Industrial Policy, Institute of Policy and Regulation, Sadie Intelligence. China Computer News. 2020-09-14 (008).
- [19] Li He Wu, Wu Xi, Xu Sc, et al. Progress and trends in research on integrated networks between heaven and earth [J]. Science and Technology Herald, 2016, 34(014): 95-106.
- [20] Nie Xiaoxue. Research on SDN data center traffic scheduling algorithm based on link state [D]. Inner Mongolia University of Technology, 2021. doi:10.27225/d.cnki.gnmgu.2021.000456.
- [21] Xu XY, Dai XF, Xia J, Li WC. A SDN-based load balancing routing algorithm for data centers [J]. Ship Electronics Engineering, 2021, 41(07): 129-132.
- [22] Benson T, Akella A, Maltz D A. Network Traffic Characteristics of Data Centers in the Wild [C]// ACM SIGCOMM Conference on Internet Measurement. ACM, 2010: 267-280.
- [23] FENG X B, YANG M C, GUO Q, et al. A novel distributed routing algorithm based on data driven in GEO/LEO hybrid satellite network [C]// Proc. of the International Conference on Wireless Communications and Signal Processing, 2015.
- [24] Dijkstra E W. A Note on Two Problems in Connexion with Graphs [J]. Numerische Mathematik, 1959, 1(1): 269-271.
- [25] Liu L S, Lin J F, Yao J X, et al. Path Planning for Smart Car Based on Dijkstra Algorithm and Dynamic Window Approach [J]. Wireless Communications and Mobile Computing, 2021, 2021(4): 1-12.
- [26] Li Chenxi. Research on network representation learning and routing optimization for software-defined networks [D]. Beijing University of Posts and Telecommunications, 2021. DOI:10.26969/d.cnki.gbydu.2021.000385.
- [27] Kong Xiangbin, Shen Subin, Li Li. A QoS routing scheme based on software defined networking [J]. Computer Technology and Development, 2018, 28(2): 102-106.
- [28] Mao Yan. Research on SDN-based traffic scheduling technology [D]. Chengdu: Southwest Jiaotong University, 2018.
- [29] Barros B M D, Jr M A S, Rojas M A T, et al. Applying Software defined Networks to Cloud Computing [M]// 33rd Brazilian Symposium on Computer Networks and Distributed Systems. 2015.
- [30] Guo Z, Liu C, Feng Y, et al. CCSA: A Cloud Computing Service Architecture for Sensor Networks [C]// International Conference on Cloud and Service Computing, 2012: 25-31.
- [31] PARSAEIMR, MOHAMMADIR, JAVIDANR, et al. A new adaptive traffic engineering method for telesurgery using ACO algorithm over software defined networks [J]. European Research in Telemedicine, 2017, 6(3/4): 173-180.

- [32] LI Honghui, YANG Guang, LU Hailiang, *et al.* Research on SDN data center network elephant flow scheduling based on ant colony algorithm[J]. Computer Application Research, 2019, 36(12): 3837-3841.
- [33] Li M. Research on SDN routing and flow table update strategy based on ant colony optimization algorithm [D]. Hebei University, 2021. DOI:10.27103/d.cnki.ghebu.2021.001098.
- [34] Puris A, Bello R, Herrera F. Analysis of the efficacy of a Two-Stage methodology for ant colony optimization: Case of study with TSP and QAP[J]. Expert Systems with Applications, 2015, 37(7): 5443-5453.
- [35] Wang C, Zhang G, Xu H, *et al.* An ACO-based link load-balancing algorithm in SDN [C]. In 2016 7th International Conference on Cloud Computing and Big Data (CCBD). IEEE, 2016: 214-218.
- [36] Liu ZP, Zhang QW, Li M, Li ZEY, Li SF. SDN routing strategy based on ant colony optimization algorithm [J/OL]. Journal of Kunming University of Science and Technology (Natural Science Edition): 1-11 [2022-05-22]. DOI:10.16112/j.cnki.53-1223/n.2022.
- [37] 03.132.
- [38] Wang Guizhi. Research on SDN intelligent routing optimization based on data-driven traffic awareness [D]. Sichuan University, 2021. DOI:10.27342/d.cnki.gscdu.2021.000355.
- [39] Al-Fares M, Radhakrishnan S, Raghavan B, *et al.* Hedera: Dynamic flow Scheduling for data center networks [C]// Proceedings of the 7th US ENIX Symposium on Networked Systems Design and Implementation, NSDI 2010, April 28-30, 2010, San Jose, CA, USA. dblp, 2017, 451-453.
- [40] WANG J, DELAATC, ZHAO Z, *et al.* QoS aware virtual SDN network planning [C]// IEEE Symposium on Integrated Network and Service Management, 2017: 644-647.
- [41] Zhou Peng. Research on SDN routing optimization based on reinforcement learning [D]. Chongqing University of Posts and Telecommunications, 2020. doi:10.27675/d.cnki.gcydx.2020.000517.
- [42] Wang Guizhi, Lv Guanghong, Jia Wuzai, Jia Chuanghui, Zhang Jianshin. A review of research on the application of machine learning in SDN routing optimization[J]. Computer Research and Development, 2020, 57(04): 688-698.
- [43] Li Shenhao. Research and implementation of traffic classification-based routing optimization technology in SDN [D]. Beijing University of Posts and Telecommunications, 2021. doi:10.26969/d.cnki.gbydu.2021.000431.
- [44] Canziani A, Paszke A, Culurciello E. An analysis of deep neural network models for practical applications [EB/OL]. (2017-04-14). An analysis of deep neural network models for practical applications [EB/OL]. <https://arxiv.org/abs/1605.07678>.
- [45] Azzouni A, Boutaba R, Pujolle G. NeuRoute: Predictive dynamic routing for software-defined networks [C]// Proc of the 13th Int Conf on Network and Service Management (CNSM). Piscataway, NJ: IEEE, 2017: 1-6.
- [46] Li Yanjun, Li Xiaobo, Osamu Y. Traffic engineering framework with machine learning based meta layer in software defined networks [C]// Proc of the 4th IEEE Int Conf On Network Infrastructure and Digital Content. Piscataway, NJ: IEEE, 2014: 121-125.
- [47] Kato N, Fadlullah Z M, Mao B, *et al.* The deep learning vision for heterogeneous network traffic control: proposal, challenges, and future perspective[J]. IEEE Wireless Communications, 2016, 24(3): 146-153.
- [48] Zhuang Zirui. Research on key technologies of routing for knowledge-defined networks [D]. Beijing: Beijing University of Posts and Telecommunications, 2020.
- [49] Azzouni A, Boutaba R, Pujolle G. NeuRoute: Predictive dynamic routing for software defined networks [C]. International Conference on Network and Service Management, IEEE Computer Society, 2017.
- [50] Zhang ZQ. Design and implementation of QoS optimization for SDN networks based on ONOS [D]. University of Electronic Science and Technology, 2021. doi:10.27005/d.cnki.gdzku.2021.001885.
- [51] Sutton R S, Barto A G. Reinforcement learning: an introduction [M]. Cambridge, MA: MIT Press, 2018.
- [52] S. C. Lin, Akylidiz F, W Pu, *et al.* QoS-Aware Adaptive Routing in Multi-layer Hierarchical Software Defined Networks: A Reinforcement Learning Approach [C]// IEEE International Conference on Services Computing. IEEE, 2016.
- [53] Che Xiangbei, Kang Wenqian, Ouyang Yuhong, Yang Kehan, Li Jian. A reinforcement learning based SDN routing optimization algorithm[J]. Computer Engineering and Applications, 2021, 57(12): 93-98.
- [54] Farahnakian F, Ebrahimi M, Daneshmand M, *et al.* Q-learning based congestion aware routing algorithm for onchip network [C]// Proc of the 2nd IEEE International Conference on Networked Embedded Systems for Enterprise Applications. Piscataway, NJ: IEEE Press, 2011: 1-7.
- [55] Liu Kechi. Intelligent routing algorithm based on reinforcement learning under SDN [D]. Harbin Institute of Technology, 2021. doi:10.27061/d.cnki.ghgdu.2021.004897.
- [56] JIANG J, HU L, HAO P, *et al.* Q-FDBA: improving QoE fairness for video streaming [J]. Multimedia Tools and Applications, 2018, 77(9): 10787-10806.
- [57] Wu, Lingling. Research on traffic classification methods and scheduling optimization in SDN [D]. Chongqing University of Posts and Telecommunications, 2021. DOI:10.27675/d.cnki.gcydx.2021.000909.
- [58] Xu Zhiyuan, Tang Jian, Meng Jingsong, *et al.* Experience-driven networking: a deep reinforcement learning based approach [C]// Proc of IEEE Conference on Computer Communications. Piscataway, NJ: IEEE Press, 2018: 1871-1879.
- [59] BOUTABAR, SALAHUDDINMA, LIMAMN, *et al.* A comprehensive survey on machine learning for networking: evolution, applications and research opportunities [J]. Journal of Internet Services and Applications, 2018, 9(1): 1-99.
- [60] FADLULLAH ZM, TANG F, MAOB, *et al.* State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems [J]. IEEE Communications Surveys & Tutorials, 2017, 19(4): 2432-2455.
- [61] Hao X.Y., Lv G.H.. A review of SDN traffic engineering research based on machine learning[J]. Computer Application Research, 2022, 39(04): 961-967+977. DOI:10.19734/j.issn.1001-3695.2021.09.0394.

# TIRec: Transformer-based Invoice Text Recognition

Yanlan Chen\*

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China  
chen\_yl@bupt.edu.cn.

## ABSTRACT

A novel invoice text recognition model is proposed. In the past few years, researchers have explored text recognition methods with RNN-like structures to model semantic information. However, RNN-based approaches have some obvious drawbacks, such as the level-by-level decoding approach and the one-way serial transmission of semantic information, which greatly limit semantic information's effectiveness and computational efficiency. In contrast, invoice text has obvious contextual relationships due to its fixed text pattern, the text font in the invoice is more fixed and the complexity of the background is much lower than that of natural scenes. To further exploit these contextual relationships and adapt to the characteristics of invoice text, we propose a new text recognition framework inspired by Transformer [1]. Self-attention-based architectures, in particular Transformer, have been successful in natural language processing (NLP). It has demonstrated powerful semantic information modeling capabilities in NLP. Inspired by its success, we try to apply Transformer to invoice text recognition. Unlike the RNN-based approach, we reduce the parameters of the vision network used to extract image features, use the Convolutional Vision Transformer Attention module to capture the semantic information, and use the Transformer decoding module to decode all characters in parallel. We hope that this Transformer-based architecture can better model the semantic information in invoices while remaining lightweight. Meanwhile, we collected text images of more than 40,000 train invoices, VAT invoices, rolled invoices, and cab invoices. Experiments on the collected invoice text recognition dataset show that our approach outperforms previous methods in terms of accuracy and speed.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Object detection;

## KEYWORDS

Text recognition, Invoice, Convolutional Vision Transformer

### ACM Reference Format:

Yanlan Chen. 2023. TIRec: Transformer-based Invoice Text Recognition. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*

\*Corresponding author.

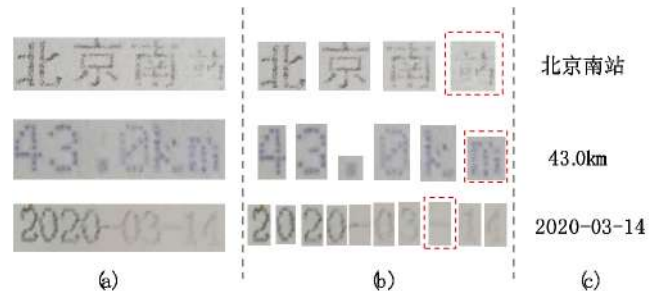
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590034>



**Figure 1: Examples of invoice text. (a) are some low-quality text images, (b) are separate character images after splitting, and (c) are the true text. The characters with dashed boxes in (b) are difficult to distinguish, only based on visual features.**

(CACML 2023), March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590034>

## 1 INTRODUCTION

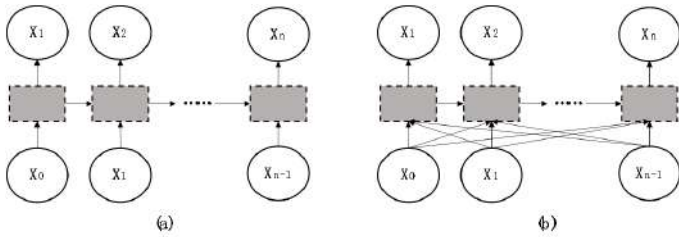
As a basic tool for human communication, texts have connotations and a wide range of uses. Text is ubiquitous in life, and with the development of recognition technology, automated tools for text play an important role in areas such as translation, retrieval, and autopilot. In traditional optical character recognition systems, text recognition is an indispensable part. Due to the wide range of text appearances, text can appear in a wide variety of forms in different scenarios, including variations in color, shape, and background factors. Although research on text recognition has been conducted for decades, the challenges in this area are still considerable.

### 1.1 Status of current text recognizer

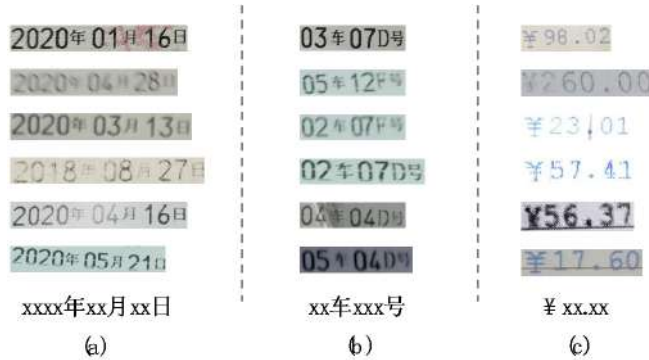
The important information on which text recognition is based is the visual features of the images, and usually convolutional neural networks are used to extract them. Some recent works have attempted to optimize the extraction, including upgrading the backbone network [6], adding another module [8], and improving the attention mechanism [9]. However, text images contain rich semantic information in addition to visual features. When reading, humans can often infer the missing words from contextual information. For example, as shown in Figure 1, if a text picture is cut into individual symbols, it is difficult to infer the characters in the dashed box.

Instead, considering semantic context information, human is easy to infer the correct characters with the total text.

Unfortunately, previous research has focused on one-way serial propagation of semantic information, such as recursively perceiving the text semantic information of the last decoding time step. There are several obvious drawbacks to this method: First, the semantic information it perceives in each decoding step is very limited, and



**Figure 2: Two different manners of semantic information encoding. (a) is unidirectional serial transmission, and (b) is the multiway parallel transmission**



**Figure 3: Some text patterns in the invoice text. (a) is the structure of the date, (b) is the structure of the train seat number, and (c) is the structure of the amount of money.**

even no useful semantic information is available in some of the earliest decoding steps. Second, the incorrect decoded information has a long-tail problem, which may propagate downward with the step length and accumulate with subsequent incorrect information, leading to biased modeling of semantic information by the model. Also, the time consumption of the serial approach increases linearly with the step length and tends to be time-consuming and inefficient compared to the parallel approach.

And invoice texts have a similar semantic structure from the human perspective. On some key characters, humans can infer the category of this character from a priori information. Especially for fixed category invoices, the information of the semantic model is more obvious, and even without visual information, some key characters can be identified based on prior knowledge. If this information can be modeled effectively, the accuracy of text recognition can be further improved.

As shown in Figure 3, we selected images with the same text patterns from the dataset and used expressions to express these text patterns, and there are many of these text patterns in the invoice text. We want the model to learn and understand these text patterns so that it can complete and correct some characters in case of poor image quality

## 1.2 Our contributions

To mitigate the limitations of the previous, we propose a text recognition model for invoice text inspired by Transformer. Our model directly encodes text information in 2D space and decodes by Transformer. To achieve this goal, we first use the Convolutional Vision Transformer Attention module to capture the global contextual information. Then, we map all 2D feature maps to feature vectors. At last, we use a standard Transformer Decoder to get the output probabilities.

In contrast to the previous ones, our model does not have a complex backbone such as Resnet [16] or VGG [17] to extract visual features, and we put more parameters on modeling semantic information to better adapt to the ticket text. With the Convolutional Vision Transformer Attention module, our framework can model local and global contextual information in one step to process different texts more efficiently. Our approach is also faster because the proposed Transformer-based module has a smaller number of parameters compared to previous RNN-based architectures.

To verify our model, we conduct experiments on our collected invoice text dataset. We find an effective way to synthesize data that can improve the performance of the pre-training network on real data. Meanwhile, our model achieved better results on the dataset which demonstrate the advantages of the proposed algorithm. The accuracy of our method beats the previous method [12], [11], and [13] by 0.8%, 0.3%, and 0.5%, respectively. In addition, the number of parameters of our method is significantly reduced and the inference speed is faster.

In general, the major contribution of this paper consists of two parts: 1) An effective and efficient text recognition algorithm was proposed which is designed with the Convolutional Vision Transformer Attention module and achieved outstanding results on our collected ticket text dataset. 2) We construct textual datasets for train invoices, VAT invoices, rolled invoices, and cab invoices, and propose an efficient data synthesis method.

## 2 RELATED WORK

### 2.1 Scene text recognition

In recent years, plenty of methods for scene text recognition have been proposed. The traditional methods are mostly character-based or word-based, and later sequence-based methods have been developed. The early works are mostly character-based. They first segment the text image to get a picture of individual images and then recognize the individual characters. In most cases, this approach classifies the character pictures using features produced by school workers, and then the classification results are composed into a paragraph of text. Recently, researchers tried to combine convolution-based neural networks with recurrent neural networks. Some researchers attempted to solve the text recognition problem with multi-class classification based on synthetic images, which classify common English words directly. Shi et al. [12] suggested the use of recurrent neural networks to process different numbers of characters in the text. And they first convert the input images into feature sequences using CNN and RNN and then use CTC to obtain recognition results. In [7], the sequence is generated by the attention model.

## 2.2 Vision Transformer

With the introduction of the Vision Transformer, researchers discovered the capability of the Transformer on vision tasks. ViT [4] first achieved state-of-the-art results on classification tasks (i.e. on JFT-300M [5]). In particular, ViT cuts the input image into fixed-size blocks, which are then position-encoded and used as input to the standard Transformer layer, consisting of Multi-Head Self-Attention modules and Position-wise Feed-forward module (FFN).

Some concurrent work proposes design changes to better model the local environment in visual deformers. For example, the Conditional Position Encoding Visual Transformer (CPVT) [15] introduces a convolutional block with zero-padding to implicitly encode the location information, eliminating the fixed positional embedding in the ViT, thus obtaining stable performance when the input image size changes. Swin Transformer [10] uses shift windows to restrict self-attention computations to non-overlapping local windows, leading to better efficiency. Token-to-token (T2T) [2] uses a convolutional scribing window similar to that in CNN to locally aggregate adjacent tokens, which helps to model local features. However, this operation is completely different from convolution, especially in the normalization details, and the concatenation of multiple labels greatly increases the computational and memory complexity. PVT [3] adopts a multi-scale multi-stage design (without convolution) similar to CNN in Transformer, tending to intensive prediction tasks.

## 2.3 Convolutional vision Transformer

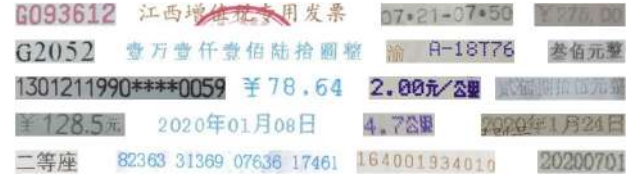
In computer vision and NLP (Natural Language Processing), convolutions have been exploited to improve Transformer blocks, either replacing multi-head attention with convolutional layers or adding additional convolutional layers in parallel or sequentially to capture additional relationships. Other previous work proposes to propagate the attention map to subsequent layers via residual connections and the attention maps are first extracted by convolution. Different from these works, Convolutional vision Transformer introduces convolution into Vision Transformer, which replaces the position-linear projection for attention operation with convolutional projection. This unique design gives it a better advantage than previous work in terms of performance and efficiency.

# 3 METHODOLOGY

## 3.1 Dataset

Due to the lack of Chinese invoice text datasets, we collect some invoices and crop images. The dataset includes approximately 40 thousand images with word-level labels from train invoices, taxi invoices, VAT invoices, and rolled invoices, as shown in Figure 4.

We also synthesized some pictures. The synthetic dataset consists of about 1 million images and about 6 million instances of synthetic words. Each image is annotated with its text. We used background images, font files, and corpus files to generate images and labels. Since the background of invoice text differs greatly from the text background of natural scenes and each type of invoice has a relatively similar background, we extracted the backgrounds of each type of ticket when synthesizing the data, as shown in Figure



**Figure 4: Samples in our collected real invoice text data. From left to right, each column shows train invoices, VAT invoices, cab invoices, and rolled invoices respectively.**



**Figure 5: Visualization of some background image samples used in data synthesis.**

**Table 1: The amount of the invoice dataset**

Type of invoice	Images
Train invoices	10684
Taxi invoices	12359
VAT invoices	11352
Rolled invoices	8785

5. At the same time, we choose fonts similar to those appearing in the ticket text to generate the text on the background map.

To enhance the robustness of the dataset, we introduced random image transformation. We analyzed the image transformations present in the real invoice text and grouped them into four categories, which are image rotation, blurring, filling noise, and adding lines, as shown in Figure 6. In addition, we randomly combine the above methods to fit more complex real data.

The invoice text dataset includes train invoices, VAT invoices, rolled invoices, and cab invoices, each of which has about 10,000 pictures and corresponding text labels.

## 3.2 Network Structure

The network structure is shown in Figure 7. Given an input image, we first use a Convolutional Vision Transformer Attention module to transform the input image into feature maps. Then, feature vectors were transformed from the identically located pixel of the feature maps. After that, the character decoder decodes the feature vectors into characters.



Figure 6: Four kinds of image transformation methods and corresponding real pictures.

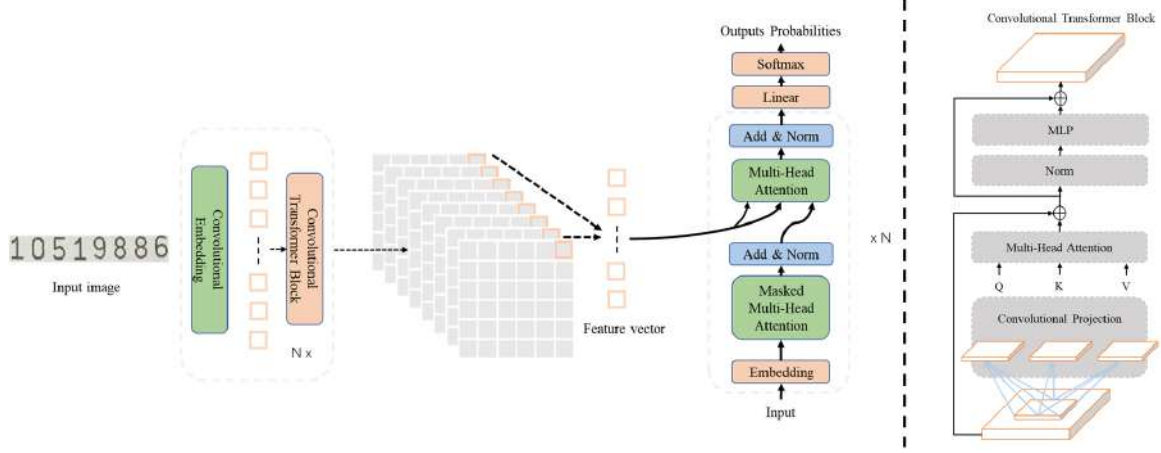


Figure 7: The architecture of our approach. (a) The overall architecture, including the Convolutional Transformer Encoder and the standard Transformer Decoder. (b) Details of the Convolutional Transformer Block, which takes a convolution layer as the first layer

**3.2.1 Convolutional Vision Transformer Attention module.** The Convolutional Vision Transformer introduces convolutional projections into the architecture of the Vision Transformer. As shown in Figure 7, the input image is submitted to a convolutional embedding layer, which is implemented as a convolution of overlapping patches with a max-pooling layer. This allows the number of feature maps reduce step by step while increasing the width (i.e., feature dimensionality), enabling spatial downsampling and increasing representation richness. Unlike other previous Transformer-based architectures [4], this model does not perform ad-style position embedding of markers. Next, the proposed stack of Convolutional Transformer blocks processes the output of the convolutional embedding layer. Figure 7(b) shows the structure of the Convolutional Transformer block, where a depthwise separable convolution operation is applied to the query, key, and value embeddings respectively, instead of the one in ViT [4]. Finally, we reshape the output feature map and permute the matrix to generate each feature vector from pixels at the same position in the feature map.

**3.2.2 Transformer Decoder module.** The decoder also consists of stacked standard Transformer Decoder layers. In addition to the Multi-Head Self-Attention module and the Position-wise Feed-forward module, the decoder performs masked multi-head attention on the output of the encoder. We also use a mask in the self-attention

sublayer of the decoder to prevent the preceding position from paying attention to subsequent positions. This mask helps the network process inputs in parallel during the training phase. Finally, the output of the decoder is passed through a linear layer and the probability distribution is obtained by a softmax layer.

## 4 EXPERIMENTS

We conduct experiments on the invoice text dataset to verify the advantages of our proposed model. The model is pre-trained on synthetic datasets, fine-tuned, and evaluated on real text datasets. Section 4.1 gives the network settings, dictionary set settings, and training parameters. In Section 4.2, the results of the comparison with the synthesis are reported.

### 4.1 Implementation Details

**4.1.1 Network Settings.** The Convolutional Vision Transformer Encoder of our model is adapted from the Vision Transformer Block [14]. Specifically, we modify the first embedding convolution layers for feature maps down-sampling to shape-invariant layers which contain four  $3 \times 3$ -stride convolutions. Meanwhile, we add max-pool layers behind each convolution layer to down-sampling the feature map.

As for the Transformer block in the encoder module, the hidden size is set to 512 and the number of self-attention heads is set to 8.

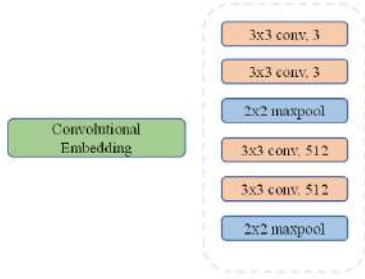


Figure 8: The convolutional embedding layer

In the decoder module, the max-length of texts is set to 42, because the lengths of Chinese sentences in invoices are shorter than 42. Meanwhile, we set the category of the characters to 6627, following the setting in [19], which consists of digits (0-9), English characters (a-z/A-Z), commonly used symbols, commonly used Chinese characters, a symbol “EOS” representing the end, a symbol “START” representing the start of the text, a symbol “PAD” representing padding of a sequence and a symbol “UNK” representing other characters.

**4.1.2 Training.** In the training phase, the input image is resized to  $32 \times 280$ . In addition, we train with the ADADELTA [20] optimizer. The batch size is set to 64. Empirically, the initial learning rate is 0.001 and it will reduce to 0.0001 at 3K iterations. The training stage ends at round 2 on synthetic images and round 100 on real images. When it comes to inference, we also adjust the input to  $32 \times 280$ . We decode the output into character sequences according to the following rules: 1) For each output vector, the character with the highest probability is used as the predicted character. 2) All “UNK” and “PAD” symbols will be deleted. 3) The decoding process will be terminated when the first “EOS” is encountered. We use CTC [18] loss for backward.

**4.1.3 Environment.** We implement our approach using PyTorch and perform experiments on a Linux server. Meanwhile, we train our model with three NVIDIA Tesla T4 GPUs and evaluate with one.

## 4.2 Experimental Results

In this section, we evaluate our model on the invoice datasets and compare its performance with other methods. The results are summarized in Tables 2 and 3



Figure 9: Hard examples successfully recognized by TIRec

On the ticket text recognition datasets, our approach outperforms the compared methods. In particular, our approach gives accuracy increases of 0.8% (88.2% to 87.4%) on CRNN, 0.5% (88.2% to 87.7%) on SRN, and 0.3% (88.2% to 87.9%) on SAR. Meanwhile, the parameter of our methods is 3.21 times less than CRNN, 1.23 times less than SRN, and 2.05 times less than SAR, which means the proposed method is lighter and more accurate than other methods.

At the same time, collecting large amounts of data with word-level labeling is expensive. Our method of generating synthetic datasets can help the pre-training of the network through any amount of annotated data and improve the performance of the network on real data.

Observing the results, we found that images with blurred appearance, occlusion, and distortion were also frequently recognized (Figure 8), which indicates that joint learning of visual and semantic information through the Transformer is a promising direction.

## 5 CONCLUSION

In this paper, we propose a new method for invoice text recognition. Since invoice text has a more distinct textual pattern, the proposed method reduces the scale of the backbone network and recognizes text images by Convolutional Vision Transformer Attention module, which can directly retain and utilize the two-dimensional spatial information and semantic background information of the text, thus better modeling the semantic information. Moreover, benefiting from the efficiency of the proposed parallel Transformer Decoder, our approach is faster and smaller than previous methods. Since the background of the invoice text is rather homogeneous, we extract background from invoice images when synthesizing the data and introduced data augmentation at the same time. We evaluated our model on the invoice text recognition dataset. The excellent performance of the model demonstrates the effectiveness and efficiency of our proposed approach. In the future, we aim to recognize text in more complex scenarios.

Table 2: Comparison of accuracy

Method	Backbone	Accuracy (%)	Parameters
CRNN [12]	Resnet34	87.4	About 250 million
SRN [13]	Resnet50	87.7	About 96 million
SAR [11]	Resnet13	87.9	About 160 million
Ours	Convolutional Vision Transformer	88.2	About 78 million

**Table 3: Comparison of classified accuracy**

Method	Train invoices (%)	Taxi invoices (%)	VAT invoices (%)	Rolled invoices (%)
CRNN [12]	85.3	89.2	87.2	87.5
SRN [13]	87.4	88.6	87.0	87.5
SAR [11]	89.1	87.1	87.1	88.4
Ours	89.2	87.0	87.7	89.5

**Table 4: Comparison of training with and without Synthetic data**

Method	Accuracy with Synthetic data (%)	Accuracy without Synthetic data (%)
CRNN [12]	87.4	80.1
SRN [13]	87.7	80.2
SAR [11]	87.9	80.0
Ours	88.2	83.8

## REFERENCES

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.
- [2] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. arXiv preprint arXiv:2101.11986, 2021.
- [3] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122, 2021.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- [5] Sun C, Shrivastava A, Singh S, *et al.* Revisiting unreasonable effectiveness of data in deep learning era[C]//Proceedings of the IEEE international conference on computer vision. 2017: 843-852.
- [6] Minghui Liao, Pengyuan Lyu, Minghang He, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. IEEE transactions on pattern analysis and machine intelligence, 2019.
- [7] C. Lee and S. Osindero. Recursive recurrent nets with attention modeling for OCR in the wild. In CVPR, 2016.
- [8] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In CVPR, pages 4168–4176, 2016.
- [9] Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz. Attention-based extraction of structured information from street view imagery. In ICDAR, volume 1, pages 844–850. IEEE, 2017.
- [10] Liu Z, Lin Y, Cao Y, *et al.* Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.
- [11] Li, Hui, *et al.* "Show, attend and read: A simple and strong baseline for irregular text recognition." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
- [12] Shi, B., B. Xiang, and Y. Cong. "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.11(2016):2298-2304.
- [13] Yu D, Li X, Zhang C, *et al.* Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 12113-12122.
- [14] Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., & Zhang, L. (2021). Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808.
- [15] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? arXiv preprint arXiv:2102.10882, 2021.
- [16] He K, Zhang X, Ren S, *et al.* Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [17] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
- [18] Graves, Alex, *et al.* "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." Proceedings of the 23rd international conference on Machine learning. 2006.
- [19] Du, Yuning, *et al.* "Pp-ocr: A practical ultra lightweight ocr system." arXiv preprint arXiv:2009.09941 (2020).
- [20] MD Zeiler. ADADELTA: An Adaptive Learning Rate Method[J]. arXiv e-prints, 2012.

# Two-channel Conformance Test Analysis of S-band Dual-polarization Radar

Yuxin Gong  
Key Laboratory for Meteorological  
Disaster Prevention and Mitigation of  
Shandong, Jinan 250031, China,  
Shandong Meteorological  
Engineering Technology Center, Jinan  
250031, China  
yvxingong@163.com

Chuancheng Ma  
Shandong Meteorological Service  
Center, Jinan 250031, China  
30347344@qq.com

Qian Zhang\*  
Key Laboratory for Meteorological  
Disaster Prevention and Mitigation of  
Shandong, Jinan 250031, China,  
Shandong Meteorological  
Engineering Technology Center, Jinan  
250031, China  
nuist\_zq@126.com

Yucheng Gong  
China University of Geosciences  
Beijing, Beijing 100083, China  
1665961892@qq.com

Xiqiang Yuan  
Shandong Meteorological Service  
Center, Jinan 250031, China  
yxq12121@163.com

Weijia Sun  
Shandong Meteorological  
Engineering Technology Center, Jinan  
250031, China  
425368535@qq.com

Juxiu Wu  
Shandong Meteorological Service  
Center, Jinan 250031, China  
gurunmin@163.com

## ABSTRACT

The consistency of the dual-channel radar plays a crucial role in the performance of the dual-polarization radar. In theory, the performance of the two channels is required to be completely consistent, but it cannot be completely consistent due to the influence of hardware errors, temperature and noise in practical applications. Therefore, it is necessary to test the consistency of horizontal and vertical channels of radar regularly in business applications. Aiming at this problem, the receiving system of Jinan S-band dual polarization radar is tested by off-line manual testing and online automatic testing. The offline measurement uses two signal sources inside and outside the machine to test separately, it is found that the output power difference between the two channels is too large. After the connection lines of the two channels and the two-channel power divider are exchanged, the output power of the two channels is basically the same. The test results of noise coefficient and echo intensity of the two channels are good and meet the requirement of consistency. CW signal source and TS signal source are used for online automatic test. The amplitude and phase standard deviation of the CW signal and the TS signal meet the requirements of the index. However, TS signal is used to calibrate the received full link, which increases the loss of azimuth rotation joint, so its amplitude

and phase standard difference are higher than CW signal. Therefore, it is necessary to test and correct the deviation caused by the rotation joint regularly after running for a long time for the dual polarization radar. The two measurement methods in this paper can effectively detect the dual-channel consistency of radar.

## CCS CONCEPTS

• **Hardware** → Hardware test; Board- and system-level test.

## KEYWORDS

S-band Dual-polarization Radar, Dual Channel Consistency, Dynamic Range, Noise, Error Analysis

### ACM Reference Format:

Yuxin Gong, Qian Zhang, Weijia Sun, Chuancheng Ma, Yucheng Gong, Juxiu Wu, and Xiqiang Yuan. 2023. Two-channel Conformance Test Analysis of S-band Dual-polarization Radar. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590035>

## 1 INTRODUCTION

Dual polarization Doppler weather radar has two transmission modes, namely simultaneous transmission and alternate transmission of electromagnetic waves in two directions (horizontal and vertical), and simultaneous reception of echoes in two directions. Bipolarized radar detects the target through the amplitude and phase changes of the echo. The type, shape, size and density of airborne precipitation particles can be obtained by using dual polarization radar [1].

On the basis of obtaining the single polarization information such as echo intensity  $Z$ , velocity spectrum width  $W$  and radial velocity  $V$ , dual-polarization radar adds the double polarization

\*Corresponding author: Zhang Qian, E-mail: nuist\_zq@126.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590035>

**Table 1: Jinan S band dual-polarization weather radar CINRAD/SA-D**

Parameter	Values
Antenna polarization mode	Double polarization (H and V)
Polarization mode	Dual polarization / Double sending and double receiving
Noise factor	$\leq 3\text{dB}$
Dynamic range	$\geq 95\text{dB}$
Factor of reflectivity $Z_H$	$\leq 1\text{dB}$
Radial velocity $V_r$ / Spectral width $S_W$	$\leq 1\text{m/s}$
Differential reflectance factor $Z_{DR}$	$\leq 0.2\text{dB}$
Differential propagation phase shift $\theta_{DP}$	$\leq 3^\circ$
Differential propagation phase shift rate $K_{DP}$	$\leq 0.2^\circ/\text{km}$
The correlation coefficient $C_C$	$\leq 0.01$

information such as differential phase shift rate  $K_{DP}$ , differential emissivity  $Z_{DR}$ , correlation coefficient  $C_C$  and differential phase shift  $\theta_{DP}$ , which is richer than the parameter information of single-polarization radar. It can more specifically analyze the size, shape, phase state, type and drop spectrum distribution of precipitation particles [6, 7]. Dual-polarization radar plays an important role in monitoring tornado, typhoon, rainstorm, hail and other strong weather systems. The dual polarization parameter is obtained from the difference between the horizontal and vertical echoes received by the radar [8]. The difference between the two directions is small [2, 5]. Therefore, in order to accurately extract the difference, external interference should be minimized to ensure that the horizontal and vertical channels of the radar are completely consistent [9-15]. However, due to the limitations of manufacturing process and hardware conditions, the actual manufacturing cannot be completely the same, so the difference can only be controlled within a certain range to reduce its impact on the accuracy of the results. In daily application, the two channels shall be tested regularly to correct errors in time.

Jinan's new generation weather radar completed the dual-polarization technology upgrade (CINRAD/SA-D) and passed the field test on May 29, 2019. This radar is the first S-band dual-polarization weather radar installed in the north of China. In order to ensure the consistency of the two channels, this paper tests the dual channels of Jinan S band dual polarization weather radar, and tests the main performance parameters of the receiver, such as dynamic range, echo intensity, noise coefficient and the amplitude and phase consistency of the receiving dual channels. In order to ensure the accuracy of the analysis results, this paper uses offline measurement analysis and online automatic calibration data analysis to test the radar dual-channel consistency. By combining offline and online, the system deviation of dual-polarization radar can be acquired and the error caused by time can be corrected.

## 2 S BAND DUAL POLARIZATION RADAR PERFORMANCE TEST TECHNOLOGY

The performance of S-wave dual-polarization radar can be tested in two ways: offline and online. Off-line test is a test of radar parameters completed when the radar is down. Online testing refers to the autonomous testing of radar links in accordance with the set time during radar operation [16].

**Figure 1: Flow chart of dual-polarization radar test**

As shown in Figure 1, Jinan CINRAD/SA radar has two calibration signal sources. The calibration signal source 1 is installed in the machine room and is responsible for checking the status of some links installed in the machine room. The calibration signal source 2 is responsible for testing the operation status of the whole receiving link and is installed on the back of the antenna reflector. The parameter accuracy of Jinan S band dual polarization Doppler weather radar is given in Table 1. [3, 4, 9, 17-19]

## 3 OFFLINE MEASUREMENT AND ANALYSIS

### 3.1 Test and analysis of receiver dynamic characteristics

**3.1.1 Internal signal source measurement.** The continuous wave test signal was injected with the internal signal source, and the dynamic range test data of narrow pulse ( $1.57\mu\text{s}$  pulse width) H (horizontal) channel and V (vertical) channel were obtained, as shown in Figure 2. The slope of the fitting line for H channel and V channel is 0.9905 and 0.9903, respectively, in the range of 0.9850~1.015. The

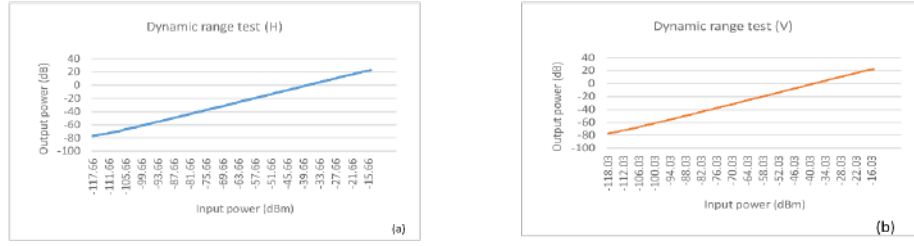


Figure 2: Dynamic range test data in the machine, (a) H channel, (b) V channel

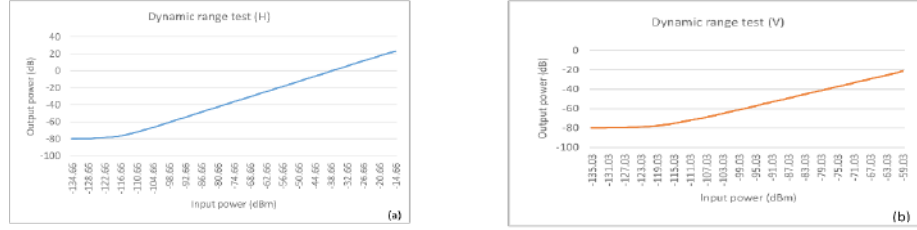


Figure 3: External dynamic range test data (a) H channel, (b) V channel

fitting root mean square error is 0.2388dB and 0.2248dB, both less than 0.5dB. The dynamic range is 100dB and 99dB, both greater than 95dB. The dynamic characteristics of the radar meet the index requirements. The difference in slope ( $\Delta S_{lope}$ ) of the fitted straight lines for the H and V channels is 0.0043. The error value ( $\Delta Z_{DR}$ ) [10] of differential reflectance  $\Delta Z_{DR}$  caused by  $\Delta S_{lope}$  is

$$\Delta Z_{DR} = Z_{DR} \cdot \Delta S_{lope} \quad (1)$$

Therefore, the error value  $\Delta Z_{DR}$  caused by  $\Delta S_{lope}$  is 50dBz and 70dBz respectively when the echo intensity is 0.01dB and 0.014dB, which is far lower than 0.2dB. The dynamic characteristics of the radar plane are good.

**3.1.2 External signal source measurement.** An external signal source (Agilent E4428C) was used to inject CW test signals, and the dynamic range test data of narrow pulse H channel and V channel were obtained, as shown in Figure 3. The slope of the fitted line for H and V channels is 0.9853 and 0.9890, the root mean square error of the fitted line is 0.3886dB and 0.2437dB, and the dynamic range is 101dB and 99dB, respectively, which are within the range required by the index. The  $\Delta S_{lope}$  is 0.0002. Therefore, the  $\Delta Z_{DR}$  is 0.185dB when the echo is 50dBz, which is lower than the index with the  $Z_{DR}$  calibration error of 0.2dB. But when the echo intensity is 70dBz,  $Z_{DR}$  is 0.259dB, slightly greater than 0.2dB. This is because there are some errors in the calibration of external signal sources and some losses caused by connecting cables, so the measurement results will be slightly larger.

**3.1.3 Offline dual-channel consistency measurement analysis.** In order to avoid large errors in the test results of the two channels under different SNR, the amplitude and phase consistency of the two channels were tested under the same SNR. After inputting CW test signal, the intensity difference  $Z_{DR}$  and phase difference  $\theta_{DP}$  of dual-channel dynamic range, H and V channel signal of

the receiver are obtained. The dynamic range test results of the two channels (Figure 4) show that the output power difference between H and V channels is too large, exceeding 3dB, and the  $Z_{DR}$  products generated by the radar are obviously too large. Due to the simultaneous problems of the two channels, the preliminary judgment is that the two-channel power splitter of the radar is damaged. After the replacement of the two-channel power splitter, the dynamic range test is carried out. The difference between the H and the V channel is still about 3dB, and the fault is not removed. After the connection lines of the two channels of the two-channel power divider were switched, the problem was solved. And the  $Z_{DR}$  product is restored (Figure 5(d)(e)). The output power of the two channels were basically the same, as shown in Figure 5(a), and the test results of  $Z_{DR}$  and  $\theta_{DP}$  are shown in Figure 5(b)(c).

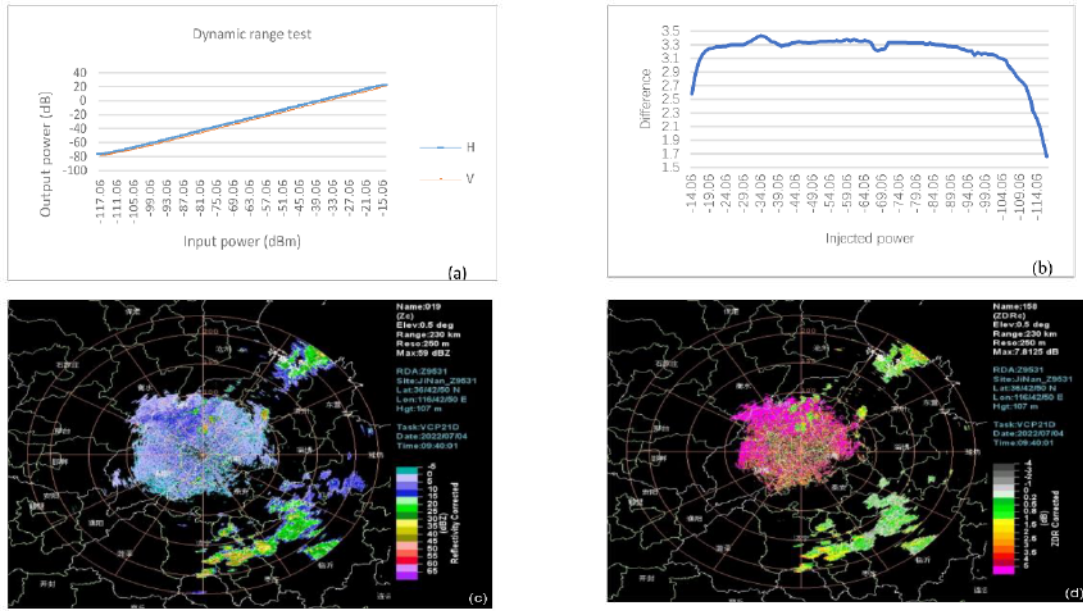
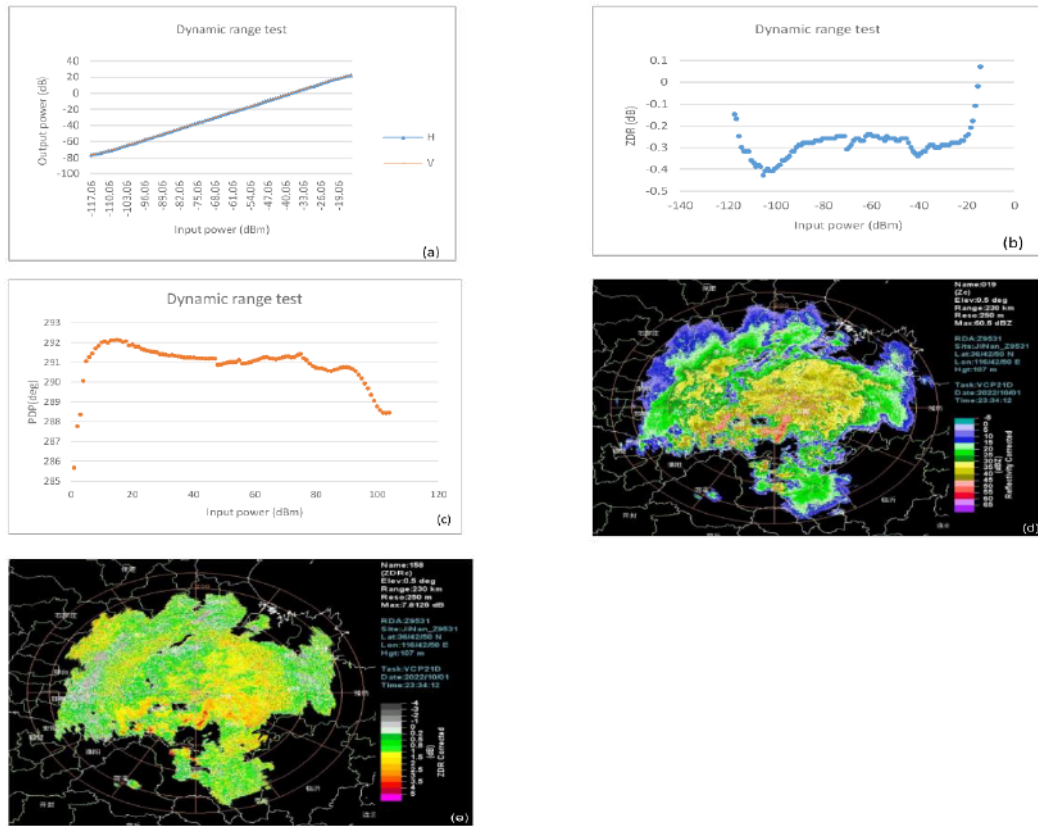
To further check the consistency of the dynamic range of the two channels, calculate the standard deviation of  $Z_{DR}$  and  $\theta_{DP}$ . From the point where the signal-to-noise ratio is not lower than 0.2dB to the end of 10 values below the starting point of the dynamic range, that is, -23.83~-95.83dBm, it can be calculated that the standard deviations of  $Z_{DR}$  and  $\theta_{DP}$  are 0.04dB and  $0.34^\circ$  respectively, which are far less than the index requirements of 0.2dB and  $3.0^\circ$ . The amplitude consistency of the two receiving channels is good in the full dynamic range, and the fluctuation is very small. Through comprehensive analysis, the radar receiving system has good consistency of dynamic range dual channel.

### 3.2 Noise coefficient measurement

Input noise signal from the front end of the receiver, test at the terminal, get the noise coefficient  $N_F$ . The calculation formula is

$$N_F = ENR - 10 \lg [(P_{hot}/P_{cold}) - 1] \quad (2)$$

Among them,  $ENR$  is the effective excess noise ratio, which is related to the type of noise source and the operating frequency of

Figure 4: (a) Two channel dynamic range test results, (b) Two channel dynamic range difference, (c) Z, (d)  $Z_{DR}$ Figure 5:  $Z_{DR}$  and  $\theta_{DP}$  dynamic range test data, (a) in-machine two-channel dynamic range, (b) in-machine  $Z_{DR}$ , and (c) in-machine  $\theta_{DP}$ , (d) Z, (e)  $Z_{DR}$

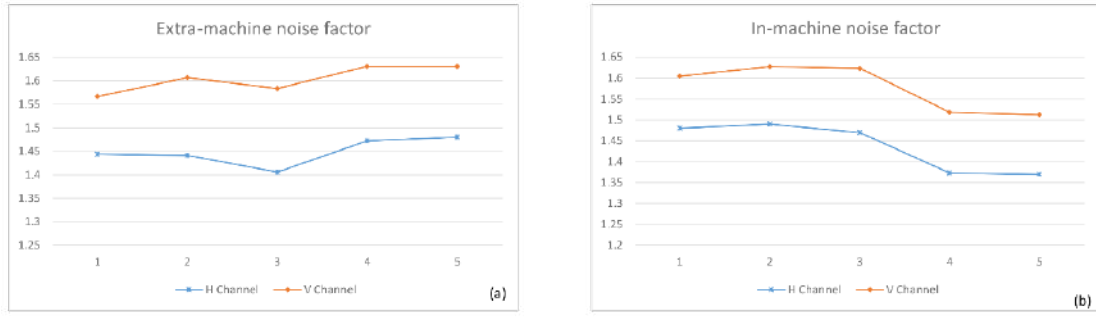


Figure 6: (a) Test data of noise coefficient outside the machine, (b) Test data of noise coefficient in the machine

Table 2: Radar parameters

Parameter	Value
wavelength $\lambda$ /cm	10.59
Antenna gain $G$ /dB	H: 45.27 V: 44.86
Transmit pulse power $P_t$ /kw	664
Pulse width $\tau$ / $\mu$ s	Narrow pulse: 1.57 Wide pulse: 4.64
Total system loss except $L_{at}L_{\Sigma}$ /dB	H: 6.49 V: 9.41
Horizontal beam width $\theta$ /( $^{\circ}$ )	0.967
Vertical beam width $\phi$ /( $^{\circ}$ )	0.911
Distance detection $R$ /km	5~200
Atmospheric loss $L_{at}$ /(dB $\cdot$ km $^{-1}$ )	0.016 (round-trip)
Input signal power $P_r$ /dBm	H: -43.83 ~ -93.83 V: -43.96 ~ -93.96

the measured radar. In this paper,  $ENR$  is 14.62dB.  $P_{hot}$  refers to the noise level of the receiver when the internal or external noise source is accessed. When the internal or external noise source is not accessed, the noise level of the receiver is  $P_{cold}$ .

Inject the internal and external noise signals respectively to obtain the noise coefficients of the two channels of the receiver (Figure 6). The mean values of external noise factor of the two channels are 1.45dB and 1.60dB respectively. The mean internal noise factor is 1.44dB and 1.58dB respectively. The noise inside and outside the machine is less than 3.0dB, meeting the consistency requirements. The maximum difference between the internal and external noise coefficients of the two channels is 0.12dB, which is far less than 0.3dB. In general, the noise coefficients of the two channels have little difference and good consistency.

### 3.3 System echo intensity calibration test and analysis

-94dBm~-44dBm signals were injected at the input respectively to obtain the measured value of the echo strength within the range of 5km~200km, and the difference between the measured value and the calculated expected value was calculated. The difference between the measured value and the expected value should be no

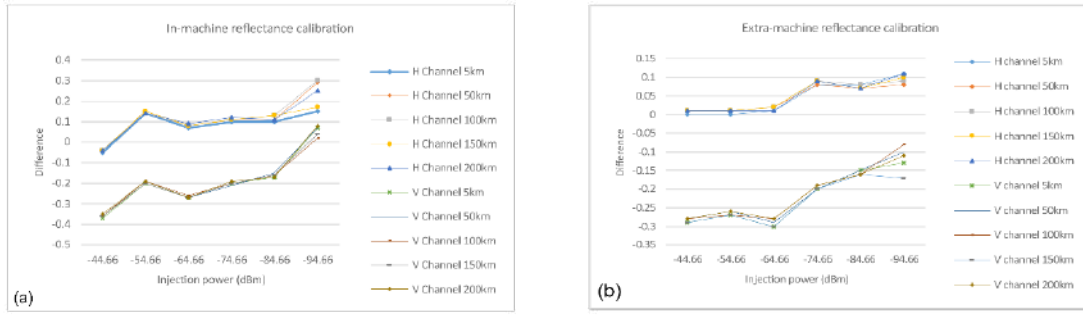
more than  $\pm 1$ dB. The expected value is calculated as follows

$$10lgZ = 10lg \left[ \left( 2.69 \times 10^6 \lambda^2 \right) / \left( P_t G^2 \theta \phi \right) \right] + P_r + 20lgR + L_{\Sigma} + RL_{at} \\ = C + P_r + 20lgR + RL_{at} \quad (3)$$

Among them

$$C = 10lg \left[ \left( 2.69 \times \lambda^2 \right) / \left( P_t \tau \theta \phi \right) \right] - 2G + 160 + L_{\Sigma} \quad (4)$$

Echo intensity calibration was carried out for the two channels of the radar system. The measured values and expected values of echo intensity of the H channel and the V channel were different. The test result internal and External machine were shown in Figure 7. When the power is large, the difference between the echo intensity values at different distances and the expected value is small. As the power decreases, the loss becomes large, and the difference between the expected value and the measured value gradually increases. The maximum difference of the horizontal channel in the aircraft is 0.25dB when the injection power is -94.66dBm and the distance is 100km, and the maximum difference of the vertical channel is -0.37dB when the injection power is -94.66dBm and the distance is 5km. The maximum difference of the horizontal channel outside the machine is 0.11dB, which appears when the injection power is -94.66dBm and the distance is 200km. The maximum difference of the



**Figure 7: (a) Calibration test of echo strength by signal source in the machine, (b) Calibration test of echo strength of external signal source**

vertical channel is -0.29dB, which appears when the injection power is -64.66dB and the distance is 50km. In general, the fluctuation range of the difference is small, the maximum difference is within the range of  $\pm 1$ dB, and the difference of the maximum difference between the two channels is not big, and the calibration consistency of the echo intensity of the two channels is good.

## 4 ONLINE AUTOMATIC CALIBRATION DATA ANALYSIS

### 4.1 Online calibration of noise figure

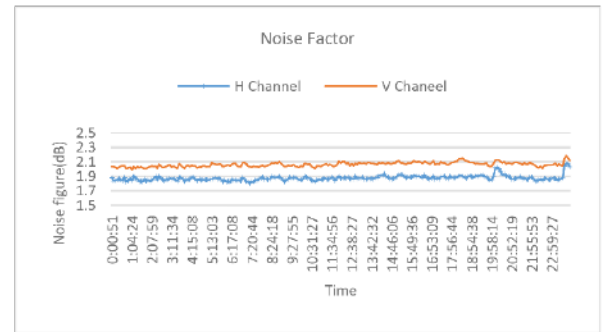
The dual-polarization radar calibrates the noise temperature ( $T_N$ ) once at each sweep interval, and the receiver noise coefficient  $N_F$  represents the ratio of the receiver input signal-to-noise ratio (SNR) to the output SNR. The noise coefficient of online calibration is obtained by the noise temperature conversion of each volume scan calibration. The conversion formula is

$$N_F = 10 \lg [T_N / 290 + 1] \quad (5)$$

Figure 8 shows the results of the continuous 24-hour online calibration of Jinan dual-polarization radar on February 10, 2021. It can be seen from the test results that the noise figure of the horizontal channel and the vertical channel of the receiver is basically stable, the fluctuation direction is consistent and the range is small, the maximum fluctuation range is , less than , and the maximum difference between the two channels is within , and the consistency of the two channels is very good.

### 4.2 Receive two-channel consistency analysis online

CW signals are calibrated at intervals of each volume scan, and the results of the 24-hour continuous calibration on February 10, 2022 are used in this article. The calibration results of the CW signal are shown in Figure 9(a)(b). The average value of the  $CW\_Z_{DR}$  is 1.017dB, the maximum value is 1.04dB, the minimum value is 1dB, and the standard deviation is 0.008dB. The mean value of the  $CW\_0_{DP}$  is  $119.38^\circ$ , the maximum value is  $119.56^\circ$ , the minimum value is  $119.13^\circ$ , and the standard deviation is  $0.084^\circ$ . The calibration results of  $CW\_Z_{DR}$  and  $CW\_0_{DP}$  are stable with small fluctuation range. TS signals are calibrated at each elevation angle



**Figure 8: Jinan dual-polarization radar online calibration noise figure on February 10, 2021**

and volumetric scan interval, and the results of the 24-hour continuous calibration on February 10, 2022 are shown in Figure 9(c)(d). The average value of the  $TS\_Z_{DR}$  is 1.35dB, the maximum value is 1.40dB, the minimum value is 1.30dB, and the standard deviation is 0.02dB. The mean value of the  $TS\_0_{DP}$  is  $239.63^\circ$ , the maximum value is  $241.43^\circ$ , the minimum is  $238.41^\circ$ , and the standard deviation is  $0.71^\circ$ . The amplitude standard deviation of both CW signal and TS signal calibration dual channels is less than 0.2dB, and the phase standard deviation is less than  $2^\circ$ , meeting the dual channel consistency requirements. The consistency of the two channels is good.

## 5 CONCLUSIONS

In this paper, the dual channel consistency of Jinan dual polarization Doppler weather radar is tested and analyzed by off-line manual test and online automatic test. In this paper, the dynamic range, noise figure, echo intensity of the system, and the consistency of the amplitude and phase of the dual channel are tested. The test and analysis conclusions are as follows:

1. In this paper, the dynamic characteristics, noise coefficient and echo intensity of Ji 'nan S band dual polarization radar are tested by off-line measurement method using two signal sources inside and outside the machine respectively. In the measurement, it is found that the output power difference between the two channels is too large. After the connection line between the two channels

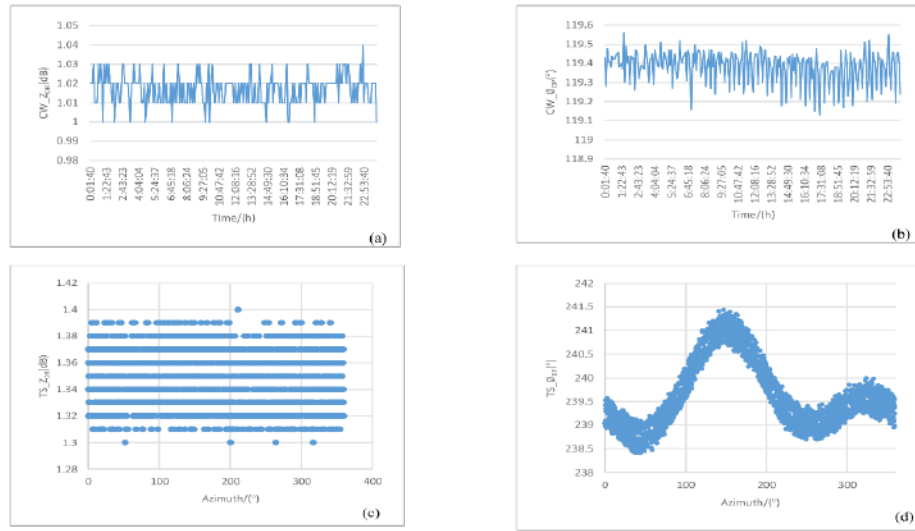


Figure 9: Calibration results of CW and TS signals, (a)  $CW\_Z_{DR}$ , (b)  $CW\_0_{DP}$ , (c)  $TS\_Z_{DR}$ , (d)  $TS\_0_{DP}$

and the two-way power divider is exchanged, the output power of the two channels is basically the same. The test results of noise coefficient and echo intensity of the two channels are good and meet the requirement of consistency.

2. The on-line test method is used to continuously calibrate the noise coefficient. The noise coefficient of the two channels is basically stable, the fluctuation direction is consistent and the range is small, and the consistency is good.

3. The amplitude and phase of the dual-channel signal of the receiver are tested online by CW signal and TS signal. The calibration results of  $CW\_Z_{DR}$ ,  $CW\_0_{DP}$ ,  $TS\_Z_{DR}$  and  $TS\_0_{DP}$  are stable, and the dual-channel consistency is good. However, since the TS signal passes through the azimuth rotation joint, the values of  $TS\_Z_{DR}$  and  $TS\_0_{DP}$  are larger than those of  $CW\_Z_{DR}$  and  $CW\_0_{DP}$ . Therefore, during the long-term operation of the dual-polarization radar, it is necessary to regularly test and correct the deviation caused by the azimuth rotation joint.

## REFERENCES

- [1] Yu Xiaoding, Yao Xiuping, Xiong Tingnan, *et al.* Principle and Application of Doppler Weather Radar [M]. Beijing: Meteorology Press, 2006. (in Chinese)
- [2] Zhang Peichang, Wei Ming, Huang Xingyou, *et al.* Principle and application of dual linear polarization Doppler weather radar [M]. Beijing: Meteorological Press, 2018. (in Chinese)
- [3] Zhang Peichang, Wei Ming, Huang Xingyou, *et al.* Principle and application of dual linear polarization Doppler weather radar [M]. Beijing: Meteorological Press, 2018. (in Chinese)
- [4] BRINGI V N, CHANDRASEKAR V. Principles and Applications of Polarization Doppler Weather Radar [M]. Li Chen, Zhang Yue, Yi, Zhang Peichang, Jiao, Beijing: Meteorological Press, 2018. (in Chinese)
- [5] LAKSHMANAN V, FRITZ A, SMITH T, *et al.* An automated technique to quality control radar reflectivity data[J]. J Appl Meteor Climatol, 2007, 46(3):288-305.
- [6] Du Muyun, Liu Liping, Hu Zhiqun, *et al.* Quality Analysis of Dual Linear Polarization Doppler Radar Data [J]. Acta Meteorology, 2013, 71 (1): 146-158. (in Chinese)
- [7] Zhao S, Qin X, Li S, *et al.* Application of CINRAD Weather Radar Common Products on Weather Modification in China[J]. Advances in Earth Science, 2012, 27(6): 694.
- [8] Chen X, Su D, Liu Y, *et al.* Comparison and Validation of Reflectivity of Dual Linear Polarization SA Radar and CINRAD SA Radar[C]//2019 International Conference on Meteorology Observations (ICMO). IEEE, 2019: 1-4.
- [9] Yang Chuanfeng, Diao Xiuguang, Zhang Qian, *et al.* Jinan Dual Polarization Doppler Weather Radar On-line Automatic Calibration Data Quality Analysis and Evaluation [J]. Journal of Marine Meteorology, 2020, 40 (4): 114-123. (in Chinese)
- [10] Hu Dongming, Zhang Yu, Fu Peiling, *et al.* Conformance test and analysis of dual-channel for Guangzhou S-band dual-linear polarization weather radar [J]. Meteorology, 2019, 47 (3): 373-379. (in Chinese)
- [11] Meng Qingchun, Shen Yonghai, Su Debin. Consistency and test method of dual-channel dual-polarization radar [J]. Plateau Meteorology, 2014,33 (5): 1440-1447. (in Chinese)
- [12] Wei Hongfeng, Xue Zhengang. Measurement error of differential reflectivity factor for dual-polarization multi-frequency weather radar [J]. Meteorology, 2008, 36 (2): 223-227.
- [13] Li Zhe, Wang Chongwen, Li Chunhua, *et al.* Engineering calibration method for differential reflectivity of dual-transmitter dual-receive dual-polarization weather radar [J]. Meteorology, 2014,42(6):951-956. (in Chinese)
- [14] Li Zhe, Li Bai, Zhao Kun, *et al.* Performance analysis of domestic dual polarization weather radar differential reflectivity measurement [J]. Meteorological science and technology, 2016,44(6):855-859. (in Chinese)
- [15] Zhao Shiyong, Li Bai, Chen Xiaohui, *et al.* Hardware calibration of dual-polarization radar differential reflectivity based on cross-parallel method [J]. Meteorology, 2015,43(5):775-782. (in Chinese)
- [16] Yu Haifeng. Application of dual linear polarization radar calibration technology [J]. Advances in meteorological science and technology, 2018,8(06):139-146. (in Chinese)
- [17] China Meteorological Administration. QX / T 464-2018 S-band dual polarization Doppler weather radar [S]. Beijing: China Meteorological Administration, 2018. (in Chinese)
- [18] Comprehensive Observation Department of China Meteorological Administration. New generation weather radar system factory acceptance test outline [Z]||Gas measuring function (2018) No. 70. Beijing: Comprehensive Observation Department of China Meteorological Administration, 2018. (in Chinese)
- [19] Comprehensive Observation Department of China Meteorological Administration. Field acceptance test outline of new generation weather radar system [Z]||Gas measuring function (2018) No. 70. Beijing: Comprehensive Observation Department of China Meteorological Administration, 2018. (in Chinese)

# Feature selection based on improved principal component analysis

Zhangyu Li

School of Economics and Management, Xiamen University  
of Technology, Xiamen, China  
lzy1173745655@outlook.com

Yihui Qiu

School of Economics and Management, Xiamen University  
of Technology, Xiamen  
China, qiuyihui@xmut.edu.cn

## ABSTRACT

**Abstract:** The filtered feature selection method has low computational complexity and less time, and is widely used in feature selection, but the filtered method only considers the importance of features for label classification and ignores the correlation between features. For this reason, a feature selection method with improved principal component analysis is proposed. The main idea of the method is that on the basis of principal components, the loadings of each indicator on different principal components and their variance contribution ratios with that principal component are considered. A number of indicators with the largest cumulative contribution rates were selected, so that the final extracted indicators retained more information. Subsequently, comparative experiments are conducted using the UCI dataset, and the results show that the approach proposed in this paper has some superiority over other methods. Finally, the features of China's green innovation efficiency are selected using the approach proposed in this paper to demonstrate the feasibility of the method.

## CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Feature selection.

## KEYWORDS

PCA, Feature selection, contribution rate

### ACM Reference Format:

Zhangyu Li and Yihui Qiu. 2023. Feature selection based on improved principal component analysis. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590036>

## 1 INTRODUCTION

Feature selection is an important pre-processing step in machine learning. Its purpose is to select some features from the original dataset while retaining as much of the underlying information as possible, so as to form the desired dataset. In this way, it reduces the dimensionality of the dataset, preventing the curse of dimensionality, and reducing training time while improving algorithm

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590036>

efficiency [1]. Currently, there are two main categories of feature selection methods. From the selection perspective, feature selection methods can be divided into filter, wrapper, and embedded. From the transformation perspective [2], the main methods are principal component analysis, independent component analysis, and popular learning algorithms, etc. These methods all aim to reduce high-dimensional data to a lower dimensional space, but they cannot extract the original and optimal subset of features. Kale [3] introduced a PCA-based optimal feature subset selection, which is capable of handling weighted classification problems. Chi G T [4] proposed a feature selection method based on principal component analysis. By retaining the indicators with large absolute factor loadings under the same criterion, the significance of the indicators to the evaluation results is ensured. Then, by eliminating one of the two highly correlated indicators, it ensures that only a small amount of information is needed to obtain an indicator that accounts for 80% of the variance of the original indicator. However, selecting indicators pairwise through correlation has subjectivity, which is not suitable for reducing or identifying important indicators [5]. Given the above situation, this paper proposes an improved method of principal component analysis for extracting the selection of evaluation indicators, using cumulative contribution rate to identify the most important variables, thus avoiding subjectivity.

## 2 FEATURE SELECTION METHOD BASED ON PRINCIPAL COMPONENT ANALYSIS

PCA is an effective method in statistical data analysis, which is mainly used for dimension reduction [6]. Its main principle is to use the transformation of the feature space of the dataset to reduce the dimensionality of the dataset that has a high dimension and is correlated. After dimension reduction by using PCA, the original dataset will be transformed into a dataset consisting of several principal components, which do not have correlation. However, after dimension reduction by using PCA, it will become a new feature to be used with the original feature. This paper proposes a feature selection method based on PCA. The pseudo code of the algorithm is shown in Table 1.

Suppose there are  $m$  input indicators, denoted as  $X = (X_1, X_2, \dots, X_m)^T$ ;  $m$  principal components are denoted as  $F = (F_1, F_2, \dots, F_m)$ , and  $(a_1, a_2, \dots, a_n)^T$  is the relative contribution value of  $m$  input indicators, namely  $X = (X_1, X_2, \dots, X_m)^T$ . The specific steps are as follows:

Step 1: Calculate the sample correlation matrix corresponding to the standardized variable  $X^* = (X_1^*, X_2^*, \dots, X_m^*)^T$ .

Step 2: Obtain the eigenvalues  $\lambda_1, \lambda_2, \dots, \lambda_m (\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0)$  and unit eigenvectors

**Table 1: Pseudo Code for Feature Selection Based on Principal Component Analysis**


---

Input: sample set $X=(X_1, X_2, \dots, X_m)^T$
Procedure:
1, Normalize the samples to obtain the correlation matrix $X^{**}$
2, Calculate the eigenvalues $\lambda$ , eigenvectors $l$ , and variance contribution rate $\sqrt{\omega}$
3, Calculate the cumulative contribution rate $a$ of each indicator
4, Take the indicator corresponding to the top 80% contribution rate
Output: the set $X^{**}$

---

**Table 2: Datasets Information**


---

Datasets	Total Number of Samples	Sample Categories	Number of Features
Iris	150	3	4
Heart-h	294	13	5
Blood	748	2	4
Segment	5456	4	2
Dermatology	385	34	6

---

$l_1, l_2, \dots, l_m$  from the sample correlation matrix, and calculate the contribution rate  $\omega_i = \lambda_i^* (\sum_{j=1}^m \lambda_j)^{-1} (i = 1, \dots, m)$ .

Step 3: In order to improve the influence of each index, the square root operation with calculated contribution rate of variance was executed. Then, the weight of the feature vector (indexes on each principal component) is multiplied by the variance contribution rate corresponding to belonging the principal component, and the results of the same index are added, i.e., the corresponding contribution value of each input variable can be calculated by equation  $(a_1, a_2, \dots, a_m)^T = |L^* (\sqrt{\omega_1}, \sqrt{\omega_2}, \dots, \sqrt{\omega_m})^T|$ , where the absolute value symbol represents taking the absolute value of the  $m$  elements in the column vector.

Step 4: Sort the elements in vector  $(a_1, a_2, \dots, a_n)$  in descending order to obtain vector  $(a_1^*, a_2^*, \dots, a_n^*)^T (a_1^* \geq a_2^* \geq \dots \geq a_m^* \geq 0)$

Step 5: According to the cumulative contribution rate criterion, select the top 80% of the  $k$  variables in terms of cumulative contribution rate to complete the feature selection.

### 3 COMPARISON EXPERIMENT

In order to verify the superiority and inferiority of the proposed feature selection method, this paper took the methods proposed by Singh [7] and Lim [8] as reference, and compared the proposed method with the Correlation-based Feature Selection (CFS), the Fisher Score, the Chi-square feature selection, the Relief, the minimum redundancy maximum relevance (mRMR) feature selection method and the Alpha-investing online stream feature selection algorithm. A series of performance metrics were used to evaluate feature selection, including the classification accuracy, the Normalized Mutual Information (NMI), the Jaccard coefficient, and F-value. The experiment used two classifiers, namely Support Vector Machine Classifier and Decision Tree Classifier, and In reference to Sahebi [9], five UCI data sets were adopted. The specific information of the datasets is shown in Table 1.

The paper measured the above indicators of the model through ten-fold cross validation. The principle of this method is: all training

samples are divided into ten parts as randomly as possible. Nine parts are selected as the training set and the remaining one as the test set for ten times. Then the average of these ten tests is taken as the result of the experiment. In order to prove the practicability of the method proposed in this paper, 10-fold cross-validation was carried out for 30 times according to the central limit theorem to eliminate the effect of randomness on the results. Therefore, the results listed in this paper are the average after 30 times of 10-fold cross-validation.

### 4 RESULTS AND DISCUSSION OF THE EXPERIMENT

In Table 3, Table 4, Table 5, Table 6, it shows the results of the evaluation metrics for the six algorithms on five datasets under two classifiers, namely SVM and decision tree. The bold text in the table represents the best results achieved by the proposed method compared to other methods. As shown in Table 3, Table 4, Table 5, Table 6, the method proposed in this paper achieves better accuracy on five data sets. The accuracy of the proposed method is superior to other methods in the decision tree (DT) classifier, and it also performs well in SVM classifier. In comparison the value of F, the method proposed in this paper is better than other methods: there are four data sets and two data sets for DT classifier and SVM classifier to obtain the optimality respectively. And it also achieves better results on other evaluation metrics. Although the performance on the SVM classifier is a bit worse than that of the decision tree classifier, it still has some advantages compared with other algorithms.

### 5 APPLICATION TO CHINA GREEN INNOVATION EFFICIENCY

As today's world evolves, green innovation is becoming an increasingly important topic. In order to better improve green innovation efficiency, it is necessary to determine which characteristics have more significant effects on green innovation efficiency in China.

**Table 3: Accuracy comparison with other algorithms in different classifiers.**

DT	This article's	CFS	Accuracy				relieff
			fisher	alpha-investing	chi square	MRMR	
iris	<b>0.95</b>	0.9292	0.9292	0.9283	0.9383	0.8917	0.933
segment	<b>0.9654</b>	0.9364	0.8889	0.9642	0.964	0.9581	0.9562
heart-h	<b>0.642</b>	0.5569	0.5882	0.5537	0.635	0.5402	0.5568
dermatology	<b>0.959</b>	0.9505	0.9285	0.6742	0.8531	0.9382	0.9004
blood data	<b>0.7609</b>	0.7136	0.7274	0.6836	0.753	0.7442	0.7071
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	relieff
iris	<b>0.955</b>	0.95	0.95	0.9523	0.95	0.9417	0.9417
segment	0.9242	0.945	0.7602	0.9118	0.9264	0.9264	0.822
heart-h	0.6594	0.6161	0.6558	0.6936	0.6344	0.6637	0.6344
dermatology	0.9216	0.976	0.93172	0.6828	0.8533	0.9624	0.9523
blood data	<b>0.7742</b>	0.7658	0.7525	0.7709	0.7575	0.7575	0.7575

**Table 4: Comparison of NMI with other algorithms in different classifiers.**

DT	This article's	CFS	NMI				relieff
			fisher	alpha-investing	chi square	MRMR	
iris	<b>0.8939</b>	0.8483	0.8546	0.8615	0.8835	0.8079	0.8631
segment	<b>0.9346</b>	0.6707	0.8323	0.933	0.9304	0.9255	0.9201
heart-h	0.411	0.424	0.3743	0.4147	0.4579	0.2799	0.3159
dermatology	<b>0.9488</b>	0.9416	0.9139	0.7564	0.8917	0.9224	0.8616
blood data	0.0321	0.0384	0.0277	0.0271	0.0699	0.0098	0.0307
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	relieff
iris	<b>0.9106</b>	0.8919	0.8914	0.9077	0.9047	0.886	0.8853
segment	0.8827	0.708	0.7419	0.86	0.8865	0.8896	0.818
heart-h	0.1929	0.4182	0.3822	0.43066	0.3533	0	0
dermatology	0.9034	0.9647	0.9081	0.7546	0.9237	0.9524	0.9323
blood data	0.013	0	0	0	0.0239	0	0.0015

**Table 5: Comparison of Jaccard coefficients with other algorithms in different classifiers.**

DT	This article's	CFS	Jaccard coefficients				relieff
			fisher	alpha-investing	chi square	MRMR	
iris	<b>0.9099</b>	0.8758	0.8748	0.874	0.8923	0.8167	0.8785
segment	<b>0.9333</b>	0.8796	0.801	0.9313	0.9299	0.9201	0.9171
heart-h	<b>0.4767</b>	0.3936	0.4244	0.3879	0.4685	0.3723	0.3899
dermatology	<b>0.9329</b>	0.9259	0.8725	0.5092	0.7417	0.8886	0.8192
blood data	<b>0.6144</b>	0.5594	0.572	0.5217	0.6063	0.5928	0.548
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	relieff
iris	<b>0.9277</b>	0.9099	0.9077	0.9231	0.9138	0.8967	0.8967
segment	0.8596	0.8977	0.6138	0.8383	0.864	0.8633	0.6979
heart-h	0.4922	0.4478	0.4931	0.5333	0.466	0.4969	0.4649
dermatology	0.8583	0.9541	0.8762	0.5186	0.7448	0.9294	0.9106
blood data	0.6057	0.6206	0.6033	0.6273	0.6098	0.6098	0.6098

**Table 6: Comparison of F-values with other algorithms in different classifiers.**

DT	This article's	CFS	F-values				
			fisher	alpha-investing	chi square	MRMR	relieff
iris	<b>0.9491</b>	0.9297	0.9312	0.9275	0.9382	0.8889	0.9306
segment	<b>0.9651</b>	0.9353	0.8887	0.9642	0.9637	0.9568	0.9565
heart-h	<b>0.6119</b>	0.5584	0.5777	0.555	0.592	0.4981	0.5293
dermatology	<b>0.9585</b>	0.9593	0.9274	0.5736	0.8222	0.9355	0.8961
blood data	0.6802	0.7074	0.6936	0.6716	0.7272	0.645	0.6972
SVM	This article's	CFS	fisher	alpha-investing	chi square	MRMR	relieff
iris	<b>0.9588</b>	0.9496	0.949	0.9576	0.9491	0.9399	0.9402
segment	0.9227	0.9441	0.7469	0.9097	0.9258	0.9257	0.778
heart-h	0.5369	0.5609	0.5867	0.629	0.565	0.5297	0.4927
dermatology	0.9208	0.9754	0.9328	0.58	0.8074	0.9602	0.9518
blood data	<b>0.6675</b>	0.6643	0.6463	0.6712	0.6589	0.6531	0.6545

**Table 7: Green Innovation Efficiency Evaluation Index System.**

innovation input	Full-time equivalent of R&D staff( $X_1$ )
	Internal expenditure on R&D expenses( $X_2$ )
	Environmental pollution treatment investment amount( $X_3$ )
	Energy saving and environmental protection expenditure( $X_4$ )
	New product development expenses( $X_5$ )
	Technology introduction and renovation expenses( $X_6$ )
expected output	Total energy consumption( $X_7$ )
	Number of patents granted( $Y_1$ )
	Technology Market Turnover( $Y_2$ )
	New product sales revenue( $Y_3$ )
	Industrial value added( $Y_4$ )

Therefore, it is necessary to make a scientific selection of indicators of green innovation efficiency. Through the typical literature, 14 indicators were selected, which contain 7 input indicators, output indicators, as shown in the Table 7. However, an excessive number of indicators will result in a lack of discriminatory efficiency in the final result. It is necessary to make a selection of indicators for subsequent work.

The empirical sample is 30 Chinese provincial-level cities excluding Hong Kong, Macao, Taiwan and Tibet, for a total of ten years from 2011 to 2020. The above input variables as well as expected output variables were screened separately using the method proposed in this paper, and the cumulative contribution rate of the five variables  $X_1, X_5, X_2, X_6, X_4$  reached 86.8%, which can be used to replace the original seven variables; the cumulative contribution rate of the variables  $Y_4, Y_3, Y_1$  has reached 99% of the original four expected output indicators. The results show that the full-time equivalent of R&D personnel and new product development expenses have a greater impact on the efficiency of green innovation, which is the same as the previous studies by scholars [10, 11]. The experiment shows the practicality of the method proposed in this paper.

## 6 CONCLUSIONS

In this paper, we propose a PCA-based feature selection method, which first uses the loadings of each indicator on different principal components and their variance contributions with that principal component to calculate a number of indicators with the largest cumulative contributions for feature selection. Then, based on the data in the UCI dataset, the proposed method is compared with other feature selection methods. Finally, the method proposed in this paper is used to select the characteristics of China's green innovation efficiency, which verified the practicability of the method proposed in this paper. The experiments show that the method performs better in the public dataset, and it may be possible to try to use other PCA (For example kernel PCA, probabilistic PCA) for subsequent studies. Finally, the features of green innovation efficiency in China were selected using the method proposed in this paper, and the feasibility of the method was verified.

## REFERENCES

- [1] Zhou H, Zhang Y, Zhang Y and Liu H. Feature selection based on conditional mutual information: minimum conditional relevance and minimum conditional redundancy. *Applied Intelligence*, 2019, 49(3) : 883-896. <https://doi.org/10.1007/s10489-018-1305-0>.
- [2] Ye X L, Lan J L and Guo T. (2014) Network traffic feature selection algorithm based on PCA and taboo search. *Computer*

- Science,41(01):187-191. [https://kns.cnki.net/kcms2/article/abstract?v=\\$3uoqIhG8C44YLtIOAiTRKgchrJ08w1e7M8Tu7YZds89NyEjIjuMbEOVYc9b1HujUP\\_e\\_JuVGMXjzV5RWNdGXPIQoT5NIm\\_MK&uniplatform\\$=\\$NZKPT](https://kns.cnki.net/kcms2/article/abstract?v=$3uoqIhG8C44YLtIOAiTRKgchrJ08w1e7M8Tu7YZds89NyEjIjuMbEOVYc9b1HujUP_e_JuVGMXjzV5RWNdGXPIQoT5NIm_MK&uniplatform$=$NZKPT).
- [3] Kale A P and Sonavane S. (2018) PF-FELM: A robust PCA feature selection for fuzzy extreme learning machine. *IEEE Journal of Selected Topics in Signal Processing*, 12(6): 1303-1312. <https://doi.org/10.1109/JSTSP.2018.2873988>.
  - [4] Chi G T and Zhao Z C. (2018) The construction of an evaluation index system of science and technology innovation with enterprises as the main body. *Scientific Research Management*,39(S1):1-10. <http://www.cqvip.com/qk/95604x/2018s1/75897176504849568349484849.html>.
  - [5] Dariush K, Wade D C and Joe Z. (2019) Number of performance measures versus number of decision making units in DEA. *Annals of Operations Research*. 303(1-2):529-562. <https://doi.org/10.1007/s10479-019-03411-y>.
  - [6] Li M, Wang H, Yang L, Liang, Y, Shang Z and Wan, H. (2020). Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. *Expert Systems with Applications*, 150, 113277. <https://doi.org/10.1016/j.eswa.2020.113277>.
  - [7] Singh N and Singh P. (2021) A hybrid ensemble-filter wrapper feature selection approach for medical data classification. *Chemometrics and Intelligent Laboratory Systems*, 217: 104396. <https://doi.org/10.1016/j.chemolab.2021.104396>.
  - [8] Lim H and Kim D W. (2021) Pairwise dependence-based unsupervised feature selection. *Pattern Recognition*, 111: 107663. <https://doi.org/10.1016/j.patcog.2020.107663>.
  - [9] Sahebi G, Movahedi P, Ebrahimi M, *et al.* GeFeS: A generalized wrapper feature selection approach for optimizing classification performance[J]. *Computers in biology and medicine*, 2020, 125: 103974. <https://doi.org/10.1016/j.combiomed.2020.103974>.
  - [10] Li Y, Huang N and Zhao Y. (2022) The Impact of Green Innovation on Enterprise Green Economic Efficiency. *International Journal of Environmental Research and Public Health*,19(24): 16464. <https://doi.org/10.3390/ijerph192416464>.
  - [11] Chen X, Liu X, Gong Z and Xie J. (2021) Three-stage super-efficiency DEA models based on the cooperative game and its application on the R&D green innovation of the Chinese high-tech industry. *Computers & Industrial Engineering*, 156: 107234. <https://doi.org/10.1016/j.cie.2021.107234>.

# A water quality parameter prediction method based on transformer architecture and multi-sensor data fusion

Bo Fang\*  
Powerchina Zhongnan Engineering  
Corporation Limited, Changsha,  
410014, China,  
1305499682@qq.com.Hunan  
Provincial Key Laboratory of  
Hydropower Development Key  
Technology, Changsha, 410014, China.

Hao Liu  
Powerchina Zhongnan Engineering  
Corporation Limited, Changsha,  
410014, China,  
330633026@qq.com.Hunan Provincial  
Key Laboratory of Hydropower  
Development Key Technology,  
Changsha, 410014, China.

Wei He  
Powerchina Zhongnan Engineering  
Corporation Limited, Changsha,  
410014, China,  
1016339389@qq.com.Hunan  
Provincial Key Laboratory of  
Hydropower Development Key  
Technology, Changsha, 410014, China.

Dexin Li  
Powerchina Zhongnan Engineering  
Corporation Limited, Changsha,  
410014, China,  
1037661085@qq.com.Hunan  
Provincial Key Laboratory of  
Hydropower Development Key  
Technology, Changsha, 410014, China.

Chengzhao Liu  
Powerchina Zhongnan Engineering  
Corporation Limited, Changsha,  
410014, China,  
498361957@qq.com.Hunan Provincial  
Key Laboratory of Hydropower  
Development Key Technology,  
Changsha, 410014, China.

## ABSTRACT

Water quality monitoring provides a basis for water quality control and water resources management. Prediction of water quality parameters can plan water use strategies, prevent further water pollution and improve water resource utilization efficiency. We propose a water quality parameter prediction method based on transformer architecture model and multi-sensor data fusion. The proposed multiple water quality parameter prediction model accepts multiple types of water quality parameter data input at the same time. The data embedding module integrates multiple types of water quality parameter information and assigns a unique position code to the data at each time step. The self-attention mechanism of the model mining the potential correlation between different time step data. The model can learn the internal relationship of the fusion data of multiple water quality parameters, and effectively predict the future trend of water quality parameters. The effectiveness of the proposed algorithm is verified by the measured data, and the advantages of the proposed method are verified by comparative experiments.

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590037>

## CCS CONCEPTS

• **Applied computing**; • **Theory of computation** → Theory and algorithms for application domains; Machine learning theory; Models of learning;

## KEYWORDS

Water quality parameter, Prediction, Transformer, Self-attention mechanism

## ACM Reference Format:

Bo Fang, Hao Liu, Wei He, Dexin Li, and Chengzhao Liu. 2023. A water quality parameter prediction method based on transformer architecture and multi-sensor data fusion. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590037>

## 1 INTRODUCTION

Water is the basis for human survival and development, and plays an irreplaceable role in human material life and social production activities. However, in the past hundred years, the cost of sustained global economic growth has been the over-exploitation and utilization of ecological resources, and the water environment in many regions of the world has been damaged to varying degrees. At present, the assessment of water quality is a very important topic. Predicting the changing trend of water quality parameters will help environmental management personnel prepare water pollution control actions in advance to prevent further water pollution; Further, it can improve the utilization efficiency of water resources and realize the transformation from post-treatment to pre-prevention.

Scholars began to study the prediction of water quality parameters decades ago. Various methods for predicting water quality parameters are divided into two categories: methods based on physical models and methods based on machine learning.

Gong, R et al. [1] established a reliable hydrodynamic-water quality model for an urban lake and studied the impact of extreme conditions such as rainstorms and sewage leakage on hydrological conditions. Liang, J et al. [2] developed a water quality parameter evaluation management tool based on MIKE 11 and simulated the trend of water quality parameters such as biochemical oxygen demand, chemical oxygen demand, and ammonia. The methods based on physical models need to be combined with multiple data such as precipitation, river shape, pollutant distribution, etc. for comprehensive analysis. The model needs to be adjusted and adapted to different hydrological environments, which requires very high professionalism.

Based on the machine learning method, the water quality parameter prediction model is established by mining the complex dependencies in the historical data of water quality parameters. Dellana et al. [3] compare the multi-period predictive ability of ARIMA models to neural network models in water quality applications. They verified the feasibility of machine learning algorithms to predict water quality parameters. Hanh et al. [4] researched the influence of climate and hydrology on the water quality of the lower Mekong River based on ARIMA. Ömer Faruk [5] proposed a hybrid ARIMA and neural network model to predict water quality parameters such as water temperature, boron, and dissolved oxygen. The deep learning model can automatically capture the key features according to the training data to learn the complex mapping relationship from input to output. It has been widely used in water quality parameter prediction in recent years. Chen et al. [6] designed a deep convolutional neural network (CNN) architecture to analyze the water pollution of agricultural irrigation. Valadkhan et al. [7] considered factors such as rainfall rate, temperature, and humidity, and used long short-term memory network (LSTM) to predict the degree of groundwater pollution. Tan et al. [8] proposed a hybrid model based on CNN and LSTM. Firstly, CNN was used to extract the local characteristics of water quality data, and then LSTM was used to predict the dissolved oxygen index in water. The transformer model [9] proposed by the Google team in 2017 is not only brilliant in the field of natural language processing but also widely used in image recognition, target detection, semantic segmentation, sequence prediction, and other fields. Li et al. [10] used Informer [11] (based transformer) to predict the wear value of machine tools, improving the reliability of machining quality. Wang et al. [12] proposed a wind farm wind forecasting method based on CNN and Informer model. Yao et al. [13] applied the existing multiple series prediction models to predict the integrated water quality index (WQI) in the Chaohu Lake area, and the test results proved the advantages of the model based on the transformer architecture.

Given the low precision of collaborative prediction of multi-class water quality parameters, this paper designs a model for multi-class water quality parameters prediction based on the classic transformer architecture and verifies the effectiveness of the proposed model through the measured data of Shima river, Guangdong province from 2020 to 2022.

## 2 THEORETICAL BACKGROUND

### 2.1 Data embedding

Data embedding includes two sub-modules: word embedding and positional encoding.

The purpose of word embedding is to map the water quality parameters of category  $M$  to the space of dimension  $d$  (usually  $d > M$ ), to achieve the fusion of different water quality parameters. Word embedding is a one-dimensional convolution operation. The schematic diagram of this process is shown in 1. Class  $M$  water quality parameters of  $C$  time steps form the input matrix. The convolution kernel moves along the direction of the time step to perform the convolution operation on the original data. The matrix with dimension  $d \times C$  is obtained by splicing the one-dimensional convolution operation results of  $d$  channels. One-dimensional convolution maps the water quality parameters in  $M$  dimension to  $d$  dimension, and realizes the fusion of multi-sensor data.

Unlike the recurrent neural network, the transformer model processes data of different time steps through parallel computing. Therefore, before the forward calculation of the model, it is necessary to give the unique position information of the water quality parameters at different time steps as the unique identification. This process is called positional encoding. Sine-cosine encoding is a common method:

$$PE(pos, 2i) = \sin(pos/10000^{2i/d})$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{2i/d})$$

where  $pos$  is the position and  $i$  is the dimension.

Data embedding is realized by adding the word embedding data with the positional encoding data.

### 2.2 Multi-head attention

Self-attention mechanism is the core module based on transformer architecture model.  $Q$ ,  $K$ , and  $V$  matrices are the three input matrices, and the operation process of self-attention mechanism is:

Attention( $Q, K, V$ ) =  $\text{softmax}(\frac{QK^T}{\sqrt{d}})V$   $\text{softmax}(\cdot)$  function normalizes  $\frac{QK^T}{\sqrt{d}}$  matrix, then multiplies with matrix  $V$  to obtain a matrix consistent with the dimension of matrix  $V$ .

For the multi-head attention, the calculation process is as follows:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

?>

where  $\text{Concat}(\cdot)$  concat multiple output matrices. The concat matrix is multiplied by the matrix  $W^O$ , and the output matrix of multiple attention is mapped to the same dimension as the input matrix  $Q$ . In the encoder,  $Q$ ,  $K$ , and  $V$  are copies of the previous calculation matrix. In the multi-head attention module of the decoder,  $K$  and  $V$  come from the output of the encoder, and  $Q$  is calculated by the decoder.

To maintain the temporal causality, the attention of the current moment associated with the future moment in the decoder should not be calculated. The masked multi-head attention module fills the values of the triangle on matrix  $\frac{QK^T}{\sqrt{d}}$  with  $10^{-9}$  to ignore the attention values that do not conform to the causal relationship.

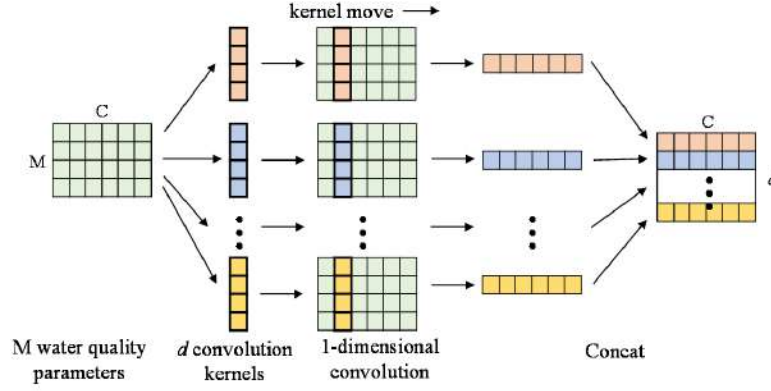


Figure 1: Implementation of word embedding.

### 2.3 Layer normalization

Transformer model contains multiple layer normalization (Norm) modules, which is a common module in the field of natural language processing. Its principle is as follows:

<?TeX

$$y = \frac{x - E[x]}{\sqrt{\text{Var}[x] + \varepsilon}} * \gamma + \beta$$

?>

where  $x$  represents the input matrix,  $E[\cdot]$  and  $\text{Var}[\cdot]$  are the mean and variance of  $x$ .  $\gamma$  and  $\beta$  are two learnable parameters, which are initialized to 1 and 0 respectively.

The layer normalization effectively avoids the data being in the saturation zone of the activation function and alleviates the disappearance of the gradient in the training stage.

## 3 PROPOSED METHOD

### 3.1 Model structure

Referring to Transformer and Informer models, a water quality parameter prediction model based on multi-sensor data fusion is proposed. The structure of the model is shown in 2. After the input passes through the data embedding module, it enters two serial encoders. The Maxpool1d inside the encoder samples the data down to maximize the retention of features while reducing the amount of model calculation. The outputs of the encoder, as K and V matrix, participate in the attention mechanism operation of the decoder. The input  $X_{de}$  of the decoder includes the known data  $X_{token}$  and the data  $X_0$  to be predicted, where  $X_0$  is filled with 0 vector. The full connection layer remaps the decoder output to a dimension consistent with  $X_{de}$ . We use the mean square error as the loss function:

<?TeX

$$MSE = \frac{1}{M \cdot S} \sum_{i=1}^M \sum_{j=1}^S (y_{i,j} - y_{i,j}^p)^2$$

?>

where  $i$  represents the category of water quality parameters,  $j$  represents the time step to be predicted,  $y$  represents the predicted value, and  $y^p$  represents the true value.

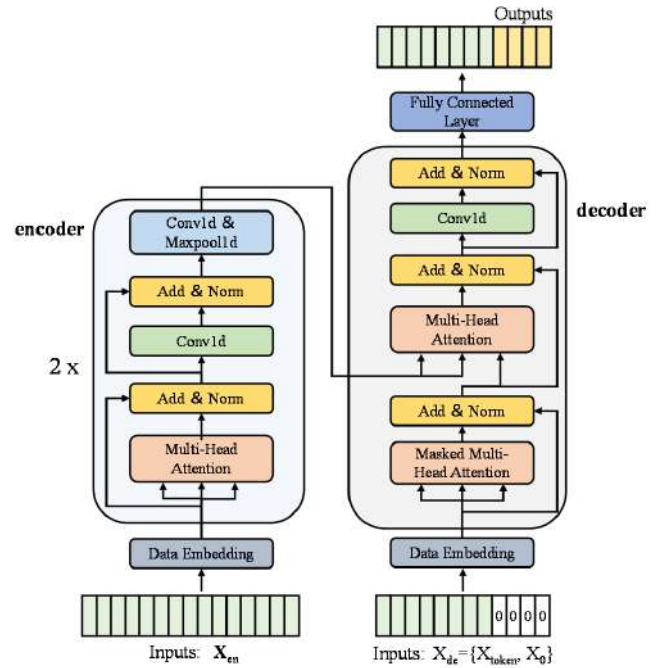


Figure 2: The proposed model.

### 3.2 Multi-sensor water quality parameter prediction method

Before training the model, it is very important to eliminate outliers, complete missing values, and normalize data. These steps will directly affect the accuracy and reliability of the prediction model. The detailed process of data pretreatment and water quality parameter prediction is as follows:

(1) Multi-sensor water quality parameter data acquisition. Install multiple types of water quality parameter sensors at the points to be measured, and collect multiple types of water quality parameters such as pH value, dissolved oxygen (DO), chemical oxygen demand

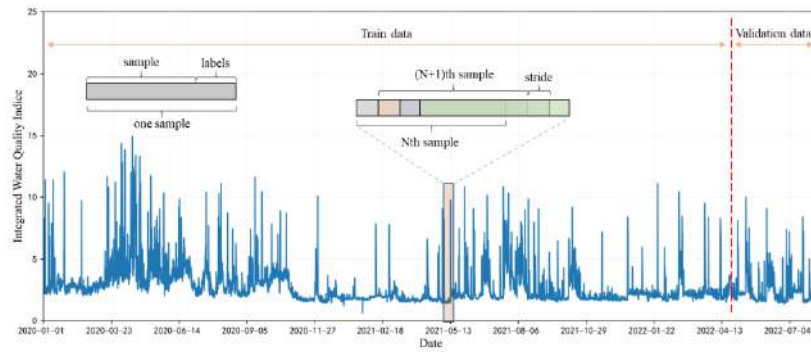


Figure 3: Intercept training set and validation set from time series data.

(COD), turbidity, conductivity, ammonia nitrogen (AN), etc. at fixed sampling intervals.

(2) Eliminate the outliers of water quality parameters. Each type of water quality parameter has a reasonable value range, for example, the value of COD is usually 10-40mg/L. Eliminate unreasonable water quality parameter values caused by sensor failure.

(3) Complete and normalize water quality parameters. For the missing water quality parameters caused by sensor offline, fault, and other reasons, linear interpolation algorithm is used to complete them. Then calculate the mean value and standard deviation of each type of water quality parameter, and normalize the water quality parameters.

(4) Prepare training dataset. Samples and labels are intercepted along the time dimension, and the training set and validation set are divided. Taking the WQI as an example, the method of intercepting the dataset from the data is shown in 3. The data is divided into two parts: training data and validation data. One sample contains sample and labels, and the distance between adjacent samples is stride.

(5) The water quality parameter prediction model is trained, and the effect of the model is evaluated by the validation set.

(6) The trained model is used to predict water quality parameters, and the prediction results of water quality parameters are obtained by inverse normalization of the prediction results. The flowchart of the proposed method is shown in Figure 4.

#### 4 EXPERIMENTAL RESULTS

We have installed multiple types of water quality parameter monitoring sensors at Shima river, Guangdong Province. The sensors are uniformly installed at the bottom of the wireless communication buoy powered by photovoltaic panels and batteries. The sensors collect water quality parameters with a depth of 0.5m, and the data is transmitted to the cloud platform in real-time through the wireless communication module. The water quality parameter acquisition device is shown in 5. The monitored water quality includes eight categories: water quality index (WQI), chemical oxygen demand (COD), conductivity, dissolved oxygen (DO), ammonia nitrogen (AN), PH, total phosphorus (TP), and turbidity. The sampling period of all sensors is set to 4 hours, and the validity of the proposed algorithm is verified by the measured data of water quality parameters from January 1, 2020 to August 9, 2022.

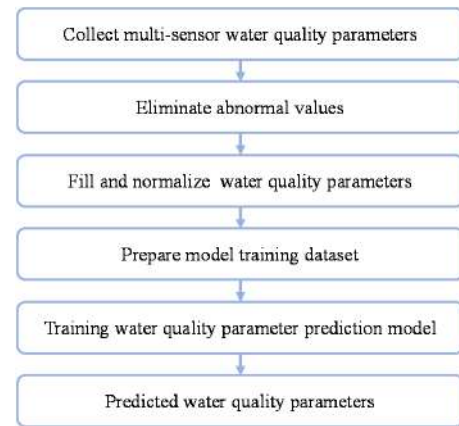


Figure 4: The flowchart of the proposed method.



Figure 5: Water quality parameter acquisition device.

We first remove some unreasonable water quality parameters and then use a linear interpolation algorithm to fill in the deleted data and missing data caused by other reasons. 6 shows the data pre-treatment results of COD, in which the green triangles are the outliers, and the orange triangles are the interpolated water quality parameters.

Intercept 12 consecutive time step data as training samples, the data of the next 6 time steps as the corresponding labels of the

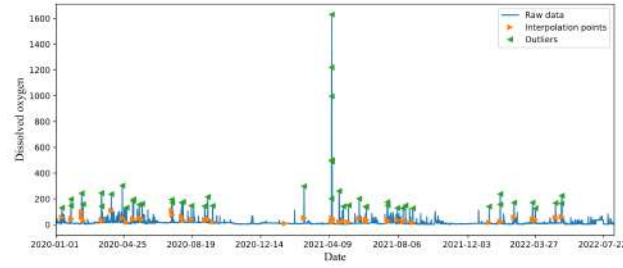


Figure 6: Data pretreatment results of COD

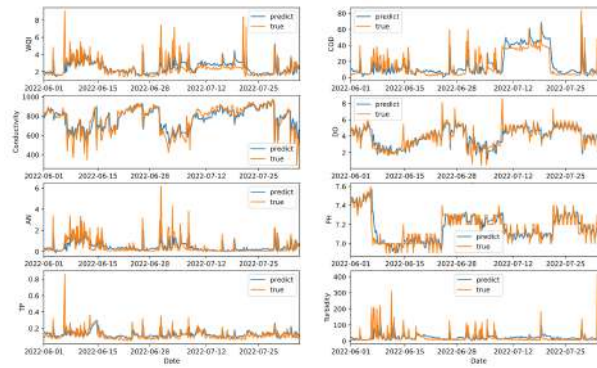


Figure 7: Results of 1-step prediction of multiple water quality parameters.

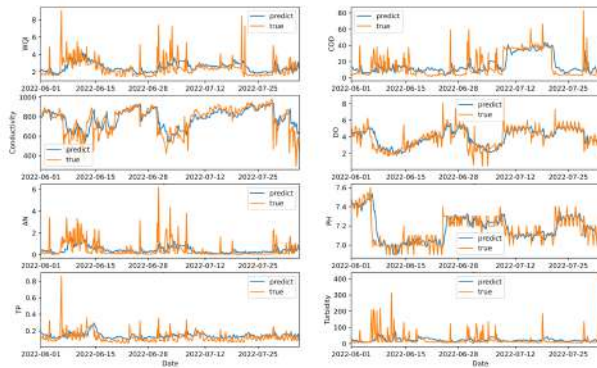


Figure 8: Results of 3-step prediction of multiple water quality parameters.

samples, and then move 1-time step to intercept the next sample and its corresponding labels. Use the data before June 1, 2022 as the training set, and the data after June 1, 2022 as the validation set. The number of heads of the attention mechanism is 8, the model dimension is 512, the number of iterations of the model is set to 10,

the learning rate is 0.00001, the optimizer is Adam, the batch size is 32, and the model is trained on the device with 1650Ti.

7 shows the 1-step prediction results of the proposed method for 8 types of water quality parameters (that is, the prediction of water quality parameters in the next four hours). The prediction accuracy of the algorithm is high, and drastic changes in water quality

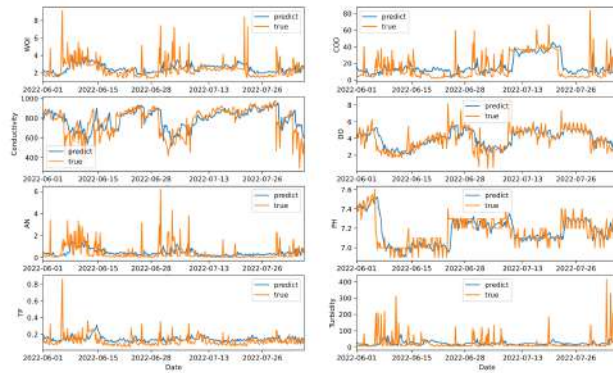


Figure 9: Results of 6-step prediction of multiple water quality parameters.

Table 1: MAE of the proposed method and LSTM

Water quality parameters	Method	MAE					
		1 step	2 step	3 step	4 step	5 step	6 step
WQI	Proposed	0.374	0.396	0.408	0.411	0.425	0.430
	LSTM	0.411	0.419	0.423	0.428	0.429	0.432
COD	Proposed	0.347	0.352	0.363	0.387	0.428	0.449
	LSTM	0.352	0.357	0.362	0.371	0.420	0.455
Conductivity	Proposed	0.237	0.282	0.355	0.382	0.439	0.452
	LSTM	0.229	0.269	0.356	0.381	0.428	0.435
DO	Proposed	0.213	0.237	0.251	0.286	0.303	0.317
	LSTM	0.219	0.241	0.255	0.297	0.324	0.350
AN	Proposed	0.401	0.415	0.432	0.457	0.462	0.477
	LSTM	0.451	0.464	0.482	0.482	0.493	0.501
PH	Proposed	0.159	0.165	0.185	0.192	0.216	0.225
	LSTM	0.154	0.167	0.189	0.197	0.220	0.249
TP	Proposed	0.317	0.356	0.402	0.429	0.451	0.488
	LSTM	0.382	0.399	0.457	0.503	0.584	0.586
Turbidity	Proposed	0.397	0.410	0.415	0.438	0.443	0.451
	LSTM	0.421	0.426	0.429	0.445	0.443	0.454

parameters are predicted. 8 shows the 3-step prediction effect. Although the method cannot predict the drastic changes in water quality parameters such as AN and turbidity, it can track the trend of water quality parameter changes. 9 shows the effect of the 6-step prediction. The model is difficult to predict the drastic change in water quality parameters, but it still effectively predicts the changing trend of water quality parameters.

Mean absolute error (MAE) and mean square error (MSE) are commonly used indicators of quantitative model accuracy. Their calculation equations are as follows:

<?TeX

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^*|$$

?>

<?TeX

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^*)^2$$

?>

To further illustrate the effectiveness of the proposed method, we use LSTM as a control experiment. 1 and 2 show the MAE and MSE of the proposed method and LSTM for eight types of water quality parameters, each of which contains 6-time steps. It can be seen from the table that the two methods have poor prediction accuracy for AN and Turbidity, which is caused by too drastic changes in the curve of AN and Turbidity. LSTM and the proposed method have good prediction performance for the remaining 6 types of water quality parameters. The comparison between the prediction results of the proposed algorithm and LSTM shows that the overall prediction accuracy of the proposed method for the remaining seven types of water quality parameters is higher except for conductivity.

**Table 2: MSE of the proposed method and LSTM**

Water quality parameters	Method	MSE					
		1 step	2 step	3 step	4 step	5 step	6 step
WQI	Proposed	0.414	0.420	0.425	0.424	0.426	0.428
	LSTM	0.421	0.423	0.451	0.458	0.469	0.467
COD	Proposed	0.313	0.312	0.306	0.339	0.340	0.351
	LSTM	0.321	0.326	0.327	0.384	0.391	0.389
Conductivity	Proposed	0.234	0.278	0.290	0.337	0.402	0.412
	LSTM	0.256	0.295	0.307	0.330	0.384	0.429
DO	Proposed	0.109	0.130	0.159	0.182	0.199	0.217
	LSTM	0.108	0.127	0.164	0.196	0.199	0.231
AN	Proposed	0.577	0.578	0.572	0.582	0.584	0.593
	LSTM	0.589	0.596	0.603	0.614	0.607	0.618
PH	Proposed	0.051	0.065	0.077	0.084	0.108	0.127
	LSTM	0.053	0.071	0.082	0.095	0.120	0.143
TP	Proposed	0.300	0.325	0.339	0.361	0.378	0.397
	LSTM	0.322	0.355	0.379	0.423	0.441	0.468
Turbidity	Proposed	0.743	0.748	0.764	0.750	0.755	0.764
	LSTM	0.756	0.751	0.773	0.769	0.782	0.787

## 5 CONCLUSION

In this paper, a water quality parameter prediction method based on transformer architecture and multi-sensor data fusion is proposed. The model based on transformer architecture accepts multiple types of water quality parameter data input at the same time. The data embedding module fuses multiple types of water quality parameter information and assigns a unique position code to the data at each time step. The self-attention mechanism of the model mining the potential correlation between data at different time steps. The proposed algorithm is tested by the water quality parameter data of Shima river, Guangdong province for 32 months. The results show that the proposed algorithm can achieve good prediction accuracy for most water quality parameters. The difficulty in predicting AN and turbidity may be caused by the drastic changes in these two types of water quality parameters. Compared with the classic LSTM algorithm, the overall prediction accuracy of the proposed method is higher.

## CONFLICTS OF INTEREST

## ACKNOWLEDGMENTS

The authors declare that they have no conflicts of interest to report regarding the present study.

## REFERENCES

- [1] R. Gong, L. Xu, D. Wang, H. Li, and J. Xu, "Water Quality Modeling for a Typical Urban Lake Based on the EFDC Model," *Environ. Model. Assess.*, vol. 21, no. 5, pp. 643–655, Oct. 2016, doi: 10.1007/s10666-016-9519-1.
- [2] J. Liang, Q. Yang, T. Sun, J. D. Martin, H. Sun, and L. Li, "MIKE 11 model-based water quality model as a tool for the evaluation of water quality management plans," *J. Water Supply Res. Technol.-Aqua*, vol. 64, no. 6, pp. 708–718, Sep. 2015, doi: 10.2166/aqua.2015.048.
- [3] S. A. Dellana and D. West, "Predictive modeling for wastewater applications: Linear and nonlinear approaches," *Environ. Model. Softw.*, vol. 24, no. 1, pp. 96–106, Jan. 2009, doi: 10.1016/j.envsoft.2008.06.002.
- [4] P. T. M. Hanh, N. V. Anh, D. T. Ba, S. Sthiannopkao, and K.-W. Kim, "Analysis of variation and relation of climate, hydrology and water quality in the lower Mekong River," *Water Sci. Technol.*, vol. 62, no. 7, pp. 1587–1594, Oct. 2010, doi: 10.2166/wst.2010.449.
- [5] D. Ömer Faruk, "A hybrid neural network and ARIMA model for water quality time series prediction," *Eng. Appl. Artif. Intell.*, vol. 23, no. 4, pp. 586–594, Jun. 2010, doi: 10.1016/j.engappai.2009.09.015.
- [6] H. Chen *et al.*, "A deep learning CNN architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources," *Agric. Water Manag.*, vol. 240, p. 106303, Oct. 2020, doi: 10.1016/j.agwat.2020.106303.
- [7] D. Valadkhan, R. Moghaddasi, and A. Mohammadinejad, "Groundwater quality prediction based on LSTM RNN: An Iranian experience," *Int. J. Environ. Sci. Technol.*, vol. 19, no. 11, pp. 11397–11408, Nov. 2022, doi: 10.1007/s13762-022-04356-9.
- [8] W. Tan *et al.*, "Application of CNN and Long Short-Term Memory Network in Water Quality Predicting," *Intell. Autom. Soft Comput.*, vol. 34, no. 3, pp. 1943–1958, 2022, doi: 10.32604/iasc.2022.029660.
- [9] A. Vaswani *et al.*, "Attention Is All You Need." arXiv, Dec. 05, 2017. Available: <http://arxiv.org/abs/1706.03762>
- [10] W. Li, H. Fu, Z. Han, X. Zhang, and H. Jin, "Intelligent tool wear prediction based on Informer encoder and stacked bidirectional gated recurrent unit," *Robot. Comput.-Integr. Manuf.*, vol. 77, p. 102368, Oct. 2022, doi: 10.1016/j.rcim.2022.102368.
- [11] H. Zhou *et al.*, "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting," arXiv, Mar. 28, 2021. Available: <http://arxiv.org/abs/2012.07436>
- [12] H.-K. Wang, K. Song, and Y. Cheng, "A Hybrid Forecasting Model Based on CNN and Informer for Short-Term Wind Power," *Front. Energy Res.*, vol. 9, p. 788320, Jan. 2022, doi: 10.3389/fenrg.2021.788320.
- [13] S. Yao, Y. Zhang, P. Wang, Z. Xu, Y. Wang, and Y. Zhang, "Long-Term Water Quality Prediction Using Integrated Water Quality Indices and Advanced Deep Learning Models: A Case Study of Chaohu Lake, China, 2019–2022," *Appl. Sci.*, vol. 12, no. 22, p. 11329, Nov. 2022, doi: 10.3390/app122211329.

# Unknown Radar Signals Deinterleaving Based on TCN Network

Liying Ma  
State Key Laboratory of High  
Performance Computing, National  
University of Defense Technology  
Changsha, China  
maliying20@nudt.edu.cn

Xueqiong Li\*  
State Key Laboratory of High  
Performance Computing, National  
University of Defense Technology  
Changsha, China  
lixueqiong13@nudt.edu.cn

Yuhua Tang  
State Key Laboratory of High  
Performance Computing, National  
University of Defense Technology  
Changsha, China  
yhtang@nudt.edu.cn

## ABSTRACT

Radar signals deinterleaving plays a critical role in electronic reconnaissance. Nevertheless, due to the extremely high density of intercepted signal trains and the unknown number of emitters, along with the low probability of interception (LPI), high loss rate, and high spurious rate, the deinterleaving task is becoming more challenging. In this paper, we propose a temporal convolutional network (TCN)-based method for deinterleaving radar signal pulse trains, using only the time of arrival (TOA) parameter without knowing how many emitters there are. Simulation results indicate that the proposed method can still achieve high accuracy in situations with high pulse loss and spurious rates.

## CCS CONCEPTS

• Computing methodologies → Neural networks; Rule learning.

## KEYWORDS

Radar signal deinterleaving, deep learning, temporal convolutional network (TCN), TOA, high noise rate, unknown emitters

### ACM Reference Format:

Liying Ma, Xueqiong Li\*, and Yuhua Tang. 2023. Unknown Radar Signals Deinterleaving Based on TCN Network. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590038>

## 1 INTRODUCTION

In modern electronic warfare (EW), it is essential for electronic support measures (ESM) to distinguish different non-cooperative radar emitters. The ESM must separate them from the received interleaved pulse trains in an increasingly complicated electromagnetic environment [20]. The radar pulse signal is usually described with pulse descriptor words (PDW). PDW consists of parametric information such as pulse duration (PD) or pulse width (PW), pulse amplitude (PA), pulse radio frequency (RF), angle of arrival (AOA) or

direction of arrival (DOA), and time of arrival (TOA) [4; 16; 19; 20]. It is essential to have all of these parameters available and precisely measured. However, it is almost impossible to achieve this condition in an actual complex electromagnetic environment, for some data will be missing as the antenna rotates or is susceptible to interferences. In contrast to other metrics, TOA is easier to measure and more precise, and its first-order difference, called Pulse Repetitive Interval (PRI), is easily identifiable. In practice, it is effective enough to accomplish the sorting task with TOA data alone [19].

Since this issue was raised by N. J. Whittall in 1985 [19], various solutions have been proposed, especially methods based on PRI, e.g., PRI-Based gating method [20], TOA difference histogram [20], cumulative difference histogram (CDIF) [11], sequential difference histogram (SDIF) [12], PRI transform-based methods and improved approaches based on it [15; 18; 21; 22]. While these conventional methods may be effective early, they must catch up in increasingly complicated real-world environments. Therefore, data-driven neural network methods for radar signal sorting have emerged since significant breakthroughs in machine learning have been made in recent years. Among them are mainly cluster methods [2; 5; 7; 13; 17] and deep learning-based methods [1; 3; 8; 9; 14].

Generally, the radar signal deinterleaving problem is one of the practical applications of the blind source separation (BSS) problem [16]. The BSS problem is fundamental and significant in signal processing, including audio signals (e.g., speech separation and enhancement), radar signals, and images. As the only known knowledge, it aims to restore several independent individual sources from a set of mixed signals [16]. Especially for audio and radar signal processing problems, there are many similarities in data between them, e.g., 1-dimension in the time domain, overlapping in time, existing missing data, and random noise from other sources. These suggest that these two categories of issues can share some similar solutions. In the current study, we investigated the possibility that, if a model is effective enough in one field, it could also succeed in a related one, with certain adjustments or finetuning included if necessary.

The TOA sequence increases drastically with time, and the PRI sequence consisting of its first-order difference has prominent timing characteristics for most PRI modulation types, e.g., constant PRI, sliding PRI, wobulated PRI, D&S PRI, and staggered PRI. Therefore, the investigation could be conducted using a time-series methodology. Due to the outstanding performance of TCN [6] in sequence tasks and with Conv-TasNet's success in speech separation task [10], this study used Conv-TasNet as a reference to extend the validation of the model in radar signal pulse trains sorting problem. Compared to previous approaches, the proposed method requires

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590038>

only one model for all scenarios, given the maximum number of possible emitters.

The rest of this paper is organized as follows. The proposed deinterleaving method for mixed pulse signal trains is introduced in Section II. Section III shows the proposed method's training and test results of simulation experiments. Finally, Section IV concludes this paper.

## 2 METHODOLOGY

### 2.1 Pre-Processing of TOA Sequence

Since the values of the TOA sequence increase drastically with time, the data range becomes excessively vast. Moreover, this is not conducive to neural networks. We have converted the TOA sequence into binary code, as shown in [8], by setting a short time window  $t_{unit}$ . Upon encoding, the original TOA sequence  $T$  will be represented as a 0-1 sequence  $T'$ , which is appropriate for machine processing.

$$T = \{t_i\}, i \in [0, n], \quad (1)$$

$$T' = \{t'_i\}, i \in [0, n], \quad (2)$$

$$t'_i = \begin{cases} 1, & \text{if } i = t_k / t_{unit}, t_k \in T \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

For instance, given a TOA sequence  $T = \{100, 250, 400, 550, 700, 850, \dots\}$ , and  $t_{unit} = 50$ , we will get  $T' = \{0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, \dots\}$  according to (1)(2)(3).

### 2.2 Training Objective

Due to the superior performance of Conv-TasNet [10], a deep learning framework for end-to-end time-domain speech separation, we have tested the task of deinterleaving radar signal emitters based on this model. Different from Conv-TasNet, using the scale-invariant source-to-noise ratio (SI-SNR) [10] as its training objective, we exploit the total MSE losses of all subsequences to evaluate the training performance as follows:

$$Loss = \min\{loss_k\}, k = N!, \quad (4)$$

$$loss = \sum_{i=1, j=1}^N MSELoss(\hat{y}_i, y_j), (i, j) \in comb(N, N). \quad (5)$$

where  $n$  is the length of  $T'$ ,  $\hat{y}$  and  $y$  are represented as the predicted and target subsequence, respectively,  $N$  stands for the maximum number of classes, and  $comb(N, N)$  denotes all the possible combinations of indexes from the predicted and target subsequences. For instance, suppose  $N = 3$ ,  $\hat{y} = [\hat{y}_1, \hat{y}_2, \hat{y}_3]$ ,  $y = [y_1, y_2, y_3]$ , then  $k = N! = 3! = 6$ . All the combinations for  $(\hat{y}, y)$  are as follows:

- (1)  $[(\hat{y}_1, y_1), (\hat{y}_2, y_2), (\hat{y}_3, y_3)]$
- (2)  $[(\hat{y}_1, y_1), (\hat{y}_2, y_3), (\hat{y}_3, y_2)]$
- (3)  $[(\hat{y}_1, y_2), (\hat{y}_2, y_1), (\hat{y}_3, y_3)]$
- (4)  $[(\hat{y}_1, y_2), (\hat{y}_2, y_3), (\hat{y}_3, y_1)]$
- (5)  $[(\hat{y}_1, y_3), (\hat{y}_2, y_1), (\hat{y}_3, y_2)]$

- (6)  $[(\hat{y}_1, y_3), (\hat{y}_2, y_2), (\hat{y}_3, y_1)]$ .

As a result,

$$loss_1 = MSELoss(\hat{y}_1, y_1) + MSELoss(\hat{y}_2, y_2) + MSELoss(\hat{y}_3, y_3)$$

$$loss_2 = MSELoss(\hat{y}_1, y_1) + MSELoss(\hat{y}_2, y_3) + MSELoss(\hat{y}_3, y_2)$$

...

$$loss_6 = MSELoss(\hat{y}_1, y_3) + MSELoss(\hat{y}_2, y_2) + MSELoss(\hat{y}_3, y_1)$$

Therefore,  $Loss = \min\{loss_1, loss_2, \dots, loss_6\}$ .

### 2.3 Algorithm Description

The significant steps of the proposed deinterleaving approach are depicted in Fig. 1. First, an interleaved TOA sequence within a given period is divided into several samples of equal duration. Then, transform these samples into 0-1 sequences according to (1)(2)(3), which will be fed into the deinterleaving model for training. Upon deinterleaving, calculate each sample's length ratios  $r_1 \sim r_N$  of  $N$  output sequences to determine which sequence(s) should be kept. Finally, calculate the accuracy of each valid sequence and restore them to TOA sequences according to  $t_{unit}$ .

## 3 EXPERIMENTS AND DISCUSSION

### 3.1 Experiment settings

Suppose that the maximum number and modulation types of emitters are both  $N$  and that the mixed signal trains are made up of  $n$  emitters from them, with noise (measurement error, missing pulses, and spurious pulses) simultaneously. In our experiments, we have performed four TOA sequences of different PRI modulation types as emitter sources. Table 1 lists the scope of main parameters for each PRI modulation type, and the corresponding parameters are created randomly under the constraints. Based on these settings, we simulate a series of various TOA mixtures of different modulation types. As shown in Table 2, there are six types of 2-mixtures, four types of 3-mixtures, and one type of 4-mixtures, respectively. In addition, loss rate  $\rho_l$  and spurious rate  $\rho_s$  for training data are set to  $[0, 70\%]$  and  $[0, 50\%]$ , respectively.

### 3.2 Performance

#### 3.2.1 Performance on Training.

In Table 2, the STD of measurement error is set about  $MeanPRI * 5\%$ , loss rate  $\rho_l \leq 70\%$ , and spurious rate  $\rho_s \leq 50\%$  at the same time.

Table 2 compares the performance of various combinations of PRI modulations. We can see that, as emitters' complexity increases, the overall accuracy trends for all modulations tend to decrease. This could be explained. As the number of emitters increases, each one experiences an increase in noise. Besides, each emitter also experiences a decrease in the proportion of exact values corresponding to oneself.

**TABLE 1: PARAMETER SETTINGS OF SIMULATED PRI MODULATIONS**

Class	PRI Modulation Type	PRI Value( $\mu$ s)
1	Constant	[300, 1000]
2	Sliding	$[200, 600] + [10, 100] \times n$
3	D&S	$\{[300, 1000] \times [3, 9] \times [3, 7]\}$
4	Wobulated	$[500, 700] + [100, 200] \times \sin([0.2, 0.8] \times n + [0, 2])$

**TABLE 2: TRAINING ACCURACY OF VARIOUS PRI MODULATION MIXTURES**

PRI Modulation Mixtures	Training Accuracy <sup>a</sup> (%)			
	class1	class2	class3	class4
Constant + Sliding	95.22	96.78	-	-
Constant + D&S	92.82	-	95.24	-
Constant + Wobulated	94.16	-	-	97.01
Sliding + D&S	-	97.47	97.81	-
Sliding + Wobulated	-	97.50	-	97.74
D&S + Wobulated	-	-	97.75	97.23
Constant + Sliding+D&S	91.16	93.71	94.71	-
Constant + Sliding + Wobulated	93.49	94.69	-	94.92
Constant + D&S + Wobulated	92.13	-	93.23	94.02
Sliding + D&S + Wobulated	-	92.41	94.52	92.91
Constant + Sliding + D&S + Wobulated	90.73	90.44	92.56	90.81

<sup>a</sup>Data is calculated on a circumstance with *measurement error* about  $\pm \text{MeanPRI} * 5\%$ ,  $\rho_I \leq 70\%$  and  $\rho_S \leq 50\%$ .

**TABLE 3: TEST ACCURACY OF VARIOUS PRI MODULATION MIXTURES**

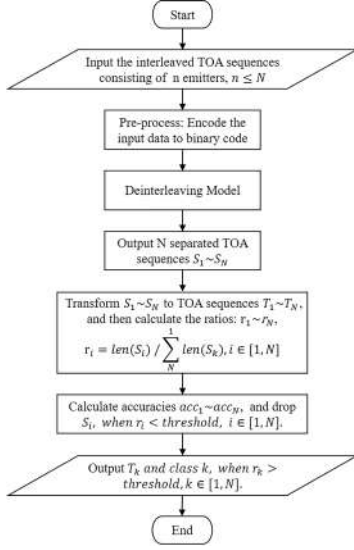
PRI Modulation Mixtures	Test Accuracy <sup>a</sup> (%)			
	class1	class2	class3	class4
Constant + Sliding	98.91	98.06	-	-
Constant + D&S	98.71	-	97.84	-
Constant + Wobulated	98.78	-	-	98.04
Sliding + D&S	-	97.79	98.37	-
Sliding + Wobulated	-	97.81	-	98.18
D&S + Wobulated	-	-	98.19	97.69
Constant + Sliding+D&S	97.56	95.57	96.38	-
Constant + Sliding + Wobulated	97.64	95.81	-	96.16
Constant + D&S + Wobulated	97.39	-	96.18	95.43
Sliding + D&S + Wobulated	-	95.52	96.76	95.87
Constant + Sliding + D&S + Wobulated	96.28	93.45	94.83	93.84

<sup>a</sup>Data is calculated on a circumstance with *measurement error* about  $\pm \text{MeanPRI} * 5\%$ ,  $\rho_I = 0$  and  $\rho_S = 0$ .

### 3.2.2 Performance on Test Under Various Conditions.

We have tested the performance of the proposed approach in a situation only with measurement error about  $\pm \text{MeanPRI} * 5\%$ , without lost noise and spurious noise, i.e.,  $\rho_I = 0$  and  $\rho_S = 0$ , as Table 3 shows.

Furthermore, we have performed a series of experiments in various noisy environments. Fig. 2 illustrates how accuracy changes with fixed  $\rho_I$  in a range of  $[0, 90\%]$  when  $\rho_S \leq 50\%$ , while Fig. 3 displays how spurious rate influences accuracy when  $\rho_S$  varies in a range of  $[0, 90\%]$  with  $\rho_I \leq 70\%$ . The actual combinations of PRI



**Figure 1: Flowchart of main deinterleaving process.** Note that it is not necessary to identify a potential emitter individually for each sequence. Only by discriminating the length ratios of all TOA sequences can we obtain the exact numbers and modulations of potential emitters. For instance, we get four TOA sequences(class1 to class4) in the length of 134, 21, 112 and 89, respectively, from the model, and then the total length of all sequences is  $134 + 21 + 112 + 89 = 356$ . Hence, the corresponding ratios are  $134/356 = 0.376$ ,  $21/356 = 0.059$ ,  $112/356 = 0.315$ , and  $89/356 = 0.250$ . Given  $threshold = 0.10$ , we can conclude that the original interleaved pulse train consists of class1, class3, and class4. Furthermore, since part of the random noise may coincide with at least one of the classes, we could explain why the length the class2 is not equal to 0.

modulations consisting of interleaved TOA pulse train in Fig. 2 and Fig. 3 are as follows:

- Constant + Sliding;
- Constant + D&S;
- Sliding + D&S;
- D&S + Wobulated;
- Constant + Sliding + D&S;
- Constant + D&S + Wobulated;
- Sliding + D&S + Wobulated;
- Constant + Sliding + D&S + Wobulated.

### 3.3 Discussion

According to the results of the experiment shown in Fig. 2 and Fig. 3, we can deduce that:

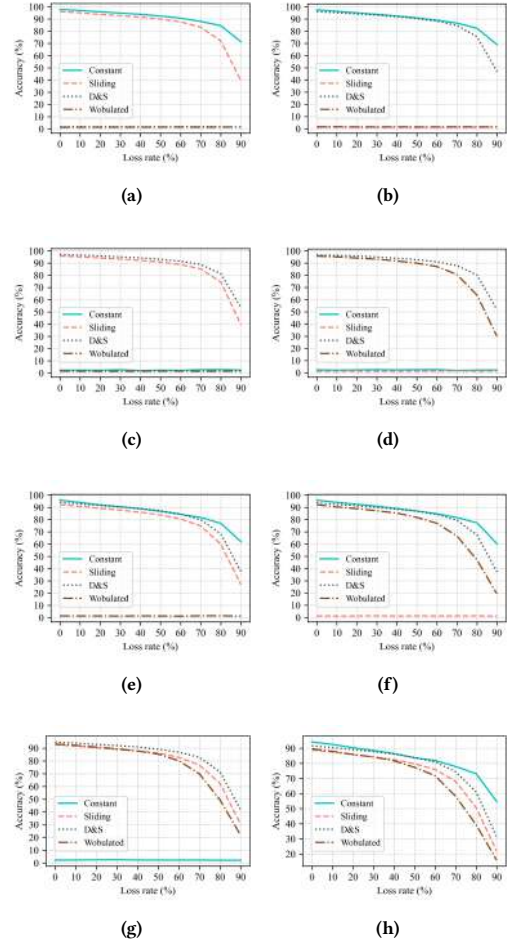
- The model has a great generalization and can adapt to various combinations of emitters without estimating the number of sources iteratively during the process.
- The performance of deinterleaving declines gradually with the increase of noise, but it can still maintain relatively high accuracy at low SNR.

- Compared with spurious noise, the lost noise has a more significant influence on the accuracy of deinterleaving in terms of TOA, for the temporal regularity of the TOA sequence is more severely disrupted as the loss rate rises due to the missing pulses.

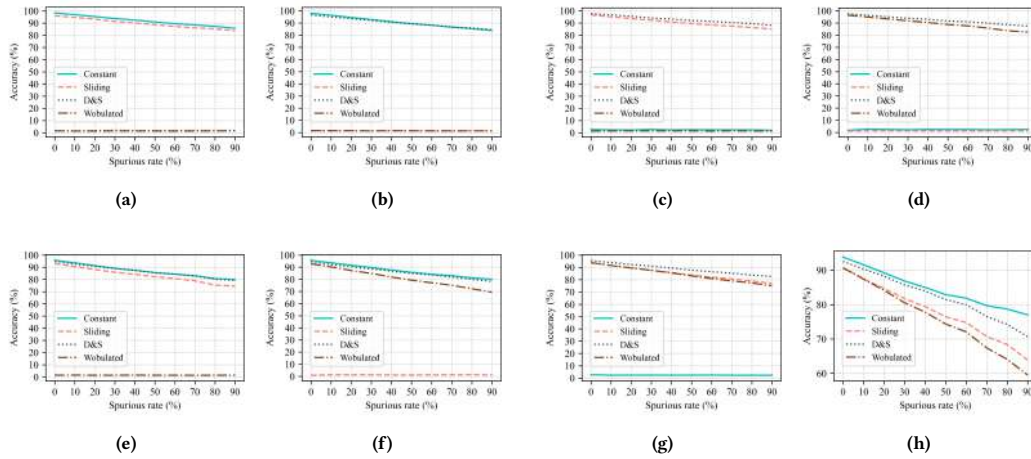
## 4 CONCLUSION

We developed a novel method that relies solely on the TOA parameter based on TCN [6] architecture for deinterleaving the radar pulse trains. The approach has been proven effective in solving the BSS problem, especially for data with prominent temporary features. Simulation results demonstrate that the suggested method has strong robustness against high noise rate and good generalization to various combinations of sources, both modulations and numbers included.

An apparent limitation of the method is that it is offline instead of online, which is more significant in an actual EW environment.



**Figure 2: Test accuracies of deinterleaving for interleaved TOA trains from various combinations of emitters at various loss rates, with a spurious rate of not more than 50%.**



**Figure 3: Test accuracies of deinterleaving for interleaved TOA trains from various combinations of emitters at various spurious rates, with a loss rate of not more than 70%.**

Furthermore, another drawback is that it is a supervised learning approach, meaning that it depends on some specific conditions, and its performance would suffer under a circumstance with multi-function radars included. Further research will continue to explore these issues based on this work.

## ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant No. 91948303-1 and No. 62101575.

## REFERENCES

- [1] S. Amari and A. Cichocki. 1998. Adaptive blind signal processing-neural network approaches. *Proc. IEEE* 86, 10 (1998), 2026–2048. <https://doi.org/10.1109/5.720251>
- [2] Wenhai Cheng, Qunying Zhang, Jiaming Dong, Chuang Wang, Xiaojun Liu, and Guangyou Fang. 2021. An Enhanced Algorithm for Deinterleaving Mixed Radar Signals. *IEEE Trans. Aerospace Electron. Systems* 57, 6 (2021), 3927–3940. <https://doi.org/10.1109/TAES.2021.3087832>
- [3] Kun Chi, Jihong Shen, Yan Li, Yunjie Li, and Sheng Wang. 2021. Multi-Function Radar Signal Sorting Based on Complex Network. *IEEE Signal Processing Letters* 28 (2021), 91–95. <https://doi.org/10.1109/LSP.2020.3044259>
- [4] Alex Erdogan and Kiran George. 2019. Deinterleaving Radar Pulse Train Using Neural Networks. In *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*. New York, NY, 141–147. <https://doi.org/10.1109/CSE/EUC.2019.00036>
- [5] Lipeng Gao, Huiyu Shan, and Fengyou Ji. 2017. A radar signal sorting algorithm based on improved k-means dynamic clustering and sub linear time algorithm. In *2017 First International Conference on Electronics Instrumentation & Information Systems (EIIS)*. Harbin, China, 1–5. <https://doi.org/10.1109/EIIS.2017.8298750>
- [6] Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. 2017. Temporal Convolutional Networks for Action Segmentation and Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, 1003–1012. <https://doi.org/10.1109/CVPR.2017.113>
- [7] Hongbo Li, Jian Zhao, and Yun Zhang. 2019. Signals Deinterleaving for ES systems using Improved CFSFD Algorithm. In *2019 IEEE Radar Conference (RadarConf)*. Boston, MA, USA, 1–5. <https://doi.org/10.1109/RADAR.2019.8835717>
- [8] Xueqiong Li, Zhangmeng Liu, and Zhitao Huang. 2020. Deinterleaving of Pulse Streams With Denoising Autoencoders. *IEEE Trans. Aerospace Electron. Systems* 56, 6 (2020), 4767–4778. <https://doi.org/10.1109/TAES.2020.3004208>
- [9] Zhangmeng Liu. 2021. Pulse Deinterleaving for Multifunction Radars With Hierarchical Deep Neural Networks. *IEEE Trans. Aerospace Electron. Systems* 57, 6 (2021), 3585–3599. <https://doi.org/10.1109/TAES.2021.3079571>
- [10] Yi Luo and Nima Mesgarani. 2019. Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 8 (2019), 1256–1266. <https://doi.org/10.1109/TASLP.2019.2915167>
- [11] H.K. Mardia. 1989. New techniques for the deinterleaving of repetitive sequences. *Radar and Signal Processing, IEE Proceedings F* 136, 4 (1989), 149–154. <https://doi.org/10.1049/ip-f-2.1989.0025>
- [12] D. J. Milojevic and B. M. Popovic. 1992. Improved algorithm for the deinterleaving of radar pulses. In *IEEE Proceedings F '92, Radar and Signal Processing*.
- [13] Manon Mottier, Gilles Chardon, and Frédéric Pascal. 2021. Deinterleaving and Clustering unknown RADAR pulses. In *2021 IEEE Radar Conference (RadarConf21)*. Atlanta, GA, 1–6. <https://doi.org/10.1109/RadarConf2147009.2021.9455272>
- [14] Jiwoo Mun, Seokhyeon Ha, and Jungwoo Lee. 2020. Automotive Radar Signal Interference Mitigation Using RNN with Self Attention. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3802–3806. <https://doi.org/10.1109/ICASSP40776.2020.9053013>
- [15] K. Nishiguchi and M. Kobayashi. 2000. Improved algorithm for estimating pulse repetition intervals. *IEEE Trans. Aerospace Electron. Systems* 36, 2 (2000), 407–421. <https://doi.org/10.1109/7.845217>
- [16] Madhab Pal, Rajib Roy, Joyanta Basu, and Milton S. Bepari. 2013. Blind source separation: A review and analysis. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. 1–5. <https://doi.org/10.1109/ICSDA.2013.6709849>
- [17] Shunqi Su, Xiongjun Fu, Congxia Zhao, Jingfang Yang, Min Xie, and Zhifeng Gao. 2019. Unsupervised k-means combined with SOFM structure adaptive radar signal sorting algorithm. In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*. Chongqing, China, 1–4. <https://doi.org/10.1109/ICSIDP47821.2019.9172926>
- [18] Haibin Wang and Qi Ma. 2013. Method of deinterleaving radar signal based on PRI transform algorithm. *Modern Electronics Technique* 36, 1 (2013), 28–31. <https://doi.org/10.16652/j.issn.1004-373x.2013.01.026>
- [19] N.J. Whittall. 1985. Signal sorting in ESM systems. *Communications, Radar and Signal Processing, IEE Proceedings F* (1985), 226–228. Issue No.4.
- [20] Richard G. Wiley. 2006. *ELINT: the interception and analysis of radar signals*. Artech House, Norwood, MA.
- [21] Yin Xi, Xiongjun Wu, Yingchun Wu, Ye Cai, and Yongwu Zhao. 2019. A Novel Algorithm for Multi-signals Deinterleaving and Two-dimensional Imaging Recognition Based on Short-time PRI Transform. In *2019 Chinese Automation Congress (CAC)*. Hangzhou, China, 4727–4732. <https://doi.org/10.1109/CAC48633.2019.8996290>
- [22] Yixiao Zhang, Wenpu Guo, Kai Kang, Yunlong Yao, Linke Zhang, and Wei Zhang. 2019. Radar signal sorting method based on data field combined PRI transform and clustering. *Systems Engineering and Electronics* 41, 7 (2019), 1510–1516. <https://doi.org/10.3969/j.issn.1001-506X.2019.07.11>

# A Component for Query-based Object Detection in Crowded Scenes

Shuo Mao

Beijing Institute of Technology  
3120201056@bit.edu.cn

## ABSTRACT

Query-based object detection, including DETR and Sparse R-CNN, has gained considerable attention in recent years. However, in dense scenes, end-to-end object detection methods are prone to false positives. To address this issue, we propose a graph convolution-based post-processing component to refine the output results from Sparse R-CNN. Specifically, we initially select high-scoring queries to generate true positive predictions. Subsequently, the query updater refines noisy query features using GCN. Lastly, the label assignment rule matches accepted predictions to ground truth objects, eliminates matched targets, and associates noisy predictions with the remaining ground truth objects. Our method significantly enhances performance in crowded scenes. Our method achieves 92.3% AP and 41.6%  $MR^{-2}$  on CrowdHuman dataset, which is a challenging objection detection dataset.

## CCS CONCEPTS

• **Human-centered computing** → Visualization; Visualization design and evaluation methods.

## KEYWORDS

End-to-end object detection, Graph convolution, Query

### ACM Reference Format:

Shuo Mao. 2023. A Component for Query-based Object Detection in Crowded Scenes. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590039>

## 1 INTRODUCTION

Object detection is a challenging task in computer vision. One of the methods used for object detection is YOLO (You Only Look Once) [1], which has achieved state-of-the-art performance through the integration of various tricks, greatly improving detection accuracy. However, object detection in crowded scenes aims to localize each object using a single bounding box. It poses a challenging problem, an accurate object detector should both detect all objects and avoid predicting duplicate boxes.

To solve this, most of the previous state-of-the-art methods [2–4] first assign one ground truth (GT) to many generated dense

candidate boxes to get a high object recall, as shown in Figure 1. (a). However, this method could result in redundant predictions. So, non-maximum suppression (NMS) is adopted to solve this, but it also results in missing targets in crowd scenes.

In recent years, Carion et al. [5] proposed a novel end-to-end object detection framework, Detection Transformer (DETR). The framework achieves competitive performance, meanwhile, it removes hand-designed parts and post-processing. Different from the traditional box-based [6–8] and point-based [9, 10] paradigms, it is classified as a query-based method. Due to the end-to-end nature, DETR produces only one bounding box for each object, and each query matches one loss. Therefore, the optimization target and the definition of target detection are consistent.

The framework of DETR is shown in Figure 1. (b). However, the shortcomings of DETR are also obvious: requiring a long training time, and the detection results on small objects are not satisfactory. After this paradigm, deformable DETR [11] was proposed to solve the above issues. It just concentrates on a small set of key sampling points, which not only solves the high computational overhead but also improves the performance. Another work is Sparse R-CNN [12], which does not operate on global features, but only calculates features extracted from RoIAlign, and also speeds up the calculation. Based on the above work, we want to design a more complex query-based detector suitable for dense scenes. The current query-based methods [10, 13] perform better in scenes where objects are not dense like COCO [14]. However, these approaches have two key issues. Firstly, in crowded scenes, a single person will produce multiple predictions; Secondly, as the number of decoding layers increases, the performance of the detector tends to saturate.

### 1.1 Motivations

As shown in Figure 2. Most of the bounding boxes can be detected correctly with a confidence score higher than 0.7. However, there are quite a few false positives that exist. These predictions can be regarded as noisy predictions. As we can see, if a target has already been detected, there is no need for the Network to detect it again. As a result of this, the noisy predictions can be filtered out to correct the result.

### 1.2 Our contributions

According to the above analysis, we propose a progressive prediction method equipped with a prediction selector, relation information extractor, query updater, and label assignment to improve the performance of query-based object detectors in handling crowded scenes.

First, we design a prediction selector to select high-score queries, which are named accepted queries, and the rest are regarded as noisy queries. Second, we use GCN to update adjacent queries, where

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590039>

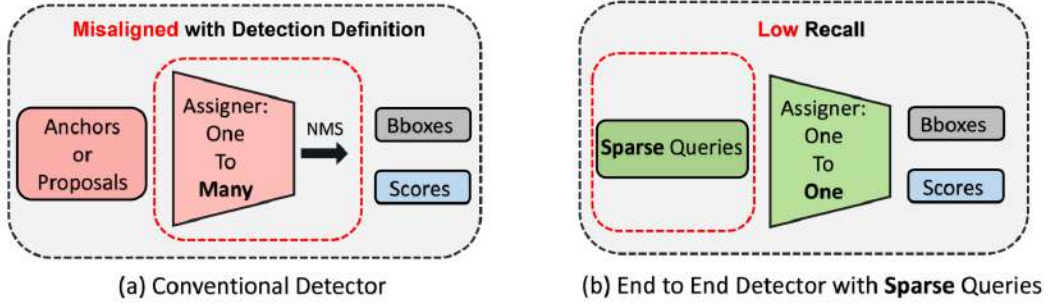


Figure 1: (a) Dense priors in traditional object detectors no longer exist in (b) end-to-end object detectors.

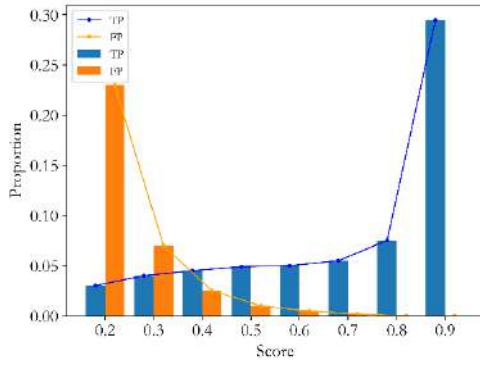


Figure 2: False positive ratio in sparse R-CNN

adjacent means that the IoU of the two detection boxes is greater than 0. Finally, we develop a new one-to-one label assignment rule to assign samples among the accepted and refined noisy queries step by step. With these methods, the issues above can be solved well.

Our framework fits into many architectures and has resulted in a significant improvement over query-based detectors. With our method, Sparse RCNN [12] with ResNet-50 [15] backbone achieves 91.4% AP, 41.4%  $MR^{-2}$  on the challenging dataset CrowdHuman [33], outperforming existing box-based method MIP. Moreover, with our method, deformable DETR [11] also obtains 91.5% AP and 41.6%  $MR^{-2}$  on CrowdHuman.

## 2 RELATED WORK

### 2.1 End-to-end Object detection

A lot of work has been done to get rid of post-processing operations. DETR [5] introduces learnable queries to represent objects and assigns a one-hot label for each ground truth in bipartite matching. As a result, DETR produces one prediction for each object. After that, deformable-DETR [11] uses a simple and effective iterative bounding box refinement mechanism. This framework replaces the Transformer attention module with deformable attention modules, which limits the attention field of each query to a local area. So it improves the detection performance and speeds up the convergence.

Sparse R-CNN applies RoIAlign to limit the attention field of each query to a local region. Meanwhile, the work uses a set of learnable queries to derive the instances instead of anchors. UP-DETR [17] and TSP [18] analyze the slow convergence of DETR and propose faster convergence methods. The above methods perform worse than traditional methods.

### 2.2 Object detection in crowded scenes

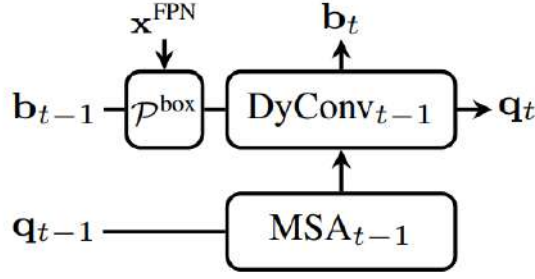
The occlusion problem in pedestrian detection is of great concern. Researchers have proposed serious methods to address this problem, including detecting a part to judge the whole person [19–21] and improving the hand-crafted rules in the detection framework. Recently, CNN-based detection methods have become the mainstream for target detection in dense scenes and have achieved good results. Several works propose new loss functions [22, 23] to address this problem. Since the premise of using NMS is that the instances are relatively scattered, It is not suitable for crowd scenes. Soft-NMS [24] and Softer-NMS [25] replace hard removal of nearby proposals with score decay. Several works propose to use of a neural network to simulate the function of NMS for duplicate removal [26, 27]. Others explore NMS-aware training, such as NMS with adaptive threshold [28, 29], feature embedding [30], and multiple predictions with set suppression [16, 31], to tackle the problem of object detection in crowded scenes.

## 3 METHODOLOGY

In this section, we first review Sparse RCNN. Secondly, we will explain how to deploy our approach to Sparse RCNN. Finally, we will discuss the differences in detector designs.

### 3.1 Query-Based Object Detector Framework

Our approach can be deployed on most query-based object detectors, e.g., DETR, Sparse RCNN, and deformable-DETR. We choose Sparse RCNN to illustrate our approach. Figure 3 explains the pipeline of Sparse RCNN, which can also be formulated as:



**Figure 3: The diagram of the decoding stage. P – RoIAlign-Pool, DynConv – Dynamic Convolution, MSA – Multi-head Self-Attention, S – Prediction Selector, R – Relation Information Extractor, QU – Query Updater**

$$\begin{aligned}
 x_{t-1} &\leftarrow \mathcal{P}^{box}(x^{FPN}, b_{t-1}), \\
 q_{t-1}^* &\leftarrow MSA_{t-1}(q_{t-1}), \\
 q_t &\leftarrow DynConv_{t-1}(q_{t-1}^*, x_{t-1}), \\
 b_t &\leftarrow B_{t-1}(q_t),
 \end{aligned}$$

where  $q \in \mathbb{R}^{N \times d}$  denotes the learnable object query.  $N$  and  $d$  denote the number and dimension of query  $q$ , respectively. At stage  $t$ , a RoIAlign  $P$  box extracts RoI features from FPN features  $x^{FPN}$ , under the guidance of bounding box  $b_{t-1}$  predicted by the previous stage. After that, a multi-head self-attention module  $MSA_{t-1}$  is applied to the input query  $q_{t-1}$  to get the transformed query  $q_{t-1}^*$ . Then, a dynamic convolution module  $DynConv_{t-1}$  takes both  $x_{t-1}$  and  $q_{t-1}^*$  as inputs and performs dynamic convolution to generates  $q_t$  for the next stage. Simultaneously,  $q_t$  is fed into the box prediction branch  $B_{t-1}$  for current bounding box prediction  $b_t$ , which is the input of the next stage  $t$ .

### 3.2 Our method

We propose a progressive prediction method that is equipped with a prediction selector, GCN query updater, and label assignment. Figure 4 illustrates our pipeline.

**3.2.1 Prediction selector.** The prediction selector is designed to select queries with high scores as accepted queries, leaving the rest

as noisy queries. We set a threshold  $s$ , and the queries with scores greater than  $s$  are accepted, the others are treated as noisy queries.

**3.2.2 GCN Query updater.**

$$adj \in \mathbb{R}^{N \times N}$$

$$edge \leftarrow \{(i, j) | IoU(\mathcal{P}_i, \mathcal{P}_j) > 0\}, 0 \leq i, j < N$$

$$\mathcal{F}_{new} \leftarrow GCN(\mathcal{F}, adj)$$

We develop a GCN query updater to refine the features of noisy queries, which is formulated in Equ. (2). The input of this process is all proposal features and the adjacency matrix  $adj$ , where  $adj$  is an  $N \times N$  matrix, representing whether there is an intersection (i.e.,  $IoU > 0$ ) between the detection boxes corresponding to the  $N$  queries. GCN updates the noisy predictions through graph convolution operations.

**3.2.3 Label assignment.** We develop a new label assignment rule to suit our framework. Firstly, we match the accepted predictions  $D_{t-1}^h$  with the ground truth set of objects. Secondly, we remove the targets that have been matched. Thirdly, we match the noisy predictions and the rest ground truth objects.

## 4 EXPERIMENTS

### 4.1 Datasets

We adopt CrowdHuman [33] to evaluate our framework. The CrowdHuman dataset has an average of 22.64 objects per image and overlaps with IoU greater than 0.5 has an average of 2.40.

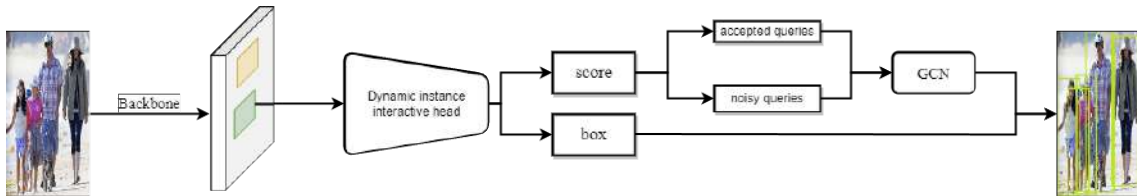
### 4.2 Evaluation metrics

We mainly take three criteria as our evaluation metrics:

**Averaged Precision (AP).** AP reflects both the precision and recall ratios of the detection results. A larger AP means better performance.

**$MR^{-2}$ ,** which is short for log-average Miss Rate on False Positive Per Image (FPPI) in  $[10^{-2}, 100]$ .  $MR^{-2}$  is very sensitive to false positives (FPs). Smaller  $MR^{-2}$  means better performance.

**Jaccard Index (JI)** is used to evaluate the counting ability of a detector. JI evaluates how much the prediction set overlaps the ground truths. We use the official SDK of CrowdHuman for JI calculation. A larger JI indicates better performance.



**Figure 4: The diagram of the proposed progressive end-to-end object detection framework. First, the prediction selector selects queries associated with high confidence scores as accepted queries, leaving the rest as noisy queries. Then the GCN model updates the noisy queries.**

**Table 1: Comparisons of different methods on CrowdHuman**

Method	Queries	AP	$MR^{-2}$	Jl
<i>Box-based</i>				
RetinaNet [6]	-	85.3	55.1	73.7
ATSS [8]	-	87.0	51.1	75.9
ATSS + MIP	-	88.7	51.6	77.0
FPN [7]+NMS	-	85.8	42.9	79.8
FPN +soft NMS [24]	-	88.2	42.9	79.8
FPN+MIP [16]	-	90.7	41.4	82.4
<i>Point-based</i>				
FCOS [32]	-	86.8	54.0	75.7
FCOS +MIP	-	87.3	51.2	77.3
PHOTO [10]	-	89.1	47.8	79.3
<i>Query-based</i>				
DETR [5]	100	75.9	73.2	74.4
PEDR	1000	91.6	43.7	83.3
deformable DETR [11]	1000	91.5	43.7	83.1
Sparse RCNN [12]	500	90.7	44.7	81.4
Sparse RCNN	750	91.3	44.8	81.3
Ours	500	91.8	<b>41.4</b>	83.2
Ours	750	<b>92.3</b>	41.6	83.3

### 4.3 Implementation details

We take Sparse RCNN as our default instantiation and use ResNet-50 as the backbone, which is pre-trained on ImageNet. We use Adam optimizer to train our model, whose momentum is 0.9 and weight decay is 0.0001. Models are trained for 50, 000 iterations. The initial learning rate is 0.00005 and reduced by a factor of 0.1 at iteration 35,000.  $\lambda_{cls} = 2$ ,  $\lambda_{L1} = 5$ ,  $\lambda_{giou} = 2$ . The default number of proposal boxes, proposal features, and stages are set to 500, 500, and 6, respectively. Moreover, the dimension of intermediate features in relationship extractor R is 256. The gradients are detached at proposal boxes from the second stage to stabilize training. Besides, those negative samples, whose intersection-over-area (IoA) between any ignore region is higher than a threshold of 0.7, are not involved in training. Further, the hyper-parameters  $s$  and  $\theta$  are 0.7 and 0.4 by default in the different query-based detectors.

+MIP represents multiple instance prediction with set NMS as post-processing.

The result is shown in Table 1. Our approach performs better than Box-based, point-based, and query-based methods, which means that our method shows excellent performance in crowd scenes. Equipped with our framework, Sparse RCNN is 1.1%, 3.3%, and 1.8% better than its original version. When we increase the number of queries to 750, Sparse RCNN achieves 92.3% AP, 41.6%  $MR^{-2}$  and 83.3% Jl. This is because more queries can cover more patterns of objects in the image, such as scale, size, position, and other characteristics.

The improvement result from the relation information extractor can reduce false positives and recall false negatives. Additionally, when equipped with the new local self-attention LMSA, the performance is further boosted, since the local self-attention can reduce duplicates effectively. Finally, we reinitialize the embeddings to

approximate the new data distribution of noise predictions, which can slightly improve  $MR^{-2}$ .

### 4.4 Ablation study

The results of the ablation experiment are shown in Table 2, equipped with our framework, the performance of Sparse R-CNN has been significantly improved, which proves the effectiveness of our method.

## 5 CONCLUSION

In this paper, we propose advanced an end-to-end object detection in crowded scenes. Equipped with our framework, two representative query-based methods, Sparse RCNN and deformable DETR yield a remarkable boost in crowded scenes. Since these two methods are inherently computationally intensive, our method is difficult to apply to practical scenarios. We believe that the performance of our method can be further improved and adopted with a better loss function or a feature extraction method.

## REFERENCES

- [1] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, realtime object detection [C]. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779–788.
- [2] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018)
- [3] Gao, P., Zheng, M., Wang, X., Dai, J., Li, H.: Fast convergence of detr with spatially modulated co-attention. arXiv preprint arXiv:2101.07448 (2021)
- [4] Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European Conference on Computer Vision, pages 213–229, 2020.
- [6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2):318–327, 2020.

Table 2: Ablation study

Method	Queries		$AP$	$MR^{-2}$	$JI$
Sparse RCNN	500	90.7		44.7	81.4
Sparse RCNN + Ours	500	91.8		<b>41.4</b>	83.2
Sparse RCNN	750	91.3		44.8	81.3
Sparse RCNN +Ours	750	<b>92.3</b>		41.6	83.3

- [7] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 936–944, 2017.
- [8] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. arXiv preprint arXiv:1912.02424, 2019.
- [9] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 9626–9635, 2019.
- [10] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. arXiv preprint arXiv:2012.03544, 2020.
- [11] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In ICLR 2021: The Ninth International Conference on Learning Representations, 2021.
- [12] Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al.: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14454–14463 (2021)
- [13] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In CVPR, 2017.
- [14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In European Conference on Computer Vision, pages 740–755, 2014.
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.
- [16] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. pages 12214–12223, 2020.
- [17] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1601–1610, June 2021.
- [18] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris Kitani. Rethinking transformer-based set prediction for object detection. CoRR, abs/2011.10881, 2020.
- [19] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In AAAI 2020: The Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.
- [20] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Li, and Xudong Zou. Relational learning for joint head and human detection. In AAAI 2020: The Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.
- [21] Kevin Zhang, Feng Xiong, Peize Sun, Li Hu, Boxun Li, and Gang Yu. Double anchor r-cnn for human detection in a crowd. arXiv preprint arXiv:1909.09998, 2019.
- [22] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. arXiv preprint arXiv:1711.07752, 2017.
- [23] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision (ECCV), pages 637–653, 2018.
- [24] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-nms – improving object detection with one line of code. arXiv preprint arXiv:1704.04503, 2017.
- [25] Yihui He, Xiangyu Zhang, Marios Savvides, and Kris Kitani. Softer-nms: Rethinking bounding box regression for accurate object detection. 2018.
- [26] J. Hosang, R. Benenson, and B. Schiele. Learning non-maximum suppression. In CVPR, 2017.
- [27] Lu Qi, Shu Liu, Jianping Shi, and Jiaya Jia. Sequential context encoding for duplicate removal. In Advances in Neural Information Processing Systems, volume 31, pages 2049–2058, 2018.
- [28] Jan Hendrik Hosang, Rodrigo Benenson, and Bernt Schiele. A convnet for non-maximum suppression. In 38th German Conference on Pattern Recognition, pages 192–204, 2016.
- [29] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6459–6468, 2019.
- [30] Niels Ole Salscheider. FeatureNMS: Non-maximum suppression by learning feature embeddings. In 2020 25th International Conference on Pattern Recognition (ICPR), pages 7848–7854, 2021.
- [31] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10750–10759, 2020.
- [32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: A simple and strong anchor-free object detector. 2021.
- [33] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting humans in a crowd. arXiv preprint arXiv:1805.00123, 2018.

# Explainable Deep Learning for Medical Image Segmentation With Learnable Class Activation Mapping

Kaiyue Wang

KaiyueWang@bupt.edu.cn  
Beijing University of Posts and Telecommunications  
Beijing, China

Yining Wang

Chinese Academy of Medical Sciences and Peking Union  
Medical College  
Beijing, China  
yiningpumc@163.com

Sixing Yin

Beijing University of Posts and Telecommunications  
Beijing, China  
yinsixing@bupt.edu.cn

Shufang Li

Beijing University of Posts and Telecommunications  
Beijing, China  
lisf@bupt.edu.cn

## ABSTRACT

Medical image segmentation is crucial for facilitating pathology assessment, ensuring reliable diagnosis and monitoring disease progression. Deep-learning models have been extensively applied in automating medical image analysis to reduce human effort. However, the non-transparency of deep-learning models limits their clinical practicality due to the unaffordably high risk of misdiagnosis resulted from the misleading model output. In this paper, we propose a explainability metric as part of the loss function. The proposed explainability metric comes from Class Activation Map(CAM) with learnable weights such that the model can be optimized to achieve desirable balance between segmentation performance and explainability. Experiments found that the proposed model visibly heightened Dice score from 69.5% to 72.2%, Jaccard similarity from 59.32% to 61.82% and Recall from 64.21% to 82.61% respectively compared with U-net. In addition, results make clear that the drawn model outdistances the conventional U-net in terms of explainability performance.

## CCS CONCEPTS

• Computing methodologies → Image segmentation; Image segmentation.

## KEYWORDS

deep learning, medical image segmentation, image segmentation interpretability, Class Activation Mapping

### ACM Reference Format:

Kaiyue Wang, Sixing Yin, Yining Wang, and Shufang Li. 2023. Explainable Deep Learning for Medical Image Segmentation With Learnable Class Activation Mapping. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590040>

## 1 INTRODUCTION

Medical image analysis is essential for facilitating pathology analysis, ensuring reliable diagnosis and monitoring disease progression[8]. As a research upsurge, deep-learning models have been extensively applied in automating medical image analysis to reduce human effort[13]. Compared with traditional methods, deep Convolutional Neural Networks (CNNs) have higher representation ability and

can use a large number of feature channels to automatically learn the most useful features. However, the non-transparency of deep-learning models limits their clinical practicality due to the unaffordably high risk of misdiagnosis resulted from the misleading model output, which makes explainable deep-learning models that can be understood by experts imperative[10, 18, 19, 21]. There have been prior research[6, 16, 24, 25] on understanding how deep-learning models works by visualizing feature maps or convolutional kernels in distinct layers of a convolutional neural network[23]. Recent works like class activation mapping/map (CAM) and gradient-weighted class activation mapping (Grad-CAM) study how different parts of input (e.g., pixels in an image) contribute to the model output[26]. However, most of the prior works focus only on analyzing the explainability of well-trained deep-learning models. We argue that it is of great importance to involve the explainability in algorithm design (e.g., loss function definition), which necessitates a metric to reasonably evaluate the explainability of a model[2]. With such motivation, in this paper, we investigate medical image segmentation with explainable deep learning in a CNN architecture and propose a explainability metric as part of the loss function. The proposed explainability metric comes from CAM(class activation mapping/map) with learnable weights such that the model can be trained to strike a balance between segmentation performance and explainability. We propose a CNN architecture with class activation mapping/mapping to achieve more accurate and interpretable medical image segmentation. At the same time, we understand the most important spatial positions and channels in image segmentation through interpretable heat maps.

The framework for medical image segmentation and explainable metric is shown in Fig. 2, where medical images both segmentation loss and explainability loss are taken into account. Through convolutional neural networks (feature extractors), we obtain high-level feature maps from the input medical images and we introduce the trainable weights, which represent the importance of the corresponding feature map. On this basis, medical image segmentation and interpretable heat maps are obtained.

The contributions of this paper are summarized as follows:

1) Without reducing the segmentation performance of the network model, we introduce a new module to make the proposed network interpretable. We enhance the structure of the CNN(U-net) model and introduce trainable weights to the high-level feature

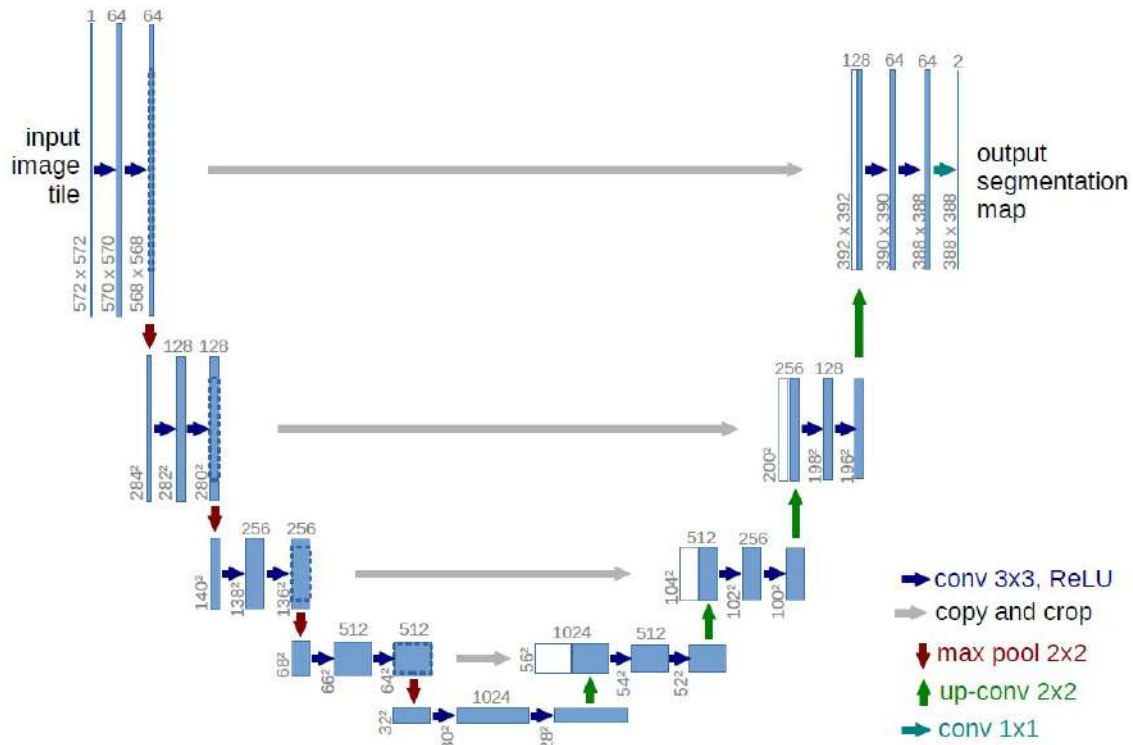


Figure 1: U-Net

maps with the most semantic information, and make the proposed model interpretable by highlighting important pixels and visualizing.

2) By optimizing and improving the loss function and setting interpretability measures, the image segmentation network model and interpretability network model are alternately optimized during the training process, which enhances the interpretability and the performance of image segmentation consequences.

3) Through comprehensive experiments, the results show that the image segmentation and interpretability performance of the proposed model outperforms the existing U-net model.

## 2 RELATED WORK

### 2.1 Neural network model

In medical image segmentation, several popular models have good results, such as FCN[14], U-net[17], Mask RCNN[9] and SegNet[1]. They all adopt the coder-decoder structure. For example, Mask RCNN, which integrates detection, classification and segmentation functions, also uses the FCN structure as the segmentation part of its network structure. The difference between FCN and U-net is that the decoder part of the former is simpler, only uses deconvolution operation and uses addition operation when connecting with the structure of each part of the encoder, while the decoder of U-net structure not only uses convolution structure and stack operation.[4]. Many experiments have proved that the segmentation effect of these models have great performance[12, 27].

“U-shaped” coding-decoding formation are not only uncomplicated and plain, but also have great effects and expression in segmentation result. In this paper, we use the typical U-net structure[17] to extract the feature map. The u-net network structure will be introduced below. U-Net structure is divided into encoder and decoder parts, and realizes fast connection between different resolutions. Different encoders are seen as feature extractors to obtain distinct high-dimensional features. The role of the decoder is to use these encoded features to recover the segmented object. In the u-net network structure, the two core parts are the encoder (down-sampling) and the decoder (up-sampling) [11]. In addition, down-sampling increases robustness through image transformation, such as image translation and rotation, so as to lessen over-fitting and reduce the quantity of computation while increasing the receptive field. Up-sampling can restore the size of the extracted abstract features, so as to obtain segmentation results consistent with the size of the original image. In this experiment, four down-sampling and four up-sampling are selected. In the feature extraction stage, with the deepening of the network structure, more abstract features can be captured. U-net is distinguished by the elongated connection between the encoder and decoder. The chief function of long connection is to lessen the information loss caused by down-sampling and the method used is to integrate the information of the input image. Therefore, this paper believes that the semantic information of the last group of characteristic graphs is the strongest in the u-net network structure, so our newly added interpretability

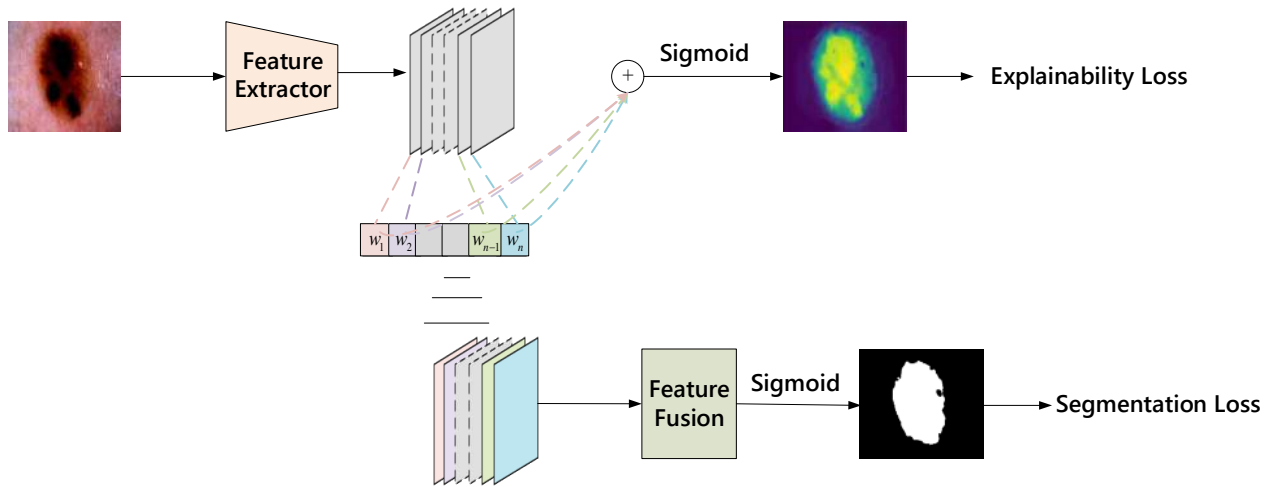


Figure 2: Overall framework of segmentation and explainable heatmap.

module is also based on the last group of characteristic graphs. The interpretability module in this model will be described in detail below.

## 2.2 Attention mechanism model architecture

The innate character of attention mechanism [15, 20] is to locate interested information and suppress useless information. Attention mechanisms can be divided into different types, such as spatial attention model, channel attention model, channel and spatial mixed attention model [7, 22]. In image processing, the contribution of different regions to the task is different. Only the regions related to the task are needed to be concerned [3]. For instance, in the image segmentation task, the spatial attention model is to seek the most significant piece of the network processing. Its function is to enhance the region of interest in the feature map and restrain the background or irrelevant parts.

## 3 METHODOLOGY

In the medical sphere, it is not dependable to hand over the decision-making procedure to the neural network. Consequently, for the sake of enhancing the efficiency of diagnosis and remedy, medicine will work with the assistance of artificial intelligence. In order to allow physicians to comprehend as much as possible how the network makes such decisions and understand the regions that the neural network model focuses on, the study of in-depth learning interpretability becomes increasingly necessary. It can be made out that it is extremely significant to show how the network makes policy and follow with interest to the interpretability of the network structure. We define a measurement method and make the results as close as possible to the good interpretability we defined. Therefore, a network model is proposed, which generates visual interpretation based on convolutional network. The baseline used in this paper is U-net network. On the premise of not reducing the segmentation effect of the original network model, we add a new interpretability module. Visualization is achieved by highlighting important pixels.

On this basis, the criterion of interpretability is drawn into and used as a loss function to train the network model to expect the content of visualization to be similar to the content of specialist note to a certain extent. The study procedure is essential divided into two sections, one is the structure of neural network model structure for medical image segmentation, the other is the structural design of interpretability display and interpretability measurement.

Feature extraction model: we use U-Net as the basis, which is shown in Figure 1.

Explainability model: In order to make the model interpretable, we show its heat map by weighting each channel in the final feature map with the strongest semantic information. Its interpretability is that the brighter the color of the heat map, the greater the effect of the final segmentation decision, and creatively introduce a mean square error loss function to optimize its interpretability heat map. The specific implementation is to introduce a random initialization parameter Linear layer to represent the weight of the corresponding channel. The network model used to obtain features is the u-net network, which can sample and fuse lower-level features with high resolution and higher-level features with powerful semantic information. Segmentation network model: feature fusion is performed on the weighted channel, and the segmentation result can be obtained through a sigmoid function.

As a weighted sum of feature maps, CAMs have been found useful in understanding how different parts of the model input contribute to the output. However, those weights are trained to minimize the segmentation loss (e.g., pixel-wise cross entropy) without considering model explainability. Thus, a metric to evaluate model explainability is necessary as well in order to supervise model training. Intuitively, in image segmentation tasks, CAMs should be closely related to the ground-truth segmentation label, i.e., the contour of the reign of interest should be also identified in CAMs. Therefore, we consider the difference between the CAM and the ground-truth segmentation label as the explainability metric. The proposed model is depicted in Fig. 2, where both segmentation loss

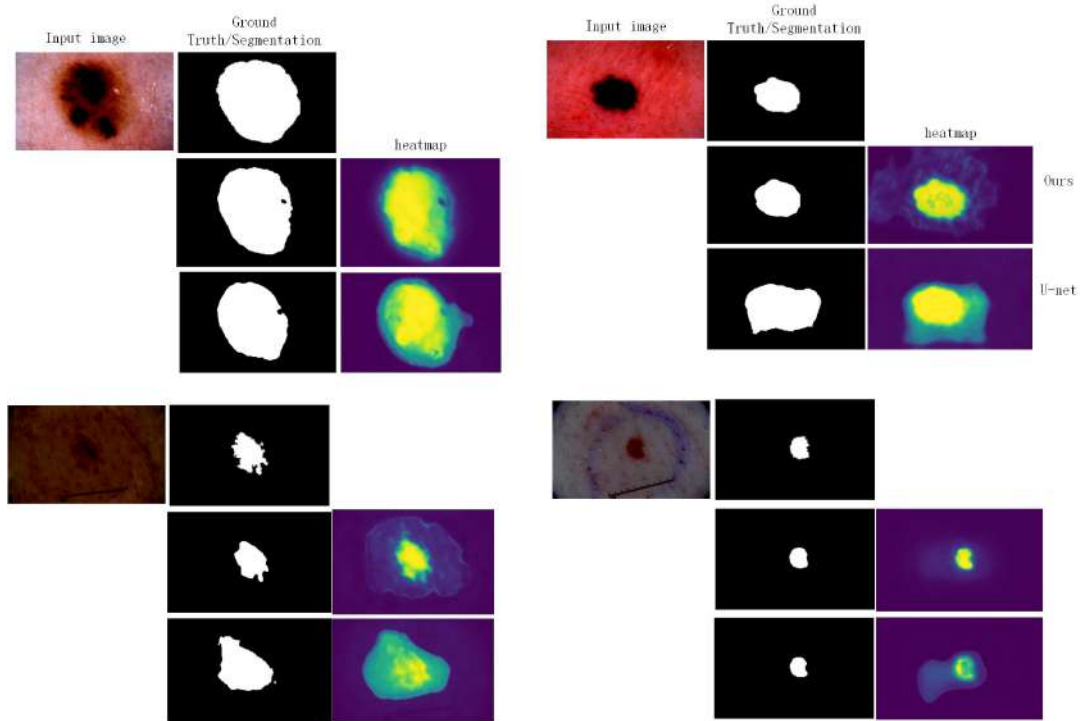


Figure 3: Visual comparison for segmentation performance and heatmap.

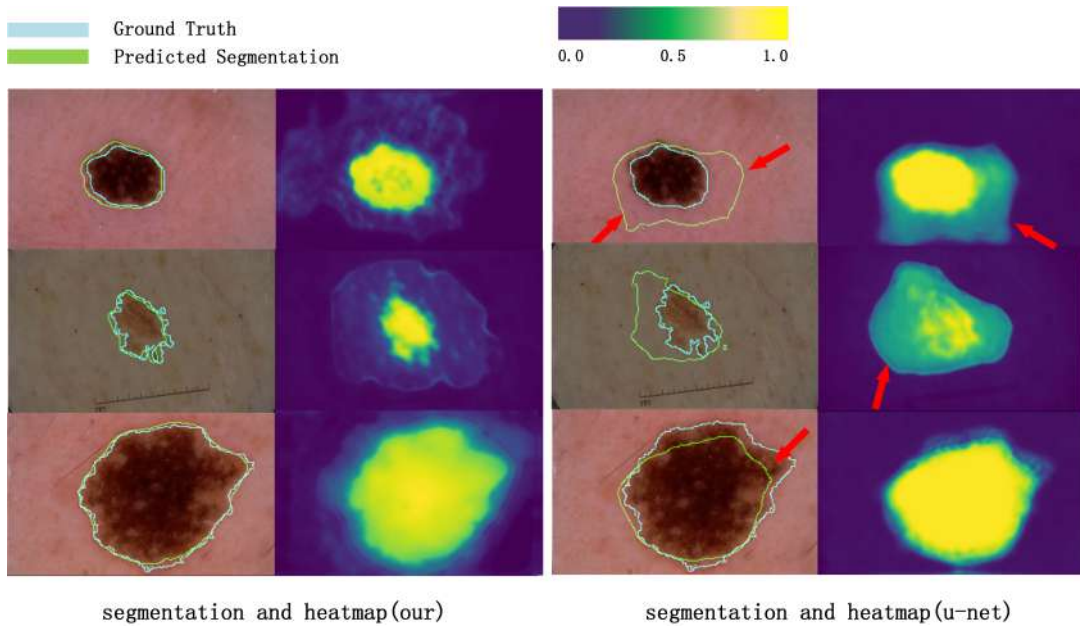


Figure 4: Visual comparison for segmentation performance and heatmap.

and explainability loss are taken into account. As part of the model loss, the proposed explainability metric then makes the weights in computing CAM ( $w_1, \dots, w_n$ ) learnable. We simply use binary cross

entropy for segmentation loss and mean square error for explainability loss. The overall loss of the proposed model is weighted sum of the two with a weight as a hyper-parameter.

**Table 1: image segmentation Performance Comparison**

Metrics	Accuracy	Sensitivity	Jaccard Similarity	Dice Coefficient
Ours	0.9310	0.8901	0.7480	0.8385
U-net	0.8788	0.6421	0.5903	0.6902

#### 4 EXPERIMENT

Our dataset is the public ISIC dataset[5], which contains 2594 the binary skin lesion segmentation from dermoscopic images. We stochastic break the dataset into 1815, 259, 518 for training set, validation set and testing set respectively. During the training, we conducted data improvement by stochastic cropping, flipping horizontal and vertical and random rotation with a angle in  $(-\frac{\pi}{6}, \frac{\pi}{6})$ . Our model is carried out and trained in the Python framework, in which we use adaptive moment estimation (Adam) and the learning rate is set to 10, and the weight decay is 10-8; The batch size of each training is 1 and the number of iterations is 250. The feature extraction model is mainly based on the U-net model, which mainly contains three convolution layers. In the convolution layer, we set the size, number and sliding step of the convolution core, and set the maximum pooling layer after the first two convolution layers, and finally two complete full-connection layers. The first full connection layer has a dimension of 256, the second full connection layer has a dimension of 3, and finally the Softmax layer. In the proposed network model, we use ReLU as the activation function of all intermediate layers, and then perform batch normalization for optimization. For the image segmentation model, the final output is based on the prediction of each pixel with a range of 0 and 1, which represents background and the segmentation target respectively, and we use the cross entropy loss as the segmentation module loss. The algorithm principle of the BCEloss function is based on the comparison between each pixel and the label, so it is also called the pixel by pixel cross loss. The predicted value of each pixel vector is evaluated separately, and then all pixels are averaged. The numerical value of the loss function indicates whether the two distributions are similar, and a small value represents similarity. We used BCEloss function in the chief modules which is defined as

$$MSEloss = \frac{1}{N} \sum_i (y_i - p_i)^2 \quad (1)$$

and MSEloss function in the explainable module for the training which is defined as

$$BCEloss = \frac{1}{N} \sum_i -(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)) \quad (2)$$

where  $y_i$  and  $p_i$  denote the pixel sets of the predicted and ground-truth segmentations, respectively. The overall loss of the proposed model is defined as

$$loss = 0.65 \cdot BCEloss + 0.35 \cdot MSEloss \quad (3)$$

#### 5 RESULTS

We compare the proposed model with the conventional U-net models[17] with the public ISIC dataset[5]. Typical results of visual comparison for segmentation performance and heatmaps are shown in Figure 3 and Figure 4. It can be observed that the proposed model presents not only high segmentation performance but also satisfactory explainability since the heatmaps are consistent with the ground-truth contours of the target lesions. In contrast, heatmaps generated by The U-net model are ambiguous to identify the contour of target lesion and the brighter of the color of the heatmaps, the greater impact on the final segmentation decision. Compared to the conventional U-Net model, it can be shown in Table 1, the proposed model improves the average Dice score from 69.5% to 83.8%, Jaccard similarity from 59.32% to 74.8%, respectively.

#### 6 CONCLUSION

In our work, the focus is to design an interpretable network model through learnable class activation mapping. The whole network consists of feature extraction module, interpretability module and image segmentation module, which gives weight to the last set of feature maps and introduces a loss function to optimize its interpretable heat map. During training process, our proposed network can virtually obtain the features and optimize loss function of the image segmentation module and loss function of the interpretability module, which not only enhances the segmentation effect of the network model, but also has good interpretability. The comparative experiments on ISIC2018 dataset verify that the algorithm in this paper has certain advantages. The next step is to further analyze medical image segmentation and try to combine new methods such as more advanced network models with interpretable modules, and test and improve different networks to achieve a better effect.

#### ACKNOWLEDGMENTS

The research was supported by Beijing Natural Science Foundation (Z210013).

#### REFERENCES

- [1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 39, 12 (2017), 2481–2495.
- [2] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.
- [3] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. 2016. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3640–3649.
- [4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part VII* (Munich, Germany).

- Springer-Verlag, Berlin, Heidelberg, 833–851. [https://doi.org/10.1007/978-3-030-01234-2\\_49](https://doi.org/10.1007/978-3-030-01234-2_49)
- [5] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kallou, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *CoRR* abs/1902.03368 (2019). arXiv:1902.03368 <http://arxiv.org/abs/1902.03368>
  - [6] Alexey Dosovitskiy and Thomas Brox. 2015. Inverting Visual Representations with Convolutional Networks. <https://doi.org/10.48550/ARXIV.1506.02753>
  - [7] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. 2019. Dual Attention Network for Scene Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
  - [8] Geert Litjens, Thijs Kooi, Babak Ehteshami, Bejnordi, Arnaud, Arindra, and Adiyoso. 2017. A survey on deep learning in medical image analysis. *Medical Image Analysis* (2017).
  - [9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.
  - [10] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923* (2017).
  - [11] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-Excitation Networks. *CoRR* abs/1709.01507 (2017). arXiv:1709.01507 <http://arxiv.org/abs/1709.01507>
  - [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* 60, 6 (may 2017), 84–90. <https://doi.org/10.1145/3065386>
  - [13] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. 2017. A survey on deep learning in medical image analysis. *Medical image analysis* 42 (2017), 60–88.
  - [14] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.
  - [15] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 375–383.
  - [16] Aravindh Mahendran and Andrea Vedaldi. 2014. Understanding Deep Image Representations by Inverting Them. *CoRR* abs/1412.0035 (2014). arXiv:1412.0035 <http://arxiv.org/abs/1412.0035>
  - [17] O. Ronneberger, P. Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Springer International Publishing* (2015).
  - [18] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. 2018. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part I*. Springer, 421–429.
  - [19] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services 1* (10 2017), 1–10.
  - [20] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2018. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. <https://doi.org/10.48550/ARXIV.1808.08114>
  - [21] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. 2019. Attention gated networks: Learning to leverage salient regions in medical images. *Medical image analysis* 53 (2019), 197–207.
  - [22] Yi Wang, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni. 2018. Deep attentional features for prostate segmentation in ultrasound. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part IV 11*. Springer, 523–530.
  - [23] MD Zeiler and R. Fergus. 2014. Visualizing and Understanding Convolutional Networks. *Springer, Cham* (2014).
  - [24] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2901 (2013). arXiv:1311.2901 <http://arxiv.org/abs/1311.2901>
  - [25] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2014. Object Detectors Emerge in Deep Scene CNNs. <https://doi.org/10.48550/ARXIV.1412.6856>
  - [26] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. Learning Deep Features for Discriminative Localization. *CVPR* (2016).
  - [27] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning Deep Features for Scene Recognition using Places Database. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.),

# An Emotion Recognition Method Based On Feature Fusion and Self-Supervised Learning

Xuanmeng Cao

Department of Computer Science and Engineering,  
University of Electronic Science and Technology of China  
Chengdu, China  
cao\_xuanmeng@std.uestc.edu.cn

Ming Sun\*

Department of Computer Science and Engineering,  
University of Electronic Science and Technology of China  
Chengdu, China  
sunm@uestc.edu.cn

## ABSTRACT

Emotional diseases being represented in many kinds of human mental and cardiac problems, demanding requirements are imposed on accurate emotion recognition. Deep learning methods have gained widespread application in the field of emotion recognition, utilizing physiological signals. However, many existing methods rely solely on deep features, which can be difficult to interpret and may not provide a comprehensive understanding of physiological signals. To address this issue, we propose a novel emotion recognition method based on feature fusion and self-supervised learning. This approach combines shallow features and deep learning features, resulting in a more holistic and interpretable approach to analyzing physiological signals. In addition, we transferred the self-supervised learning method from processing images to signals, which learns sophisticated and informative features from unlabeled signal data. Our experimental results are conducted on WESAD, a publicly available dataset and the proposed model shows significant improvement in performance, which confirms the superiority of our proposed method compared to state-of-the-art methods.

## CCS CONCEPTS

• **Computing methodologies** → *Philosophical/theoretical foundations of artificial intelligence.*

## KEYWORDS

emotion recognition, physiological signals, self-supervised learning, feature fusion

### ACM Reference Format:

Xuanmeng Cao and Ming Sun. 2023. An Emotion Recognition Method Based On Feature Fusion and Self-Supervised Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590041>

## 1 INTRODUCTION

Emotional problems, such as mental diseases, can take a significant toll on both the physical and emotional well-being of individuals. In fact, almost 55% of people report feeling overwhelmed in their daily lives due to emotional disorders. Moreover, intense emotional pressure can even lead to sudden cardiac death[1]. It is widely recognized that emotional states can be evaluated using various data modes or their aggregation. These modes include visual data, such as facial expression, motion data, sound data, electroencephalogram (EEG), electrocardiogram (ECG), and others, which have been

shown to be correlated with emotions and used for emotion state recognition. Physiological signals have become one of the most important tools in recognizing emotions, being a primary means of detecting and recording human physiological activities[2]. Emotion recognition has greatly benefited from advanced computer technologies, such as deep learning methods. These methods have shown promising results in extracting meaningful features from physiological signals.

In the task of emotion recognition, feature extraction, and selection play a critical role in determining the classification results. The organization of data and selection of features are particularly important, as they can greatly affect the accuracy of the classification.

Recent research on emotion recognition has demonstrated the reliability of physiological signals in classifying human emotional states. For instance, one study used electrocardiogram (ECG) signals to classify the mental stress of drivers, where time and frequency domain features were extracted from the ECG signals and used for emotion classification[3]. Traditional machine learning models were also adopted by Schmidt et al.[4] to classify three emotional states, achieving a maximum accuracy of 0.80. Meanwhile, Cui et al.[5] proposed the regional-asymmetric CNN method, which showed remarkable performance achieving average accuracies of 0.9688 and 0.9628 on the DEAP and DREAMER datasets, respectively. In addition, Lin et al.[6] proposed a deep learning model that supported simultaneous batch training for different sampling rate signals, achieving an 83% accuracy on the WESAD dataset. Furthermore, Pritam Sarkar et al.[7] utilized a self-supervised deep learning framework, which achieved an accuracy of over 90% on the emotion recognition task.

Most deep learning methods for emotion recognition only utilize the features extracted by neural networks such as CNNs or RNNs, which limits the performance of emotion recognition and poses a challenge in obtaining comprehensive information necessary for a better understanding of physiological signals. Moreover, solely relying on traditional deep learning structures ignores the potential of high-level abstract representations in physiological signal data.

As a result, in this paper, we proposed a model combined with a feature fusion technique and a self-supervised learning method. Our main contributions can be summarized as follows:

- We propose a novel method for emotion recognition tasks that utilizes feature fusion and self-supervised learning, providing a fresh perspective on emotion classification using physiological signals.
- The simplicity and interpretability of our proposed method make it well-suited for integration with other models, and

\*Corresponding author.

easy to manipulate. This ease of integration provides possibilities for practical applications in emotion recognition.

- We conducted experiments on the publicly available WESAD dataset and observed that it performs better than existing deep learning models in terms of classification accuracy. Our results indicate that the model has good generalization capabilities and can be effectively applied to other datasets in the emotion recognition field.

To provide a more detailed account of our work, this paper is structured as follows. In Section 2, we introduce the methodological details of our proposed approach. In Section 3, we present the results of our experiments, showcasing the effectiveness and applicability of our method in emotion recognition. Finally, in Section 4, we draw conclusions from our work.

## 2 PROPOSED METHOD FOR EMOTION RECOGNITION

We propose an architecture named FFSS-CNN model that combines feature fusion with self-supervised learning for emotion recognition. The proposed model is designed to make the most effective use of both deep and shallow features and to learn high-level abstract representations, leading to improved accuracy and interpretability in emotion recognition tasks.

### 2.1 FFSS-CNN model

We propose an FFSS-CNN (Feature Fusion and Self-Supervised learning - Convolutional Neural Network) model using ECG signal data. The architecture of our proposed FFSS-CNN model is illustrated in Figure 1, which consists of four main parts: 1) Shallow Feature Extraction; 2) Deep Feature Extraction; 3) Feature Fusion; 4) Emotion Recognition[11]. The detail of the modules are as follows:

- **Module 1 : Shallow Feature Extraction**  
The purpose of extracting shallow features from raw physiological signals is to capture the basic characteristics of the signals in different wavebands. The shallow feature extraction module typically uses traditional signal processing techniques, such as Fourier transform and frequency analysis, to extract features that are easy to interpret and analyze. To ensure that the size of the shallow features matches that of the deep features, we use one flatten layer and two fully connected layers to generate the actual input of the shallow features. This helps to enhance the representation power of the shallow features and improve their compatibility with the deep features.
- **Module 2 : Deep Feature Extraction:**  
Module 2 utilizes convolutional neural networks (CNNs) and self-supervised learning to extract higher-level, abstract features from the raw physiological signals. While the shallow feature extraction module captures the basic characteristics of the signals, the number of shallow features is relatively small with low numerical values, and the upper limits of the shallow feature parameters are close to the lower limits. To refine the feature representation and address this limitation, we use module 2 to extract deep features that are expected to have stronger discriminative power for emotion recognition tasks.

**Table 1: Shallow features of physiological signal**

Shallow Feature	Definition
statistic feature	$\mu = \frac{1}{n} \sum_{i=1}^n S(i)$
statistic feature	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (S(i) - \mu)^2}$
time-domain feature	$AVNN = \frac{1}{N} \sum_{i=1}^N RR_i$
time-domain feature	$SDNN = \sqrt{\frac{\sum_{i=1}^N (RR_i - \mu_{RR})^2}{N}}$
frequency-domain feature	LF
frequency-domain feature	VF

- **Module 3 : Feature Fusion:**  
Module 3 is responsible for fusing the deep and shallow features extracted in the previous modules. The aim of feature fusion is to combine the complementary information provided by deep and shallow features to improve the performance of the emotion recognition task. In this module, we transform the channels of deep and shallow features to the same dimension and concatenate them by column to create a unified feature representation.
- **Module 4 : Emotion Recognition:**  
In module 4, the fused features generated by module 3 are processed through a dropout layer, which randomly discards a portion of the features to prevent overfitting. The remaining features are then supplied to a classifier to classify the emotional states.

Module 2 in our proposed FFSS-CNN model adopts a self-supervised learning method to learn high-level abstract representations from the physiological signal data. Modules 1, 2, and 3 work together to perform feature fusion and improve the accuracy of the model. In the paper, we take ECG signals as an example to explain our FFSS-CNN model.

### 2.2 Shallow feature extraction

For the analysis of emotion recognition, we extract three kinds of shallow features that are listed in Table 1.

In Table 1, we have listed the shallow features that we used in the proposed model. Specifically,  $\mu$  represents the average of  $n$  samples of signal amplitude,  $\sigma$  means the standard deviation of the

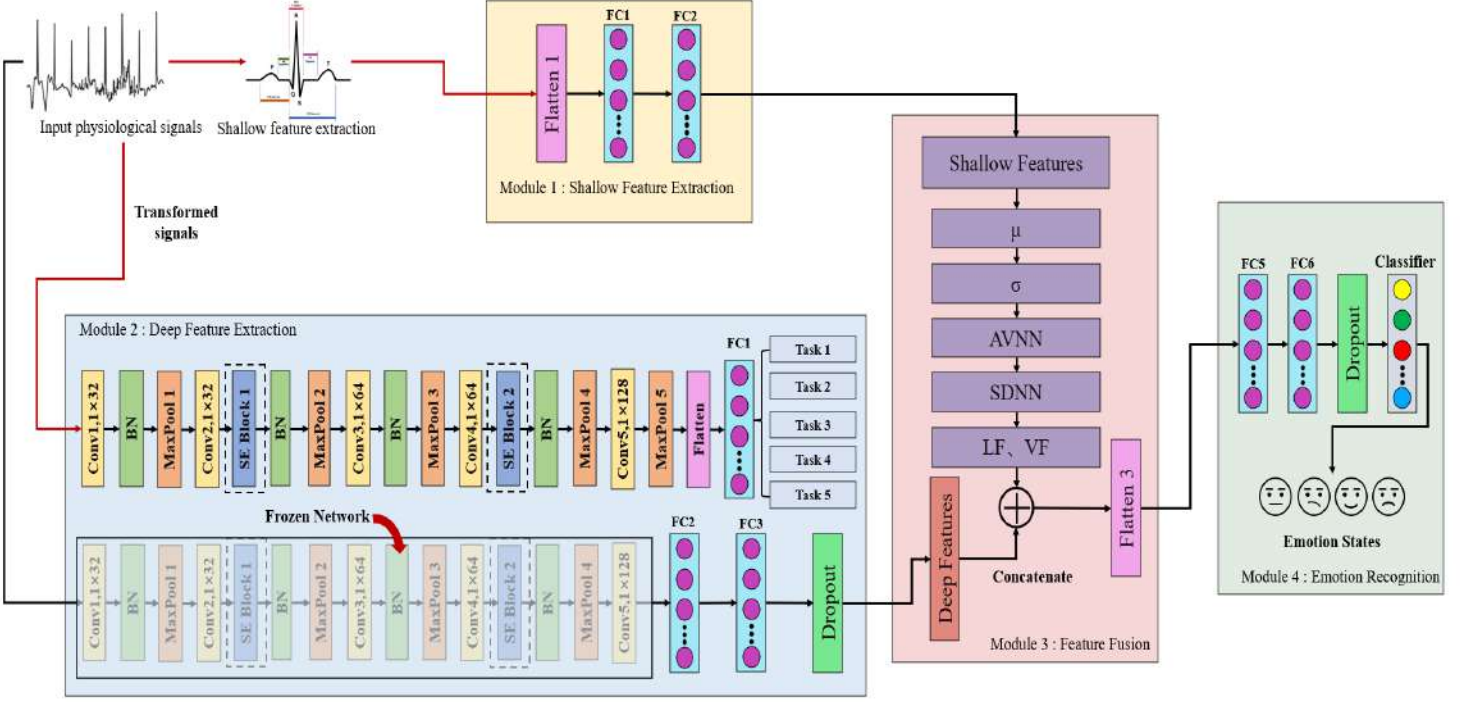


Figure 1: The architecture of the FFSS-CNN Model for emotion recognition.

signal amplitudes over a window of  $n$  samples, AVNN represents the average time interval between successive R-peaks in the ECG signal, SDNN represents the variability of the time interval between successive R-peaks, LF represents the spectral component in the frequency range of 0.04 to 0.15 Hz, and HF means the spectral component in the frequency range of 0.15 to 0.4Hz. Additionally, Table 2 presents the values of the structure parameters used in this module.

Table 2: The parameters of Shallow Feature Extraction.

Layers	Filter Size	Activation	Strides	Padding
Flatten 1	/	/	/	/
FC 1	32	ReLU	/	/
FC 2	8	ReLU	/	/

### 2.3 Deep feature extraction with self-supervised learning

The deep feature extraction module in our proposed model is based on a 1-D convolutional neural network (CNN), which is well-suited

for processing time-series data such as physiological signals. The use of a 1-D CNN allows for automatic feature extraction and representation learning, making it a popular choice in the area of emotion recognition.

In addition, we have also incorporated the attention mechanism into our proposed model. The attention mechanism has gained much attention in recent years and has been successfully applied in various domains. Specifically, we utilize the Squeeze-Excitation (SE) Attention Mechanism[9], which has shown to be effective in signal processing tasks. The SE block plays a vital role in recalibrating the learned features of our model by selectively emphasizing informative features and de-emphasizing less useful ones. By incorporating the SE block, our model can effectively capture discriminative features from the combined shallow and deep features, leading to improved performance in emotion recognition.

The Squeeze-Excitation Attention Mechanism is composed of three main components: squeeze, excitation, and scale steps[10]. Given an input vector  $U \in \mathbb{R}^{H \times W \times M}$ , the squeeze operation aims to obtain a global descriptor of the input by aggregating spatial information.

The squeeze part can be formulated refers to Eq. (1) as:

$$z_m = F_{sq}(u_m) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W u_m(i, j). \quad (1)$$

where  $F_{sq}(\cdot)$  means the squeeze function,  $z_m$  is the output of squeeze operation,  $u_m$  is the  $m$ -th feature map of  $U$ ,  $i$  and  $j$  are the parameters of the feature.

The excitation part can be formulated as Eq. (2):

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_{1z})). \quad (2)$$

where  $F_{ex}(\cdot)$  represents the excitation function,  $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$  and  $W_2 \in \mathbb{R}^{\frac{C}{r} \times C}$  are the fully connected layers for reducing and increasing dimension.  $r$  denotes the reduction ratio,  $\sigma(\cdot)$  refers to the ReLU function.

The scale part referred to Eq. (3).

$$\hat{x}_m = F_{scale}(u_m, s_m) = s_m u_m. \quad (3)$$

where  $F_{scale}(\cdot)$  represents the scale function,  $\hat{x}_m$  is the output of the SE block.

As previously mentioned, we utilize a self-supervised deep learning method to extract deep features of physiological signal representatives. The method involves recognizing five different transformations applied to raw signal data, generating pseudo-labels based on the type of transformation applied, and training the model on the transformed signal data and corresponding labels[7]. By creating various signal transformation tasks, the model learns useful representations of the input data that can be further utilized for emotion recognition. Let's assume the transformed signal data and pseudo-labels for  $T_p$  as  $(X_j, P_j)$  where  $X_j$  is  $j^{th}$  transformation and  $j \in [0, N]$  where  $N$  denotes the kinds of transformations. We create different signal transformation tasks to learn  $T_p$ .

Five pretext tasks are created to learn signal transformations from the unlabeled signals[8]:

- Permutation: The signal data  $S(t)$  is divided into  $m$  segments and the segments are shuffled to perturb the temporal location of each segment, resulting in a new signal.
- Negation: The original signal data  $S(t)$  is multiplied by  $-1$ .
- Scaling: The original amplitude of the signal data  $S(t)$  is transformed as  $\beta \times S(t)$ , where  $\beta > 0$  and  $\beta$  is the artificial assigned scaling factor.
- Noising: A Gaussian noise with random amplitudes is added to the original signal data  $S(t)$ .
- Temporal Inversion: Make the original signal data  $S(t)$ , where  $t = 1, 2, \dots, N$  and  $N$  is the number of the time-window, the temporal inversion of the data is defined as  $S'(t)$ , where  $t = N, N-1, \dots, 1$ .

Table 3 lists the structure parameter values in the deep feature extraction module.

## 2.4 Feature fusion

After extracting shallow and deep features from the original physiological signal data in previous modules, the feature fusion module is designed to combine these features effectively[12].

Typically, the dimensions of extracted deep features are larger than those of extracted shallow features, making it impossible to use them together directly. Therefore, in the feature fusion module, we first transform the channels of the deep and shallow features to the same dimension.

After the transformation steps, the feature fusion module uses the early fusion technique to combine the extracted shallow and deep features. In this case, we adopt the concatenate layer to fuse the features by column. This approach enables us to merge the

**Table 3: The parameters of Deep Feature Extraction.**

Layers	Filter Size	Activation	Strides	Padding
Conv 1	$32 \times 32 \times 1$	ReLU	1	valid
BN	/	/	/	/
MaxPool 1	$8 \times 1$	/	1	same
Conv 2	$32 \times 32 \times 1$	ReLU	1	valid
SE Block 1	16	ReLU	/	/
BN	/	/	/	/
MaxPool 2	$8 \times 1$	/	1	same
Conv 3	$64 \times 16 \times 1$	ReLU	1	valid
BN	/	/	/	/
MaxPool 3	$8 \times 1$	/	1	same
Conv 4	$64 \times 16 \times 1$	ReLU	1	valid
SE Block 2	16	ReLU	/	/
BN	/	/	/	/
MaxPool 4	$8 \times 1$	/	1	same
Conv 5	$128 \times 8 \times 1$	ReLU	1	valid
Flatten 2	/	/	/	/
FC 3	128	ReLU	/	/
FC 4	128	ReLU	/	/

transformed deep and shallow features into a single feature vector, which is then used as the input to the final classification layer.

We let  $F_d$  and  $F_f$  be the deep features and fused features, respectively, which can be defined as Eq. (4):

$$\begin{aligned} F_d &= FFSS - CNN(Raw \text{ Signals}) \\ F'_d &= Transform(F_d) \\ F_f &= \{F'_d, F_h\} \end{aligned} \quad (4)$$

where  $FFSS-CNN(\cdot)$  represents the proposed feature fusion model and  $Transform(\cdot)$  means the integration of shallow and deep features.

This multi-scale method helps to preserve the discriminative information present in the original signal data.

### 3 EXPERIMENTS AND RESULTS

#### 3.1 Dataset

We use the WEable Stress and Affect Detection (WESAD) dataset[4] for conducting experiments on emotion recognition tasks. This dataset includes ECG data collected from 17 participants. To adapt it for emotion recognition tasks, we encode the affective labels into four categories: baseline (stage I), meditation (stage II), amusement (stage III), and stress (stage IV).

#### 3.2 Data Preprocessing

In this paper, we utilized a sliding time-window method for the segmentation of physiological signals. We experimented with various time-window sizes to extract more feature information and determined that a 700Hz segmentation window for the raw signal data was the optimal choice. To utilize all the available features in each recording, we uniformly sampled all the recordings to a fixed length.

#### 3.3 Training and evaluating

The proposed method is established using TensorFlow 2.5 on a GPU platform with an NVIDIA RTX3070. We used the adam optimizer and cross-entropy loss function. We adopted a 10-fold cross-validation approach to evaluate the model. The batch size was set to 32, with 50 epochs for training. The initial learning rate was set to 0.0001 and was divided by 5 after every 10 epochs. For evaluation, we use standard classification metrics listed as Eq. (5) to Eq. (8) below:

$$accuracy = \frac{1}{m} \sum_{i=1}^m \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (5)$$

$$recall_i = \frac{TP_i}{TP_i + FN_i} \quad (6)$$

$$precision_i = \frac{TP_i}{TP_i + FP_i} \quad (7)$$

$$F1 = \frac{1}{m} \sum_{i=1}^m 2 \times \frac{recall_i \times precision_i}{recall_i + precision_i} \quad (8)$$

where  $m$  means the number of classes,  $TP_i$  represents the number for which the  $i^{th}$  class is correctly classified,  $TN_i$  stands for the number of the correctly classified data in the absence of the  $i^{th}$  class,  $FP_i$  means the number of the  $i^{th}$  class misclassified, and  $FN_i$  represents the number of misclassified data instances that belong to the  $i$ -th class.

The training evaluation values are shown in Fig. 2. The training loss curve is present in Fig.3.

To better understand the impact of each component in our proposed FFSS-CNN model, we conducted ablation experiments by separating the model into three parts. The first part is the baseline

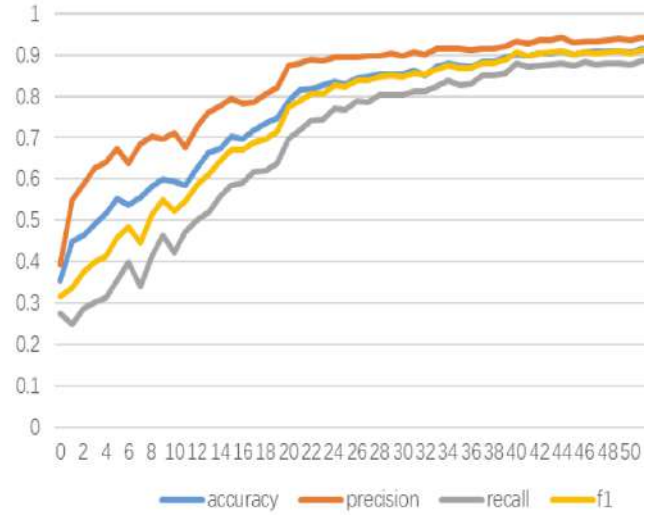


Figure 2: Evaluation values vs. training epochs of 3 emotion classes recognition.

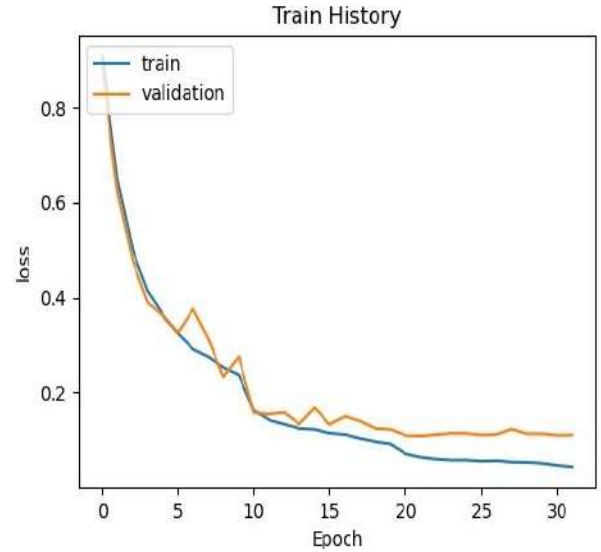


Figure 3: Training loss curve vs. training epochs of 3 emotion classes recognition.

CNN model without any feature fusion, the second part is the baseline CNN model with self-supervised learning, and the third part is the complete FFSS-CNN model. By comparing the performance of these three models, we can analyze the contribution of each component in the proposed method. Table 4 presents the results for the recognition of three classes of emotional states.

Therefore, we can conclude that the self-supervised learning and feature fusion method we used in our proposed model proved to be effective.

**Table 4: Recognition results of ablation experiments on FFSS-CNN model**

Model	Emotion State	accuracy	F1 Score
Baseline CNN Model	3	80.58	79.84
+ Self-Supervised Learning	3	92.96	91.87
<b>FFSS-CNN(Complete)</b>	3	<b>93.41</b>	<b>92.90</b>

To further evaluate our proposed model, we conducted experiments using various deep learning models to compare their performance on our own devices. The results are presented in Table 5, where the best results are highlighted.

**Table 5: Recognition results of existing methods (On the same device)**

Model	Emotion State	accuracy	F1 Score
KNN	3	54.76	47.77
LDA	3	66.29	56.03
CNN	3	74.01	73.09
CNN+LSTM	3	78.00	76.00
Fully-Supervised CNN	3	91.14	90.20
Self-Supervised CNN	3	92.96	91.87
<b>FFSS-CNN(Ours)</b>	3	<b>93.41</b>	<b>92.90</b>
<b>FFSS-CNN(Ours)</b>	4	<b>92.66</b>	<b>91.55</b>

Based on the test results, our proposed FFSS-CNN model outperforms other deep learning models in most emotion recognition tasks. Furthermore, when compares to existing models, the average

accuracy of our proposed model is found to be higher. Specifically, our proposed model achieves an accuracy of 93.41%, which is 0.45% higher than the state-of-art model, and the F1 score for each emotion state is above 0.91, which suggests that the combination of shallow and deep features results in a higher quality feature representation and the self-supervised learning method extracted more high-level information from the raw physiological signal data.

## 4 CONCLUSIONS

In this paper, we propose an emotion recognition method based on feature fusion and self-supervised learning. We evaluate our method on the public WESAD dataset and compare it with six state-of-the-art reference methods. Experimental results confirm the effectiveness of the FFSS-CNN model, which outperforms the reference methods in terms of accuracy and F1 score. Our analysis indicates that our proposed approach is effective in learning improved physiological signal data, which leads to better accuracy performances of the recognition tasks. In the future, we plan to explore the use of our model for larger and more complex scenes in the emotion recognition field.

## ACKNOWLEDGMENTS

This work is supported by the Science and Technology Department of Sichuan Province of China (grant no. 2021YFG0331).

## REFERENCES

- [1] Maria TeresaLa Rovere, Alessandra Gorini, Peter J.Schwartz. 2022. Stress, the autonomic nervous system, and sudden death. *Autonomic Neuroscience*. Volume. 237.
- [2] Y. Hsu, J. Wang, W. Chiang, and C. Hung. 2020. Automatic ECG-based emotion recognition in music listening. *IEEE Transactions on Affective Computing*. Volume. 11(1), pp. 85–99.
- [3] Jennifer Healey and Rosalind W Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*. Volume. 6(2), pp. 156–166.
- [4] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th International Conference on Multimodal Interaction*. pp. 400–408.
- [5] H. Cui, A. Liu, X. Zhang, X. Chen, K. Wang, X. Chen. 2020. EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowl. Based Syst.*. Volume 205, pp. 106243.
- [6] Jionghao Lin, Shirui Pan, Cheng Siong Lee, Sharon Oviatt. 2019. An Explainable Deep Fusion Network for Affect Recognition Using Physiological Signals. In *Proceedings of the 28th International Conference on Information and Knowledge Management*. pp. 2069–2072.
- [7] Pritam Sarkar, Ali Etemad. 2021. Self-supervised ECG Representation Learning for Emotion Recognition. *IEEE Transactions on Affective Computing*. Volume 13(3), pp. 1541–1554.
- [8] Terry Taewoong Um, Franz Michael, Josef Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche et al. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. *arXiv preprint arXiv:1706.00527*.
- [9] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. 2020. Squeeze-and-excitation networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, Volume. 42(8), pp. 2011–2023.
- [10] A. Alshehri, Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan. 2019. Deep attention neural network for multi-label classification in unmanned aerial vehicle imagery. *IEEE Access*. Volume. 7, pp. 119873–119880.
- [11] Pritam Sarkar, Kyle Ross, Aaron J Ruberto, Dirk Rodenburg, Paul Hungler, and Ali Etemad. 2019. Classification of cognitive load and expertise for adaptive simulation using deep multitask learning. In *8th IEEE International Conference on Affective Computing and Intelligent Interaction*, pp. 1–7.
- [12] Henry Friday Nweke, Teh Ying Wah, Ghulam Mujtaba. 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Information Fusion*, Volume 46, pp. 147–170.

# A Modified Fuzzy K-nearest Neighbor Using the Improved Sparrow Search Algorithm for Two-classes and Multi-classes Datasets

Chengfeng Zheng\*

Mohd Shareduwan Mohd Kasihmuddin\*

chengfengzheng@student.usm.my

shareduwan@usm.my

Universiti Sains Malaysia

Penang, Penang, Malaysia

Ju Chen

Chengdu University of Traditional Chinese Medicine

Chengdu, China

Yuan Gao

Universiti Sains Malaysia

Penang, Malaysia

gaoyuan@student.usm.my

Mohd.Asyraf Mansor

Universiti Sains Malaysia

Penang, Malaysia

## ABSTRACT

The Sparrow search algorithm is a new and effective swarm intelligence method proposed in recent years and studied in many publications. Based on the basic principle of sparrow search algorithm, this paper combines the inverse learning algorithm with the refined inverse solution to form an improved sparrow search (SSA) algorithm. Combining the fuzzy k-nearest neighbor method and the improved SSA, the numerical simulation of two-classes datasets and multi-classes datasets is carried out, and many numerical results are obtained, and the results are analyzed. At the same time, this paper lists the data comparison results and tables with other models. The hybrid SSA-FKNN proposed in this paper has a clear advantage in terms of accuracy (ACC).

## CCS CONCEPTS

• Theory of computation → Theory of randomized search heuristics.

## KEYWORDS

swarm intelligence method, modified sparrow search algorithm, Fuzzy K-Neighbour Classifier, multiple application scenario datasets

## ACM Reference Format:

Chengfeng Zheng, Mohd Shareduwan Mohd Kasihmuddin, Yuan Gao, Ju Chen, and Mohd.Asyraf Mansor. 2023. A Modified Fuzzy K-nearest Neighbor Using the Improved Sparrow Search Algorithm for Two-classes and Multi-classes Datasets. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590042>

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org). CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-9944-9/23/03...\$15.00 <https://doi.org/10.1145/3590003.3590042>

## 1 INTRODUCTION

As a meta-heuristic algorithm, swarm intelligence algorithm has increasingly attracted the attention of researchers all over the world [6, 8–12]. It has a very special relationship with artificial life, especially evolutionary strategy and genetic algorithm, such as PSO [10], ABC [12], GSA [14] and WOA [6]. The Sparrow Search Algorithm, proposed by Xue and Shen in 2020, mainly simulates the predation and anti-predation behavior of sparrows [15]. This method is a new and efficient meta-heuristic algorithm, at the same time, like the general heuristic intelligent optimization algorithm, it also has problems such as low convergence accuracy and easy to fall into local convergence

In order to improve the effect of various meta-heuristic algorithms, many scholars have put forward different ideas in recent years, [18] proposed to use spiral search method and Gaussian transform method to improve the effect of the algorithms, which has achieved great results in numerical calculation. The paper [17] introduces in detail the use of reverse learning and parameter adjustment mechanism to optimize the SCA algorithm, and combines the classifier FKNN to carry out numerical simulation, and obtains good numerical results. Due to the endless emergence of various meta-heuristic algorithms, the existing optimization and upgrading can not exhaust all the algorithms.

This paper improves the SSA model based on reverse learning, and combines with the very mature classifier FKNN to proposed the new method Hybrid SSA FKNN. The Hybrid SSA FKNN method is used to carry out numerical calculations on the two and multi-class data sets, and a large number of numerical results are obtained. The contributions of this paper are as follows:

- (1) This paper proposes a new Hybrid SSA algorithm by integrating reverse learning into SSA algorithm.
- (2) Combining the newly constructed Hybrid SSA algorithm with the classifier, this paper constructs the Hybrid SSA FKNN algorithm.
- (3) In this paper, Hybrid SSA FKNN algorithm is used to solve the classification problem of two-class and multi-class data sets, and good numerical results are obtained.

The rest of this paper is as follows. Section 2 introduces the sparrow search algorithm and reverse learning mechanism in detail,

and propose the hybrid SSA. Section 3 introduces the classifier FKNN in detail. Section 4 introduces the combination of hybrid SSA algorithm and classifier FKNN, and proposes Hybrid SSA FKNN algorithm. Section 5 introduces and analyzes the numerical results of Hybrid SSA FKNN algorithm in the two-classes dataset and multi-classes datasets in detail, and compares it with other classical algorithms. Section 6 describes the relevant conclusions and the next research direction.

## 2 THE MODIFIED SPARROW SEARCH ALGORITHM

This section will introduce sparrow search algorithm and reverse learning mechanism in detail, and propose Hybrid SSA algorithm on this basis. This method integrates the reverse learning mechanism into the sparrow search algorithm, improving the ability of the algorithm to jump out of local convergence.

### 2.1 Sparrow Search Algorithm (SSA)

This section provides a detailed explanation of the operational mechanism of the sparrow search algorithm, which is inspired by the foraging and anti-predation behavior of sparrows. Prior to presenting the mathematical model, six rules are outlined.

(1) Sparrows with high energy reserves, referred to as explores, are responsible for locating areas with abundant food sources. Sparrows with low energy reserves are called followers.

(2) Sparrows will sound the alarm to alert others of predators, and followers must be directed to safe locations if the alarm value exceeds a safety threshold.

(3) Any sparrow can become an explorer if it finds a better food source, but the ratio of explores and followers remains the same.

(4) As an explorer, sparrows with higher energy and several hungry followers are more likely to fly elsewhere to obtain more energy.

(5) Followers will follow explorers providing the best food while also competing for food to increase predation rates.

(6) When sparrows sense danger, they move quickly to safety for a better position, while sparrows in the middle of a flock move randomly to get closer to other sparrows.

### 2.2 Explores

In SSA, the best individual in the group will get food first. As an explorer, it has access to a larger search range than its followers. The position of explorer is updated as follows:

$$X_{i,j}^{t+1} = \begin{cases} X_{i,j}^t \cdot \exp\left(\frac{-i}{\alpha \cdot iter_{max}}\right) & \text{if } R_2 < ST \\ X_{i,j}^t + Q \cdot \mathcal{L} & \text{if } R_2 \geq ST \end{cases} \quad (1)$$

The position of each sparrow is denoted by  $X_{i,j}$ , where  $i$  represents the sparrow index and  $j$  represents the position index. The algorithm runs for a maximum of  $iter_{max}$  iterations, and  $\alpha$  is a random number between 0 and 1. The early warning values,  $R_1$  and  $R_2$ , both fall within the range of 0 to 1, while the safety value,  $ST$ , falls within the range of 0.5 to 1. The variable  $Q$  follows a normal distribution, and each element of the list  $\mathcal{L}$  is equal to 1.

When  $R_2$  is less than  $ST$ , it implies that there are no predators in the vicinity, allowing the Explorer to conduct an extensive search. On the other hand, if  $R_2$  is greater than or equal to  $ST$ , certain

sparrows have spotted predators and all sparrows should take appropriate measures.

### 2.3 Followers

The formula for followers to update the location is as follows:

$$X_{i,j}^{t+1} = \begin{cases} Q \cdot \exp\left(\frac{X_{worst}^t - X_{i,j}^t}{i^2}\right) & \text{if } i > n/2 \\ X_p^{t+1} + |X_{i,j}^t - X_p^{t+1}| \cdot A^+ \cdot L & \text{others} \end{cases} \quad (2)$$

Where  $X_p$  is the location of the optimal explorer and  $x_{worst}$  is the current worst position for all sparrows;  $N$  is the population size.  $A$  is a  $1 \times D$  matrix, and the random amplitude of each element is 1 or -1. Here,  $A^+$  is defined as follows:

$$A^+ = A^T (AA^T)^{-1}$$

When  $i > \frac{n}{2}$ , it represents that the  $i$ -th follower who has the low fitness value, is in poor condition and has to go to other places for feeding.

### 2.4 Scouters

In the algorithm, the original text assumes that 10% ~ 20% of the individuals in the population (20% in the paper) will be aware of the danger, and the initial locations of sparrows are randomly generated:

$$X_{i,j}^{t+1} = \begin{cases} X_{best}^t + \beta \cdot \left| X_{i,j}^t - X_{best}^t \right| & \text{if } f_i > f_g \\ X_{i,j}^t + K \cdot \left( \frac{X_{i,j}^t - X_{worst}^{t+1}}{(f_i - f_w) + \varepsilon} \right) & \text{if } f_i = f_g \end{cases} \quad (3)$$

The current global optimal position is denoted as  $X_{best}$ , while the control parameter  $\beta$  takes on a random value following the normal distribution  $N(0, 1)$ . Additionally,  $K$  is a random number between -1 and 1,  $f$  represents the fitness value,  $f_g$  stands for the current best fitness value, and  $f_w$  represents the worst fitness value. Avoiding division by zero, we introduce a constant  $\varepsilon$ .

In short,  $f_i > f_g$  means that sparrow is in the outermost position of the population. When  $f_i = f_g$  represents that parts of sparrows sense the danger and close to other sparrows to avoid being preyed.  $K$  is the control parameter indicating the movement direction of the sparrow.

### 2.5 Reverse Learning

This section details the principle of reverse learning and the process of using reverse learning to increase elite solutions. Relying only on the current optimal location to update the state of the next step will easily lead the algorithm to fall into local optimization and cannot achieve the desired results. Reverse learning the current optimal solution to increase the possibility of escaping from the local convergence. The relevant formula is:

$$X_i^* = X_i^s + \omega \otimes (X_i^s - X_i^t) \quad (4)$$

Where:  $X_i^s$  means the position of population  $i$ ;  $X_i^*$  means the position after executing Eq. 1, Eq. 2 and Eq. 3;  $X_i^*$  represents the new position after reverse learning;  $\omega \in [-1, 1]$ ;  $\otimes$  represents dot multiplication.

In order to further avoid the negative impact of reverse learning process on the results, greedy learning is used to get the best result according to the current optimal state and after reverse learning state.

$$\omega = C^{(-t/MaxFEs)} \times \cos(r_5) \quad (5)$$

Where:  $r_5$  is in  $[0, \pi]$ ;  $C$  is a constant, and  $C = 100$  will be better.

In order to further improve the efficiency of the algorithm, this paper adds the elite reverse learning strategy in the optimization process. The following is the elite reverse learning strategy[11] :

1. The populations are ranked after formula *fitness*, the top 10% populations form the elite solution  $\check{\nu}_{best}$ ;
2.  $X_{best}^i \in \check{\nu}_{best}$  boundary  $[lb_j^i, ub_j^i]$ , and get  $[min(lb_j^i), max(ub_j^i)]$ ;
3. The dynamic elite reverse population  $\check{\nu}_{best}'$  is gotten according to Eq.6;
4. If  $\check{\nu}_{best}'$  exceeds the boundary  $[min(lb_j^i), max(ub_j^i)]$ , it is replaced by a new location randomly generated;
5. Get top 50% from  $[\check{\nu}_{best}, \check{\nu}_{best}']$  to the next loop according to *fitness*;
6. Repeat the progress 2 and 5 until the algorithm ends.

$X_{best}' = (x_1', x_2', \dots, x_D')$  is the inverse solution of the selected elite sparrows..  $X_{best} = (x_1, x_2, \dots, x_D)$  is the current solution. And the inverse solution can be expressed as

$$x_i' = k(lb_i + ub_i) - x_i \quad (6)$$

where  $k \in [0, 1]$  is a random number and is subject to uniform distribution.

The elite solution generated increases the useful information for population convergence and improves the ability of the method to escape from local optimum.

### 3 FUZZY K-NEAREST NEIGHBOR

This section details a mature and adaptive classifier (FKNN). Fuzzy FKNN is proposed on the basis of KNN algorithm. It has the advantages of high accuracy and few parameters. It is a relatively mature classifier[4, 7, 13].

$$U_{i,k} = \begin{cases} 0.51 + \left(\frac{n_k}{K}\right) \cdot 0.49, k = Y_k \\ \left(\frac{n_k}{K}\right) \cdot 0.49, k \neq Y_k \end{cases} \quad (7)$$

where:  $k$  means the  $k$ -th class,  $i$  means the  $i$ -th sample from 1 to  $N$ , and  $N$  stands for the samples number.  $K$  represents the neighbors nearby number.  $U_{i,k}$  indicates the correlation degree of the  $i$ -th individual corresponding to the  $k$ -th class.  $Y_k$  stands for the real class of the current sample,  $n_k$  stands for the adjacent neighbors number belonging to the  $k$ -th class among the nearest  $K$  neighbors. The relationship of each training sample here corresponding to each class is described in the following fomula:

$$U_k(x) = \frac{\sum_{j=1}^K U_{I_j,k} (x - x_{I_j})^{\frac{2}{m-1}}}{\sum_{j=1}^K (x - x_{I_j})^{\frac{2}{m-1}}} \quad (8)$$

where,  $x$  represents the test sample,  $U_k(x)$  stands for the test sample weight to the  $k$ -class. The  $j$  stands for the  $j$ -th nearest neighbor sample from 1 to  $K$ .  $U_{I_j,k}$  denotes membership degree calculated by Eq. 7. The  $x - x_{I_j}$  is used to calculate the distance difference. The

value of  $m$  is from 1 to  $\infty$ , which is used to control the influence of distance on membership degree.

$$C(x) = \arg \max_k U_k(x) \quad (9)$$

### 4 HYBRID SSA FKNN

After combining modified SSA and FKNN, the flow chart is shown below

#### Algorithm 1 The Hybrid SSA-FKNN

---

```

while  $t < MaxFEs$  do
  Update  $X_{explore}, X_{follower}, X_{scouter}$ 
  calculate  $f(X_{explore}), f(X_{follower}), f(X_{scouter})$ 
  if  $f(X_{explore}) > BF$  then
     $X_{new} = X_{explore}$ 
    reverse learning
    generated elite inverse solution
     $BF = f(X_{explore})$ 
  end if
  if  $f(X_{follower}) > BF$  then
     $X_{new} = X_{follower}$ 
    reverse learning
    generated elite inverse solution
     $BF = f(X_{follower})$ 
  end if
  if  $f(X_{scouter}) > BF$  then
     $X_{new} = X_{scouter}$ 
    reverse learning
    generated elite inverse solution
     $BF = f(X_{scouter})$ 
  end if
end while

```

---

### 5 NUMERICAL RESULTS FOR THE HYBRID SSA FKNN

The datasets used in this paper can be obtained from the following websites <https://archive.ics.uci.edu>.

Tab.1 lists indicators such as the number of data categories, sample size and data selectable characteristic quantity of six two-calses and multi-classes datasets.

#### 5.1 Two-classes datasets

For the above three two-classes datasets, Figure 1, 2 and 3 are obtained by Hybrid SSA-FKNN according to numerical results in paper [17]. As shown in Figure 1, Figure 2 and Figure 3, the Hybrid SSA-FKNN model gets better results on ACC and other indicators, about 5.0%-14.1% higher than the comparison models for ACC.

#### 5.2 Multi-classes datasets

As shown in Figure 4, 5 and 6, for multi-class datasets, the modified SSA FKNN method can still achieve good results. It has good data prediction results for datasets with multiple or few eigenvalues, multiple or few samples.

In the numerical results of the above six secondary and multi-class data sets, Hybrid SSA FKNN is superior to SSA FKNN in the

Table 1: The datasets information.

	Categories	Samples	Features	Positive	Negative
Bupa	2	345	6	145	200
Hepatitis	2	155	19	32	123
SPECT	2	267	22	212	55
Glass	7	214	9	69	145
User hamdi	4	403	5	102	301
Breast tissue	6	106	9	20	86

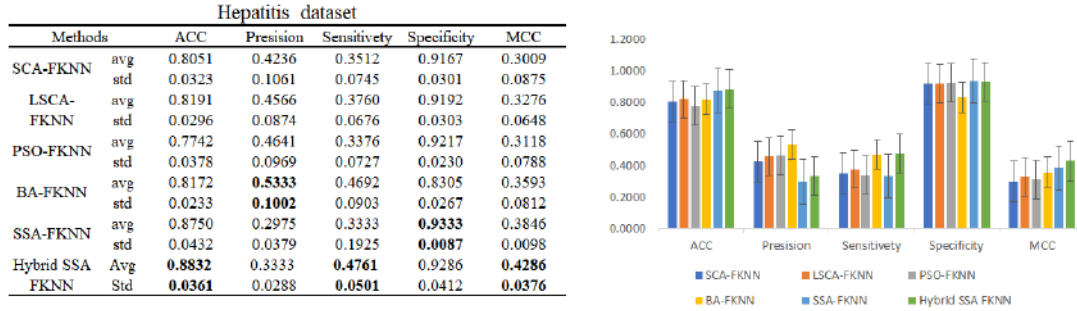


Figure 1: Comparison results in the in hepatitis dataset.

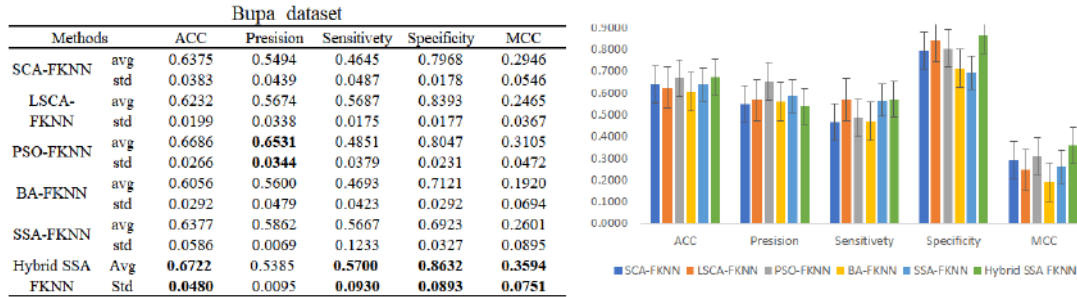


Figure 2: Comparison results in the in Bupa dataset.

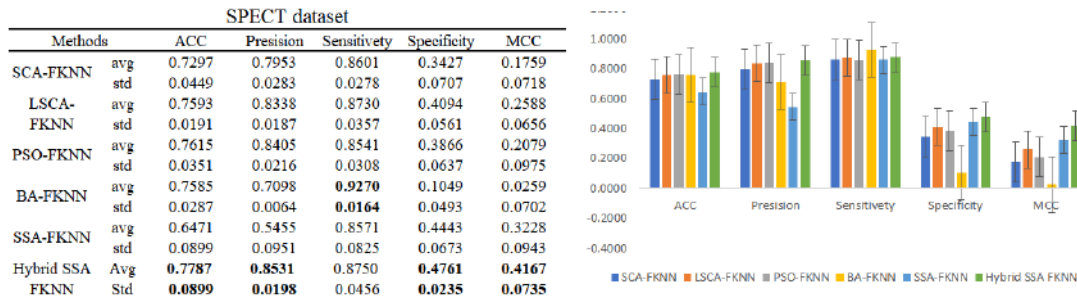


Figure 3: Comparison results in the in SPECT dataset.

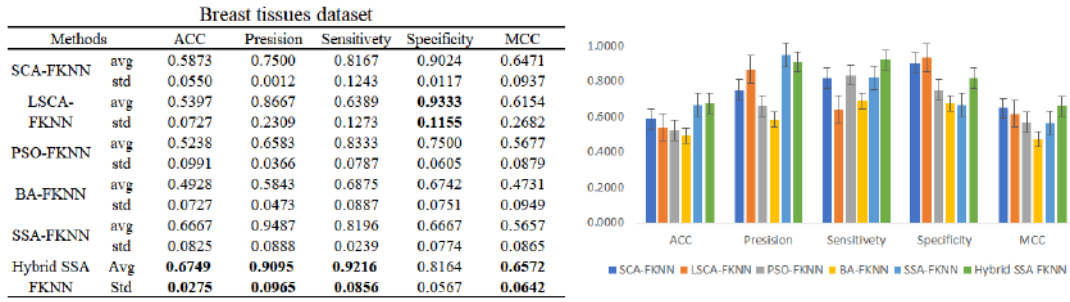


Figure 4: Comparison results in the in Breast tissue dataset.

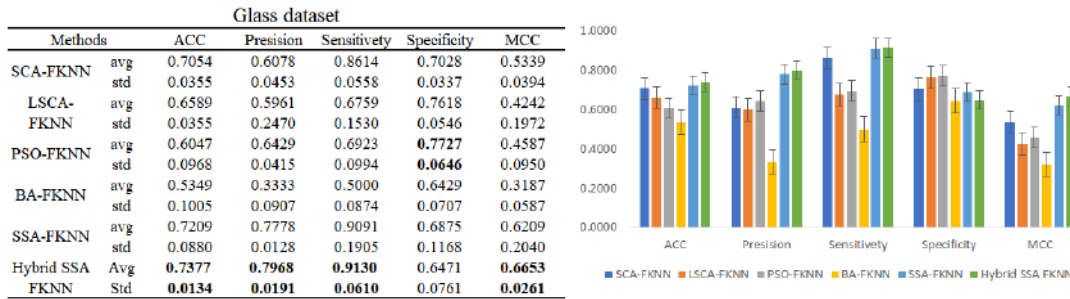


Figure 5: Comparison results in the in Glass dataset.

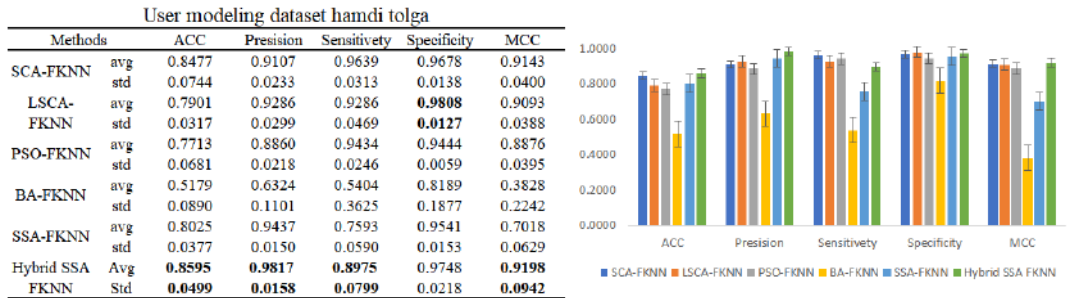


Figure 6: Comparison results in the in User modeling dataset hamdi tolga dataset.

accuracy of ACC in five evaluation indicators, which further proves the effectiveness of the improvement measures of this method. For ACC, the relevant numerical results increased by 1% - 20%. As for the Specificity index, there are three data sets that the SSA FKNN results are better than the Hybrid SSA FKNN algorithm. The characteristics of the dataset itself lead to this situation.

The results of this study suggest that the Hybrid SSA FKNN algorithm is a promising approach for classification tasks, as it achieved the best performance in terms of accuracy, precision, sensitivity, and MCC evaluation indicators across most of the datasets tested. The high ACC values obtained by this algorithm indicate that it can correctly classify instances with a high degree of accuracy, which is a crucial factor in many real-world applications.

Furthermore, Hybrid SSA FKNN also performed well in terms of Precision, achieving the best results in four out of six datasets. This indicates that the algorithm has a good ability to identify positive instances accurately, which is essential in scenarios where false positives can have serious consequences. Additionally, the fact that the remaining two datasets still resulted in competitive Precision scores suggests that Hybrid SSA FKNN is a robust method that can handle different types of data effectively. In terms of sensitivity, Hybrid SSA FKNN achieved the best results in five out of six datasets, demonstrating its ability to identify true positive instances effectively. Although the sensitivity results for the SPECT dataset were lower than those of BA-FKNN, it is important to note that this was due to the inherent characteristics of the algorithm rather

than any inherent limitations of Hybrid SSA FKNN. The Specificity evaluation index, on the other hand, did not yield optimal results except for the Bupa and SPECT datasets. This indicates that Hybrid SSA FKNN may not be the best method for identifying negative instances in certain contexts. However, it is important to note that Specificity is often less critical than other evaluation metrics in many classification tasks. Overall, the results of this study suggest that Hybrid SSA FKNN is a promising method for classification tasks, particularly in scenarios where accuracy, precision, and sensitivity are critical factors. Future research could focus on further optimizing the algorithm or applying it to other types of data to explore its potential applications in various fields.

The datasets used in this paper covers two-classes and multi-classes classification problems. The selection of data sets fully considers that the eigenvalues are at different scales, the samples of data are at different scales, the balance of positive and negative samples is different. The selection of datasets is relatively comprehensive, all of which are open source case datasets. At the same time, the algorithm proposed in this paper compares the numerical results of the other five algorithms under the same sample and parameter conditions, and has achieved good numerical results. It is further proved that this method is effective in dealing with complex classification problems and has good adaptability.

## 6 CONCLUSIONS

The Hybrid SSA FKNN method presented in this paper is a modified version of the SSA method that combines reverse learning and classifier FKNN. The method has been applied to both two-class and multi-class datasets, yielding promising numerical results. Its high adaptability makes it a method with great potential for future applications. However, there is still room for improvement, and this will be the main focus of future research. Moreover, the proposed method can be extended with the use of other rules such as Satisfiability Logic [1–3, 5] and Logic Mining [16].

## REFERENCES

- [1] Suad Abdeen, Mohd Shareduwan Mohd Kasihmuddin, Nur Ezlin Zamri, Gaeithry Manoharam, Mohd Asyraf Mansor, and Nada Alshehri. 2023. S-Type Random k Satisfiability Logic in Discrete Hopfield Neural Network Using Probability Distribution: Performance Optimization and Analysis. *Mathematics* 11, 4 (2023), 984.
- [2] Ju Chen, Mohd Shareduwan Mohd Kasihmuddin, Yuan Gao, Yueling Guo, Mohd Asyraf Mansor, Nurul Atiqah Romli, Weixiang Chen, and Chengfeng Zheng. 2023. PRO2SAT: Systematic Probabilistic Satisfiability logic in Discrete Hopfield Neural Network. *Advances in Engineering Software* 175 (2023), 103355.
- [3] Yueling Guo, Mohd Shareduwan Mohd Kasihmuddin, Yuan Gao, Mohd Asyraf Mansor, Habibah A Wahab, Nur Ezlin Zamri, and Ju Chen. 2022. YRAN2SAT: A novel flexible random satisfiability logical rule in discrete hopfield neural network. *Advances in Engineering Software* 171 (2022), 103169.
- [4] Z. Hao, A. C. Berg, M. Maire, and J. Malik. 2006. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*.
- [5] Siti Zulaikha Mohd Jamaludin, Nurul Atiqah Romli, Mohd Shareduwan Mohd Kasihmuddin, Aslina Baharum, Mohd Asyraf Mansor, and Muhammad Fadhil Marsani. 2022. Novel logic mining incorporating log linear approach. *Journal of King Saud University-Computer and Information Sciences* 34, 10 (2022), 9011–9027.
- [6] A. Kaveh and A. Dadras. 2017. A novel meta-heuristic optimization algorithm: Thermal exchange optimization. *Advances in Engineering Software* 110 (2017), 69–84.
- [7] J. M. Keller, M. R. Gray, and J. A. Givens. 2012. A fuzzy K-nearest neighbor algorithm. *IEEE Transactions on Systems Man & Cybernetics SMC-15*, 4 (2012).
- [8] S. Mirjalili. 2016. SCA: A Sine Cosine Algorithm for Solving Optimization Problems. *Knowledge-Based Systems* 96 (2016).
- [9] N. Singh and S. B. Singh. 2017. A novel hybrid GWO-SCA approach for optimization problems. *Engineering Science and Technology, an International Journal* 20, 6 (2017), 1586–1601.
- [10] O. P. Verma, S. Gupta, S. Goswami, and S. Jain. 2017. Opposition based modified particle swarm optimization algorithm. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.
- [11] M. P. Wachowiak, R. Smolíkova, Y. Zheng, J. M. Zurada, and A. S. Elmaghraby. 2004. An approach to multimodal biomedical image registration utilizing particle swarm optimization. *Evolutionary Computation IEEE Transactions on* 8, 3 (2004), 289–301.
- [12] X. Wang, Z. Y. Li, G. Y. Xu, and W. Yan. 2012. Artificial bee colony algorithm based on chaos local search operator. *Journal of Computer Applications* 32, 4 (2012), 1033–1036.
- [13] W. Wong, D. Cheung, B. Kao, and N. Mamoulis. 2009. Secure kNN computation on encrypted databases. *ACM* (2009).
- [14] Y. Xu, J. Zhou, X. Xue, W. Fu, and C. Li. 2016. An adaptively fast fuzzy fractional order PID control for pumped storage hydro unit using improved gravitational search algorithm. *Energy Conversion & Management* 111 (2016), 67–78.
- [15] Jiankai Xue and Bo Shen. 2020. A novel swarm intelligence optimization approach: sparrow search algorithm. *Systems Science & Control Engineering An Open Access Journal* 8, 1 (2020), 22–34.
- [16] Nur Ezlin Zamri, Siti Aishah Azhar, Mohd Asyraf Mansor, Alyaa Alway, and Mohd Shareduwan Mohd Kasihmuddin. 2022. Weighted random k satisfiability for k= 1, 2 (r2SAT) in discrete hopfield neural network. *Applied Soft Computing* 126 (2022), 109312.
- [17] Chengfeng Zheng, Mohd Shareduwan Mohd Kasihmuddin, Mohd. Asyraf Mansor, Ju Chen, and Yueling Guo. 2022. Intelligent Multi-Strategy Hybrid Fuzzy K-Nearest Neighbor Using Improved Hybrid Sine Cosine Algorithm. *Mathematics* 10, 18 (2022). <https://doi.org/10.3390/math10183368>
- [18] Wei Zhou, Pengjun Wang, Ali Asghar Heidari, Xuehua Zhao, and Huiling Chen. 2022. Spiral Gaussian mutation sine cosine algorithm: Framework and comprehensive performance optimization. *Expert Systems with Applications* 209 (2022), 118372. <https://doi.org/10.1016/j.eswa.2022.118372>

# An Analysis Software for Visual Position and Attitude Measurement Algorithm

Xu Tao

CSSC Systems Engineering Research Institute  
18911990671@189.CN

Cai Bin

CSSC Systems Engineering Research Institute  
acaibin@sina.com

Zhang Jing

University of Electronic Science and Technology of China  
zjing@uestc.edu.cn

Wang Yafei

CSSC Systems Engineering Research Institute  
18911990744@163.com

## ABSTRACT

Visual position and attitude measurement (VPAM) system has been widely used in obtaining space target information. In order to better obtain different target information and meet the requirements, it is particularly important to select a correct and effective measurement algorithm. In this paper, a performance evaluation software of VPAM algorithm is designed, which can compare and analyze the accuracy and complexity of algorithms used by different VPAM models, and help users select appropriate position models to obtain more accurate target information. Finally, the software is verified by using the dual photogrammetric model in the shipborne helicopter landing system, and the validity of the analysis software is verified by comparing the calculation results with the theoretical value of the algorithm accuracy analysis. The main contribution of this paper is that, as far as we know, it is the first time to try to evaluate the complexity and accuracy of the algorithm by building analysis software instead of theoretical analysis.

## CCS CONCEPTS

• Software and its engineering;

## KEYWORDS

Position attitude measurement, Analysis software, Double photography

### ACM Reference Format:

Xu Tao, Zhang Jing, Cai Bin, and Wang Yafei. 2023. An Analysis Software for Visual Position and Attitude Measurement Algorithm. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590043>

## 1 INTRODUCTION

VPAM is an important part of aerospace technology, such as space rendezvous and docking, space robot navigation, space robotic arm

capture, unmanned aerial vehicle takeoff and landing, aerial refueling, aircraft status monitoring, etc. It is also widely used in civil fields such as automatic production equipment, automatic detection equipment, face recognition, etc. Therefore, many researchers and research teams have studied the technology of VPAM and developed many VPAM systems. According to the number of cameras used in the measurement, these VPAM systems are divided into monocular system, binocular measurement system and multi-cular system; According to the target feature information used in the measurement, it can be divided into the system based on cooperative targets and the one based on non-cooperative targets. Measuring moving objects based on monocular system [1, 2] and obtaining 3D information of objects based on binocular system [3-5] have been widely used in many fields. The VPAM algorithms based on cooperative targets and the ones based on non-cooperative targets have been reported in the literature [2, 6, 7]. The number of feature points on the target and the relationship between them are the key to obtain the target position and attitude. For these problems, a series of methods and algorithms have been proposed [8–11]. Given the 3D feature points of the target and its 2D projection image, the problem of solving the projection equation in the measurement process can be converted into the classic problem of machine vision: PnP problem, that is, given the 3D coordinates of object space and corresponding 2D coordinates of image space of several feature points, the attitude angle information and 3D distance information of the measured target can be obtained through measurement algorithms. Now, many algorithms have emerged in the discussion of PnP, among which P3P, P4P and P5P have been further studied [12–15].

In the face of so many VPAM algorithms, which method should be chosen? When constructing each algorithm, researchers will evaluate the performance of the algorithm, which may adopt theoretical analysis methods, simulation analysis methods, or experimental verification methods. Different algorithms have different evaluation methods and indicators. For system designers and engineering application personnel, if they have a large number of visual measurement algorithms available, how should they make a choice? Considering that they do not pay much attention to the specific implementation details, it is unrealistic to carefully compare the advantages and disadvantages of each algorithm from the perspective of theory and implementation process, so a unified analysis software is needed to help them make correct choices quickly and effectively.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590043>

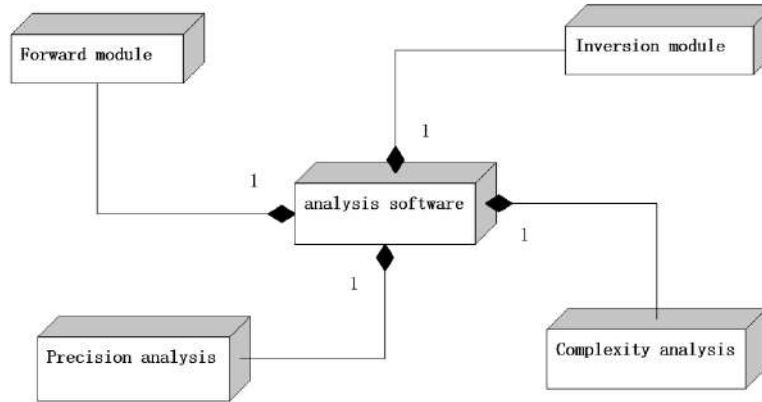


Figure 1: Composition of analysis software

To solve the above problem, this paper designs an analysis software to evaluate the VPAM algorithm. The analysis software adopts modular design, which can analyze the time complexity and accuracy of the visual pose calculation algorithm, and help the system designers and engineering application personnel select the appropriate visual pose algorithm to better meet the requirements. In this paper, the accuracy of the dual photogrammetric algorithm in the shipborne helicopter landing system is calculated by the analysis software. The calculation results are compared with the theoretical value of the algorithm accuracy analysis, which verifies the effectiveness of the analysis software.

## 2 DESIGN OF THE SOFTWARE

### 2.1 Design method

Comparisons can be made both theoretically and practically. The so-called theoretical comparison method is to count the process complexity of the algorithm in the form of pseudocode of the algorithm to be analyzed whose work flow has been fully understanding, and use the error theory to analyze and derive the error formula of the test algorithm, and then make a horizontal comparison [16]. The so-called experimental method is to repeatedly change the target parameters, system parameters and condition parameters of the algorithm, repeatedly execute the algorithm count its execution time, and use the reversible nature of the forward model, that is simulation model, and inversion model, that is the algorithm, to repeatedly execute the process of simulation, parameter noise, measurement, and error calculation, and statistically analyze the accuracy of the test model. The latter is essentially a Monte Carlo simulation process. For an engineering application software, the experimental method is preferable. Therefore, this software uses this practical method to design the comparison software.

### 2.2 Software system composition

As shown in Figure 1, the analysis software is composed of four main modules: forward modeling module, inversion module, complexity analysis and precision analysis module. The simulation model is loaded and executed by the forward module, and the VPAM algorithm loaded and executed by the inversion module. The complexity analysis module and precision analysis module are used to analyze the complexity and precision of the position model respectively. The analysis program is used for human-computer interaction, including determination of measurement model, parameter setting and comparative analysis.

As shown in Figure 2, the analysis software consists of software interface, forward modeling module, inversion module, complexity analysis and accuracy analysis module. The user interacts with the analysis software through software interface, and the relevant instructions are transmitted to the corresponding modules to achieve the corresponding functions. Each module is independent of each other, but also related to each other and interacts with each other to complete all functions of the comparison software.

### 2.3 Software deployment

As shown in Figure 3, the software interface is developed in Matlab and compiled into an executable file. Modules such as forward modeling, inversion, complexity analysis and precision analysis are integrated into the software in the form of dynamic library. The analyzer module interacts with other modules in the way that the executable file calls the dynamic library. The algorithms to be analyzed are either evaluated in an independent operation mode, or in a dynamic library function call mode.

## 3 MODULE FUNCTION

### 3.1 Forward module

The forward modeling module is used to load the simulation model (such as the dual camera model) in which the imaging position of the cooperative feature points on the target in the image plane are

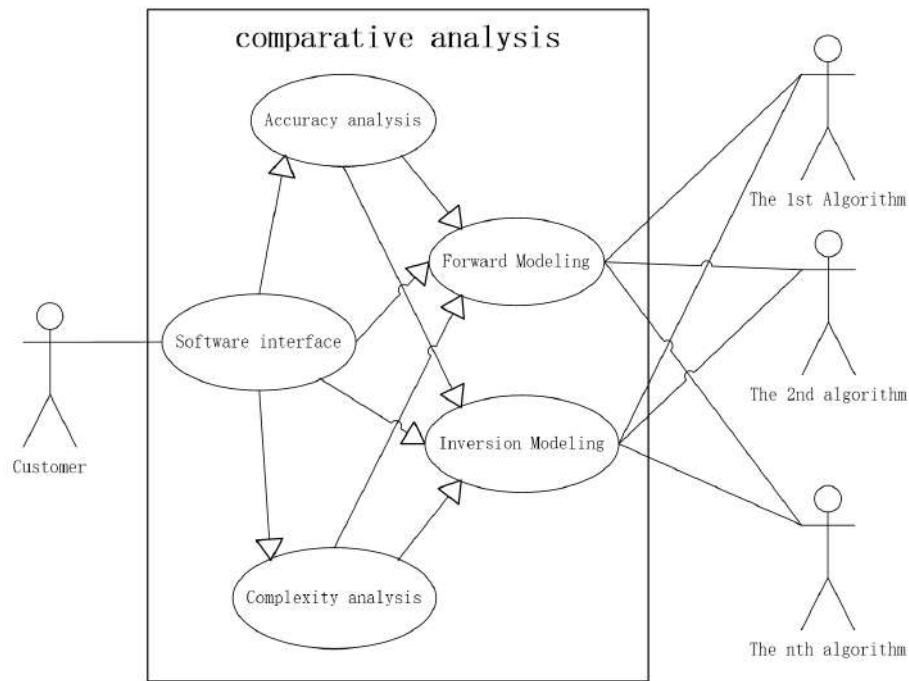


Figure 2: Use case of analysis software

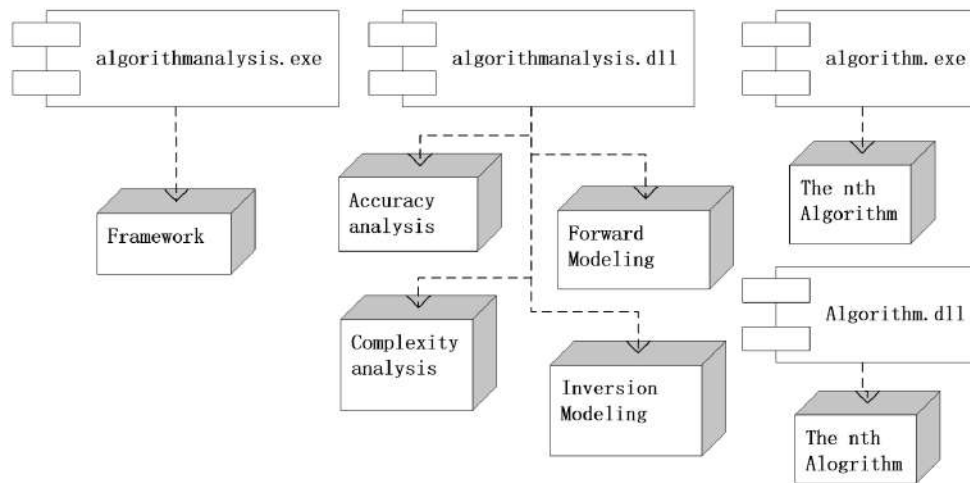


Figure 3: Software deployment

calculated. On the premise that the system parameters (position relationship between cooperative targets, camera position and attitude, camera internal parameters and other parameters) are fixed and the target parameters (space position, attitude and other state parameter information) have been set, the module generates conditional parameters (image coordinates of cooperative beacons in the corresponding camera on target). Multiple measurement algorithm can have a common simulation model or multiple simulation models. Information about the loading process comes from the analyzer.

### 3.2 Inversion module

The inversion module is used to load the measurement model (such as dual photogrammetric algorithm). The comparison object can calculate the target parameters from the simulation data, that is, the position of the cooperative points. The comparison between the observed value and the true value of the target parameter is used to analyze the complexity and accuracy of the comparison object. On the premise that the system parameters, including the position relationship between cooperative targets, camera space

settings, camera internal parameters, etc., are fixed, the conditional parameters of the target feature points, such as the target space position and attitude, are generated according to the conditional parameters, such as the image coordinates corresponding to the cooperative feature points on the target in the camera).

### 3.3 Complexity analysis module

The complexity analysis module is used to analyze the time complexity of the measurement algorithm. The module first sets the system parameters, then randomly generates a series of target parameters, then calls the forward modeling module to calculate the condition parameters, and then adds noise to the condition parameters to obtain the noisy condition parameters. The module then takes the noise-containing condition parameters and system parameters as input parameters, calls the inversion module to calculate the test algorithm, and records the running time of the VPAM algorithm. Repeat the process of generating target parameters until recording the running time of the algorithm to obtain a set of running time of the VPAM algorithm. Through statistical analysis of this set of data, we can obtain statistical information such as average calculation time, maximum calculation time and minimum calculation time.

### 3.4 Precision analysis module

The accuracy analysis module is used to analyze the measurement accuracy of the measurement algorithm. The module first sets the system parameters, randomly generates the true values of the target parameters, and then calls the forward module to calculate the condition parameters. Then add noise to the condition parameter to get the noise-containing condition parameter, and then take the noise-containing condition parameter and system parameter as the input parameters, call the inversion module to calculate the VPAM algorithm to calculate the observed value of the target parameter, and calculate the error between the true value and the observed value and record it. The process of repeatedly generating the true value of the target parameter until calculating and recording the test error is repeated to obtain a set of test errors of the VPAM algorithm. Through statistical analysis of this group of data, statistical information such as average test error, maximum test error and minimum test error are obtained.

### 3.5 main program

The main program is designed with Matlab, which is responsible for the human-computer interaction between the software and the user, and the relevant instructions are transmitted to the corresponding modules through the interface between the analysis program and other modules to achieve the corresponding functions. When the analyzer is configuring, it can input the configuration information directly from the graphical interface or read the configuration information from the specified XML file. After obtaining the configuration information, according to these information, call the creation model function of the forward module to configure the forward module, call the creation model function of the inversion module to configure the inversion module, set the complexity analysis module through the input parameters, system parameters, output parameters and random number seed function, and also set the precision analysis module through the corresponding input

parameters, system parameters, output parameters and random number seed function. When the analyzer receives the complexity analysis, it calls the running function of the complexity analysis module, and the analysis results are given in the form of report. When the analyzer receives the precision analysis, it calls the running function of the precision analysis module, and the analysis results are given in the form of report.

## 4 EXAMPLE PROGRAM VERIFICATION

It is easy for helicopters to land on land, but it is relatively difficult to land on ships. Affected by complex sea conditions, the ship may make irregular bumps and swings at any time. The helicopter hovers over the ship and makes random motion relative to the ship, which adds many uncertainties to the smooth landing of the helicopter. In order to improve the safety of helicopter ship landing, the helicopter landing system can be equipped on the ship. The landing system needs to determine the position and attitude of the helicopter relative to the ship in real time, which can be realized by double photogrammetry. Next, we use our analysis software to analyze the dual photography algorithm.

### 4.1 Double photogrammetry

The dual photogrammetric algorithm [17] is based on two camera coordinate systems, one target coordinate system and one reference coordinate system. The key of the algorithm is to determine the relationship between the target coordinate system and the reference coordinate system, that is, a translation vector and three rotation angles, by which the target coordinate system is transformed into of the reference coordinate system.

The geometric model of the dual photogrammetric system is as shown in Figure 4, where  $Oxyz$  is the ship coordinate system,  $O_Lx_Ly_Lz_L$  is the left camera coordinate system,  $O_Rx_Ry_Rz_R$  is the right camera coordinate system,  $O_Hx_Hy_Hz_H$  is the helicopter coordinate system,  $O_1, O_2, O_3, O_4$  are the cooperative points installed on the left side of the helicopter, and  $O_5, O_6, O_7, O_8$  are the cooperative points installed on the other side.

Assume the coordinates of the cooperative points in the target coordinate system, the rotation matrix  $r$  and translation matrix  $p$  from the reference coordinate system to the camera coordinate system, the rotation matrix  $R$  and translation matrix  $P$  from the target coordinate system to the reference coordinate system, and the internal parameters of the camera.

$A = \begin{bmatrix} f_u & 0 & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}$  have been all known, the image coordinates  $(u, v)$  of the cooperative target can be determined by formula

$$z[v] = \begin{bmatrix} u & f_u & 0 & u_0 \\ 0 & f_v & v_0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r & p \end{bmatrix} \begin{bmatrix} R & P \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_H \\ y_H \\ z_H \\ 1 \end{bmatrix},$$

which is realized by the forward module of the double photogrammetric algorithm analysis.

Using the camera pinhole model, the dual photogrammetric algorithm can establish two nonlinear equations from each cooperative target. A total of 16 equations can be established from eight points,

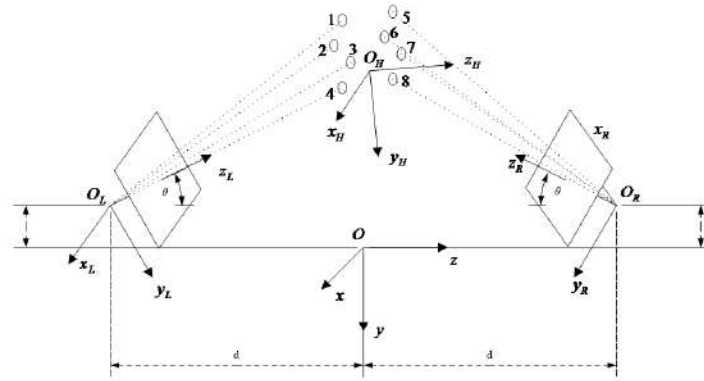


Figure 4: Geometric model of dual photogrammetric system

and a six-variable nonlinear equation system composed of 16 equations is obtained. This is an overdetermined form of equations. First, an initial solution is obtained by geometric method, and then the least square solution is obtained by Newton-Simpson iterative method. This is the inversion module of dual photogrammetric algorithm analysis.

## 4.2 Software workflow

The analysis software involves three workflows, including environment setting, precision analysis and complexity analysis. The specific process of setting environmental parameters is: start main program of the software to enter the analysis program, use the forward module to load the dual photogrammetric simulation model, and use the inversion module to load the dual photogrammetric algorithm, according to SEA HAWK helicopter set the position relationship of cooperative points (The horizontal distance between points on the same side width is 1m, the vertical distance between points height is 1m; the horizontal distance between beacons on both sides thickness is 1.5m, the distance between the bottom points and the coordinate origin lander is 0.5m, and the coordinates are shown in Table 1), according to camera parameter and layout, camera pose (rotation angle from left camera to reference coordinate system is  $\{-\theta, \phi, 0\}$ , translation vector from left camera to reference coordinate system is  $\{b, -a, -d\}$ , rotation angle from right camera to reference coordinate system is  $\{180 - \theta, \phi, 180\}$ , translation vector from right camera to reference coordinate system is  $\{b, -a, d\}$ , The z-direction distance from the lens center to the reference coordinate system coordinate origin is 10m, The y-direction distance from the lens center to the reference coordinate system coordinate origin is 0.1m, The x-direction distance from the lens center to the coordinate origin of the reference coordinate system is 5; Relative to the camera elevation in the reference coordinate system  $\theta$  is 12 Degree; The horizontal angle of camera in the landing coordinate system is  $\arctan(b/d)/\pi * 180$ ), camera internal parameters (the internal parameters of the left camera is  $\{506, 0, 358\}$ ,  $\{0, 503, 368\}$ ,  $\{0, 0, 1\}$ ; the internal parameters of the right camera is  $\{502, 0, 378\}$ ,  $\{0, 510, 348\}$ ,  $\{0, 0, 1\}$ ) as shown in Figure 4, enter the complexity analysis module to set the analysis

parameters, enter the accuracy analysis module to set the analysis parameters, return to the analysis program, and the environment setting is complete.

The process of precision analysis and complexity analysis is similar. Enter the precision analysis module/complexity analysis module of analysis software, call forward modeling module to set the mean value of helicopter position and attitude parameters (rotation angle to be  $\{0, 10, 0\}$ , translation vector to be  $\{0, 2, 0\}$ ) and parameter variance (point pose noise variance to be 0.01, camera internal parameter noise variance to be 1, camera translation vector noise variance to be 0.1 degrees, camera translation vector noise variance to be 0.01 m), then calculate condition parameters, then use the condition parameter (noise variance of image coordinate to be 0.5) to calculate the measured value by calling the inversion module, compare the error results of the measured value and the true value, return to the accuracy analysis module/complexity analysis module, set the number of test cycles N (100000) and set the corresponding convergence conditions, and count the algorithm error and the running time to determine whether the convergence or the number of cycles has been reached. Finally, the precision analysis module/complexity analysis module carries out comprehensive statistical analysis and gives the results in the form of reports.

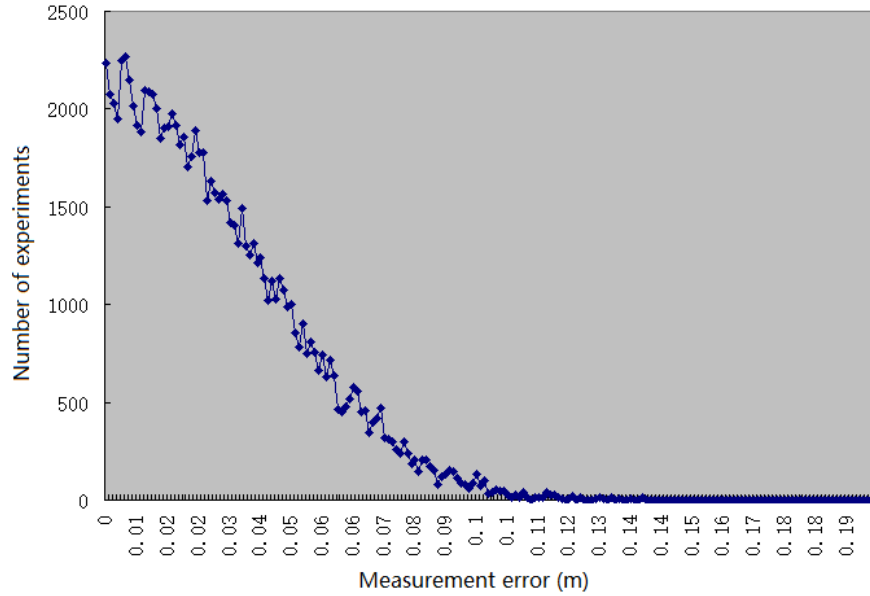
## 4.3 Algorithm analysis results

**4.3.1 Accuracy analysis of analysis software.** The error distribution obtained by the precision analysis module of the analysis software is shown in Figure 5, and the statistical error is 0.029862 meters.

**4.3.2 Influence of horizontal position of target on measurement accuracy.** In addition to the true value of the helicopter's x coordinate, the other parameters are still set according to the parameters described in the accuracy analysis section of Section 4.2. The analysis software obtains the relationship between the measurement accuracy of the x coordinate and the position error of the cooperative feature point at different true values the helicopter's x coordinate, as shown in Figure 6. The results show that the smaller the true value of the x coordinate is, the more accurate the measurement results

**Table 1: Position of beacon in helicopter coordinate system (unit: m)**

Location	Beacon No	xH	yH	zH
left	1	- width/2	Lander+ height	- thickness/2
left	2	width/2	Lander+ height	- thickness/2
left	3	- width/2	Lander	- thickness/2
left	4	width/2	Lander	- thickness/2
right	5	- width/2	Lander+ height	thickness/2
right	6	width/2	Lander+ height	thickness/2
right	7	- width/2	Lander	thickness/2
right	8	width/2	Lander	thickness/2

**Figure 5: Parameter measurement and analysis**

will be under the same position error, and this trend is consistent with the measured situation.

**4.3.3 Influence of camera distance on measurement accuracy.** In addition to setting the distance between the two cameras, the other parameters are still set according to the parameters described in the accuracy analysis section in Section 4.2. The relationship between the measurement accuracy and the position error of the cooperative feature points is obtained by the analysis software when the distance between the cameras is different, as shown in Figure 7. The results show that the smaller the distance between cameras, the more accurate the measurement results will be under the same position error, and this trend is consistent with the actual situation.

#### 4.4 Theoretical estimation of algorithm accuracy

The error of the algorithm measurement results  $w = [\alpha \ \beta \ \gamma \ P_1 \ P_2 \ P_3]$  is analyzed with theoretical methods. This paper mainly analyzes the error caused

by the measurement error of parameters. Any component of  $w$  is represented by  $g$ . Obviously,  $g$  is a nonlinear function of all components of  $w$ . By using the error transfer formula of the nonlinear function, the formula relationship between the mean square error  $\delta_g^2$  of  $g$  and the mean square error of each parameter can be obtained. After a series of calculations, a parameter expression can be obtained. The difference method is used to construct the parameter equation. According to the error estimation formula of the least squares method, the estimated accuracy of the measured value should be

$$\sigma_g = \sqrt{\frac{\delta_g^2}{2m-6}}.$$

The measurement error of the distance between the target coordinate origin and the reference system coordinate origin is

$$\sigma_d = \sqrt{\frac{\delta_{w_4}^2 + \delta_{w_5}^2 + \delta_{w_6}^2}{2m-6}}$$

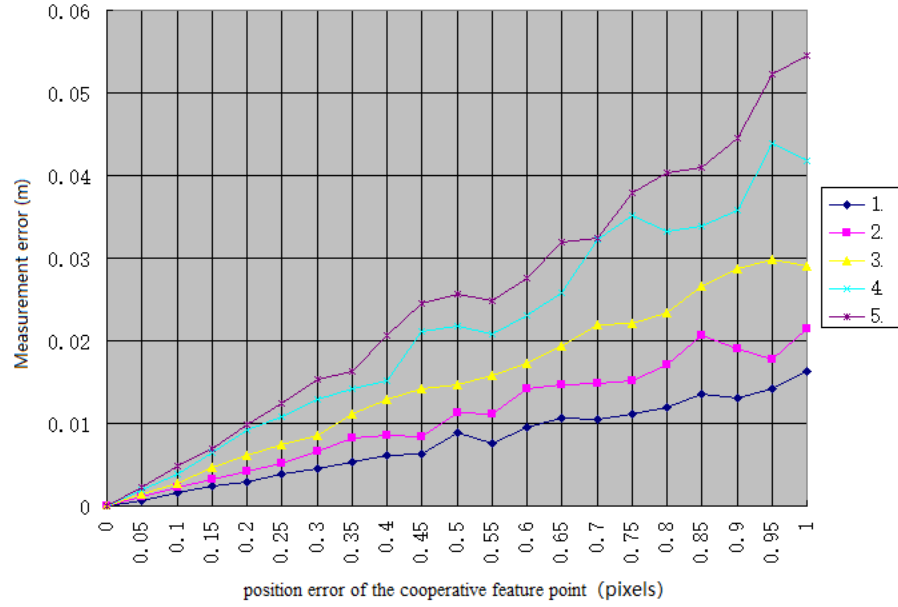


Figure 6: Influence of horizontal position of target on measurement accuracy

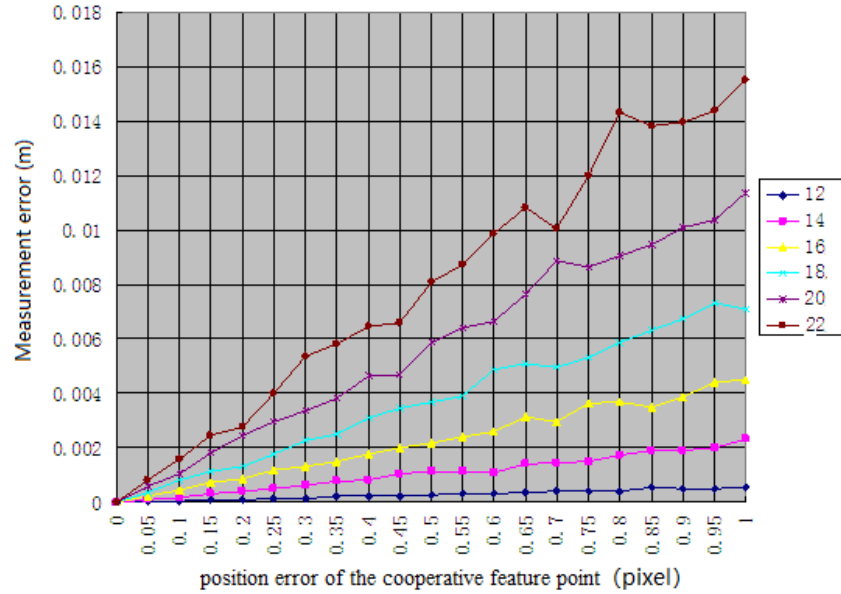


Figure 7: Influence of camera distance on measurement accuracy

Estimate the simulation experiment, and the result is  $\sigma_\alpha=2.78^\circ, \sigma_\beta=3.16^\circ, \sigma_\gamma=5.04^\circ, \sigma_{P_1}=0.056346, \sigma_{P_2}=0.053708, \sigma_{P_3}=0.054349$ .

Because the number of beacon points is 8,

$$\sigma_d = \sqrt{\frac{0.056346^2 + 0.053708^2 + 0.054349^2}{2 \times 8 - 6}} = 0.0300\text{m}.$$

It is consistent with the experimental result of 0.029862 meters.

## 5 CONCLUSION

This paper designs an analysis software to evaluate the visual pose algorithm, explains the design methods and tools of the analysis software, and introduces the overall analysis of the software system,

the composition modules and the workflow in detail. Finally, the analysis software is verified by the dual photogrammetric model in the shipboard helicopter landing system, and the accuracy analysis results of the algorithm are obtained, it is compared with the theoretical value of the algorithm accuracy. It provides an accurate and effective software tool for analyzing different pose measurement system models in the future, which makes it efficient and convenient for the system application personnel to select the appropriate visual pose algorithm to better meet the requirements.

## REFERENCES

- [1] Yuan Ye. 2011. Methods and System Implementation On Three-Dimensional Attitude Measurement Based On Monocular Vision. Master's thesis. Harbin Institute of Technology, Harbin, China.
- [2] Niu Hao. 2014. Pose Estimation of Non-Cooperative Object Based On Monocular Vision. Master's thesis. Harbin Institute of Technology, Harbin, China.
- [3] Wu Bin, Xue Ting, Zhu Ji-gui, Ye Sheng-hua. 2006. Calibrating Stereo Visual Sensor with Free-position Planar Pattern. *Journal of Optoelectronics -Laser*, 17(11), 1293-1296.
- [4] Yang Jing-Hao, Liu Yang, Wang Fu-ji, Jia Zhen-yuan. 2016. Calibration of Binocular Vision Measurement System. *Optics and Precision Engineering*, 24(2), 300-308.
- [5] Dong Feng, Sun Li-ning, Ru Chang-hai. 2014. Measurement Method of Medical Robot Positioning System Based on Binocular vision. *Journal of Optoelectronics -Laser*, 25(5), 1027-1034.
- [6] Shang Yang. 2006. Researches on Vision-Based Pose Measurements for Space Targets. Ph.D. Dissertation. Graduate School of National University of Defense Technology, ChangSha, China.
- [7] Zhao Lianjun. 2014. Research on Mono-vision Pose Measurement Based on Features of Target. Ph.D. Dissertation. Institute of Optics and Electronics Chinese Academy of Sciences, Chengdu, China.
- [8] J. C. Tietz, L. M. German. 1982. Autonomous Rendezvous and Docking. *Proceeding of the American Conference*, 2, 460-465. O
- [9] W. B. Jatko, J. S. Goddard, *et al.* 1996. Crusader Automated Docking System Phase III Report. Tech. Rep. ORNL/TM-13177, Oak Ridge National Laboratory. 140 pages.
- [10] J. L. Junkin, D. Hughes, *et al.* Vision-Based Navigation for Rendezvous. 1999. Docking and Proximity Operations. 22nd Annual AAS Guidance and Control and Control Conference, *Advances in Astronautical Sciences*. 335 pages.
- [11] L. Quan, Z. D. Lan. 1999. Linear N-Point Pose Determination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(7), 1-7.
- [12] SUN Feng-Mei, WANG Bo. 2010. A Note on the Roots Distribution and Stability of the PnP Problem. *ACTA AUTOMATICA SINICA*, 36(9), 1213-1219.
- [13] Leng D, Sun W. 2009. Finding all the solutions of PnP problem. *Proceedings of the Imaging Systems and Techniques, IEEE International Workshop on Imaging Systems and Techniques; IST '09*. IEEE. Shenzhen, China.
- [14] Moreno Noguera F, Lepetit V, Fua P. 2007. Accurate Non-Iterative O (n) Solution to the PnP Problem. *IEEE 11th International Conference on Computer Vision*. 1-8.
- [15] Gao X-S, Hou X-R, Tang J. 2003. Complete solution classification for the perspective-three-point problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 930-943.
- [16] CATHERINE C. MCGEOCH, 2012. *A Guide to Experimental Algorithms*. Cambridge University Press.
- [17] Zhao Zhenqing, Ye Dong, *etc.* 2014. Binocular Vision Method of Measuring Pose Based on Perpendicular Lines. *ACTA OPTICA SINICA*, 34(10), 193-199.

# Heuristic Search for DNN Graph Substitutions

FeiFei Deng

School of Computer Science and Artificial Intelligence  
Wuhan University of Technology  
Wuhan, China  
dff@whut.edu.cn

HongKang Liu

School of Computer Science and Artificial Intelligence  
Wuhan University of Technology  
Wuhan, China  
lhhkk@whut.edu.cn

## ABSTRACT

The research and development of deep learning cannot be separated from deep neural networks (DNNs). DNNs become deeper and more complex in pursuit of accuracy and precision, leading to significantly increasing inference time and training cost. Existing deep learning frameworks optimize a DNN to improve its runtime performance by transforming computational graphs based on hand-written rules. It is hard to scale when adding some new operators into DNNs. TASO can automatically generate graph substitutions that solve maintainability problems. An optimized graph will be explored by applying a sequence of graph substitutions. However, TASO only considers the runtime performance of the model during the search, which may lose potential optimization. We propose HeuSO, a fine-grained computational graph optimizer with heuristics to handle this problem. HeuSO extracts the type and number of operators of the computational graph and classifies them into four abstract types as high-level features, which facilitate subsequent heuristic search and pruning algorithms. HeuSO generates a better sequence of graph substitutions and finds a better-optimized graph by the heuristic function, which integrates the cost and high-level features of the model. To further reduce the time of searching, HeuSO implements a pruning algorithm. Through high-level specifications, HeuSO can quickly determine whether subgraphs of the original graph match the substitution rules. Evaluations on seven DNNs demonstrate that HeuSO outperforms state-of-the-art frameworks with 2.35 $\times$  speedup while accelerating search time by up to 1.58 $\times$ .

## CCS CONCEPTS

• **Computing methodologies**  $\rightarrow$  **Neural networks**; • **Computer systems organization**  $\rightarrow$  **Neural networks**.

## KEYWORDS

graph substitutions, deep neural network, graph-level optimization, heuristic search

## ACM Reference Format:

FeiFei Deng and HongKang Liu. 2023. Heuristic Search for DNN Graph Substitutions. In *2023 2nd Asia Conference on Algorithms, Computing and*

*Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590044>

## 1 INTRODUCTION

Deep learning has been successful in many applications, such as natural language processing, speech recognition, medical applications, computer vision, and intelligent transportation system [5]. The success of deep learning cannot be achieved without high-performance deep learning models. For Example, ResNet [6] won the first prize in the ILSVRC 2015 competition with 152 layers, and BERT [4] achieved state-of-the-art performance on several natural language understanding tasks When it was published. BERT<sub>BASE</sub> has 110 million parameters and BERT<sub>LARGE</sub> has 340 million parameters.

The DNN model with a complex structure and enormous parameters gains great performance and well-pleasing accuracy, but such complexity results in time-consuming training and inference. Exist deep learning frameworks like TensorFlow [1], TensorRT [13], PyTorch [11], and TVM [2] usually handle the DNN model as a computational graph. It is usually a directed acyclic graph (DAG) consisting of inputs, outputs, points, and edges. Edges represent data dependency between every two operators, and points represent mathematical math operators (e.g., convolution, add, relu, etc.), which includes many opportunities for graph optimization like operator fusion or graph rewriting. A common way to optimize a graph is to use a sequence of graph substitutions generated by a set of fixed rules that need domain experts to make.

In general, a substitution rule is also a computational graph consisting of several points and edges. A typical example is the fusion of a convolution layer with a relu activation layer that satisfies the property of equation (1).

$$\begin{aligned} &Conv(input, weight, strides, padding, A_{Relu}) = \\ &Relu(Conv(input, weight, strides, padding, A_{None})) \end{aligned} \quad (1)$$

Suppose we have a convolution operator  $Conv$  with input, weight, padding, and activation function of  $relu$   $A_{Relu}$  and a relu operator  $Relu$ .  $A_{None}$  means no activation function. As Figure 1 depicts, subgraphs are compared with the source operators in each rule, and if they match, they are replaced with the destination operators in the rule to complete a graph substitution process. Mapping of the original graph to the target graph is realized by substitution rules.

Manually designed graph substitutions need an experienced engineer with lots of work. It cannot be scaled when you meet a new operator and is also error-prone because of hand-written rules from experience or instinct. MetaFlow [9] proposes relaxed graph substitution by relaxing the strict performance improvement constraint. MetaFlow breaks the strict rules of graph substitution but is subject to predefined substitution rules. TASO [8] automatically generates

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

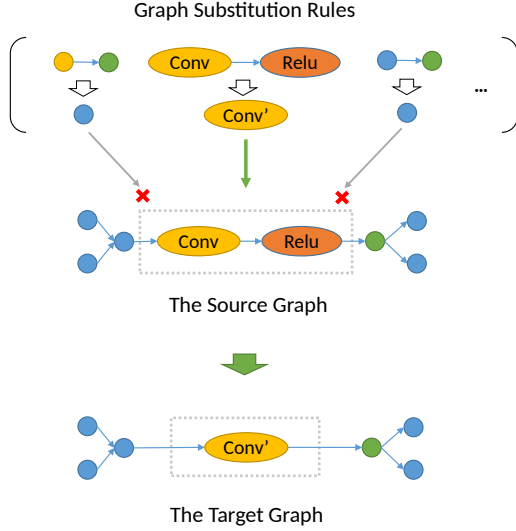
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590044>

graph substitutions and finds an optimized graph by applying substitutions. However, TASO only considers the cost of the model, which may miss the possible optimization.



**Figure 1: A Graph Substitution Example of Conv-Relu Fusion.**

In this paper, we observed that classifying operator type plays an important role in graph-level optimization. We leverage features and abstractions from the graph and benefit subsequent heuristic search and pruning algorithms. We mapped each operator into four types depending on the relationship between the inputs and the outputs similar to DNNFusion [10]: One-to-One, Many-to-Many, Reorganize, and Shuffle. Each operator type will be scored by its profitability for operator fusion and probable subsequent operator-level optimization. Then we collect all operators in the graph and calculate the whole score by summing up each type as a part of the heuristic function. We make a balance between optimization efficiency and search time because it is difficult to find the globally optimal optimized graph in a limited time. The heuristic function with high-level features guides the search for better-optimized graph substitution in each round, a good tradeoff between efficiency and time.

However, optimizing computational graphs using graph substitutions can formally define a minimum vertex cut problem [3], which is NP-hard and Poly-APX-complete. Heuristic search can find the better-optimized graph in the limited search times and replace the original graph but may cause high latency due to large search space. We propose a pruning algorithm with high-level features that quickly determine the feasibility of a substitution rule applied in a graph.

We find the relatively best-optimized graph in a large search space by heuristic function and accelerate the search procedure with the pruning algorithm. Without changing the functionality of the graph (ensuring that the original graph and the optimized graph have the same output for the same input), the inference time of the model is reduced while the search time is greatly reduced,

and facilitates subsequent end-to-end deployment. The optimized graph HeuSO found achieved state-of-the-art performance and better search efficiency.

To summarize, our main contributions include:

- We design a novel heuristic function that considers both the overall performance of the graph and the importance of the mapping of operator types.
- We use a pruning algorithm with high-level features to reduce large search space and accelerate the optimization process.
- We take advantage of high-level abstract features of graphs, greatly accelerating the time of finding the better optimized computational graph with heuristics.

HeuSO is extensively evaluated on seven cutting-edge DNNs, ResNet-50 [6], ResNext-50 [14], Inception-v3 [12], NasRNN [15], NasNet-A [16], Squeezenet [7] and BERT [4]. HeuSO will be compared to the state-of-the-art framework TASO in the following three metrics, cost, runtime, and search time. For ResNet-50, Inception-v3, and NasNet-A, HeuSO is not as good as TASO in time and cost metrics, but the difference is at a maximum of 5%. In other models, HeuSO outperforms TASO by up to 2.35×. Moreover, HeuSO found optimizations of TASO omission. As for the search time metric, HeuSO outperforms TASO with a 1.58× speedup.

## 2 HEURISTIC SEARCH ON GRAPH SUBSTITUTIONS

TASO uses a cost-based backtrack algorithm to generate candidate substitutions which optimize the inference time of the model. However, the search for optimized graphs based on cost alone might still lose potential optimization opportunities and miss the possibility of finding better graphs.

In this paper, we propose a heuristic function that jointly takes into account the cost and operator characteristics of the model. In a sequence of graph substitutions, we think the candidate graph with both lower cost and more fusion-prune operators has a better opportunity for optimization. Then we will prioritize backtracking these graphs to get better search results.

### 2.1 Heuristic Function

The most crucial part of HeuSO is the heuristic function. It is not only the core of the heuristic search algorithm but also directly determines searched graph substitutions. Given a graph  $\mathcal{G}$ , total cost and mapping cost of graph  $\mathcal{G}$  are denoted as  $Cost(\mathcal{G})$  and  $MappingCost(\mathcal{G})$ .  $Type(op)$  represents the mapping type of a op.  $Weight(Type(op))$  and  $Number(Type(op))$  represent weight and number of type of an operator.

We define a heuristic function  $H(\mathcal{G})$  as following:

$$H(\mathcal{G}) = \frac{Cost(\mathcal{G})}{MappingCost(\mathcal{G}) + 1} \quad (2)$$

where  $Cost(\mathcal{G})$  equals

$$Cost(\mathcal{G}) = \sum_{op \in \mathcal{G}.ops} Cost(op) \quad (3)$$

, and  $MappingCost(\mathcal{G})$  equals

$$MappingCost(\mathcal{G}) = \sum_{op \in \mathcal{G}.ops} Weight(Type(op)) \times Number(Type(op)) \quad (4)$$

The heuristic function  $H(\mathcal{G})$  aims to prioritize the search for graphs with smaller cost and larger mapping cost in each round of the optimal graph search. We let  $MappingCost(\mathcal{G}) + 1$  not  $MappingCost(\mathcal{G})$  as the denominator of  $H(\mathcal{G})$  because we consider the case that there may be a graph with its  $MappingCost$  equals to 0 (i.e. all operators in a graph are Many-to-Many type which weight is zero). This case may cause an exception and crush the system. The graph with a smaller cost means some operators in the graph have less runtime and thus total runtime is smaller. It will have better inference performance than the original graph. A graph with a larger mapping cost means more optimization opportunities because it contains more One-to-One operators, Reorganize operators, or Shuffle operators rather than Many-to-Many operators, and easier to find the better-optimized graph.

Therefore, the smaller the value of the heuristic function, the better the graph substitution search and the better the chance of finding the optimal graph. If two graphs have the same cost, we prefer to search for the graph with a larger mapping cost. The larger the mapping cost, the smaller result of the function  $H(\mathcal{G})$ ; If two graphs have the same mapping cost, we prefer to search on the graph with a smaller cost. The smaller cost, the smaller result of the function  $H(\mathcal{G})$ . Ideally, according to the heuristic function, the graph with both the minimum cost and the maximum mapping cost is the most worthy of search graph substitution.

## 2.2 Heuristic Search Algorithm

Algorithm 1 describes details of the heuristic search. First of all, we prepare a candidate graph substitution sequence  $\mathcal{S}$  and add the original graph  $\mathcal{G}$  into it. Sequence  $\mathcal{S}$  contains all the candidate graphs found in the search phase and is sorted according to the heuristic function. Then, continue to search for possible replacements on the selected optimal graph to generate a new optimal graph, and keep repeating this process until the number of searches is exhausted or there is no more optimization possible for the graph. Sequence  $\mathcal{S}$  will be continuously filled by optimized graph  $g'$  in the process of searching. Finally, we generate an optimized graph with a different model structure.

We set  $\alpha = 1.05$  as TASO did. The value of  $\alpha$  is not 1.0 prevent reduces the search to a simple greedy algorithm that easily obtains local optimal solution, and making  $\alpha$  a little higher makes the search explore more possible candidates and causes more backtracking. Considering the limited computing resources of the server and the performance of the algorithm, we set a constant THRESHOLD to limit search times. The search should stop immediately when  $N$  exceeds the threshold.

**Time Complexity.** For the simplicity of the analysis, we assume that each graph is weakly connected and the scale of the graph in the replacement rule is small (in fact, it does). After  $N$  graph substitution, there are  $O(|V| + N)$  nodes in the computational graph, where  $|V|$  is the number of nodes in the original computation graph. There can be  $O(|V| + N)$  graph substitutions to be checked in line

---

### Algorithm 1 Heuristic Search Algorithm

---

**Input:** A computation graph  $\mathcal{G} = (V, E)$ , a set of substitution rules  $\mathcal{P}$ , an integer  $N$ , a candidate graph substitution sequence  $\mathcal{S}$

**Output:** A optimized graph with minimum cost

```

1:  $N \leftarrow 0$ 
2:  $\mathcal{S} \leftarrow \{\mathcal{G}\}$ 
3:  $\mathcal{G}_{best} \leftarrow \mathcal{G}, C_{best} \leftarrow cost(\mathcal{G})$ 
4: while  $\mathcal{S} \neq \emptyset$  do
5:    $g \leftarrow$  selected by heuristic function in  $\mathcal{S}$ 
6:    $\mathcal{S} \leftarrow$  remove  $g$  from  $\mathcal{S}$ 
7:   if  $cost(g) < C_{best}$  then
8:      $C_{best} \leftarrow cost(g)$ 
9:      $\mathcal{G}_{best} \leftarrow g$ 
10:  end if
11:  if  $N > THRESHOLD$  then
12:    return  $\mathcal{G}_{best}$ 
13:  end if
14:   $N \leftarrow N + 1$ 
15:  for each  $\rho \in \mathcal{P}$  do
16:    if  $\rho$  can apply in  $g$  then
17:       $g' \leftarrow$  apply  $\rho$  to  $g$ 
18:      if  $cost(g') < \alpha \times cost(g)$  then
19:         $\mathcal{S} \leftarrow \mathcal{S}$  add  $g'$ 
20:      end if
21:    end if
22:  end for
23: end while
24: return  $\mathcal{G}_{best}$ 

```

---

5 of Algorithm 1, and we need  $O(|V| + N)$  time to find them all. Applying  $\rho$  to  $g$  (line 17) and calculating the cost of a computational graph (line 3) require  $O(|E| + N)$  time. In total, algorithm 1 takes  $O(|\mathcal{S}| |\mathcal{P}| (|V| + N) + |E| + N)$  time, where  $|\mathcal{P}|$  is the number of rules and  $|\mathcal{S}|$  is the length of graph substitutions sequence.

**Space Complexity.** Algorithm 1 needs to store  $O(N)$  substitution sequences (each needs  $O(|V| + |E| + N)$  space), so in total we need  $O(|\mathcal{P}| + N (|V| + |E| + N))$  space.

## 2.3 Feasibility Pruning

Applying a rule to a graph and changing a part of the structure from the rule may be time-consuming because it need to traverse the whole graph at least once to match operators from the graph and the rule. It is not easy for matching because of complexity and inefficiency, especially for some large computational graphs.

Algorithm 2 shows the pruning algorithm used in heuristic search. We can quickly judge whether a graph can apply a substitution rule to it. We count the number of operators of the graph and grasp high-level features of the graph. We will count the type and number of input operators for each optimization rule. By comparing it with the data of the previously extracted graph, as long as it is less than the number of any of the types in the replacement rule, it is judged to be infeasible. It is meaningless and time-consuming to continue searching on an infeasible graph.

**Algorithm 2** Feasibility Pruning Algorithm**Input:** A candidate graph  $g$ , a substitution rule  $\rho$ **Output:** Whether  $g$  can be optimized

```

1:  $g_{can\_opt} \leftarrow True$ 
2:  $\phi(g) \leftarrow$  high-level features from  $g$ 
3: for each  $op \in$  the original graph of  $\rho$  do
4:    $type(op) \leftarrow$  type of  $op$ 
5:   if  $type(op)$  can not be found in  $\phi(g)$  then
6:      $g_{can\_opt} \leftarrow False$ 
7:   return  $g_{can\_opt}$ 
8: end if
9: end for
10: return  $g_{can\_opt}$ 

```

### 3 EVALUATION

#### 3.1 Evaluation Setup

We use seven DNNs in the experiments, and all DNNs are state-of-the-art models in real life. Details of the models are listed in Table 1. ResNet-50 [6] is a widely used convolutional neural network for image classification and achieved the best classification performance in the ILSVRC competition. ResNeXt-50 [14] improves the model accuracy and runtime efficiency of ResNet-50 by introducing a new grouped convolution operator. Inception-v3 [12] is the third edition of Google’s Inception Convolutional Neural Network, it was intended to allow deeper networks while also keeping the number of parameters from growing too large. NasNet-A [16] and NasRNN [15] are two DNN architectures automatically discovered by machines through neural architecture search. NasNet-A and NasRNN exceed the best human-designed DNN architectures for image classification and language modeling tasks, respectively. SqueezeNet [7] is a small neural network with fewer parameters that can more easily fit into computer memory and can more easily be transmitted over a computer network. Finally, BERT [4] is a new language representation architecture that obtained state-of-the-art model accuracy on a spectrum of language tasks.

We implemented the heuristic search algorithm based on the code provided by Jia et al. [8] and ran all experiments on a Linux server with CentOS 7 system, a 14-core Intel Xeon Gold 5117 CPU, 64G DRAM, and one NVIDIA Tesla P100 GPU. We use 151 substitution rules identified by TASO [8], whose correctness has been verified by TASO. Therefore, HeuSO does not find graph substitutions outside the rules, but rather optimizations missed by TASO. To utilize GPU resources and accelerate computation, we use the CUDA toolkit and NVIDIA cuDNN library as operators’ backend, and we compute the elapsed time as the cost metrics by the cuDNN interface of all operators in a graph. The cost metrics reflect the performance of a graph and is used as the core of the heuristic function search. We did an average of 1000 experiments for each DNN to ensure the results are stable and reasonable. Considering their search speed and influence on the final results, we set the algorithm parameter  $budget = 1000$  on each model used in our experiments except Inception-v3 and NasNet-A. We set  $budget = 100$  on the Inception-v3 and NasNet-A models because of efficiency, which has a negligible effect on the result. We also set algorithm parameter  $\alpha = 1.05$  to get better performance.

**Table 1: Model used in our experiment.**

Model	Structure
ResNet-50	50 layers deep, 16 blocks of 2 types, 83 operators
ResNext-50	51 layers deep, 16 blocks of 2 types, 83 operators
Inception-v3	48 layers deep, 11 blocks of 5 types, 119 operators
NasRNN	8 layers deep, 5 blocks of 1 type, 230 operators
NasNet-A	66 layers deep, 18 blocks of 2 types, 283 operators
SqueezeNet	21 layers deep, 8 blocks of 4 types, 64 operators
BERT	16 layers deep, 8 blocks of 1 type, 113 operators

We chose TASO as the baseline and first tested the optimization performance of the heuristic search algorithm, then evaluated the optimization’s acceleration performance by comparing the search time. We will compare the TASO-optimized and our optimized cost and runtime with the original separately on seven DNNs. We also use our algorithm to compare the search time when finding the graph with the optimal cost among substitutions sequence and the entire search time with TASO. The effectiveness of the heuristic search is further demonstrated by comparing the search time for finding the relative optimal graph.

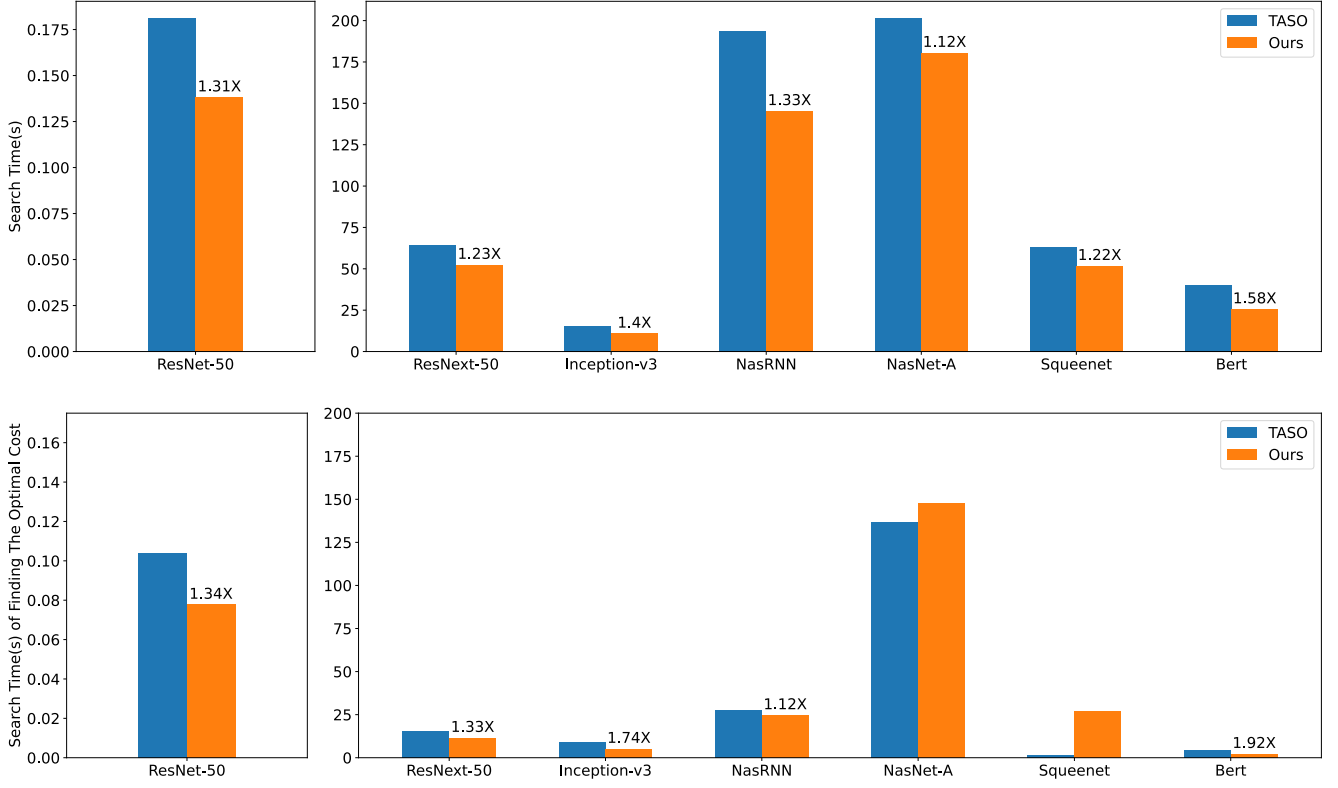
We choose the cost and the runtime of DNNs as two metrics to measure the efficiency of heuristic search. HeuSO will compare the performance of the optimized graph with TASO in these two metrics. Then, we measured the search time to verify the effectiveness of the pruning algorithm. Last, we show that HeuSO found graph optimizations on ResNext-50 and BERT that TASO did not find.

#### 3.2 Efficiency of Heuristic Search

It is extremely important to prove the effectiveness of the heuristic search. The whole work seems to make no sense if the heuristic function cannot find some optimized graph whose cost is fewer than the source graph. We compare the cost of DNNs before and after optimized by TASO and ours respectively. It is worth noting that due to the unstable experimental environment (sometimes there are multiple users on a server, and sometimes there are multiple programs running on GPU hardware), the cost calculated for the same DNN model may be different. Assuming that a given computation graph  $\mathcal{G}$ , the original cost and the optimized cost of  $\mathcal{G}$  are  $cost_\alpha$  and  $cost'_\alpha$  for TASO,  $cost_\beta$  and  $cost'_\beta$  for Ours. In the fact,  $cost_\alpha$  can hardly be exactly equal to  $cost_\beta$ , so we first calculate a certain ratio of our cost optimization,  $cost'_\beta/cost_\beta$ , then this ratio will times  $cost_\alpha$  as estimated optimized cost value. We use  $cost_\gamma$  as equation (5) not  $cost'_\beta$  as optimized cost result because this can be compared with TASO under the same cost so that the comparison is more intuitionistic and more convincing.

$$cost_\gamma = cost_\alpha \times \frac{cost'_\beta}{cost_\beta} \quad (5)$$

According to Table III(a), heuristic search has positive optimization on all seven DNNs with speedup ranging from 1.0 $\times$  to 4.0 $\times$ . Compared to TASO, we get better performance on ResNext-50, NasRNN, Squeezenet and BERT with 1.264 $\times$ , 1.057 $\times$ , 1.153 $\times$ , 2.353 $\times$  speedup but failed on ResNet-50, Inception-v3, and NasNet-A with



**Figure 2: Search time: TASO v.s. HeuSO.** The above graph is the total search time comparison between TASO and HeuSO, and the below graph is the time spent searching for the best cost in the whole search process. The number on the bar graph indicates the acceleration times of HeuSO compared to TASO, if not, it means no acceleration effect.

**Table 2: the cost and runtime(ms) of DNNs before and after optimization**

(a) cost

DNN	Original	TASO	Ours
ResNet-50	9.347	<b>9.306</b>	9.321
ResNext-50	10.065	7.789	<b>6.161</b>
Inception-v3	88.495	<b>81.197</b>	83.415
NasRNN	4.114	1.376	<b>1.302</b>
NasNet-A	30.352	<b>23.647</b>	24.905
SqueezeNet	1.806	1.339	<b>1.161</b>
BERT	24.699	14.200	<b>6.036</b>

(b) runtime(ms)

DNN	Original	TASO	Ours
ResNet-50	6.415	<b>6.269</b>	6.290
ResNext-50	11.173	8.763	<b>8.629</b>
Inception-v3	90.544	<b>86.635</b>	86.976
NasRNN	1.956	1.909	<b>1.839</b>
NasNet-A	22.802	<b>21.020</b>	21.440
SqueezeNet	1.723	1.793	<b>1.709</b>
BERT	16.648	14.391	<b>13.549</b>

a difference of 1.6%, 2.7% and 5.3%, which are within the normal error range and acceptable. For ResNet-50 and NasRNN, We achieved the same optimization effect whether the optimized cost is better or worse, and consider it as a reasonable error. We check the model with the ONNX format exported after optimization and found there is no difference between the two optimized models.

Exactly the same method as the cost calculation, assuming that the runtime of the graph  $\mathcal{G}$  before and after TASO optimization is  $runtime_{\alpha}$ ,  $runtime'_{\alpha}$  and ours is  $runtime_{\beta}$ ,  $runtime'_{\beta}$ . We also use  $runtime_{\gamma}$  as equation (6) not  $runtime'_{\beta}$  as optimized runtime result.

$$runtime_{\gamma} = runtime_{\alpha} \times \frac{runtime'_{\beta}}{runtime_{\beta}} \quad (6)$$

According to Table III(b), we achieved the same runtime performance as TASO did, and the results do not fluctuate more than 1%. This proves that in terms of runtime, heuristic search is no less efficient than TASO

### 3.3 Evaluation on Search Time

We further test the search time of heuristics after verifying its effectiveness. Figure 2 shows the total search time and the search time to best cost among whole graph substitutions of the algorithm on

seven DNNs. For total search time, HeuSO outperforms TASO with a speedup ranging from 1.12 $\times$  to 1.58 $\times$  in all deep learning models. HeuSO achieved better results on the search performance due to the pruning algorithm, which takes advantage of the extracted high-level abstractions and avoids many unnecessary searches.

As for the search time of best, HeuSO improves the search performance over TASO, speeding up by 1.34 $\times$ , 1.33 $\times$ , 1.74 $\times$ , 1.12 $\times$ , 1.92 $\times$  on ResNet-50, ResNext-50, Inception-v3, NasRNN, BERT, respectively. But HeuSO failed to get improvement on NasNet-A and Squeezenet. For NasNet-A, HeuSO spends 5 seconds more but saves 21 seconds in total search time which still has better performance than TASO. Same reason for Squeezenet and actually HeuSO discovering better graphs with lower cost than TASO. In general, HeuSO is able to find better-optimized graphs faster than TASO on most models, which indicates that HeuSO is able to find a better sequence of graph substitutions with the help of the heuristic function.

## 4 CONCLUSION

HeuSO is able to find the optimized graph with better performance and accelerate the search process. HeuSO generates an optimized graph by applying a sequence of graph substitutions with heuristics. HeuSO leverages the high abstract features of a computational graph and classifies operators into four types. Mapped operator types not only benefit heuristic function but prompt pruning. The heuristic function helps HeuSO to find a better candidate sequence that can generate an optimized graph better than the-state-of-art framework with 2.35 $\times$  speedup and the pruning algorithm will greatly reduce the search time by up to 1.58 $\times$ .

## REFERENCES

- [1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, Kimberly Keeton and Timothy Roscoe (Eds.). USENIX Association, 265–283.
- [2] Tianqi Chen, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Q. Yan, Haichen Shen, Meghan Cowan, Leyuan Wang, Yuwei Hu, Luis Ceze, Carlos Guestrin, and Arvind Krishnamurthy. 2018. TVM: An Automated End-to-End Optimizing Compiler for Deep Learning. In *13th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2018, Carlsbad, CA, USA, October 8-10, 2018*, Andrea C. Arpaci-Dusseau and Geoff Voelker (Eds.). USENIX Association, 578–594.
- [3] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. 2022. *Introduction to algorithms*. MIT press.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [5] Shi Dong, Ping Wang, and Khushnood Abbas. 2021. A survey on deep learning and its applications. *Comput. Sci. Rev.* 40 (2021), 100379. <https://doi.org/10.1016/j.cosrev.2021.100379>
- [6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [7] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size. *CoRR abs/1602.07360* (2016). [arXiv:1602.07360](https://arxiv.org/abs/1602.07360)
- [8] Zhihao Jia, Oded Padon, James Thomas, Todd Warszawski, Matei Zaharia, and Alex Aiken. 2019. TASO: optimizing deep learning computation with automatic generation of graph substitutions. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles, SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019*, Tim Brecht and Carey Williamson (Eds.). ACM, 47–62. <https://doi.org/10.1145/3341301.3359630>
- [9] Zhihao Jia, James Thomas, Todd Warszawski, Mingyu Gao, Matei Zaharia, and Alex Aiken. 2019. Optimizing DNN Computation with Relaxed Graph Substitutions. In *Proceedings of Machine Learning and Systems 2019, MLSys 2019, Stanford, CA, USA, March 31 - April 2, 2019*, Ameet Talwalkar, Virginia Smith, and Matei Zaharia (Eds.). mlsys.org.
- [10] Wei Niu, Jiexiong Guan, Yanzhi Wang, Gagan Agrawal, and Bin Ren. 2021. DNNFusion: accelerating deep neural networks execution with advanced operator fusion. In *PLDI '21: 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation, Virtual Event, Canada, June 20-25, 2021*, Stephen N. Freund and Eran Yahav (Eds.). ACM, 883–898. <https://doi.org/10.1145/3453483.3454083>
- [11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8024–8035.
- [12] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- [13] TensorRT. 2018. NVIDIA TensorRT: Programmable inference accelerator. <https://developer.nvidia.com/tensorrt>
- [14] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.
- [15] Barret Zoph and Quoc V. Le. 2017. Neural Architecture Search with Reinforcement Learning. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- [16] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. 2018. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8697–8710. <https://doi.org/10.1109/CVPR.2018.00907>

# Discrimination of seismic and non-seismic signal using SCOUTER

Kang Wang

School of Earth and Space Sciences,  
University of Science and Technology  
of China, Hefei, Anhui, P. R. China  
wk1997@mail.ustc.edu.cn

Ji Zhang

School of Earth and Space Sciences,  
University of Science and Technology  
of China, Hefei, Anhui, P. R. China  
lolitazj@mail.ustc.edu.cn

Jie Zhang

School of Earth and Space Sciences,  
University of Science and Technology  
of China, Hefei, Anhui, P. R. China  
jzhang25@ustc.edu.cn

## ABSTRACT

**Abstract**— For areas with potential occurrence of blasting events, it is essential to distinguish them from natural earthquakes. An efficient processing method is needed to save manpower, especially under the current large amount of data records by seismic stations. We apply a SCOUTER algorithm to distinguish between the two types of events. The recognition precision of the trained model for natural earthquakes and blasts can reach 95% and 92.8%, respectively, and the recall can reach 93.4% and 94.6%, respectively. The testing results of data with different epicentral distances and SNR show that our method is stable, independent on regional waveform characteristics and insensitive to data of different SNR. The explanations for each classification at the final confidence also give us a profound enlightenment.

## CCS CONCEPTS

• **General and reference** → Cross-computing tools and techniques; Performance.

## KEYWORDS

quarry blasts, natural earthquakes, SCOUTER, seismology

### ACM Reference Format:

Kang Wang, Ji Zhang, and Jie Zhang. 2023. Discrimination of seismic and non-seismic signal using SCOUTER. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590045>

## 1 INTRODUCTION

With the rise of human activities such as city construction, industrial development and production, bomb test and so on, more and more non-seismic signals (some caused by blasting events) are recorded by seismic stations. These non-seismic events are generally low in amplitude and are sometimes recorded as small magnitude earthquakes, adding false information to the earthquake catalog. An earthquake catalog always plays a decisive role in the study of seismic activity and the upcoming hazard analysis. A clean earthquake catalog is a prerequisite for reliable seismological research. Therefore, it is necessary to distinguish blasting events from natural earthquakes.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590045>

Natural earthquakes and blasts may show similar waveforms recorded at the seismic stations, and expert experience is needed to distinguish them. Several automated methods have been developed to replace the cumbersome manual identification tasks. Such as the maximum amplitude ratio of different phase waves in the time domain (Hartse et al., 1997; Horasan et al., 2009; Tibi et al., 2018; Wang et al., 2020;), the spectral ratios in frequency or time-frequency domain (Allmann et al., 2008; Bennett and Murphy, 1986; Yilmaz et al. 2013;), the waveform complexity (Gitterman and Shapira, 1993; Badawy et al., 2019) and so on. Several algorithms are also widely used in the identification of natural earthquake and blasts, such as wavelet transform (Beccar-Varela et al., 2016), discrete wavelet transform (Gendron et al., 2000), genetic algorithm (Orlic and Loncaric, 2010). For blasting events, their vibration often corresponds to small events, thus, it is difficult to extract features such as seismic phase or waveform, which leads to certain limitations of these methods.

The development of artificial intelligence is advancing technologies in many fields including seismology (Perol et al., 2018). In terms of applications in earthquake classification, Rabin et al. use a graph-based machine learning tool combined with diffusion maps to distinguish between earthquakes and explosions. The method based on support vector machine are also applied in this aspect (Saad et al. 2019, Kim et al., 2020). Convolutional neural networks are most widely used to distinguish between seismic and non-seismic events (Linville et al., 2019; Miao, et al., 2020; Kong et al., 2022; Tian et al., 2022). Their main distinction is whether the input object is time domain information, frequency domain information or a combination of them. These methods can only tell us the result of the final classification, but lack the explanation of the classification process. Existing interpretable methods are mostly based on gradient changes or feature maps in the middle layer of the network, but they do not directly participate in the decision-making process of classifiers. In this study, we use a slot attention-based classifier SCOUTER (Li et al., 2021) to classify natural earthquakes and blasts. We apply the method to a dataset collected in Southern California, and the results show that the method can effectively distinguish blasts from natural earthquakes. Meanwhile, the analysis of the visual explanation maps may bring some enlightenment to future researches.

## 2 METHODOLOGY

### 2.1 SCOUTER network structure

Slot attention can extract object-centric representations and generalize to invisible combinations when trained on unsupervised object discovery and supervised attribute prediction tasks (Locatello et al., 2020). We adopt SCOUTER (Li et al., 2021), a classifier based

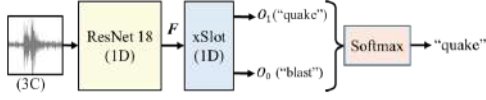


Figure 1: The network structure based on SCOUTER.

on slot attention, which involves the explanation to each category in the final confidence. Here we modify the workflow and apply it to one-dimensional seismic data (single station). The structure of the model used in our study is shown in Figure 1. The input is three-component (3C) seismic waveform data (quake waveform or blast waveform), and the output is the corresponding label 0 (blast) or 1 (quake). ResNet18 (He et al., 2016) is one of the networks commonly used in the field of image recognition and has a strong feature extraction ability. The main idea of ResNet18 is to transfer shallow information to deep layers, which can improve the learning ability of the network. We modify ResNet18 to apply to feature extraction of one-dimensional seismic data, allowing the network to extract features in time series. The main training strategies are as follows: 1. the ResNet18 model is pre-trained for a total of 50 epochs to obtain a powerful feature extraction network. 2. Using the xSlot module to obtain an explicable and visualized model.

We set the convolution kernel size to  $3 \times 1$ , and the size of down-sampling is set to (2, 1) in ResNet18. The modified ResNet18 is used to extract the features  $F \in R^{c \times d \times 1}$  of the input data. In the slot attention mechanism, a slot is a representation of a local region of attention aggregation based on a feature map. For  $n$  classification problems, slot attention can obtain  $n$  features as outputs. The xSlot module is based on the slot attention mechanism to establish the relationship between each slot and category. Each slot gives the confidence that the input waveform data falls into the category. We denote the slots for all categories by  $W \in R^{n \times c'}$ ,  $c'$  is the size of the weight vector.  $F$  goes through the convolutional layer and the ReLU layer in xSlot to get  $F' \in R^{c' \times d}$ .  $\tilde{F}$  is augmented by adding the position embedding  $PE$ .

$$\tilde{F} = F' + PE$$

Two multilayer perceptrons  $Q$  and  $K$  are used to deal with  $W$  and  $\tilde{F}$ . We obtain the attention  $A \in R^{n \times d}$  using sigmoid  $\sigma$  as

$$A = \sigma(Q(W)K(\tilde{F}))$$

$$U = AF'^T$$

For the traditional recognized network, the extracted features need to be classified based on the full connection (FC) layer. The output of xSlot module is directly used as the confidence value of each category, so the commonly used classifier based on the FC layer is no longer needed. The output  $O$  of the xSlot attention module is the sum of all elements for each category in  $U$ .

$$O = e \cdot U 1_{c'}$$

$e$  is a hyper-parameter that selects the positive and negative explanation. We consider the positive explanation here.

The cross-entropy loss can be used to classify the classification problem of quake events and blast events. To make each category relate to the xSlot feature, we use a new loss function:

$L_{scouter} = L_{ce} + a * L_{area}$  where the  $L_{ce}$  is commonly used in the neural network classification problem of the cross-entropy loss

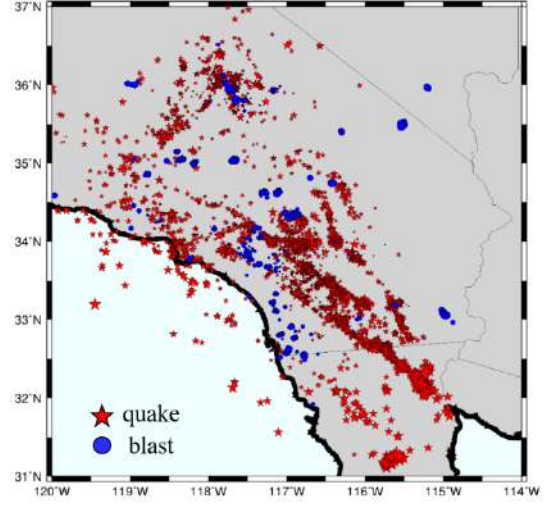


Figure 2: The red pentagrams and blue dots represent the location of the natural earthquake and blasts, respectively.

function,  $L_{area}$  is the area loss function,  $a$  is a super parameter, used to adjust the size of area loss weight.

$$L_{area} = 1_n^T A 1_d$$

With the softmax cross-entropy loss, the model learns to give the largest confidence  $O_i$  corresponding to ground-truth category  $i \in (0, 1)$  and a smaller value  $O_{i'}$  to wrong category ( $i' \neq i$ ). Both  $A$  and  $F'^T$  are non-negative, some elements in the  $i$  row vector in  $A$  will be close to 1 corresponding to the ground-truth category, whereas a smaller is given when all elements in  $i'$  row vector in  $A$  are close to 0.

## 2.2 Data and training

The Southern California Earthquake Data Center (SCEDC) provides a comprehensive data-sharing platform from which researchers can download raw waveforms recorded by Southern California Seismic Network (SCSN), as well as high-quality seismic catalog information. A great deal of previous work including the classification of the types of quakes, such as quarry blasts or natural earthquakes, is well listed in the catalog. We preprocess the raw data and obtain about 36,312 3C waveform records of blasting events (magnitudes range from 0.5 to 2.5) occurring in Southern California from 2011 to 2020, and about 37,000 3C waveform records of natural earthquakes (magnitudes range from 0 to 4.5) occurring in Southern California from 2011 to 2013, as shown in Figure 2. For all records, the distance between the stations and events (epicentral distance) are within 100 km and the signal to noise ratio (SNR) ranges from 10 dB to 60 dB. We allocate these data, among which the training dataset contains 27,000 blasts and 27,000 earthquakes, the validation dataset contains 3,000 blasts and 3,000 earthquakes, and the testing dataset contains 6,312 blasts and 7,000 earthquakes (See Table 1).

In the training process, we adopt the AdamW (Loshchilov and Hutter, 2017) as the optimizer with the initial learning rate of 0.0001. The input size is  $2750 \times 1$  and the category  $n$  is 2 (quake 1 or blast 0). The batch size is 8, which means that 8 waveform data are input

**Table 1: The number of samples in different datasets**

Dataset	Training	Validation	Testing
Quake	27000	3000	7000
Blast	27000	3000	6312

**Table 2: The confusion matrix on the testing dataset**

	Predicted Quake (1)	Predicted Blast (0)
True Quake (1)	TP (6539)	FN (461)
True Blast (0)	FP (344)	TN (5968)

into the model for training each time. The model is pre-trained on the training set for 50 epochs without using the xSlot module. The SCOUTER based the pre-trained model is trained on the training set for 50 epochs and the performance results are computed on the testing set with the trained model after the last epoch. All the experiments are conducted on the local GPU servers equipped with 160 Intel Xeon Platinum 8380 (@2.30GHz) CPUs, four NVIDIA A100 (40GB) GPUs.

### 3 RESULT AND DISCUSSION

#### 3.1 Testing data performance

We evaluate the performance of our trained model with confusion matrix on the testing dataset which contains records of 6,312 blasts and 7,000 earthquakes. The confusion matrix consists of four parts. True positives (TP) are samples (earthquakes) correctly predicted as positives (earthquakes); True negatives (TN) are samples (blasts) correctly predicted as negative (blasts); False positive (FP) are samples (blasts) incorrectly predicted as positives (earthquakes); False negative (FN) are samples (earthquakes) incorrectly predicted as negative (blasts). The statistics of the predicted results are shown in Table 2. Precision and recall are further used to evaluate the performance of the model. The testing results show that the precision of the trained model for natural earthquakes and blasts can reach 95% and 92.8%, respectively, and the recall can reach 93.4% and 94.6%, respectively.

$$Precision = \frac{TP}{TP+FP} \text{ or } = \frac{TN}{TN+FN}$$

$$Recall = \frac{TP}{TP+FN} \text{ or } = \frac{TN}{TN+FP}$$

As mentioned in previous section, before deciding the categories of the samples, we can output a visualization attention map to explain the classification basis of our model. Here we show the visualizations (A) of an earthquake and a blast before the final decision-making process in Figure 3. As we can see, for a quake sample, there are obvious bright spots (value close to 1) near the P and S waves in the visualization judged as quake; for the blast sample, the most bright spot only appears near P wave or with multiple small bright spots in the visualization judged as blast. We can conclude that our method can provide classification results and help us understand the main characteristics of earthquakes and blasts in the time series domain.

#### 3.2 Performance on different epicentral distance

The epicentral distance represents the horizontal distance between a station and an event location. Scientists have compiled this information into earthquake catalogs and waveform data headers. Generally, for events of the same magnitude, the farther away the epicentral distance is, the attenuation of the wave signal will be stronger, resulting in lower signal quality recorded by the seismic stations. We show the recall of earthquakes and blasts identification vary with epicentral distance in Figure 4. The blue histogram shows the true number of samples per epicentral distance and the orange histogram is the predicted numbers. The red pentagram represents the recalls for each data of different epicentral distance. For the blasting data, the recall is relatively stable with the increase of the epicenter distance. For earthquake data, the recall decreased slightly with the increase of epicenter distance. In general, however, a relatively high level of identification is maintained.

#### 3.3 Performance on different SNR

The SNR is also used as a means of dividing data. Some traditional classifying methods, such as using the time delay of P and S wave based on arrival time information, are greatly affected by SNR, because it is impossible to accurately pick the phase arrival time of the waveform. Here we also show the performance of our model on different SNR data in Figure 5. The blue histogram shows the true number of data samples per SNR and the orange histogram is the predicted counts. The red pentagram represents the recalls for each data of different SNR. The results show that our model prediction is less affected by the SNR.

### 4 CONCLUSION

In this study, we combine an advanced classifier SCOUTER to achieve an automatic and effective discrimination between natural earthquakes and blasts. We implement our approach based on event waveforms recorded by the SCSN. The precision for natural earthquakes and blasts can reach 95% and 92.8%, respectively, and the recall can reach 93.4% and 94.6%, respectively. The testing results of data with different epicentral distances and SNR show that our method is stable, independent on regional waveform characteristics and insensitive to data of different SNR. Our method can capture the characteristics of event signals, and visually display them to explain the underlying basis of classification. This information is very important for researchers to conduct subsequent theoretical analysis. The blasting events appear to occur repeatedly in the waveform for a short duration time, and the amplitude of P wave is generally larger than that of S wave. However, the earthquakes often have both P and S phase characteristics. In future studies, multiple types of information can be taken into account in the feature extraction process to increase the generalization ability of the model when applied in different monitoring areas.

### REFERENCES

- [1] Allmann, B. P., Shearer, P. M., & Hauksson, E. (2008). Spectral discrimination between quarry blasts and earthquakes in southern California. Bulletin of the Seismological Society of America, 98(4), 2073-2079.

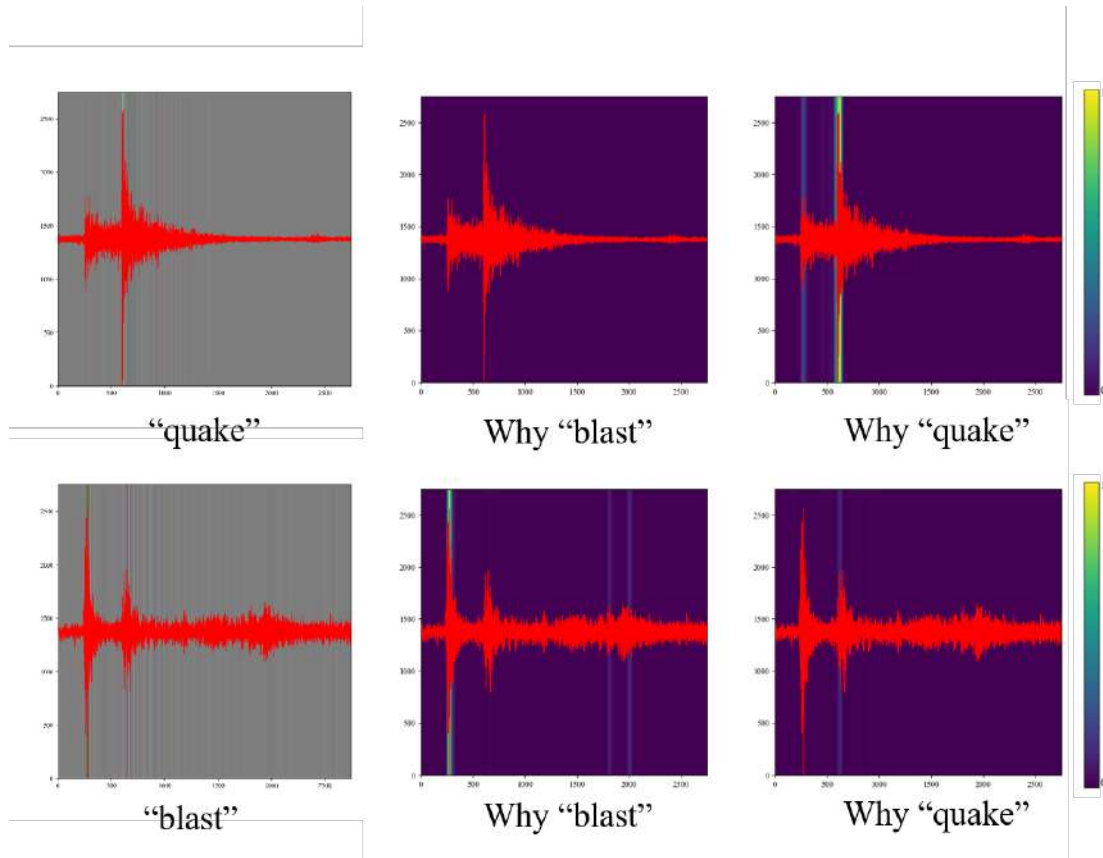


Figure 3: The attention maps of two testing examples (an earthquake and a blasts). The red solid lines represent the waveform of the vertical component of the original data.

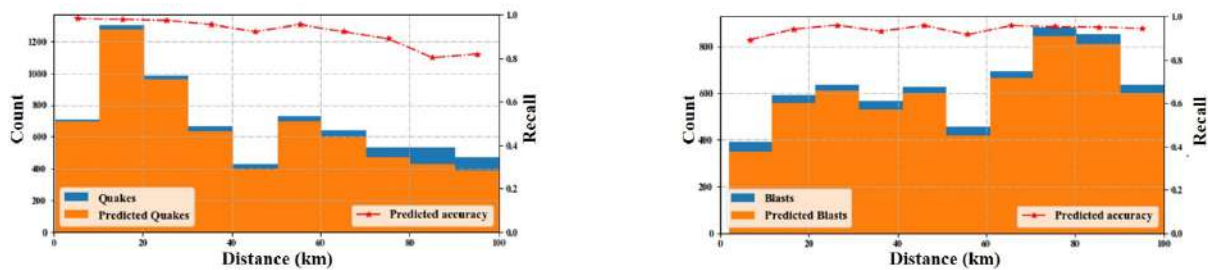


Figure 4: The performance of the model varies with epicentral distance on earthquake data (left) and blast (right) data.

- [2] Badawy, A., Gamal, M., Farid, W., & Soliman, M. S. (2019). Decontamination of earthquake catalog from quarry blast events in northern Egypt. *Journal of Seismology*, 23, 1357-1372.
- [3] Beccar-Varela, Maria P., et al. "Use of wavelets techniques to discriminate between explosions and natural earthquakes." *Physica A: Statistical Mechanics and its Applications* 457 (2016): 42-51.
- [4] Bennett, T. J., & Murphy, J. R. (1986). Analysis of seismic discrimination capabilities using regional data from western United States events. *Bulletin of the Seismological Society of America*, 76(4), 1069-1086.
- [5] Gendron, Paul, John Ebel, and Dimitris Manolakis. "Rapid joint detection and classification with wavelet bases via Bayes theorem." *Bulletin of the Seismological Society of America* 90.3 (2000): 764-774.
- [6] Gitterman, Y., & Shapira, A. (1993). Spectral discrimination of underwater explosions. *Israel Journal of Earth-Sciences*, 42(1), 37-44.
- [7] Hartse, H. E., Taylor, S. R., Phillips, W. S., & Randall, G. E. (1997). A preliminary study of regional seismic discrimination in central Asia with emphasis on western China. *Bulletin of the Seismological Society of America*, 87(3), 551-568.
- [8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
- [9] Horasan, G., Güney, A. B., Küsmezer, A., Bekler, F., Ögütçü, Z., & Musaoğlu, N. (2009). Contamination of seismicity catalogs by quarry blasts: an example from Istanbul and its vicinity, northwestern Turkey. *Journal of Asian Earth Sciences*, 34(1), 90-99.
- [10] Kim, Sangkyeum, Kyunghyun Lee, and Kwanho You. "Seismic discrimination between earthquakes and explosions using support vector machine." *Sensors* 20.7 (2020): 1879.

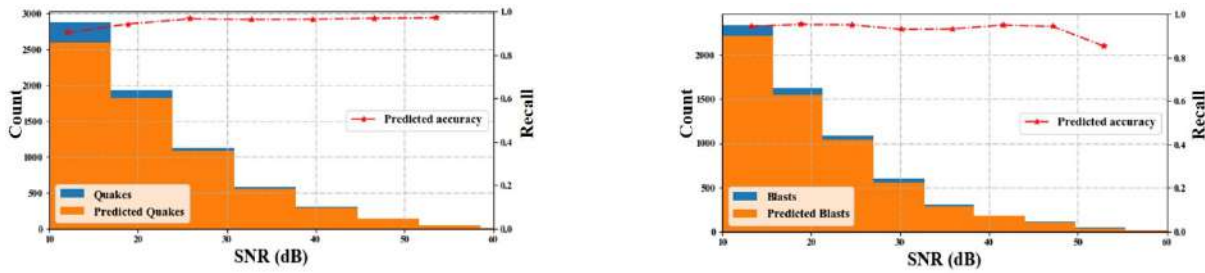


Figure 5: The performance of the model varies with SNR on earthquake data (left) and blast (right) data.

- [11] Kong, Qingkai, *et al.* "Combining Deep Learning With Physics Based Features in Explosion-Earthquake Discrimination." *Geophysical Research Letters* 49.13 (2022): e2022GL098645.
- [12] Li, L., Wang, B., Verma, M., Nakashima, Y., Kawasaki, R., & Nagahara, H. (2021). SCOUTER: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1046-1055).
- [13] Linville, Lisa, Kristine Pankow, and Timothy Draelos. "Deep learning models augment analyst decisions for event discrimination." *Geophysical Research Letters* 46.7 (2019): 3643-3651.
- [14] Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J. & Kipf, T. (2020). Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33, 11525-11538.
- [15] Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- [16] Miao, Fajun, *et al.* "High-accuracy discrimination of blasts and earthquakes using neural networks with multiwindow spectral data." *Seismological Research Letters* 91.3 (2020): 1646-1659.
- [17] Orlic, Niksa, and Sven Loncaric. "Earthquake–explosion discrimination using genetic algorithm-based boosting approach." *Computers & geosciences* 36.2 (2010): 179-185.
- [18] Perol, Thibaut, Michaël Gharbi, and Marine Denolle. "Convolutional neural network for earthquake detection and location." *Science Advances* 4.2 (2018): e1700578.
- [19] Saad, Omar M., Ahmed Shalaby, and Mohammed S. Sayed. "Automatic discrimination of earthquakes and quarry blasts using wavelet filter bank and support vector machine." *Journal of Seismology* 23 (2019): 357-371.
- [20] TIAN, X., WANG, M., ZHANG, X., WANG, X., SHENG, S., & LÜ, J. (2022). Discrimination of earthquake and quarry blast based on multi-input convolutional neural network. *Chinese Journal of Geophysics*, 65(5), 1802-1812.
- [21] Tibi, R., Koper, K. D., Pankow, K. L., & Young, C. J. (2018). Depth Discrimination Using Rg-to-Sg Spectral Amplitude Ratios for Seismic Events in Utah Recorded at Local Distances. *Depth Discrimination Using Rg-to-Sg Spectral Amplitude Ratios for Seismic Events in Utah*. *Bulletin of the Seismological Society of America*, 108(3A), 1355-1368.
- [22] Wang, R., Schmandt, B., & Kiser, E. (2020). Seismic discrimination of controlled explosions and earthquakes near Mount St. Helens using P/S ratios. *Journal of Geophysical Research: Solid Earth*, 125(10), e2020JB020338.
- [23] Yılmaz, Ş., Bayrak, Y., & Çınar, H. (2013). Discrimination of earthquakes and quarry blasts in the eastern Black Sea region of Turkey. *Journal of seismology*, 17, 721-734.

# ENOSE Performance in Transient Time and Steady State Area of Gas Sensor Response for Ammonia Gas: Comparison and Study

Geng kuan

Henan International Joint Laboratory  
Laser Technology in Agriculture  
Sciences, College of Mechanical &  
Electrical Engineering, Henan  
Agricultural University  
kuankuan0205@163.com

Ata Jahangir Moshayedi

School of Information Engineering,  
Jiangxi University of Science and  
Technology  
ajm@jxust.edu.cn

Chen Jing

College of Mechanical & Electrical  
Engineering, Henan Agricultural  
University  
chenbaobaogui@126.com

Hu Jiandong\*

Henan International Joint Laboratory  
Laser Technology in Agriculture  
Sciences, College of Mechanical &  
Electrical Engineering, Henan  
Agricultural University  
jdhu@henau.edu.cn

Zhang Hao

Henan International Joint Laboratory  
Laser Technology in Agriculture  
Sciences, College of Mechanical &  
Electrical Engineering Henan  
Agricultural University  
hao.zhang@henau.edu.cn

## ABSTRACT

This paper proposed an electronic nose system that utilized a SnO<sub>2</sub> semiconductor sensor array to detect volatile ammonia gas in farmland. All sensors were controlled by the Arduino development board. The system could collect data during both the steady-state and transient phases of sensor operation. The collected data was analyzed using PCA (principal component analysis) and MLP (Multi-layer perceptron) neural networks. The experiment was divided into two parts: The first part analyzed four concentrations of ammonia (100ppm, 200ppm, 400ppm, and Air) using PCA and MLP, which successfully distinguished the concentrations with an identification rate of over 95%. In the second part, four gases (air mixed with ammonia, pure ammonia gas, air mixed with ethanol, and pure ethanol) were analyzed using PCA and MLP, with the electronic nose system successfully distinguishing between the four types of gases. The system could read and process data during the transient phase of the sensor, and the constructed sensor array electronic nose system and acquisition method has significant potential for ammonia detection in agricultural environments.

## CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural networks**;

\* Corresponding author. Department of Electrical Engineering, Henan Agricultural University, Zhengzhou, 450002, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590046>

## KEYWORDS

Electronic nose, Volatile ammonia, PCA, MLP, Neural network

## ACM Reference Format:

Geng kuan, Ata Jahangir Moshayedi, Chen Jing, Hu Jiandong, and Zhang Hao. 2023. ENOSE Performance in Transient Time and Steady State Area of Gas Sensor Response for Ammonia Gas: Comparison and Study. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590046>

## 1 INTRODUCTION

Ammonia (NH<sub>3</sub>) is a colorless gas known for its pungent odour, which poses a serious threat to human and animal health and the environment. When exposed to moist mucous surfaces such as those found in the respiratory tract, skin, and eyes, ammonia reacts to form a corrosive alkaline solution (ammonium hydroxide), leading to liquefaction necrosis [1]. Inhaling ammonia can cause damage to airways and, in extreme cases, even result in acute death [2]. In pigs, ammonia exposure can trigger lung damage and inflammation, which can adversely affect meat quality [3]. Furthermore, ammonia emissions have been linked to an increase in PM<sub>2.5</sub> concentrations in the air [4], a type of atmospheric particulate matter that poses significant health risks, including the potential for severe illness and death. With modern industrial advancements, producing ammonia has become more cost-effective. Ammonia is widely used as a nitrogen fertilizer in agriculture, but the use of such fertilizers can lead to increased ammonia emissions into the atmosphere [5]. Research from years ago suggests that the majority of atmospheric ammonia is attributed to agriculture, accounting for 55% of emissions. In China, 29% of ammonia emissions come from livestock and 47% from agricultural fertilizer usage [6]. Therefore, detecting volatile ammonia levels in farmland is critical for managing and mitigating environmental pollution caused by agriculture. The ammonia can be detected over various methods as follows:

**The electrochemical method** involves an ammonia sensor capable of detecting gas concentrations as low as 1 ppm. Different materials adsorb  $\text{NH}_3$ , leading to changes in electrical signals such as resistance, voltage, and current. These alterations allow for the calculation of ammonia concentration [7, 8].

**The chemiluminescence method** offers a detection range of 0.25–100ppm for ammonia gas at the ppb level, making it suitable for field detection. Notably, this method boasts a fast response speed. Hu, et al. [9] have devised a catalytic luminescent gas sensor that accurately measures the ammonia concentration by analyzing the light emitted through the reaction between the gas and the catalyst surface [7]. **The passive collector method** involving the collection of gas via a membrane can accurately measure ammonia volatilization levels within a range of 0.2 ppb to 100 ppb for up to a month [7]. However, it has a drawback in that it necessitates frequent manual replacement of collection materials, resulting in excessive consumption of manpower and material resources in monitoring large areas. Additionally, the method does not provide real-time measurements of ammonia concentration, as the collected materials must be analyzed to determine the concentration of ammonia. **The photoacoustic method** is an accurate way to detect ammonia concentrations ranging from 0.3–10ppm in a laboratory setting. Laser radiation, modulated in either frequency or amplitude, is absorbed in wavelengths that align with the absorption characteristics of the target substance. This produces sound waves that can be monitored with minimal noise [10]. Though the photoacoustic method is resistant to environmental interference, it necessitates expensive and cumbersome optical equipment, which renders it unsuitable for large-scale ammonia detection in agriculture. **The fluorescence method** can detect ammonia ranging from 0.5–50ppm in the laboratory. Zhang and Lim [11] created a colorimetric array comprising  $4 \times 4$  dyes using fluorescence technology to analyze the medium's color before and after exposure to gas and determine ammonia concentration. However, this method is time-consuming and unsuitable for detecting ammonia concentration in agricultural environments. But over than mentioned methods, enose over its benefit attracts more researchers. Electronic noses have several benefits for gas ammonia detection, including their ability to detect low levels of ammonia, their fast response time, and their cost-effectiveness compared to traditional methods of ammonia detection [12]. To analyze the enose data various methods were implemented but the most common can be named PCA, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) [13], Partial Least Squares (PLS), and Fuzzy Logic [12]. This paper proposes an electronic nose sensor array system for the detection of volatile ammonia in farmland. The system employs PCA and MLP to differentiate between various gases and varying concentrations of ammonia. The research contribution can be listed below: 1) Proposed an electronic nose system that utilizes a  $\text{SnO}_2$  semiconductor sensor array for detecting volatile ammonia gas in various concentrations. 2) Collected data during both the steady-state and transient time of sensor operation and show the transient time area ability for gas concentration detection. 3) Analyzed the collected data using PCA and MLP neural networks and successfully distinguished between four concentrations of ammonia (100ppm, 200ppm, 400ppm, and Air) with an identification rate of over 95% using PCA and MLP. 4) Distinguished between four types of gases,

including Air mixed with ammonia and pure ammonia gas, using PCA and MLP. 5) Demonstrated the system's ability to read and process data during the transient area of the sensor. The paper is organized as follows: The first section covers the gas sensor array, chamber structure, and data acquisition, while section III delves into the data processing methodology. In section IV, the paper concludes by analyzing and contrasting the results obtained in two modes: identifying different concentrations of ammonia and distinguishing between various odours.

## 2 MATERIALS AND METHODS

### 2.1 Gas Sensor Array

To experiment, a sensor array was created using seven different models of TGS and MQ semiconductor sensors as shown in Table 1. All sensors are controlled by the Arduino-Mega development board, Arduino is easier to control and use than other microcontrollers and can monitor sensor data directly from the IDE (Integrated development environment).

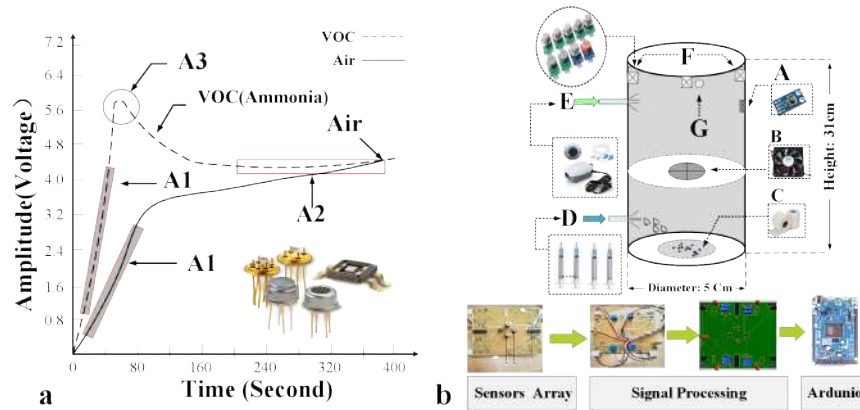
As it is shown in Table 1 the sensor array can sense the various PPM amount of VOC. In comparison to other types of gas sensors, the  $\text{SnO}_2$  metal oxide semiconductor gas sensor offers low production cost, a simple manufacturing process, and high sensitivity [14]. However, the MQ series sensors are sensitive to changes in moisture, temperature, and humidity, which can affect sensor response. To maintain consistency in the test results, the experiment strictly controlled environmental temperature and humidity.

### 2.2 Chamber structure and Data Acquisition

The data acquisition system utilized in this research is the system proposed in our previous paper [15]. This system incorporates a sensor array consisting of the gas sensors listed in Table 1. To measure temperature, the LM35 sensor is utilized, while the SHT11 sensor is employed for measuring temperature and humidity. In addition to these sensors, the system also incorporates an Aquarium Air pump, power supply, and DC fan, as shown in Figure 1 (a). As the figure depicted, the sampling chamber is comprised of a 31cm high and 5cm diameter cylinder, fitted with a DC fan at one end and an electronic PCB containing a sensor array at the other. A syringe for sample and air input is connected to the wall of the chamber, with a paper towel substrate placed at the bottom to collect ammonia liquid samples and allow for evaporation. The ambient temperature and relative humidity are detected by an SHT11 sensor and LM35. Then in order to volatilize ammonia gas in the sampling room, ammonia with a known concentration is injected into the sample syringe and then into the tissue in the sampling chamber, causing the liquid to evaporate and form ammonia gas. The sensors data in the next step sensor data acquired with the help of Arduino Mega and stored in a Personal computer (PC) followed to have some advantages like flexibility, low cost, ability to connect and communicate with external parts such as sensors, etc monitors, no need for additional software or other compilers with a simple software environment. The proposed enose assembly utilizes an aquarium pump to deliver fresh air and maintain a clean sampling chamber. The pump has a compact size of  $98 \times 66 \times 20$ mm and requires only 4W of power. Additionally, a DC electric fan with dimensions of  $10 \times 10$ cm and a voltage of 12v has been chosen to

**Table 1: The sensor parameters used in the experiment**

Sensor	Main sensing gas	Measurement range
MQ2	LPG, propane, methane, hydrogen, carbon monoxide, and alcohol	300-10000ppm
MQ3	Ethanol Vapor	10-1000ppm
MQ5	LPG, propane, hydrogen, methane, and other combustible gases	300-10000ppm
MQ6	LPG, butane, propane, methane, alcohol, hydrogen, and smoke	300-10000ppm
MQ7	Carbon Monoxide	10-1000ppm
MQ137	Ammonia	5-200 ppm.
TGS813	Methane, Propane, Butane	500-10000ppm



**Figure 1: a: The second-order response curve of the sensor (TGS and MQ gas sensors) A1: The transient ascending stage A2: The overshoot stage A3: The steady-state stage; b: Electronic nose sensor array data sampling system A: Temperature/Humidity sensor (SHT11) B: DC fan C: Tissue paper. D: Injection syringe E: Inlet fresh air by Aquarium pump**

expedite the cleaning process of the chamber. This fan will further enhance the efficiency of the enose assembly [16]. In order to calculate the ammonia concentration Equation 1) proposed by Wang, et al. [17] was used.

$$C = \frac{22.4\rho TV_s}{273MV} \times 1000 \quad (1)$$

Equation 1) relates the concentration of ammonia in parts per million (ppm), denoted by C, to the density of ammonia ( $\rho$ , measured in  $\text{g}\cdot\text{mL}^{-1}$ ), the temperature inside the experimental vessel (T, in Kelvin), and the volume of the sampling chamber (V, in liter). The equation also incorporates the volume of liquid ammonia ( $V_s$ , measured in microliters) and the molecular weight of ammonia (M, in  $\text{g}\cdot\text{mol}^{-1}$ ).

## 2.3 Data Processing method

As mentioned [15, 18, 19], the enose sensor response can be divided into three stages: the transient ascending stage, the overshoot stage, and the steady-state stage. While the transient ascent stage contains more information and data about the sensor response than the steady-state stage [20, 21], Figure 1 (a). displays the second-order response curve of the TGS and MQ gas sensors. In this experiment, both the data from the transient ascending stage and the steady-state stage were collected, processed, and analyzed as part of the dataset. Then, after using the sensor array to extract the gas

characteristic information, MATLAB software is utilized for data processing, incorporating PCA and the construction of an ANN for enose data analysis in the Transient rise time area(A1) as well as steady-state(A2). These mentioned areas along with Both methods are investigated for different concentrations in terms of PPM and different odours.

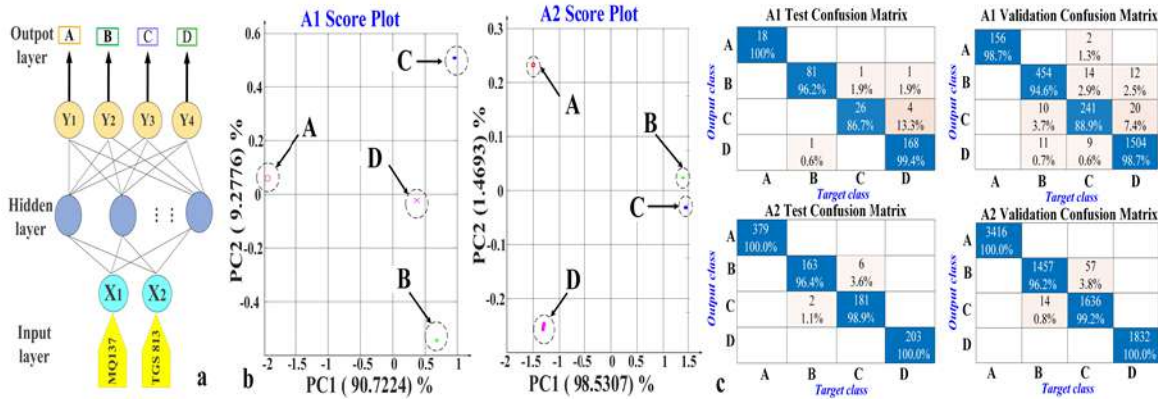
**2.3.1 Principal component analysis (PCA).** PCA is a frequently utilized statistical method that can reduce the dimension of data and extract its characteristic information. With an increase in the number of sensors, the characteristic dimension of gas also increases [22].

**2.3.2 Multilayer Perceptron (MLP).** MLP is a type of Artificial Neural Network (ANN) that operates similarly to human neurons. It consists of an input layer, an output layer, and one or more hidden layers, with multiple neurons in each layer. The Figure 2 (a) shows a multilayer perceptron network topology assign for ammonia recognition at different concentrations using MQ37 and TGS813.

Table 2 illustrates that in mode 1, both areas have an input layer of 2, while the input layer for the second mode (M2) is set to 6. The output layer for all modes is set to 4, with 25 hidden layer neurons assigned. The proposed MLP structure utilizes the Rectified Linear Unit (ReLU) as the activation function (Equation 2), which enables MLP to perform nonlinear prediction and classification. This makes

**Table 2: THE MLP Network Setting Over Different Mode, M1: Different Concentration (PPM), M2: Different Odor;**

Mode	Area	Input layer(Neurons number)	Output layer(Neurons number)	Hidden layer(Neurons number)
M1	A1	2	4	25
	A2	2	4	25
M2	A1	6	4	25
	A2	6	4	25

**Figure 2: a: MLP network topology used; Results of four concentrations of ammonia b: PCA results and c: MLP results Ammonia concentration: A:100ppm B:200ppm C:400ppm D: AIR (0ppm)**

it a useful tool for distinguishing and predicting gas types.

$$f(x) = \max(0, x) \quad (2)$$

As Equation 2) shows if the input to the function (x) is greater than zero, then the output will be equal to x. If the input is less than or equal to zero, then the output will be zero. The ReLU function is popular in neural networks because it is computationally efficient and has been shown to produce good results in many different types of problems. Additionally, the ReLU function is easier to optimize using gradient-based methods, which are commonly used for training neural networks. One potential downside of the ReLU function is that it can "die" or "vanish" when the input to the function is negative, which can lead to problems with learning. In this paper, based on the proposed structure, the MLP networks for transient and steady-state phases are constructed, sensor data are used to train them respectively, and optimization models for different phases are obtained.

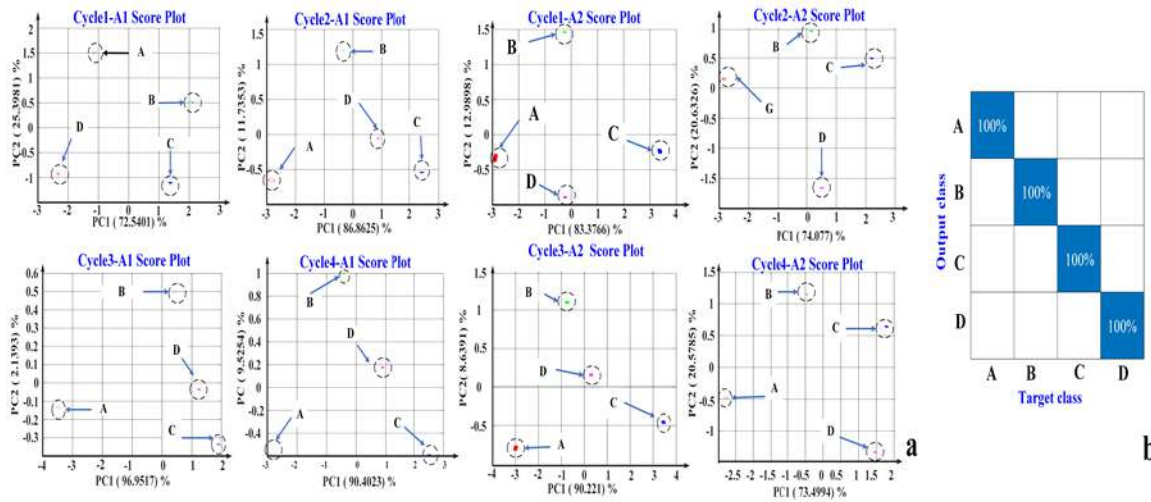
### 3 RESULT AND DISCUSSION

As mentioned before the data was collected and investigated in two areas of Transient rise time area (A1) and steady-state (A2) in two parts of Identification of different concentrations (PPM) of ammonia and odour identification which are described below.

#### 3.1 Identification of different concentrations of ammonia

In this experiment, MQ137 and TGS813 sensors were used to collect data on four different concentrations of ammonia (100ppm, 200ppm, 400ppm, and pure Air). The steady-state stage contains a total of 9,346 data of MQ137 and TGS813 sensors, forming a matrix with a size of 9,346×2. The transient phase is 2,703 pieces of data. After normalizing the data, the PCA function is used to analyze the data. Using graph-related functions, score plots can be plotted to visually show the concentrations and types of gases.

The data used for MLP training was the same as that used for PCA, 10% of the data was reserved for testing the accuracy of the model, i.e. the data set used for training was an independent part of the whole data set. The parameters of the MLP model are shown in Table 2. The number of iterations is limited to 1000. The extracted data for both methods is shown in Figure 2. As the result shown in the PCA method for both A1 and A2 (transient response or the steady-state response), the sensors can distinguish gas of different concentrations from each other without concentrations overlapping. In both areas, indicating that the sensor array has a remarkable recognition effect on different concentrations of gas the PC1 is greater than 90%. By using the MLP method the classification of the steady-state stage (A2) by neural network is good, and the accuracy of discrimination of four concentrations is more than 95%, but the classification accuracy results in the transient-state stage (A1) are not as good as that in the steady-state stage around 88.9%. Besides the results shows, the resolution accuracy at 400ppm is less



**Figure 3: The obtained results in a: The PCA score chart of the transient stage(A1) and the steady-state stage(A2), b: MLP classification confusion matrix. A: Air1 (a mixture of air and Ethanol) B:ammonia (pure ammonia gas) C: Air 2(a mixture of air and Ammonia) D: ethanol (pure ethanol gas)**

**Table 3: The odour Identification over PCA and MLP, A1: Transient rise time area A2: Steady-state stage area**

Method/ Area	Parameter		Cycle1	Cycle2	Cycle3	Cycle4
PCA-A1	PC1		72.51%	86.86%	96.95%	90.40%
PCA-A2	PC1		83.38%	74.08%	90.22%	73.50%
MLP-A1	Training	data volume	159	185	159	197
		Accuracy	100%	100%	100%	100%
	Test	data volume	17	20	17	21
		Accuracy	100%	100%	100%	100%
	Training	data volume	720	360	360	360
		Accuracy	100%	100%	100%	100%
	Test	data volume	80	40	40	40
		Accuracy	100%	100%	100%	100%

than 90%, against the resolution accuracy at other concentrations with 95%.

### 3.2 Different Odour Identification

As the second result to detect the different odours, a sensor array changed, and use the MQ2, MQ3, MQ5, MQ6, MQ7, and MQ137 sensors are used to identify four kinds of gases with gas concentrations of 200ppm. considering the odour the four types of gas with the fixed ppm values are used as follows; Air1 (a mixture of air and Ethanol), Air 2(a mixture of air and Ammonia), ethanol (pure ethanol gas), and ammonia (pure ammonia gas); A complete data acquisition cycle of the sensor includes the transient response stage and the steady-state response stage. The experiment is carried out four times with four cycles. At the end of each cycle, fans and air pumps are used to ventilate the sampling room and exhaust the sampling room. PCA and MLP were used to analyze the more than 10,000 samples of data in the experiment. 10% of the data was

reserved for testing the accuracy of the model, The obtained results are shown in Figure 3 and Table 3.

As the result shows, PCA score plots of steady-state and transient-stage under four cycles of data acquisition, all gas types are distinguished from each other in pairs, and the sensor array can distinguish different types of gas. Detailed results of PCA and MLP are shown in Table 3.

The accuracy of PCA in identifying the four gases is not as good as the accuracy in identifying the four ammonia concentrations. However, PC1 is greater than 70%, which can effectively distinguish the types of gas. The recognition accuracy of MLP is 100%, and the recognition effect is very excellent. In general, the accuracy of PCA is not as high as MLP in the identification of the four gases.

## 4 CONCLUSION

Detecting ammonia gas is critical for a variety of reasons, including protecting occupational safety by identifying this toxic gas and its impact on human health, monitoring ammonia as an environmental

pollutant, and ensuring food safety by detecting its use in refrigeration and the production of certain foods, such as baked goods and chocolate. Studying gas sensor response in both transient and steady-state regions offer numerous benefits. Figure 1 illustrates that data collected during the transient response period, which typically lasts around 20 seconds, is much faster than the 40 seconds required during the steady-state response phase. This delay allows the sensor response to stabilize, resulting in a 50% reduction in sensor performance time when operating in the transient region. This study utilized MQ and TGS series semiconductor sensors to construct an electronic nose system sensor array. By analyzing data with PCA and MLP neural networks, the researchers discovered that the electronic nose system could differentiate various concentrations of ammonia in both the steady-state and transient response stages, as well as distinguish ammonia from other gases. The MLP was successful in distinguishing gas concentrations except for concentrations exceeding 200 ppm and 400 ppm. The sensor array demonstrated the ability to work in both transient and steady-state response phases, and data could be read without the need to enter a steady state, resulting in significant time savings. However, the accuracy of estimating transient gas concentration may be compromised due to various reasons. For instance, the insufficient training data with fewer data points for transient analysis as compared to steady-state analysis, rapid changes, and shorter duration of the transient phase, and concentration identification sensor array consisting of only two sensors may fail to produce significant differences in sensor responses at similar high concentrations such as 200ppm and 400ppm. Future work will entail investigating additional machine learning methods to analyze and study sensor response, exploring the performance of E-nose in steady-state and transient time while examining the impact of temperature and humidity.

## ACKNOWLEDGMENTS

**Funding** The authors gratefully acknowledge the financial supports from: National Key Research and Development Program of China (No.2021YFD1700904), National Natural Science Foundation of China (NSFC) (No.32071890), Project of Foreign Scientist Studio of Agricultural Biological Resources Engineering Technology (GZS2021007), and acknowledge the CAAOR research group of HENAU.

## REFERENCES

- [1] R. B. Swotinsky and K. H. Chase, "Health effects of exposure to ammonia: scant information," *Am J Ind Med*, vol. 17, no. 4, pp. 515-21, 1990, doi: 10.1002/ajim.4700170409.
- [2] R. E. de la Hoz, D. P. Schlueter, and W. N. Rom, "Chronic lung disease secondary to ammonia inhalation injury: a report on three cases," *Am J Ind Med*, vol. 29, no. 2, pp. 209-14, Feb 1996, doi: 10.1002/(SICI)1097-0274(199602)29:2<209::AID-AJIM12>3.0.CO;2-7.
- [3] X. Wang et al., "Ammonia exposure causes lung injuries and disturbs pulmonary circadian clock gene network in a pig study," *Ecotoxicol Environ Saf*, vol. 205, p. 111050, Dec 1 2020, doi: 10.1016/j.ecoenv.2020.111050.
- [4] L. Cheng, Z. Ye, S. Cheng, and X. Guo, "Agricultural ammonia emissions and its impact on PM(2.5) concentrations in the Beijing-Tianjin-Hebei region from 2000 to 2018," *Environ Pollut*, vol. 291, p. 118162, Dec 15 2021, doi: 10.1016/j.envpol.2021.118162.
- [5] M. A. Sutton, J. W. Erisman, F. Dentener, and D. Moller, "Ammonia in the environment: from ancient times to the present," *Environ Pollut*, vol. 156, no. 3, pp. 583-604, Dec 2008, doi: 10.1016/j.envpol.2008.03.013.
- [6] S. M. McGinn and H. H. Janzen, "Ammonia sources in agriculture and their measurement," *Canadian Journal of Soil Science*, vol. 78, no. 1, pp. 139-148, 1998, doi: 10.4141/s96-059.
- [7] M. Insausti, R. Timmis, R. Kinnersley, and M. C. Rufino, "Advances in sensing ammonia from agricultural sources," *Sci Total Environ*, vol. 706, p. 135124, Mar 1 2020, doi: 10.1016/j.scitotenv.2019.135124.
- [8] D. Li, X. Xu, Z. Li, T. Wang, and C. Wang, "Detection methods of ammonia nitrogen in water: A review," *TrAC Trends in Analytical Chemistry*, vol. 127, 2020, doi: 10.1016/j.trac.2020.115890.
- [9] J. Hu, L. Zhang, and Y. Lv, "Recent advances in cataluminescence gas sensor: Materials and methodologies," *Applied Spectroscopy Reviews*, vol. 54, no. 4, pp. 306-324, 2018, doi: 10.1080/05704928.2018.1464932.
- [10] M. B. Pushkarsky, M. E. Webber, and C. K. N. Patel, "Ultra-sensitive ambient ammonia detection using CO<sub>2</sub>-laser-based photoacoustic spectroscopy," *Applied Physics B*, vol. 77, no. 4, pp. 381-385, 2003, doi: 10.1007/s00340-003-1266-8.
- [11] Y. Zhang and L.-T. Lim, "Colorimetric array indicator for NH<sub>3</sub> and CO<sub>2</sub> detection," *Sensors and Actuators B: Chemical*, vol. 255, pp. 3216-3226, 2018, doi: 10.1016/j.snb.2017.09.148.
- [12] A. J. Moshayedi, M. Kukade, and D. C. Gharpure, "Electronic-nose (E-nose) for recognition of Cardamom, Nutmeg and Clove oil odor," 2014.
- [13] A. Solórzano et al., "Early fire detection based on gas sensor arrays: Multivariate calibration and validation," *Sensors and Actuators B: Chemical*, vol. 352, p. 130961, 2022/02/01/ 2022, doi: https://doi.org/10.1016/j.snb.2021.130961.
- [14] A. Miquel-Ibarz, J. Burgués, and S. Marco, "Global calibration models for temperature-modulated metal oxide gas sensors: A strategy to reduce calibration costs," *Sensors and Actuators B: Chemical*, vol. 350, p. 130769, 2022/01/01/ 2022, doi: https://doi.org/10.1016/j.snb.2021.130769.
- [15] A. J. Moshayedi, E. Kazemi, M. Tabatabaei, and L. Liao, "Brief modeling equation for metal-oxide; TGS type gas sensors," *Filomat*, vol. 34, pp. 4997-5008, 2020.
- [16] Q. Zheng, D. Zhao, J. Deng, X. Xu, and Z. Ou, "Front-end Electronics Design for Micro-Pattern Gas Detectors Based on VA140," in *2021 5th International Conference on Vision, Image and Signal Processing (ICVISP)*, 18-20 Dec. 2021 2021, pp. 167-170, doi: 10.1109/ICVISP54630.2021.00038.
- [17] W. Wang et al., "SnO<sub>2</sub> nanoparticles-modified 3D-multilayer MoS<sub>2</sub> nanosheets for ammonia gas sensing at room temperature," *Sensors and Actuators B: Chemical*, vol. 321, 2020, doi: 10.1016/j.snb.2020.128471.
- [18] R. Gutierrez-Osuna, A. Gutierrez-Galvez, and N. Powar, "Transient response analysis for temperature-modulated chemoresistors," *Sensors and Actuators B: Chemical*, vol. 93, no. 1-3, pp. 57-66, 2003/08/01/ 2003, doi: 10.1016/s0925-4005(03)00248-x.
- [19] A. J. Moshayedi, A. Toudeshki, and D. C. Gharpure, "Mathematical modeling for SnO<sub>2</sub> gas sensor based on second-order response," *2013 IEEE Symposium on Industrial Electronics & Applications*, pp. 33-38, 2013.
- [20] M. P. Gherman, Y. Cheng, A. Gomez, and O. Saukh, "Compensating Altered Sensitivity of Duty-Cycled MOX Gas Sensors with Machine Learning," in *2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 6-9 July 2021 2021, pp. 1-9, doi: 10.1109/SECON52354.2021.9491586.
- [21] A. J. Moshayedi and D. Gharpure, "Implementing Breath to Improve Response of Gas Sensors for Leak Detection in Plume Tracker Robots," in *Proceedings of the Third International Conference on Soft Computing for Problem Solving*, New Delhi, M. Pant, K. Deep, A. Nagar, and J. C. Bansal, Eds., 2014/ 2014: Springer India, pp. 337-348.
- [22] F. Kherif and A. Latypova, "Chapter 12 - Principal component analysis," in *Machine Learning*, A. Mechelli and S. Vieira Eds.: Academic Press, 2020, pp. 209-225.

# MM-UNet: Multi-attention mechanism and multi-scale feature fusion UNet for tumor image segmentation

Yaozheng Xing  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
yzxing@bupt.edu.cn

Jie Yuan\*  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
yuanjie@bupt.edu.cn

Qixun Liu  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
2317297031@bupt.edu.cn

Shihao Peng  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
shihapeng@bupt.edu.cn

Yan Yan  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
2020211959yy@bupt.edu.cn

Junyi Yao  
Beijing University of Posts and  
Telecommunications, Beijing, 102206,  
China  
yaojunyimerk11@163.com

## ABSTRACT

To address the problems of many parameters and loss of spatial information in traditional Unet networks, this paper proposes a U-Net-based brain tumor segmentation model named MM-UNet to solve the problem of 3D image segmentation. Firstly, the U-Net model performs three times downsampling to extract the image features for the changing characteristics of brain tumor 3D images, which reduces the number of model parameters while maximally preserving the target edge features; then, a structure similar to FPN was used to achieve the fusion of multi-scale predictions; we introduce the channel attention mechanism and pixel attention mechanism to establish the relationship between global features; meanwhile, to improve the generalization ability of the model, data augmentation techniques are used to enhance the information. The experimental results show that the model proposed in this paper has improved the accuracy of brain tumor segmentation compared with U-Net, PSPNet, ICNet, and Fast-SCNN, suggesting 3.9%, 1.3%, 5%, and 3.9%, respectively.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Image segmentation;

## KEYWORDS

Brain tumor segmentation, Unet model, Attention mechanisms, Deep learning

\*Key Laboratory of Trustworthy Distributed Computing and Service (BUPT), Ministry of Education.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590047>

## ACM Reference Format:

Yaozheng Xing, Jie Yuan, Qixun Liu, Shihao Peng, Yan Yan, and Junyi Yao. 2023. MM-UNet: Multi-attention mechanism and multi-scale feature fusion UNet for tumor image segmentation. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590047>

## 1 INTRODUCTION

Brain tumor segmentation includes traditional segmentation algorithms based on threshold, region, pixel classification [1], and segmentation algorithms based on convolutional neural networks (CNNs). In recent years, algorithms based on deep learning have been widely used in the field of medical image segmentation.

The concept of deep learning was proposed by Hinton et al. [2]. One type of the models based on convolutional neural networks that is trained to extract image features automatically is widely used for medical image segmentation because of its greater feature extraction capability, more excellent feature representation and simpler pre-processing. Ronneberger et al. [3] designed a U-Net network based on the fully convolutional network proposed by Long [4]. The U-Net network consists of a U-shaped channel and skip connections. The U-shaped channel is similar to the encoder-decoder structure of SegNet. The skip connection concatenates feature maps of the same size, that is, different feature maps of the same size are spliced in their channel domains. The reason why U-Net is suitable for medical image segmentation is that its structure can combine shallow and deep information at the same time. [5] Shallow information helps to improve accuracy, and deep information helps to extract complex features. Pereria et al. [6] presented an improved U-Net with novel feature reorganization and recalibration modules and segmented individual substructures of tumors by cascading. Myronenko et al. [7] discarded the symmetric encoder-decoder structure of U-Net and used an asymmetric encoder-decoder, using a deeper encoder for extracting more feature information. In addition, U-Net has been widely applied to 3D architectures to extract more information and achieve higher accuracy. Özgün et al. [8] proposed 3DUnet, and Fausto et al. [9] proposed 3DVnet. 3D network structures are basically similar to 2D

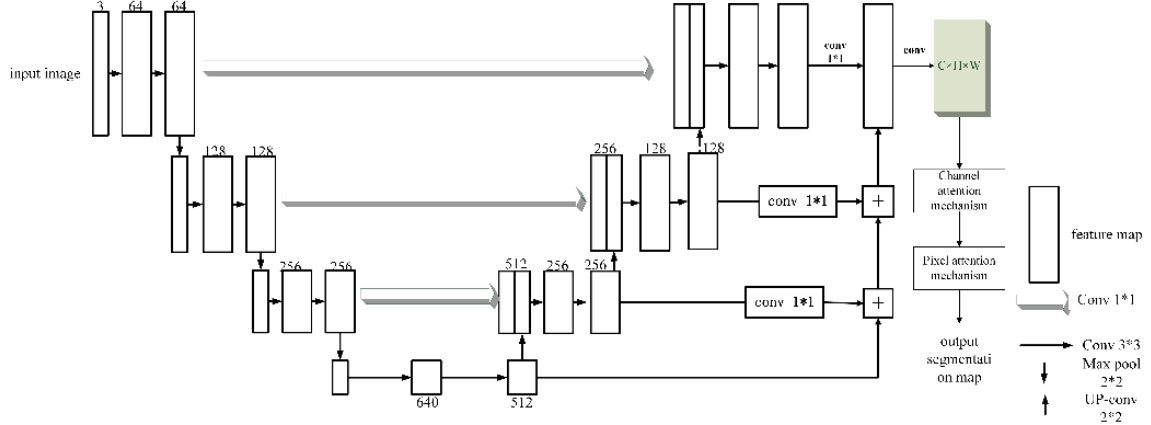


Figure 1: Structure of the improved U-Net model.

networks, but they mainly replace 2D methods such as convolution and pooling with 3D methods.

Deep learning has achieved good results, but it is still a long way from practical medicine. Gliomas in the brain vary in phenotype, size, and location, and tumor boundaries are blurred, making it challenging to automatically segment different regions of the glioma simultaneously.

The transposed convolutional upsampling used in U-Net gives relatively smooth structural features, but after the transformation of the convolutional plus non-linear activation layers, the original features are not well preserved, so the original features are preserved by passing in features from the shallow layers through the skip layer connection in U-Net. However, U-Net has a single skip layer structure and cannot fully utilize the features from the intermediate layers.

The contributions of this paper to the above issues are summarized as follows:

1. Using the original UNet image semantic segmentation algorithm as the basis, we first use 3 downsampling layers to maximize the retention of brain tumor image edge features, which improves the computational speed while reducing the number of model parameters.
2. Embedding the channel attention mechanism and pixel attention mechanism modules, to improve the accuracy of the model and establish the link between features and attention mechanisms.
3. Using multi-scale fusion, to make full use of deep and shallow features.

## 2 BRAIN TUMOR SEGMENTATION MODEL

### 2.1 Lightweight and multi-scale information enhancement of brain tumor segmentation model

To meet these challenges, this chapter proposes to design a lightweight semantic segmentation algorithm based on the U-Net model with high accuracy, and fewer model parameters. The model can be applied to real-time detection and analysis of equipment, and its

structure is shown in the figure below. MM-UNet also adopts the network structure of encoding and decoding.

To address the problem of the single hopping-layer structure, we borrowed the idea of FPN and used multi-scale prediction fusion. The fusion of feature maps at different scales makes full use of the structure in U-Net while using the global information at different scales contained in the intermediate convolutional layers.

To further optimize its feature extraction capability, the attention modules are constructed in the network based on the idea of attention in vision, allowing the model to focus more on learning information about valid regions in the image, thus improving the robustness of the model. We explicitly construct dependencies between different channels by establishing a channel attention mechanism, followed by a pixel attention mechanism, further characterizing the spatial correlation between each pixel in the scene features and aggregating the features of the two attention modules.

### 2.2 Channel attention mechanism

For the channel-level attention mechanism, we directly use SE-Net, which has two parts: global information embedding and adaptive recalibration. The input features  $C_i$  are compressed to obtain their weight for each feature channel which are then used to recalibrate the input feature by multiplying them with the corresponding channel values to complete the channel-level attention work.

In a specific implementation, we first use global average pooling (GAP) to compress the  $H \times W \times C$  input features into  $1 \times 1 \times C$ . We use two fully connected layers to transform the resulting vectors into their weight values, followed by sigmoid-activated weight values multiplied back to the original input features according to the channels, with a channel compression ratio  $r$  of 16 for the first fully connected layer, which reduces the computational cost and avoids overfitting.

### 2.3 Pixel attention mechanism

After the channel-level attention module, we use a pixel-level attention module that can be used for target detection tasks to limit the functionality. The idea is borrowed from Scaled Dot Product Attention (SDPA) in Transformer. For input features of dimensions

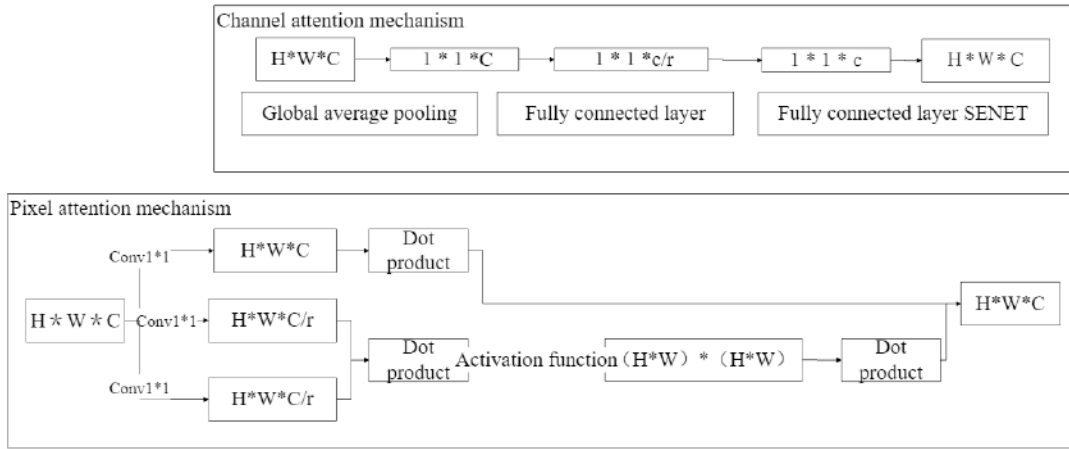


Figure 2: Multi-attention mechanism.

$H \times W \times C$ , the matrices  $Q, K, V$  corresponding to Query, Key, Value are first generated using a three-layer  $1 \times 1$  convolution. When the channel compression ratio  $r$  for  $Q$  and  $K$  is 8, computing the dot product of  $Q$  and  $K$  and applying Softmax, we obtain a weight matrix of dimension  $(H \times W) \times (H \times W)$ . The correlation between the pixels in the center of the feature, along with the weight matrix and the  $V$  dot product, give a final weighted feature map of  $H \times W \times C$ . The calculation process can be expressed as follows.

$$f_{attention}(Q, K, V) = \text{softmax}(Q, K^T) \cdot V$$

## 2.4 Activation functions

The most commonly used activation function for traditional neural networks is ReLU, which has an output equal to the input when the input is positive and a value of 0 when the input is negative. However, a disadvantage of this activation function is that if the input data is negative, the ReLU function has a zero gradient and suffers from the dead ReLU problem, i.e., when  $x < 0$ , it does not activate and causes vanishing gradients. Since some of the values in MRI images are negative, using ReLU may not be optimal. In this paper, we use Mish as the activation function, which will have a smooth curve when the data is negative, avoiding the problem of vanishing gradient, and more feature information can enter the network, providing generalisation ability and accuracy.

$$\text{Mish}(x) = x \cdot \tanh(\ln(1 + \exp(x)))$$

## 3 EXPERIMENT

### 3.1 Experimental setup

The main model evaluation metrics used in semantic segmentation models include mean pixel accuracy (mPA), mean intersection over union (mIoU), and Kappa coefficient. Pixel accuracy (PA) is the ratio of the number of pixels correctly predicted by a category to the total number of pixels, and a higher value of accuracy indicates better model quality. Assuming that the image has  $k+1$  categories (including  $k$  target classes and 1 background class),  $p_{ij}$  means that

the pixels belonging to class  $i$  are misjudged as the number of pixels of class  $j$ , and  $p_{ii}$  is the number of pixels that are correctly classified, so the accuracy expression is as follows:

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}}$$

In this paper, mPA is used as the overall model evaluation metric, which is calculated by first calculating the PA value for each class and then averaging over all classes, so the mPA expression is shown in the following equation.

$$mPA = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij}}$$

mIoU is the ratio of the intersection of the predicted and actual regions to the union of the predicted and actual regions, computed for each class individually, and then averaged over all classes. The expression for the mean intersection-to-union ratio is shown in the following equation:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}}$$

The Kappa coefficient is a consistency test index used to measure the effectiveness of pixel classification, where consistency refers to whether the model prediction results are consistent with the actual classification results. The Kappa coefficient can compensate for the defects of excessive pixel accuracy due to uneven categories, thereby punishing the bias of the model. The value is between -1 and 1, and is usually greater than 0. The higher the coefficient value, the better the quality of the model. The calculation formula is as follows:

$$Kappa = \frac{p_o - p_e}{1 - p_e}$$

$p_o$  refers to the pixel accuracy of the classifier and  $p_e$  refers to the pixel accuracy of the random classifier.

### 3.2 Dataset description

This paper uses the dataset BraTS2018, which contains 285 cases. Each case has four modalities that are segmented into three parts: whole tumor, enhanced tumor, and tumor core. The input image shape is  $4 \times 160 \times 192 \times 128$ , where 4 represents the images of modalities.

In order to improve the generalisation performance of the model to unknown images, we applied the following image enhancement methods: horizontal rotation, vertical rotation, image scaling from 80% to 110%, cropping 0%–25% per side, translation  $-20\%$ – $+20\%$ , cropping  $-18\%$ – $+18\%$ , rotation  $-45^\circ$ – $+45^\circ$ .

### 3.3 Comparative experiments

Table 1 shows the comparison of the segmentation evaluation metrics between the brain tumour segmentation model proposed in this paper and the mainstream semantic segmentation models on the BraTS2018 brain tumour dataset, where the mainstream semantic segmentation models include U-Net, PSPNet [10], ICNet [11] and Fast-SCNN [12].

The final segmentation results show that the average accuracy of the model in this paper reached 90.61%, the average intersection ratio reached 73.59%, and the kappa coefficient reached 92.13%, all of which were higher than other segmentation models. In addition, we also show a greater improvement in all evaluation metrics compared

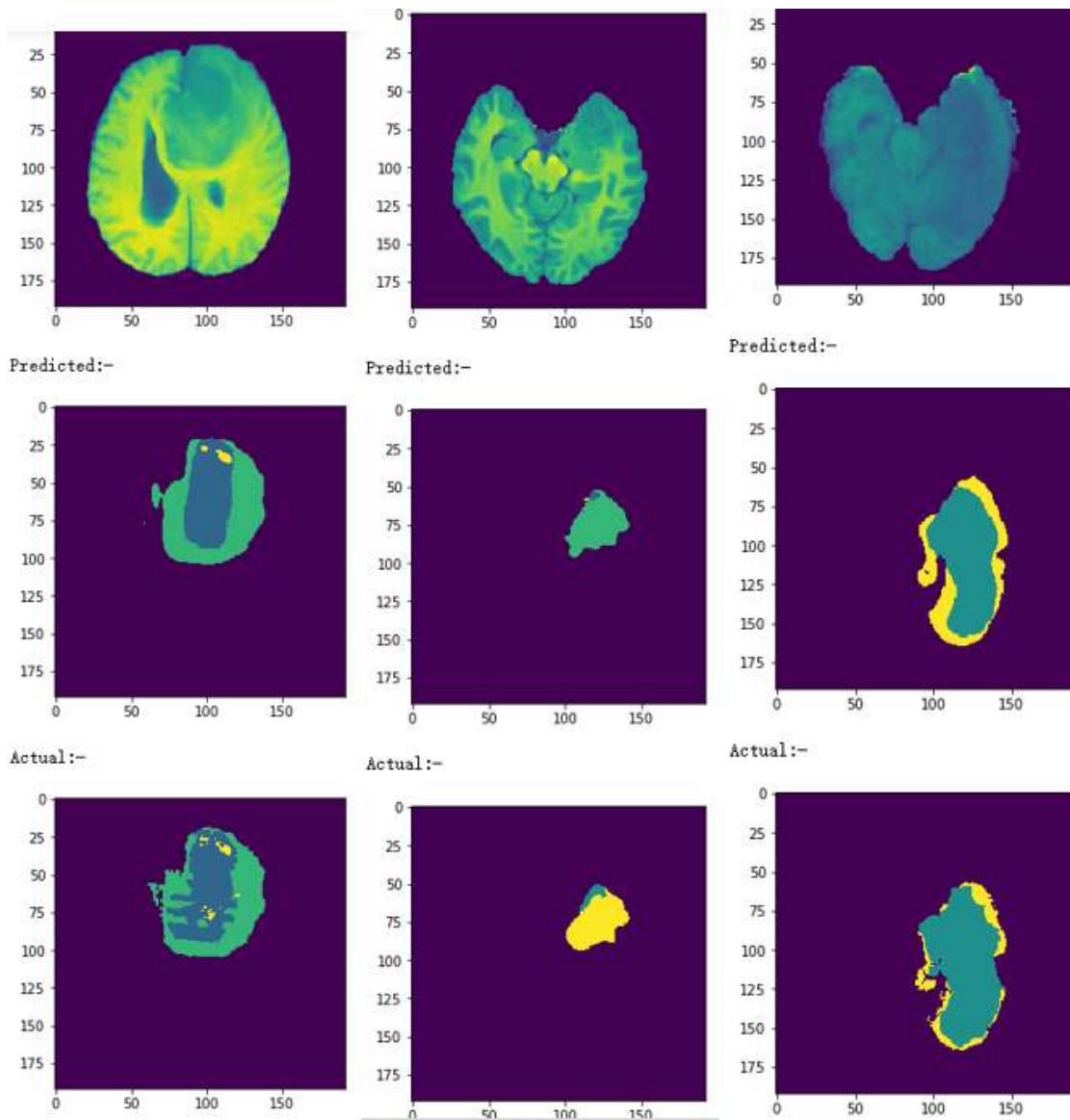


Figure 3: brain tumor segmentation.

**Table 1: Comparison of the accuracy rate of brain tumor recognition by different methods**

Method	mPA	mIoU	Kappa
UNet	0.8671	0.6459	0.8557
PSPNet[10]	0.8932	0.7159	0.8841
ICNet[11]	0.8561	0.7289	0.8918
Fast-SCNN[12]	0.8675	0.7239	0.9073
MM-UNet	0.9061	0.7359	0.9213

to the original U-Net model, which also confirms the effectiveness of the improved method.

In order to compare the segmentation results more visually, some data from the BraTS2018 dataset is selected for segmentation in this paper and generated a visual result as shown in Figure 3.

#### 4 CONCLUSION

The accurate and fast segmentation of medical images of brain tumors is of great significance for human health since brain tumors is a disease that endangers human health.

Starting from the structure of U-Net, we embed channel attention mechanism and pixel attention mechanism modules to improve the accuracy of the model, establish the connection between features and attention mechanism, and enhance the sensitivity to details. Besides, we establish global dependencies of different features by using multi-scale feature fusion with fewer parameters. The final segmentation results showed that our models achieved an average accuracy of 90.61%. The model proposed in this paper is an improvement over the traditional U-Net network in terms of efficiency and accuracy.

Considering the limitation of the sample data in this experiment, we acknowledge that our model may not be able to process enough details for segmentation, which in turn may affect its segmentation performance. Subsequent work would use methods such as weak supervision to further improve the detection and segmentation of small sample datasets. We will also try to test the model in other medical scenarios, such as segmenting medical images of the liver and other parts of the body.

#### ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China under Grant62002028 and Research Innovation Fund for College Students of Beijing University of Posts and Telecommunications.

#### REFERENCES

- [1] LI Qiang, BAI Kexin, ZHAO Liu, *et al.* Progresss and challenges of MRI brain tumor image segmentation [J].Journal of Chinese Image Graphics, 2020, 25 (3): 419-431 (in Chinese).
- [2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. nature, 2015, 521(7553): 436-444.
- [3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
- [4] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.
- [5] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
- [6] Pereira S, Pinto A, Amorim J, *et al.* Adaptive feature recombination and recalibration for semantic segmentation with fully convolutional networks[J]. IEEE transactions on medical imaging, 2019, 38(12): 2914-2925.
- [7] Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization[C]. International MICCAI Brainlesion Workshop. Springer, Cham, 2018: 311-320.
- [8] SHELHAMER, EVAN, LONG, JONATHAN, DARRELL, TREVOR. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 640-651. DOI:10.1109/TPAMI.2016.2572683.
- [9] Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation [C]//4th International Conference on 3D Vision.IEEE, 20 16:565-571.
- [10] Zhao H, Shi J, Qi X, *et al.* Pyramid Scene Parsing Network[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
- [11] Zhao H, Qi X, Shen X, *et al.* ICNet for Real-Time Semantic Segmentation on High-Resolution Images[C]. Cham: Springer International Publishing, 2018: 418-434.
- [12] Poudel R P K, Liwicki S, Cipolla R. Fast-SCNN: Fast Semantic Segmentation Network[J]. Computer Science, 2019.

# A hybrid Aquila Optimizer sine cosine Algorithm for Numerical Optimization

Fei Chu, Jiayang Wang, Fulin Tian

## ABSTRACT

To address the shortcomings of the Aquila optimizer algorithm (AO), this paper proposes a novel hybrid Aquila Optimizer sine cosine Algorithm (AO-SCA). Firstly, Singer chaotic mapping is used for initialization, so that the initial solution position distribution was more homogeneous, and increased the richness of the population. Secondly, in the exploration phase of AO, the concept of sine and cosine algorithm is integrated and the nonlinear sine learning factor is introduced to balance the local and global digging ability and accelerate the convergence speed. Finally, through the numerical experiment simulation of 8 benchmark functions, the results show that the optimization ability and convergence speed of the proposed algorithm is better.

## CCS CONCEPTS

• **Computing methodologies**; • **Theory of computation** → **Mathematical optimization**; *Design and analysis of algorithms*;

## KEYWORDS

Aquila optimizer, Singer mapping, sine and cosine algorithm, swarm intelligence

## ACM Reference Format:

Fei Chu, Jiayang Wang, Fulin Tian. 2023. A hybrid Aquila Optimizer sine cosine Algorithm for Numerical Optimization. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590048>

## 1 INTRODUCTION

In the past decades, meta-heuristic algorithms have been successfully applied to solve optimization problems in science, engineering and other fields. Since Holland proposed genetic algorithm (GA) [1], researchers have proposed multiple heuristic algorithms including differential evolution algorithm (DE)[2], particle swarm optimization (PSO)[3], fireworks algorithm (FW)[4], Harris hawks optimizer (HHO)[5], whale optimization algorithm (WOA)[6], Grey wolf optimizer (GWO)[7], water cycle algorithm (WCA)[8], salp group algorithm (SSA)[9] and so on. At present, meta-heuristic algorithms have gradually been widely used in robot subgrade planning[10], cluster vehicle routing problem[11] and other fields.

The Aquila optimizer (AO) is the latest developed algorithm proposed by Abualigah et al.[12], and its inspiration comes from Aquila's behavior of catching prey in nature. As a stochastic optimizer, AO can be easily implemented, has strong global exploration ability, and has better competitiveness compared with traditional optimization algorithms. However, AO also has obvious shortcomings such as slow convergence speed and easy to fall into local optimum. Aiming at the deficiency of AO, Yu et al. [13] proposed a new hybrid algorithm that combines arithmetic optimization

algorithm named AOAAO, which performs well in structural optimization problems. Zhang et al.[14] introduced chaotic mapping to improve the randomness of the algorithm search and mixes the Hunger Games search algorithm with AO to enhance the exploration capability of the algorithm named CHGSAO, which performs well in solving optimal reactive power dispatch problem.

Compared with the standard AO, the improved versions of AO proposed in the above documents have improved in terms of optimization accuracy. However, these improved versions of AO still have problems such as weak global search ability, weak ability to avoid falling into local optimization, premature convergence, etc., which need to be further improved. Generally, search processes with similar properties may lead to a loss of population diversity and may fall into local optima. However, combining different search methods from different algorithms can greatly improve their ability to escape from local convergence. Whereas AO has better exploration capability, its exploitation capability is relatively weak. sine cosine Algorithm (SCA)[15] is the opposite. By combining the exploration ability of AO and the exploitation ability of SCA, the proposed hybrid algorithm can make up for the defects of a single algorithm and maximize the advantages of a single algorithm to balance exploration and exploitation.

The rest of this paper is organized as follows: The second part introduces the basic principles and defects of Aquila optimizer algorithm. The third part introduces the improved algorithm. The fourth part gives the experimental results to illustrate the performance of the proposed algorithm. The fifth part summarizes this paper and looks forward to the future work.

## 2 BASIC AQUILA OPTIMIZER ALGORITHM

Aquila optimizer algorithm is inspired by four flight predator-prey behaviors of Aquila. It has four different hunting behaviors. The specific description of each behavior is as follows.

### 2.1 Expanded exploration ( $X_1$ )

In this strategy, Aquila can explore the target area extensively through high altitude flight. Once he found his prey, he would dive vertically towards it. This behavior is expressed as:

$$X(t+1) = X_{best}(t) \times \left(1 - \frac{t}{T}\right) + (X_M(t) - X_{best}(t) \times rand) \quad (1)$$

$$X_M(t) = \frac{1}{N} \sum_{i=1}^N X_i(t) \quad (2)$$

Where  $X(t+1)$  represents the individual's position in the  $t+1$  iteration, and  $X_{best}(t)$  represents the current global best position in the  $t$  iteration;  $t$  and  $T$  represent the current number of iterations and the maximum allowed number of iterations;  $X_M(t)$  is the current average position of individuals in the current iteration;  $rand$  is a random number falling into the range of 0 and 1 in the

**Table 1: Information of benchmark functions.**

Function	Dimension	Range	$f_{min}$
$F_1(x) = \sum_{i=1}^n x_i^2$	30	[-100,100]	0
$F_2(x) = \sum_{i=1}^n  x_i  + \prod_{i=1}^n  x_i $	30	[-10,10]	0
$F_3(x) = \sum_{i=1}^n \left( \sum_{j=1}^i  x_j  \right)^2$	30	[-100,100]	0
$F_4(x) = \max_i \{  x_i , 1 \leq i \leq n \}$	30	[-100,100]	0
$F_5(x) = \sum_{i=1}^n \left[ 100(x_i^2 - x_{i+1})^2 + (x_i - 1)^2 \right]$	30	[-30,30]	0
$F_6(x) = \sum_{i=1}^n ([x_i + 0.5])^2$	30	[-100,100]	0
$F_7(x) = \sum_{i=1}^n ix_i^4 + random[0, 1)$	30	[-1.28,1.28]	0
$F_8(x) = \left( \sum_{i=1}^n  x_i  \right) \exp \left( - \sum_{i=1}^n \sin(x_i^2) \right)$	30	$[-2\pi, 2\pi]$	0

**Table 2: Results of benchmark functions**

Fn	Stats	GWO	SCA	SSA	WCA	WOA	HHO	AO	AO-SCA
F1	Ave	3.07E-36	0.852131	2.57E-08	6.13E-18	2.81E-92	5.25E-19	<b>0.0E+00</b>	<b>0.0E+00</b>
	Best	1.52E-36	0.55677	1.03E-08	4.14E-19	7.08E-94	2.03E-19	<b>0.0E+00</b>	<b>0.0E+00</b>
	Std	2.15E-37	0.41722	1.17E-09	9.36E-18	3.97E-92	4.23E-19	<b>0.0E+00</b>	<b>0.0E+00</b>
F2	Ave	4.30e-108	1.03e-51	0.00293	55.3618	421.271	2.59e-52	0.00017	<b>1.25e-169</b>
	Best	4.22e-137	1.65e-57	0.00222	21.4631	400.873	2.04e-57	4.47E-05	<b>0</b>
	Std	1.36e-107	3.07e-51	0.00049	22.1837	12.0743	4.38e-52	2.73E-06	<b>0</b>
F3	Ave	1.04E-127	2.01E-07	2.11E-07	6.11E-28	4.43E-143	2.43E-66	<b>0.0E+00</b>	<b>0.0E+00</b>
	Best	5.03E-130	4.21E-06	3.14E-07	3.21E-32	3.12E-143	1.02E-69	<b>0.0E+00</b>	<b>0.0E+00</b>
	Std	2.81e-63	1.06e-51	66793.9	724731	425035	8.21e+06	3.26e-56	<b>0</b>
F4	Ave	26.4269	4289.73	103.807	9.45568	27.2305	38.0473	4.80E-192	<b>0.0E+00</b>
	Best	25.7264	246.468	75.9896	0.125462	27.1183	0.85238	6.16E-213	<b>0.0E+00</b>
	Std	7.33e-81	1.77e-52	4.35353	0.200647	1.93088	24.995	2.22e-112	<b>5.69e-153</b>
F5	Ave	1.15682	28.1853	39.3008	42.7932	1.89E-15	62.9476	<b>0.0E+00</b>	<b>0.0E+00</b>
	Best	0.0E+00	0.006367	12.8123	15.8936	0.0E+00	41.7614	<b>0.0E+00</b>	<b>0.0E+00</b>
	Std	2.59919	27.8158	16.2666	14.015	1.04E-15	16.7692	<b>0.0E+00</b>	5.44e-32
F6	Ave	2.15E-19	0.063249	0.816566	1.52E-09	6.73E-58	7.13E-06	1.37E-202	<b>0.0E+00</b>
	Best	8.13E-21	0.035709	0.663169	5.17E-10	1.83E-58	2.35E-06	2.27E-221	<b>0.0E+00</b>
	Std	3.17E-18	0.035371	0.184172	1.28E-09	6.57E-57	7.74E-06	<b>0.0E+00</b>	<b>0.0E+00</b>
F7	Ave	1.16e-04	<b>0</b>	0.03088	15365.1	105.976	0.0016101	432.101	<b>0</b>
	Best	7.50e-06	<b>0</b>	0.02207	10922.9	81.9039	7.42e-05	345.724	<b>0</b>
	Std	1.01e-04	<b>0</b>	3.71e-04	3001.99	15.1657	0.00195	71.4461	<b>5.09e-05</b>
F8	Ave	-194184	-209337	-63151.7	-15689	-66501.3	-197740	-116335	<b>-209478</b>
	Best	-208742	-209491	-72509.3	-18421.5	-77048.7	-209489	-121790	<b>-209490</b>
	Std	19636.3	465.602	3686.89	1413.99	5152.31	19893.9	3978.16	<b>14.6611</b>

Gaussian distribution. When  $\text{rand} < 0.5$ , Aquila performs extended search, and vice versa.

## 2.2 Narrowed exploration ( $X_2$ )

This strategy corresponds to the equal altitude flight of the Aquila short glide attack. At this time, the Aquila locks the prey target from high air and hovers above the target to prepare for the attack. Its mathematical model is as follows:

$$X(t+1) = X_{best}(t) \times LF(D) + X_R(t) + (y - x) \times rand \quad (3)$$

Where  $X_R(t)$  is the random position of the Aquila;  $D$  is the size;  $LF$  represents Levy flight function;  $x$  and  $y$  are the shapes of the search, denoted as:

$$\begin{cases} x = (r_1 + 0.00565 \times D_1) \times \sin\left(-\omega \times D_1 + \frac{3 \times \pi}{2}\right) \\ y = (r_1 + 0.00565 \times D_1) \times \cos\left(-\omega \times D_1 + \frac{3 \times \pi}{2}\right) \end{cases} \quad (4)$$

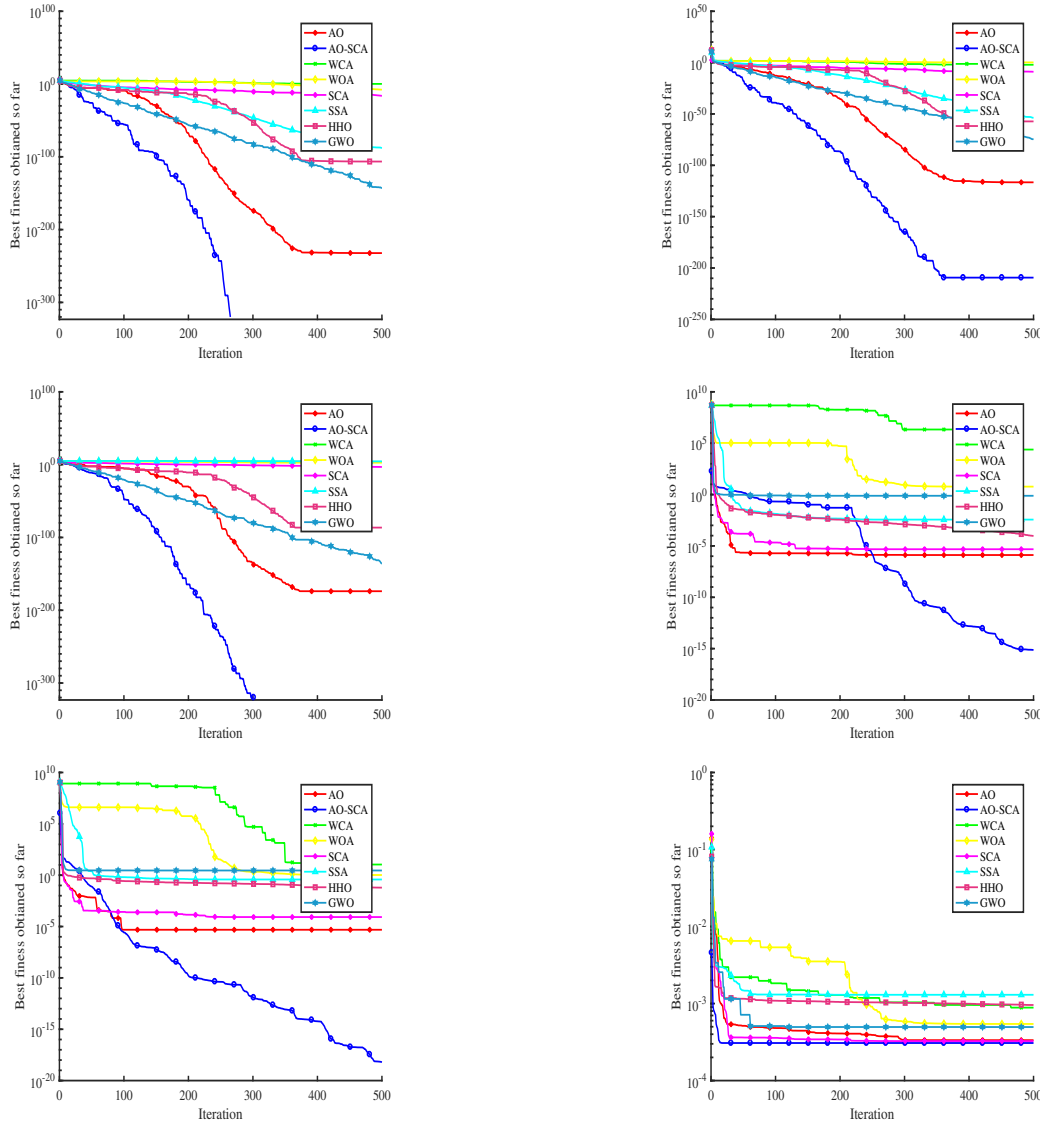


Figure 1: Qualitative results of F1-F3, F6- F8

$$LF(x) = 0.01 \times \frac{\mu \times \sigma}{|v|^{\frac{1}{\beta}}} \quad (5)$$

$$\sigma = \left( \frac{\Gamma(1+\beta) \times \sin\left(\frac{\pi\beta}{2}\right)}{\Gamma\left(\frac{1+\beta}{2}\right) \times \beta \times 2 \left(\frac{\beta-1}{2}\right)} \right)^{\frac{1}{\beta}} \quad (6)$$

Where  $r_1$  is a fixed value from 1 to 20, indicating a fixed number of search cycles;  $D_1$  is an integer value, ranging from 1 to  $D$ ;  $\omega = 0.005$ .

### 2.3 Expanded exploitation ( $X_3$ )

In this strategy, Aquila begin to attack the target through low altitude flight. This initial predatory behavior can be expressed as:

$$X(t+1) = (X_{best}(t) - X_M(t)) \times \alpha - \text{rand} + ((UB - LB) \times \text{rand} + LB) \times \delta \quad (7)$$

Where  $\alpha$  and  $\delta$  are the adjustment parameters for the exploitation stage, fixed at 0.1;  $UB$  and  $LB$  are the upper and lower limits of the search respectively;  $\text{rand}$  is a random number from 0 to 1. When  $\text{rand} < 0.5$ , Aquila performs expanded exploitation, and vice versa.

### 2.4 Narrowed exploitation ( $X_4$ )

In this strategy, Aquila land on the ground and launch a precise attack on the target. This predatory behavior can be expressed as:

$$X(t+1) = QF \times X_{best}(t) - (G_1 \times X(t) \times \text{rand}) - G_2 \times LF(D) + \text{rand} \times G_1 \quad (8)$$

$$QF(t) = t^{\frac{2 \times \text{rand} - 1}{(1-T)^2}} \quad (9)$$

$$\begin{cases} G_1 = 2 \times \text{rand} - 1 \\ G_2 = 2 \times \left(1 - \frac{t}{T}\right) \end{cases} \quad (10)$$

**Table 3: CEC2019 test functions**

F <sub>n</sub>	Range	Dim	F <sub>min</sub>
CEC01	[-8192,8192]	9	1
CEC02	[-16384,16384]	16	1
CEC03	[-4,4]	18	1
CEC04	[-100,100]	10	1
CEC05	[-100,100]	10	1
CEC06	[-100,100]	10	1
CEC07	[-100,100]	10	1
CEC08	[-100,100]	10	1
CEC09	[-100,100]	10	1
CEC10	[-100,100]	10	1

Where  $QF$  is the quality function used to balance the search strategy;  $G_1$  is the random motion parameter in the process of Aquila tracking prey, between  $[-1, 1]$ ;  $G_2$  indicates the flight slope when tracking prey, which decreases linearly from 2 to 0.

### 3 PROPOSED ALGORITHM:AO-SCA

In this section, we will introduce a novel hybrid Aquila Optimizer sine cosine Algorithm, namely AO-SCA.

#### 3.1 Singer chaotic mapping

The initial population position is uniformly distributed in the search space, which helps to enhance the algorithm's global search ability and improve the search efficiency. The initial AO initializes the population randomly, which has the risk of reducing the diversity of the population. The chaotic sequence generated by the chaotic map has the characteristics of ergodicity, nonlinearity and unpredictability, and is often used to replace the random sequence to initialize the population. This paper uses Singer mapping to initialize the population. As a typical form of chaotic map Singer mapping has the advantages of simple parameters and uniform distribution [16]. Its mathematical expression is as follows:

$$s_{k+1} = \lambda \left( 7.86s_k - 23.31s_k^2 + 28.75s_k^3 - 13.302875s_k^4 \right) \quad (11)$$

where  $\lambda \in [0.1, 1.08]$ , Singer mapping has chaotic behavior.

The generated chaotic sequence  $s_k \in [0, 1]$  is used to initialize the SSA population Position  $X_k$ , as shown in the following formula:

$$X_k = ub + s_k(ub - lb) \quad (12)$$

where  $ub$ ,  $lb$  are the upper and lower bounds of the search space.

#### 3.2 Sine and cosine algorithm

In the exploration stage of AO, it randomly inhabits the  $X(t)$  position, and uses the two strategies of Equation (1) and (3) to update the position. It selects different search strategies according to the random number of  $rand$ , but in this case, the population will gather quickly. The diversity of population decreases rapidly, and the algorithm can easily fall into a local optimum. Here, the sine-cosine strategy is introduced to control the motion range of the individual to avoid the individual from converging to the local optimal region. To address this issue, we fused the concept of sine and cosine algorithm in the exploration phase, and introduced the nonlinear

sine learning factor, which has a large value early in the search to facilitate global exploration and a small value late in the search to facilitate local exploitation. The improved exploration stage formula are as follows:

$$\omega = \omega_{\min} + (\omega_{\max} - \omega_{\min}) \cdot \sin(t\pi / \text{iter}_{\max}) \quad (13)$$

$$X_{i,j}^{t+1} = \begin{cases} (1 - \omega) \cdot X_{i,j}^t + \omega \cdot \sin(r1) \cdot |r2 \cdot X_{\text{best}} - X_{i,j}^t|, q < 0.5 \\ (1 - \omega) \cdot X_{i,j}^t + \omega \cdot \cos(r1) \cdot |r2 \cdot X_{\text{best}} - X_{i,j}^t|, q \geq 0.5 \end{cases} \quad (14)$$

Where  $r1$  is a random number in  $[0, 2\pi]$ ;  $r2$  is a random number in  $[0, 2]$ ;  $X_{\text{best}}$  represents the current global best position in the  $t$  iteration;  $\omega$  is the learning factor.

## 4 EXPERIMENTS

For the fairness of the experimental results, all tests were conducted in the same environment. The experimental environment is Windows 10, 64-bit operating system, processor is Intel(R) Core(TM) i5-9300H CPU @ 2.40GHz, RAM is 8.0GB, and programming software is MATLAB R2019a. 8 benchmark functions were selected, as shown in Table 1.

Moreover, to verify the effectiveness and robustness of the proposed AO-SCA in solving optimization problems, it is compared with other well-established optimization algorithms, including the original AO, SCA, GWO, HHO, SSA, WOA and WCA. In order to reduce the contingency of the experiment and increase the persuasiveness of the experimental results, each algorithm runs independently for 30 times.

### 4.1 Numerical Experiment

Table 2 shows the comparison of all compared algorithms for the 10 benchmark functions tested. From the experimental results, the proposed algorithm has achieved the best results in terms of convergence speed and optimization accuracy on other benchmark functions except for multimodal function  $F_7$ . The corresponding variance is also very small, mostly 0 or close to 0, which shows that the robustness of AO-SCA is excellent. The convergence curves of AO-SCA (for all functions) are located at the bottom of the convergence curves of all the compared algorithms in Fig.1, therefore, AO-SCA has the fastest convergence speed among the 8 compared algorithms.

In order to verify the superiority of the improved algorithm over other improved algorithms, the two improved algorithms AOAAO and CHGSAO and original AO are selected as the comparative experiments. CEC2019 benchmark function is selected in this experiment, which has the characteristics of large-scale and complex optimization. The CEC2019 benchmark function details are shown in the Table 3. Accordingly, CEC2019 test results are shown in Table 6. From the data, it can be seen that the average value of AO-SCA on CEC01, CEC03 and CEC09 is the closest to the theoretical optimal value and the algorithm stability is better, while the optimization results of all algorithms on CEC02 and CEC03 are the same, but the AO-SCA standard deviation is the smallest, that is, the algorithm stability is the best, and the average value of HAO on CEC10 is the closest to the theoretical optimal value, but its standard deviation

**Table 4: CEC2019 test results**

Fn	Stats	AO-SCA	CHGSAO	AOAAO	AO
CEC01	Ave	3.24E+04	4.31+E04	2.57E+07	2.57E+04
	Std	1.52E+03	2.55E+03	2.93E+07	4.87E+03
CEC02	Ave	1.73E+01	1.73E+01	1.73E+01	1.73E+01
	Std	4.22E-15	1.65E-08	1.22E-07	2.57E-04
CEC03	Ave	1.27E+01	1.27E+01	1.27E+01	1.27E+01
	Std	5.03E-16	4.21E-06	3.14E-07	2.57E-08
CEC04	Ave	1.07E+01	1.85E+01	1.97E+01	1.57E+01
	Std	1.52E+00	2.55E+00	4.03E+00	2.57E+00
CEC05	Ave	4.30E-108	1.03E-51	0.00293	2.57E-08
	Std	1.32E-02	5.05E-02	2.52E-01	1.27E-01
CEC06	Ave	1.04E+00	2.01E+00	2.11E+00	2.57E+00
	Std	5.03E+00	4.21E-01	3.14E-01	2.57E-01
CEC07	Ave	1.07E+02	2.85E+02	2.57E+02	2.76E+02
	Std	1.52E+01	4.55E+01	2.56E+02	4.59E+02
CEC08	Ave	2.30E+00	4.03E+00	3.93E+00	5.22E+00
	Std	4.22E-137	1.65E-57	0.00222	2.57E-08
CEC09	Ave	2.38E+00	2.93E+00	3.27E+00	2.57E+00
	Std	1.12E-02	1.21E-01	1.54E+00	1.12E-02
CEC10	Ave	2.00E+01	1.93E+01	2.05E+01	2.04E+01
	Std	3.72E-04	3.65E+00	4.54E-02	1.07E-03

is several orders of magnitude different from AO-SCA, and the algorithm performance is not stable enough. To sum up, AO-SCA has advantages over other improved AO algorithms.

## 5 CONCLUSION

To address the shortcomings of the AO algorithm, we proposed an Aquila optimizer algorithm(AO) using a hybrid search strategy named AO-SCA. Compared with the AO algorithm, the improved algorithm generates the initial population through Singer chaotic mapping method, so that the initial population can cover the solution space uniformly and comprehensively. The combination of the exploration phase of AO and the sine cosine strategy enhances the local search ability of AO algorithm and improves the accuracy of AO. In order to improve the convergence speed and global exploration capability of AO algorithm, the nonlinear sine learning factor is also introduced. Through numerical experiments, it is verified that the algorithm in this paper has been significantly improved in terms of optimization accuracy and convergence speed, and the ability to jump out of local optimum has been enhanced. In the subsequent research, we will consider applying the algorithm in feature selection, node location in wireless sensor networks, image segmentation and other applications.

## REFERENCES

- [1] Ralf Salomon. 1996. Re-evaluating genetic algorithm performance under coordinate rotation of benchmark functions. A survey of some theoretical and practical aspects of genetic algorithms. *BioSystems* 39, 3 (1996), 263–278. [https://doi.org/10.1016/0303-2647\(96\)01621-8](https://doi.org/10.1016/0303-2647(96)01621-8)
- [2] Kenneth V Price. 1996. Differential evolution: a fast and simple numerical optimizer. In *Proceedings of North American fuzzy information processing*. IEEE, 524–527. <https://doi.org/10.1109/NAFIPS.1996.534790>
- [3] James Kennedy and Russell Eberhart. 1995. Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks*, Vol. 4. IEEE, 1942–1948. <https://doi.org/10.1109/ICNN.1995.488968>
- [4] Ying Tan and Yuanchun Zhu. 2010. Fireworks algorithm for optimization. In *International conference in swarm intelligence*. Springer, Heidelberg, 355–364. [https://doi.org/10.1007/978-3-642-13495-1\\_44](https://doi.org/10.1007/978-3-642-13495-1_44)
- [5] Ali Asghar Heidari, Seyedali Mirjalili, Hossam Faris, Ibrahim Aljarah, Majdi Mafarja, and Huiling Chen. 2019. Harris hawks optimization: Algorithm and applications. *Future generation computer systems* 97 (2019), 849–872.
- [6] Seyedali Mirjalili and Andrew Lewis. 2016. The whale optimization algorithm. *Advances in engineering software* 95 (2016), 51–67. <https://doi.org/10.1016/j.advengsoft.2016.01.008>
- [7] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. 2014. Grey wolf optimizer. *Advances in engineering software* 69 (2014), 46–61.
- [8] Ali Sadollah, Hadi Eskandar, Ho Min Lee, Joong Hoon Kim, et al. 2016. Water cycle algorithm: a detailed standard code. *SoftwareX* 5 (2016), 37–43.
- [9] Seyedali Mirjalili, Amir H Gandomi, Seyedeh Zahra Mirjalili, Shahrzad Saremi, Hossam Faris, and Seyed Mohammad Mirjalili. 2017. Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems. *Advances in engineering software* 114 (2017), 163–191.
- [10] Zhibin Nie, Xiaobing Yang, Shihong Gao, Yan Zheng, Jianhui Wang, and Zhanshan Wang. 2016. Research on autonomous moving robot path planning based on improved particle swarm optimization. In *2016 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2532–2536. <https://doi.org/10.1109/CEC.2016.7744104>
- [11] Md Anisul Islam, Yuvraj Gajpal, and Tarek Y ElMekkawy. 2021. Hybrid particle swarm optimization algorithm for solving the clustered vehicle routing problem. *Applied Soft Computing* 110 (2021), 107655.
- [12] Laith Abualigah, Dalia Yousri, Mohamed Abd Elaziz, Ahmed A Ewees, Mohammed AA Al-Qaness, and Amir H Gandomi. 2021. Aquila optimizer: a novel meta-heuristic optimization algorithm. *Computers & Industrial Engineering* 157 (2021), 107250.
- [13] Yu-Jun Zhang, Yu-Xin Yan, Juan Zhao, and Zheng-Ming Gao. 2022. AOAAO: The hybrid algorithm of arithmetic optimization algorithm with aquila optimizer. *IEEE Access* 10 (2022), 10907–10933. <https://doi.org/10.1109/ACCESS.2022.3144431>
- [14] Yujun Zhang, Yuxin Yan, Juan Zhao, and Zhengming Gao. 2021. Chaotic map enabled algorithm hybridizing Hunger Games Search algorithm with Aquila Optimizer. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*. VDE, 1–5.
- [15] Seyedali Mirjalili. 2016. SCA: a sine cosine algorithm for solving optimization problems. *Knowledge-based systems* 96 (2016), 120–133.
- [16] Wenbo Zhang, Xiaoteng Yang, Kaiguang Wang, et al. 2022. An Improved Gray Wolf Optimization Algorithm Based on Levy Flight and Adaptive Strategies. In *2022 International Conference on Networking and Network Applications (NaNA)*. IEEE, 448–453.

# Comparative Research on Embedding Methods for Video Knowledge Graph

Zhihong Zhou

School of Artificial Intelligence and Big Data, Hefei  
University Anhui, China  
hfuuzhzhou@163.com

Hui Ding

School of Artificial Intelligence and Big Data Hefei  
University Anhui, China  
hugharchive@163.com

Qiang Xu\*

School of Artificial Intelligence and Big Data Hefei  
University Anhui, China  
xuqiang@hfu.edu.cn

Shengwei Ji

School of Artificial Intelligence and Big Data Hefei  
University Anhui, China  
jisw@hfu.edu.cn

## ABSTRACT

In the video recommendation scenario, knowledge graphs are usually introduced to supplement the data information between videos to achieve information expansion and solve the problems of data sparsity and user cold start. However, there are few high-quality knowledge graphs available in the field of video recommendation, and there are many schemes based on knowledge graph embedding, which have different effects on recommendation performance and bring difficulties to researchers. Based on the streaming media video website data, this paper constructs knowledge graphs of two typical scenarios (i.e., sparse distribution scenarios and dense distribution scenarios). Moreover, six state-of-the-art knowledge graph embedding methods are analyzed based on extensive experiments from three aspects: data distribution type, data set segmentation method, and recommended quantity range. Comparing the recommendation effect of knowledge graph embedding methods. The experimental results demonstrate that: in the sparse distribution scenario, the recommendation effect using TransE is the best; in the dense distribution scenario, the recommendation effect using TransE or TranD is the best. It provides a reference for subsequent researchers on how to choose knowledge map embedding methods under specific data distribution.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Knowledge representation and reasoning;

## KEYWORDS

Video recommendation, Knowledge graph, Graph embedding

\*Corresponding author. E-mail address: xuqiang@hfu.edu.cn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590049>

## ACM Reference Format:

Zhihong Zhou, Qiang Xu, Hui Ding, and Shengwei Ji. 2023. Comparative Research on Embedding Methods for Video Knowledge Graph. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590049>

## 1 INTRODUCTION

With the development of modern digital technology and Internet technology, the production cost and dissemination cost of video content is reduced. Because the user traffic base is large, making video recommendation gradually occupy an important position in the recommendation field. Video recommendation[1] can obtain resources that are more in line with user interests at the individual level, can increase user traffic at the enterprise level to promote economic growth, and can provide guidance for online public opinion and foster a positive and healthy network culture at the social level. Therefore, it is the most urgent core problem to be solved in today's video industry to use the exact recommendation method to obtain videos that fit the user's personal preferences from massive resource data.

In the development process of the recommendation algorithm [2], it mainly includes the following commonly used algorithms: the content-based recommendation algorithm [3] collects relevant information about users, items and user-item interactions, calculates the vector similarity between users and items and Sorting to generate a recommendation list. Collaborative filtering recommendation algorithms [4] collects the preference data of users and items and calculates the similarity between users and users or items and items, and provides recommendations based on similarity and user history information. The knowledge-based recommendation algorithm [5] collects user-related needs and preference information, and then uses the similarity measurement standard to provide users with recommended solutions. The recommendation algorithm based on association rules [6] mines frequent item sets from the user's purchase collection data, obtains strong association rules according to frequent item sets, and finally performs other items in the strong association rules. The model-based recommendation algorithm [7] establishes a recommendation algorithm model for target users by integrating machine learning algorithms, and finally applies the algorithm model to obtain recommendation results.

In recent years, utilizing knowledge graphs as auxiliary information for recommendation tasks has attracted much attention. The introduction of knowledge graphs can not only alleviate problems such as data cold start, but also provide interpretability for the recommended results. The recommendation algorithm based on the knowledge map [8] mainly includes the following methods: the Embedding-based method [9] uses the information-rich knowledge map to map the entities and relationships to a low-dimensional vector space by applying the KGE method, the similarity between entity vectors is calculated to generate recommendation results; the path-based method [10] uses graph data to induce meta-paths and extract semantic paths, and uses path-based connectivity similarity to embed paths into low-dimensional vector spaces and Obtain the vector representation of the path, and calculate the semantic similarity between entities to obtain the recommendation result. The joint-based recommendation algorithm [11] combines the embedding-based method with the path-based method, and combines the vector representation of entities and relationships with the semantic information of path connectivity, mainly using items that have interacted with users in the path information.

In the video field, the knowledge map is integrated to build a recommendation model, and the user's interests and preferences are mined through the feature information of the video in the user's viewing record, to solve the problems caused by data sparsity and cold start. However, the existing video domain knowledge graph recommendation algorithms have the following problems: 1) Most video knowledge graphs are constructed from the Internet, and the knowledge graph information is missing. 2) There is a lack of specific research on how to select the appropriate method from many video knowledge map embedding methods in the case of different data distribution forms and recommendation scenarios. These problems have brought difficulties to researchers.

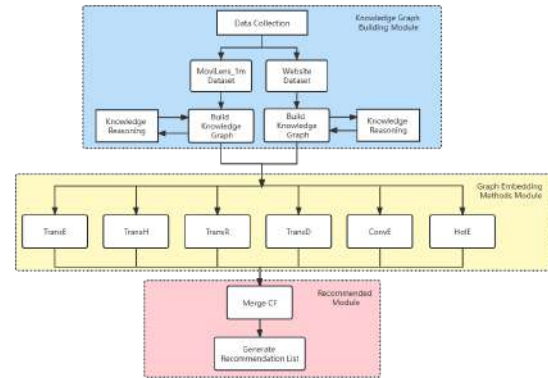
Aiming at the above problems, this paper builds a video domain knowledge map on streaming media websites by collecting, preprocessing and completing data information of the websites. Experimental comparison and analysis of the experimental results are carried out for the current mainstream knowledge graph embedding methods.

The contributions of this paper are:

1 We collect streaming video website data and build a knowledge map in the video field. The graph contains 6 node types, 10 relationship types, and the number of triples is 210,000. Graphs can also be applied to research in other fields, such as: knowledge question answering, community discovery and user category aggregation.

2 We set a variety of comparison indicators such as the distribution scene of the data set, the division method of user-idol data, the number of recommended results, and the type of algorithm. Then the experimental results can be ensured to be sufficient.

3 We summarize and analyze the experimental results, and the results show that: in the sparse data, TransE fusion random division performs best, in the dense data, when the number of recommendations is less than 10, TransE fusion proportional division performs best and the number of recommendations exceeds Beyond 10, TransD fusion random division performs best. The experimental results facilitates researchers to use the corresponding knowledge map embedding method under different data distribution conditions and conduct further research.



**Figure 1: The flow chart of the recommendation model based on the knowledge graph embedding method, including the knowledge graph building module, the knowledge graph embedding method module and the recommendation module**

## 2 RELATED WORK

### 2.1 Recommendation Model based on Knowledge Graph Embedding Method

The implementation process of the recommendation model of the knowledge graph embedding method is generally divided into three modules: (1) Build an ontology graph of the knowledge graph based on the recommended domain. Collecting data interaction and data feature information, and expressing the data information in the form of triples; (2) Using the knowledge graph embedding to embed the triplet data in the graph and obtains the vector representation of entities and relationships in the triplet; (3) Calculating the score function value between the user and the recommendation target to obtain the recommendation list.

The model implementation process is shown in Figure 1.

### 2.2 Knowledge Graph Building Module

Knowledge graph is a technical method that uses graph structure to describe knowledge and association relationship. The knowledge graph associates scattered data through relationships, contains rich semantic information, and supports data mining and analysis. Due to the powerful information storage capabilities of knowledge graphs, many types of knowledge graph platforms have emerged at home and abroad, such as Atlas knowledge graph platform, Antuo knowledge graph platform, Baidu knowledge platform, etc. These knowledge graph platforms build large-scale domain knowledge graphs by fusing, associating, annotating, and knowledge-based processing of massive multi-source heterogeneous data.

However, when these knowledge graphs are applied in specific fields, they have not exerted their due effects. The main reason is that the large-scale knowledge graph contains a lot of knowledge outside the specific domain, which will cause other knowledge to cause data interference to the knowledge of the specific domain. At present, the number of knowledge graphs in the video field is small, and there are differences in definitions between graph data, which are not suitable for data fusion or direct use. Therefore, this paper

adopts a top-down method to collect user data information and video information of streaming media websites. The missing part of data is supplemented from other data sources, eliminating the interference of irrelevant information, and building a knowledge map in the video field.

### 2.3 Knowledge Graph Embedding Method

Knowledge graph embedding method is to embed the entities and relationships in the triples into the vector representation space, and obtain the digital vector representation of the map data. Vectorization can convert complex relational texts into digital calculations in order to build mathematical models to process data, thereby realizing functions such as user recommendation and entity prediction. Commonly used knowledge graph embedding methods include the following:

**TransE**[12]: Embed all the constituent elements in the triplet  $(h, r, t)$  into a low-dimensional vector space to obtain the vector representation of the elements. The vector relationship between elements satisfies that the head vector  $h$  plus the relationship vector  $r$  is approximately equal to the tail vector  $t$ . In order to minimize the distance between  $h + r$  and  $t$ . The calculation process is shown in formula (1).

$$\mathcal{L} = \sum_{(h,r,t) \in \Delta} \sum_{(h',r',t') \in \Delta'} [f_r(h, t) + \gamma - f_{r'}(h', t')]_+ \quad (1)$$

In the above equation,  $\Delta$  represents the correct triplet, and  $\Delta'$  represents the wrong triplet, where the wrong triplet is obtained by randomly replacing the head entity or tail entity in the correct triplet.  $\gamma$  is the margin hyperparameter, which realizes the interval correction between positive samples and negative samples.

**TransH**[13]: To avoid the same vector representation of different entities on the same relation, a relation vector  $d_r$  and a hyperplane  $w_r$  are used to represent relations in triplets. Project the head and tail entities onto the hyperplane  $w_r$ , calculate the distance between the projected vectors, and make the distance close to the vector  $d_r$ . The calculation process is shown in formula (2):

$$f_r(h, t) = ||(h - w_r^T h w_r) + d_r - (t - w_r^T t w_r)||_2^2 \quad (2)$$

In the above equation,  $w_r^T h w_r$  and  $w_r^T t w_r$  represent the mapping vectors of the head vector  $h$  and the tail vector  $t$  on the hyperplane  $w_r$  respectively.

**TransR**[14]: Since the common entity features are not enough to represent the relationship as a vector, the entity space and relation space are constructed separately. Map both the head and tail entities to the relational space, so that the entities can complete the translation process in the relational space. The calculation process is shown in formulas (3).

$$f_r(h, t) = ||h M_r + r - t M_r||_2^2 \quad (3)$$

In the above equation,  $h M_r$  and  $t M_r$  represent the mapping vectors of the head vector  $h$  and the tail vector  $t$  on the relational space  $M_r$  respectively.

**TransD**[15]: Since the head and tail entities of a relation are often quite different, entities need to be mapped into different semantic spaces. The projection of entities and relationships is not enough to rely solely on relationships for reasoning, it needs to

be determined jointly by entities and relationships. The calculation process is shown in formulas (4).

$$f_r(h, t) = ||M_{rh}h + r - M_{rt}t||_2^2 \quad (4)$$

In the above equation,  $M_{rh}h$  and  $M_{rt}t$  respectively represent the projection vectors of the head vector  $h$  and the tail vector  $t$  in their respective projection matrices

**ConvE** [16]: Since the multi-layer architecture is prone to overfitting, and the information extracted by the 1-dimensional vector is limited. The use of 2-dimensional convolutions captures more feature interactions between embeddings. The widespread use of multi-layer convolutions also enables the establishment of robust methods to control overfitting problems. The calculation process is shown in formula (5).

$$\Psi_r(e_s, e_o) = f(\text{vec}(f([\bar{e}_s; \bar{r}_r] * \omega)) W) e_o \quad (5)$$

In the above equation,  $\bar{e}_s, \bar{r}_r$  represents the two-dimensional reshape of  $e_s, r_r$ ,  $*$  represents the convolution operation,  $\omega$  is the filter of the convolution layer.

**Hole**[17]: Since embedding models are always limited in scale, computationally efficient algorithms are not capable enough to capture information. Holographic embeddings with circular correlations are used to learn and obtain combinatorial vector representations of binary relations. Semantic information between entities can also be better obtained under efficient computing power.

The above six embedding methods do not take into account the impact on the recommendation effect of the embedding method due to the sparsity of the interactive data distribution, which is not conducive to the research on subsequent recommendation tasks under the premise of selecting the triple vector representation. Therefore, this paper compares the above six knowledge map embedding methods experimentally and chooses which knowledge map embedding method has the best effect under different data scenarios.

### 2.4 Recommendation Module

Commonly used embedding method recommendations mostly calculate the inner product of user embedding and item embedding, and use the calculated result as the recommended score. For example: Xiang Wang et al. [18] proposed a method of knowledge graph attention network KGAT by building a collaborative knowledge graph CKG of user-item interaction hybrid knowledge graph. The triplet data adopts the TransR structure to represent the vector of the entity, based on the attention mechanism and propagation update vector representation, calculate the inner product of the user vector and the recommended target vector as the recommendation score for recommendation. Tu K et al. [19] adopted the traditional knowledge graph representation method TransH to learn the entity representation of nodes and relations, and captured the global similarity through a graph convolutional neural network. Set the user's receptive field to capture the local preference of the target, and fuse the target-dependent representation with the global entity representation for recommendation.

The above recommendation does not take into account user interaction data, and the content of the video recommendation is to recommend the video itself. In this paper, an item-based collaborative filtering method is added to generate recommendation

results based on user interaction data. Mining the preference relationship between users and video participating stars to realize star recommendation for users.

### 3 DOMAIN KNOWLEDGE GRAPH CONSTRUCTION

This section mainly introduces the process of building a domain knowledge map. The process is divided into three steps: preprocessing the original data; designing the ontology structure for the data types and relationships involved in the field; and filling the preprocessed data into the graph structure in the form of triples. The detailed description is distributed in 3.1, 3.2, 3.3.

#### 3.1 Data Preprocessing

The composition of the data in this article is mainly divided into interactive data generated by users watching videos and various attributes of the videos themselves. Most of the data obtained through current methods have some problems such as incomplete data and unclear data correspondence, so we need to preprocess the data and analyze the type of data and whether the information is complete.

Due to the large amount of data, manual processing is time-consuming and laborious, so the script language is used to automatically extract the null value part of the data and the data that does not conform to this type of data type. For the extracted wrong data, the external data is used to import supplementary data and combined with artificial methods to complete and modify. For the problem of unclear data correspondence, we use the Jena[20] tool to perform directional data reasoning on the basis of the original data, which can effectively improve the effectiveness and accuracy of subsequent data query and reasoning, and improve the knowledge-based Application of graphs.

#### 3.2 Ontology Construction and Mapping

The construction methods of knowledge map ontology are generally divided into two types: top-down and bottom-up. The top-down construction refers to extracting ontology and pattern information from high-quality data with the help of structured data sources such as encyclopedia websites and adding them to the knowledge base. The bottom-up construction is to use certain technical means to extract information with high confidence from publicly collected data and add it to the knowledge base. The ontology construction of knowledge graphs for open domains generally adopts a bottom-up approach, using existing information to automatically extract concepts, concept types, and the relationship between concepts. For domain knowledge graphs, top-down methods are often used to create an ontology.

This paper uses a top-down approach to construct ontology and combines the general ontology construction tool Protégé [21] to standardize the ontology construction process. The specific process is as follows:

(1) Determine the core category and hierarchical relationship; the data in the field of user video interaction is constructed based on the existing video website resource data and user browsing records as knowledge sources. (2) Determine the semantic relationship between categories; carry out structural analysis on the video ontology data of the website, which is mainly divided into

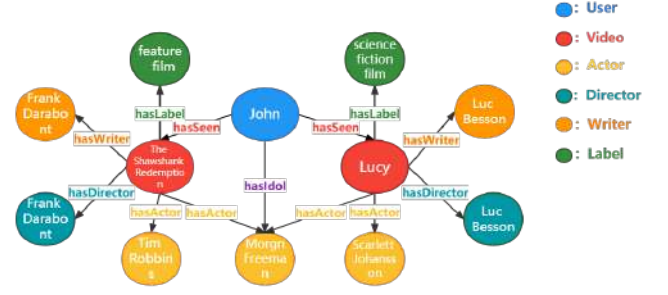


Figure 2: Partial data structure diagram of the knowledge graph in the video domain

the following semantic relationships between concepts: the whole and part relationship refer to the composition and composition relationship between concepts; the disjoint relationship refers to the relationship between concepts. It is a mutually contradictory relationship, and there is no situation where the same instance belongs to the two categories at the same time; the reverse relationship means that a certain positive relationship between concepts can obtain a reverse relationship that conforms to common sense; the operational relationship means that there is Operate or use relationships. (3) Determine the data attributes. After defining the ontology category, it is necessary to confirm the data information owned by the category.

#### 3.3 Graph Construction

According to the ontology structure, the preprocessed data is added to the relationship information contained in the data, and the data format is converted into the form of triples. The details of the data are shown in Table 1. The knowledge map uses the Neo4j tool to generate the corresponding graphic structure and display the data effect. In this paper, by calling the Neo4j operation module in python, according to the triple format information, write the corresponding Cypher statement according to the relationship category, and complete the import of entity data in the knowledge graph. Part of the map structure is shown in Figure 2.

## 4 EXPERIMENTAL DESIGN

### 4.1 Setup

*Evaluation.* The evaluation index calculation formula is shown in formulas (7)~(9). The precision rate indicates the probability that the real category is also a positive sample in the data predicted as a positive sample. The recall rate indicates the ratio of the data predicted as a positive sample to the real category positive sample data. The F1 value is used to comprehensively consider the two evaluation indicators.

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (8)$$

**Table 1: Knowledge Graph Entity and Relational Data Information**

Dataset	Entity	Number of Entity	Relation	Number of Relation	Dataset	Entity	Number of Entity	Relation	Number of Relation
MovieLens-1M	User	610	hasActor	24964	Video Site	User	41767	hasActor	28687
	Video	9748	hasSeen	56636		Video	8444	hasSeen	100832
	Actor	11291	hasWriter	771		Actor	7914	hasWriter	13173
	Director	4372	hasDirector	3498		Director	1735	hasDirector	10505
	Writer	7891	hasLabel	24860		Writer	637	hasLabel	22075
	Label	20	hasIdol	21201		Label	3402	hasIdol	47461

**Table 2: Dataset information distribution**

Dataset	Number of Users	Number of Videos	Number of User Video Interactions	Total number of triples
MovieLens-1M	610	9719	100832	375930
Streaming Video Site	41768	8445	56636	217799

*Experimental.* The hardware environment used in the experiment is: the experimental hardware processor model is Intel(R) Xeon(R) Platinum 8350C CPU @ 2.60GHz, the memory is 43GB, the GPU is RTX 3080 Ti, the video memory is 12GB, and the software environment is python3.8. The framework is Tensorflow2.5.0.

*Datasets.* The general data set MovieLens-1m and the user watching video recording data set provided by the streaming video website. The information distribution of the dataset is shown in Table 2.

*Baselines.* We conduct experimental comparisons of six state-of-the-art knowledge graph embedding methods. These methods are divided into translation-based knowledge representation models (TransE, TransH, TransR, TransD), semantic matching-based knowledge representation models (HoLE), and other embedding method models (ConvE).

## 4.2 Experimental Scheme

The experiment is mainly to compare the knowledge map embedded methods of TransE, TransH, TransR, TransD, ConvE, and HoLE. The generated recommendation list is fused with the recommendation list generated by the collaborative filtering algorithm in proportion. For the final generated recommendation result, Calculate and compare the accuracy according to the real results of users in the test set.

The factors for experimental comparison mainly include:

1) Comparison of distribution types of data sets: compare algorithms based on two scenarios of sparse user interaction data and dense user interaction data, and analyze the effects of different embedding methods in different scenarios of data distribution, which is helpful to explore the distribution types of different data sets impact on recommendation results.

2) Comparison of the division methods of the data set: the preference relationship inferred between the user and the star is randomly divided and proportionally divided into the training set and the test set, and the random division is to divide the total data set according

to the training set and The ratio of the test set is 4:1 for random segmentation, and the ratio division method divides the preference relationship data set according to the ratio of 4:1 between the training set and the test set. It is helpful to explore the impact of different partitioning schemes on the recommendation results.

3) Comparison of the number of recommendations: compare the final number of recommendations according to the five conditions of Top5, Top10, Top20, Top30, and Top50, which will help to draw the ROC curve of the corresponding algorithm by setting different recommendation numbers to further evaluate the recommendation Effect.

## 5 EXPERIMENTAL RESULTS & ANALYSIS

This section analyzes the experimental results, which are mainly divided into data-sparse scenarios and data-dense scenarios. The experimental results are shown in Tables (3, 4, 5, and 6). The detailed description is distributed in 5.1, 5.2.

### 5.1 Sparse Distribution Scenarios

It can be seen from Table 3 and Table 4 that in the segmentation method of the data set, the recommendation effect of random division is better than that of proportional division. Because there is less interaction data between users and celebrities in data sparse scenarios, the number of triples that users prefer to star is small. The number of users in the test set in the proportional division is the same as that in the training set, while the number of users in the test set in the random division is less. The base of the number of users becomes smaller, making the effect of random division better than equal proportion division.

TransE has the best performance, because the calculation method of TransE vector is simple, and the number of preferred stars in the test set is small, so there is no need to comprehensively consider other relationships. Therefore, the accuracy rate is better than other algorithms.

**Table 3: Experimental comparison of various models for streaming video websites (random division)**

Model	Metrics	Top5	Top10	Top20	Top30	Top50	Model	Metrics	Top5	Top10	Top20	Top30	Top50
TransE	P	<b>0.227</b>	<b>0.136</b>	<b>0.077</b>	<b>0.054</b>	<b>0.034</b>	TransR	P	0.219	0.125	0.067	0.046	0.029
	R	<b>0.591</b>	<b>0.709</b>	<b>0.803</b>	<b>0.842</b>	<b>0.885</b>		R	0.571	0.649	0.700	0.725	0.760
	F1	<b>0.328</b>	<b>0.229</b>	<b>0.141</b>	<b>0.101</b>	<b>0.065</b>		F1	0.317	0.209	0.123	0.087	0.056
TransD	P	0.223	0.134	0.075	0.052	0.033	ConvE	P	0.218	0.124	0.067	0.046	0.029
	R	0.582	0.699	0.781	0.814	0.854		R	0.567	0.643	0.693	0.716	0.747
	F1	0.323	0.225	0.137	0.098	0.063		F1	0.315	0.207	0.121	0.086	0.055
TransH	P	0.226	0.132	0.072	0.050	0.031	HolE	P	0.224	0.136	0.077	0.054	0.034
	R	0.588	0.686	0.747	0.774	0.809		R	0.583	0.709	0.801	0.842	0.884
	F1	0.326	0.221	0.131	0.093	0.060		F1	0.323	0.228	0.140	0.101	0.065

**Table 4: Experimental comparison of various models for streaming video websites (proportional division)**

Model	Metrics	Top5	Top10	Top20	Top30	Top50	Model	Metrics	Top5	Top10	Top20	Top30	Top50
TransE	P	<b>0.178</b>	<b>0.103</b>	<b>0.057</b>	<b>0.040</b>	<b>0.025</b>	TransR	P	0.173	0.094	0.050	0.034	0.021
	R	<b>0.626</b>	<b>0.724</b>	<b>0.807</b>	<b>0.841</b>	<b>0.880</b>		R	0.609	0.661	0.700	0.719	0.750
	F1	<b>0.277</b>	<b>0.180</b>	<b>0.107</b>	<b>0.076</b>	<b>0.049</b>		F1	0.269	0.164	0.093	0.065	0.041
TransD	P	0.176	0.101	0.055	0.038	0.024	ConvE	P	0.171	0.092	0.049	0.033	0.021
	R	0.621	0.714	0.781	0.811	0.851		R	0.603	0.652	0.691	0.708	0.735
	F1	0.274	0.177	0.103	0.073	0.047		F1	0.266	0.162	0.091	0.064	0.041
TransH	P	0.177	0.099	0.053	0.037	0.023	HolE	P	0.176	0.101	0.057	0.040	0.025
	R	0.626	0.699	0.749	0.774	0.805		R	0.620	0.712	0.799	0.840	0.880
	F1	0.277	0.173	0.099	0.070	0.044		F1	0.274	0.177	0.106	0.076	0.049

**Table 5: Experimental comparison of various models for MovieLens-1M (random division)**

Model	Metrics	Top5	Top10	Top20	Top30	Top50	Model	Metrics	Top5	Top10	Top20	Top30	Top50
TransE	P	<b>0.137</b>	0.112	0.102	0.095	0.083	TransR	P	0.127	0.109	0.101	0.094	0.082
	R	<b>0.039</b>	0.064	0.116	0.162	0.236		R	0.036	0.062	0.114	0.159	0.233
	F1	<b>0.060</b>	0.081	0.109	0.120	0.123		F1	0.056	0.079	0.107	0.118	0.122
TransD	P	0.136	<b>0.122</b>	<b>0.109</b>	<b>0.100</b>	<b>0.087</b>	ConvE	P	0.103	0.080	0.075	0.073	0.066
	R	0.038	<b>0.069</b>	<b>0.123</b>	<b>0.170</b>	<b>0.246</b>		R	0.029	0.045	0.085	0.124	0.188
	F1	0.060	<b>0.088</b>	<b>0.116</b>	<b>0.126</b>	<b>0.128</b>		F1	0.046	0.058	0.080	0.092	0.098
TransH	P	0.121	0.102	0.094	0.087	0.077	HolE	P	0.131	0.111	0.101	0.093	0.082
	R	0.034	0.058	0.107	0.149	0.219		R	0.037	0.063	0.115	0.159	0.232
	F1	0.054	0.074	0.100	0.110	0.114		F1	0.058	0.080	0.108	0.118	0.121

## 5.2 Dense Distribution Scenarios

It can be seen from Table 5 and Table 6 that in the segmentation method of the data set, when the number of recommendations is less than 10, random segmentation performs best, and when the number of recommendations is greater than 10, proportional segmentation performs best. Because in the data-intensive scene, the number of users whose favorite celebrities are more than 10 far exceeds the number of users who are less than 10, so the ratio can be fully obtained within the Top 10. Data information. Due to the decrease in the number of users outside the Top 10, the base of the number of users becomes smaller, making the effect of random division better than proportional division.

When the number of recommendations is small, there are few connection paths between users and stars, and the calculation

method of TransE can accurately obtain these star entities; but when the number of recommendations is

large, the calculation method of TransE makes a user have a preference relationship. Do not use star entities with similar vector representations, which will reduce the accuracy of recommendation. In view of the large number of recommendations, TransD, a method that can solve the many-to-many relationship, can also make the vector representation of the star entity closer to the actual meaning when there are multiple connection paths between the star and the user, and the accuracy of the recommendation has also been improved.

**Table 6: Experimental comparison of various models for MovieLens-1M (proportional division)**

Model	Metrics	Top5	Top10	Top20	Top30	Top50	Model	Metrics	Top5	Top10	Top20	Top30	Top50
TransE	P	<b>0.150</b>	<b>0.124</b>	0.105	0.093	0.080	TransR	P	0.146	0.121	0.102	0.092	0.078
	R	<b>0.047</b>	<b>0.077</b>	0.131	0.173	0.249		R	0.046	0.076	0.127	0.172	0.242
	F1	<b>0.072</b>	<b>0.095</b>	0.117	0.121	0.121		F1	0.070	0.093	0.113	0.120	0.118
TransD	P	0.148	0.122	<b>0.106</b>	<b>0.094</b>	<b>0.080</b>	ConvE	P	0.136	0.109	0.092	0.082	0.070
	R	0.046	0.076	<b>0.133</b>	<b>0.176</b>	<b>0.249</b>		R	0.042	0.068	0.115	0.154	0.219
	F1	0.070	0.094	<b>0.118</b>	<b>0.123</b>	<b>0.121</b>		F1	0.065	0.084	0.102	0.107	0.107
TransH	P	0.142	0.116	0.100	0.089	0.076	HolE	P	0.150	0.124	0.105	0.092	0.079
	R	0.044	0.072	0.125	0.167	0.238		R	0.047	0.077	0.131	0.173	0.245
	F1	0.068	0.089	0.111	0.116	0.115		F1	0.072	0.095	0.117	0.120	0.119

## 6 CONCLUSION

In this paper, we use six knowledge map embedding methods (TransE, TransD, TransH, TransR, ConvE, HolE) to fuse the collaborative filtering algorithm to generate a recommendation list, and evaluate the effect of the generated recommendation list according to the test set data. Based on the data results, we conclude that in the sparse distribution scenario, the recommendation effect of using the TransE model fusion random division method is the best. In the dense distribution scenario, when the number of recommendations is small, the recommendation effect of using TransE fusion proportional division method is the best. When the number of recommendations is large, the recommendation effect of using TransD fusion random division method is the best. When choosing a knowledge graph embedding method, for different data distribution scenarios, using the corresponding model algorithm can effectively improve the recommendation performance of the algorithm.

The shortcoming of this paper is that the data involved in the training only collected data sets with large differences in the distribution of the two data types, and the comparison results lacked multiple data comparisons. In the future, we will consider adding data sets with multiple distribution types to improve the data results.

## REFERENCES

- [1] Davidson J, Liebald B, Liu J, *et al.* The YouTube video recommendation system[C]//Proceedings of the fourth ACM conference on Recommender systems. 2010: 293-296
- [2] Dhelim S, Aung N, Bouras M A, *et al.* A survey on personality-aware recommendation systems[J]. Artificial Intelligence Review, 2022, 55(3): 2409-2454
- [3] Lops P, Jannach D, Musto C, *et al.* Trends in content-based recommendation[J]. User Modeling and User-Adapted Interaction, 2019, 29(2): 239-249
- [4] Koren Y, Rendle S, Bell R. Advances in collaborative filtering[J]. Recommender systems handbook, 2022: 91-142
- [5] Tarus J K, Niu Z, Mustafa G. Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning[J]. Artificial intelligence review, 2018, 50: 21-48
- [6] Sola D, Meilicke C, Aa H, *et al.* A rule-based recommendation approach for business process modeling[C]//International Conference on Advanced Information Systems Engineering. Springer, Cham, 2021: 328-343
- [7] Cheng H T, Koc L, Harmsen J, *et al.* Wide & deep learning for recommender systems[C]//Proceedings of the 1st workshop on deep learning for recommender systems. 2016: 7-10
- [8] Guo Q, Zhuang F, Qin C, *et al.* A survey on knowledge graph-based recommender systems[J]. IEEE Transactions on Knowledge and Data Engineering, 2020
- [9] Zhang F, Yuan N J, Lian D, *et al.* Collaborative knowledge base embedding for recommender systems[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016: 353-362
- [10] Zhao H, Yao Q, Li J, *et al.* Meta-graph based recommendation fusion over heterogeneous information networks[C]//Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017: 635-644
- [11] Wang H, Zhao M, Xie X, *et al.* Knowledge graph convolutional networks for recommender systems[C]//The world wide web conference. 2019: 3307-3313
- [12] Bordes A, Usunier N, Garcia-Duran A, *et al.* Translating embeddings for modeling multi-relational data[J]. Advances in neural information processing systems, 2013, 26
- [13] Wang Z, Zhang J, Feng J, *et al.* Knowledge graph embedding by translating on hyperplanes[C]//Proceedings of the AAAI conference on artificial intelligence. 2014, 28(1)
- [14] Lin Y, Liu Z, Sun M, *et al.* Learning entity and relation embeddings for knowledge graph completion[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015
- [15] Ji G, He S, Xu L, *et al.* Knowledge graph embedding via dynamic mapping matrix[C]//Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers). 2015: 687-696
- [16] Dettmers T, Minervini P, Stenetorp P, *et al.* Convolutional 2d knowledge graph embeddings[C]//Proceedings of the AAAI conference on artificial intelligence. 2018, 32(1)
- [17] Nickel M, Rosasco L, Poggio T. Holographic embeddings of knowledge graphs[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30(1)
- [18] Wang X, He X, Cao Y, *et al.* Kgat: Knowledge graph attention network for recommendation[C]//Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. 2019: 950-958
- [19] Tu K, Cui P, Wang D, *et al.* Conditional graph attention networks for distilling and refining knowledge graphs in recommendation[C]//Proceedings of the 30th ACM International Conference on Information & Knowledge Management. 2021: 1834-1843
- [20] Carroll J J, Dickinson I, Dollin C, *et al.* Jena: Implementing the Semantic Web Recommendations[C]// International World Wide Web Conference on Alternate Track Papers & Posters. ACM, 2004
- [21] Gennari J H, Musen M A, Fergerson R W, *et al.* The evolution of Protégé: an environment for knowledge-based systems development[J]. International Journal of Human-computer studies, 2003, 58(1): 89-123

# Performance Evaluation of an Extradosed Cable-Stayed Bridge with Corrugated Web based on Machine Learning Algorithms

Zeyu Du  
China Railway Construction  
Investment Group Co. LTD Shaanxi  
Company, Xi'an, China  
18623572476@139.com

Zhenhua Pan  
College of Water Resources and  
Architectural Engineering, Northwest  
A&F University, Yangling, China  
2022050880@nwsuaf.edu.cn

Zhihua Xiong\*  
College of Water Resources and  
Architectural Engineering, Northwest  
A&F University, Yangling, China  
xiongzhijia\_2013@126.com

Lei He  
CCCC First Highway Consultants  
Ltd., Xi'an, China  
1045651336@qq.com

Haipeng Wang  
China Railway Construction  
Investment Group Co. LTD Shaanxi  
Company, Xi'an, China  
15696262452@139.com

Houda Zhu  
College of Water Resources and  
Architectural Engineering, Northwest  
A&F University, Yangling, China  
zhuhd2020@nwfau.edu.cn

Jiangbo Wang  
CCCC First Highway Consultants  
Ltd., Xi'an, China  
36578579@qq.com

## ABSTRACT

Corrugated steel web is suitable for large-span extradosed cable-stayed bridge's design scheme. Live Load Structural Index (LLSI) is applied to evaluate the performance of the bridge with corrugated steel web. Parametric numeric models were built and investigated to explore the web height and weight's effect on the structural performance of an extradosed cable-stayed bridge. Machine learning model involving Particle Swarm Optimization BP neural network has been constructed to predict the correlation and validate the relationship between the structural variable and live load structural index.

## CCS CONCEPTS

• Applied computing; • Physical sciences and engineering; • Engineering; • Computer-aided design;

## KEYWORDS

extradosed cable-stayed bridge, machine learning, LLSI, steel corrugated web, BP

## ACM Reference Format:

Zeyu Du, Zhenhua Pan, Zhihua Xiong, Lei He, Haipeng Wang, Houda Zhu, and Jiangbo Wang. 2023. Performance Evaluation of an Extradosed Cable-Stayed Bridge with Corrugated Web based on Machine Learning Algorithms. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions.acm.org](https://permissions.acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590050>

(CACML 2023), March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590050>

## 1 INTRODUCTION

Extradosed cable-stayed bridges are widely built in China and Japan due to its economic cost and excellent structural performance [1-2]. Prestressed concrete box girder is usually applied in extradosed cable-stayed bridge, which is suitable for a span less than 200m. While as the span increases over 200m, corrugated web girder could be an alternative solution. Corrugated Steel Web (CSW) has the advantages of light weight, solid shear resistance and erection-friendly [3-4].

The cost of corrugated web scheme is an important factor which needs to be evaluated. This brings an issue that the cost of the girder and the structural integrity should keep a balance [5-6]. Previously, the authors have proposed an index called Live Load Structural Index (LLSI) to solve the similar problem in steel-concrete composite bridge [7]. This index considers both the cost of steel and the bridge's stiffness, which we will adopt in the analysis of the extradosed cable-stayed bridge with CSW. Machine learning algorithms are utilized in various fields [8-11], and they are also involved in cable-stayed bridge's deflection predictions, optimum design [12-13]. In terms of the multi-variables in extradosed cable-stayed bridge such as the cable force, girder weight and stiffness, BP neural network (BPNN) and its related optimization algorithms are found to be efficient in dealing with the nonlinear fitting problems [14-18].

This paper firstly implements the LLSI index into the evaluation of CSW of extradosed cable-stayed bridge. Parametric analysis is carried out to produce a data set for the machine learning training. Based on the obtained data, Particle Swarm Optimization BPNN (PSO-BPNN) is used to train the data set and the LLSI values of the extradosed cable-stayed bridge are calculated for the performance evaluation.



Figure 1: The elevation view of extradosed cable-stayed bridge

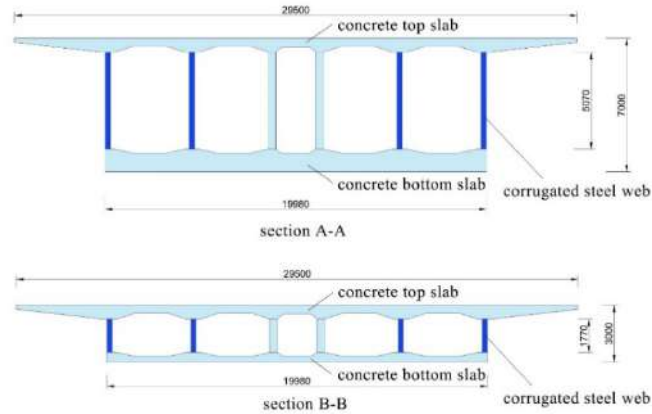


Figure 2: (a) Section A-A; (b) Section B-B

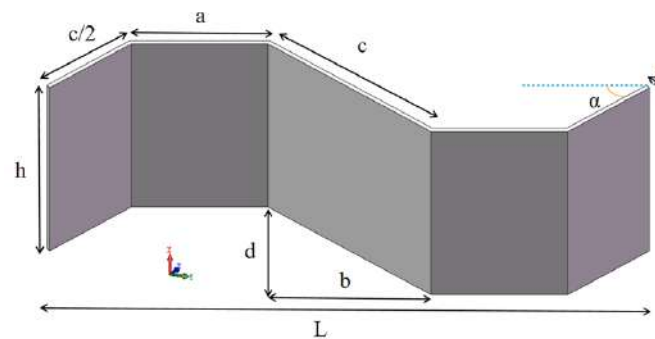


Figure 3: Parameter diagram of CSW

Table 1: The parameters of T1600 CSW

Type	a (mm)	b (mm)	c (mm)	d (mm)	L (mm)	$\alpha$ (°)	t (mm)
T1600	430	370	430	220	1600	30.7	30

## 2 ENGINEERING BACKGROUND AND DEFINITION OF LLSI

### 2.1 Introduction of the extradosed cable-stayed bridge

The bridge is a curved extra-dosed cable-stayed bridge with a span of (125+4×230+125) meters and a height of about 188.5 meters in

rigid frame system. The elevation layout is shown in Figure 1. The main girder of the bridge is in the form of variable section. The forms of approaching and midspan sections are shown in Figure 2, in which the dark blue part is the corrugated steel web. The size of corrugated steel web is shown in Figure 3 and Table 1.

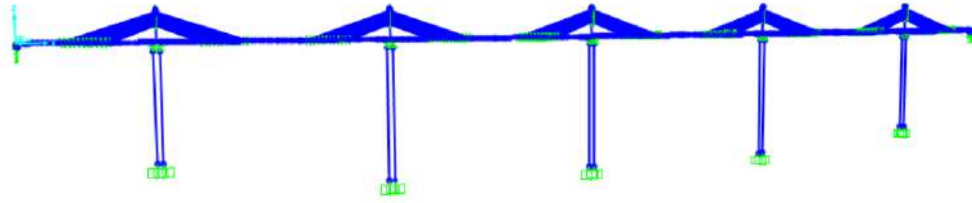


Figure 4: Finite Element Model

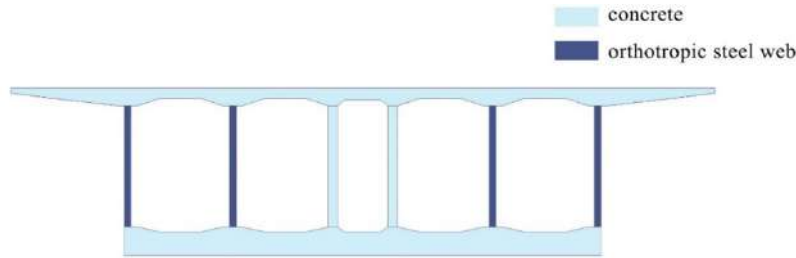


Figure 5: EOP section

## 2.2 Definition of LLSI

In order to comprehensively evaluate the structural performance and the cost of cable-stayed bridges with corrugated steel webs, Live Load Structural Index (LLSI) is introduced here. This index combines the steel consumption of the bridge structure and the deflection under the live load, and it takes the economy and overall stiffness into consideration, which is capable of a direct evaluation of the steel and steel-concrete composite bridge scheme. The equation is defined as follows:

$$I = \frac{\kappa}{\eta} \quad (1)$$

$$\kappa = \frac{L}{\Delta} \quad (2)$$

$$\eta = \frac{W}{L} \quad (3)$$

Where:  $I$  is in a unit of  $\text{m} \cdot \text{kN}^{-1}$ ;  $\kappa$  is the ratio of span to deflection, reflecting the overall stiffness of the structure;  $L$  is the calculated span. For multi-span continuous bridges,  $L$  is based on the maximum span, and the unit is meter;  $\Delta$  is the maximum deflection of the main span of the bridge under the live load, in unit m, where the live load refers to the highway - Class I-lane load according to Chinese code;  $\eta$  is the ratio of steel weight to span, indicating the weight of steel per unit span, reflecting the engineering cost of the structure;  $W$  is the gravity of the steel in the first span of the main span of the bridge, including the steel main girder, the transverse connection system, the ordinary reinforcement and the prestressed steel tendon in unit KN.

## 3 NUMERICAL CALCULATION

### 3.1 Numeric Model

A three-dimensional finite element model of the extradosed cable-stayed bridge was established by software CSiBridge. The main girder, pylons and piers of the calculation model were simulated by spatial beam elements, while the cables were simulated by truss elements. The bridge has a total of 665 elements, of which 336 are girder elements and 300 are cable elements. The finite element analysis model is shown in Figure 4.

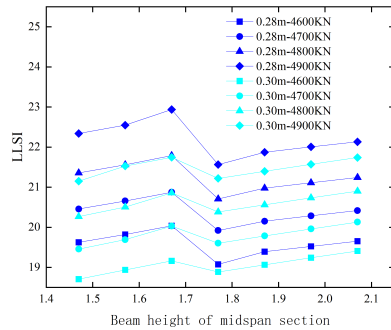
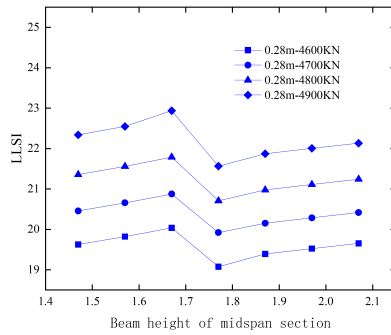
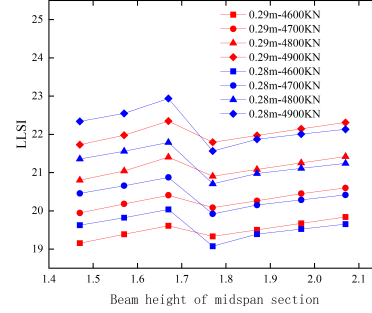
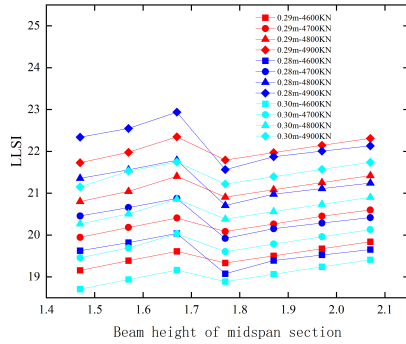
The refined model usually requires plenty of time, but if the corrugated steel web (CSW) is equivalent to the orthotropic steel plate, the model will be simplified significantly [3]. The model adopts the Equivalent Orthotropic Plate (EOP) approach here. The schematic diagram and data after equivalence are shown in Figure 5 and Table 2.

### 3.2 Influence of multi-variables on structural performance

In order to study the girder height, cable force, weight and other variables' effect on the structural performance of the extradosed cable-stayed bridge using CSW, parametric finite element analysis has been done by modifying the height and thickness of the corrugated steel web in the girder and adjusting the tension of each cable. LLSI was then obtained from the parametric results. A total of 84 groups of models are designed. The LLSI values have been categorized into each variable in Figure 6.

**Table 2: The parameters of T1600 CSW**

	Thickness $t(\text{mm})$	Elasticity modulus of x-axis $E_X(\text{MPa})$	Elasticity modulus of y-axis $E_Y(\text{MPa})$	shear modulus $G_{xy}(\text{MPa})$
CSW	30	$2.10e^5$	$2.10e^5$	$8.08e^4$
EOP	280	$3.29e^4$	195	3520

**Figure 6: Relationships between LLSI and variables: (a) 0.28m, 0.29m and 0.30m web thickness; (b) 0.28m web thickness; (c) 0.28m and 0.30m web thickness; (d) 0.28m and 0.29m web thickness**

the stiffness, so there is a falling stage after the rising. As the height increases continuously the influence of the web stiffness becomes predominant than the weight, however, LLSI will only reach the peak value in the first stage as the web height increases to almost 125%, which is obviously not an economic scheme. It is worth noting that when the thickness of EOP is 0.28m, the LLSI of the larger web height is always smaller than the smaller web height. According to Figure 6(a) and (b), the cable force is positively correlated with LLSI. When the cable force of each cable increases by 100KN, the LLSI increases by about  $0.8\text{m}\cdot\text{KN}^{-1}$ . When the thickness of EOP increases by 0.02m, LLSI wholly decreases as shown in Figure 6(c). LLSI is reduced due to the increased steel weight and the slight change in structural stiffness.

According to the LLSI analysis, the optimum solution accounting for both economic and structural needs is in the range that the web height is 1.67–4.97m and thickness of EOP is 0.28m. At the same time, with consideration of the relationship between LLSI and cable force, the cable force should be kept as large as the anchorage system in girder can resist.

#### 4 MACHINE LEARNING MODEL BASED ON PSO-BPNN

The structure of the BPNN applied in this paper is described as follows: 1) input layer with 8 input nodes corresponding to 8 indexes affecting LLSI: corrugated steel web equivalent thickness, main beam height in mid-span section, main beam height in top-pier section, steel volume in single variable section area, single steel weight, total steel weight, cable force, maximum deflection in main

Based on the relationship in Figure 6, the deflection of the bridge is within the allowable range. In terms of the overall stiffness, Figure 6(a) shows that LLSI fluctuates with the height of the web. In Figure 6(a), the first rising stage proves that the web stiffness is the main influencing factor at this time, and after reaching a certain weight, the influence of the web weight is gradually greater than

span; 2) output layer with 1 node corresponding to LLSI value ( $y$ ) and 1 hidden layer.

The initial parameters of the particle swarm optimization seeking algorithm are set as follows: the maximum number of particle iterations is 10, the particle swarm size is 10, the crossover probability is 0.2, the learning factors are all 2, the inertia weight is 0.6, the velocity range  $[-0.8, 0.8]$ , and the variation range of individual positions  $[-5, 5]$ .

The number of nodes in the hidden layer  $n$  directly affects the convergence speed and accuracy of the model, which can generally be determined by the empirical Eq. 5) as follows.

$$n = \sqrt{l + m} + c \quad (4)$$

where  $l$  is the number of nodes in the input layer,  $m$  is the number of nodes in the output layer, and  $c$  is a constant that takes values in the range of 1 to 10. Based on the empirical formula (5), it can be concluded that the number of nodes in the hidden layer of the model takes values in the range of 3 to 13.

In order to explore the most suitable number of hidden layer nodes, 70% of the whole data is taken as training data and 30% as training data.  $n=3, 5, 7, 9, 11, 13$ , and 8, respectively, several machine learning models for have been built for comparison, and the model with the smallest error is taken as the machine learning model to study the degree of influence of eight indicators on LLSI. The neural network training effect is shown in Figure 7. In order to quantify the error, the error quantification index  $Err$  is applied as:

$$Err = \sum_{i=1}^N |sim_i - test_i| (i = 1, \dots, N) \quad (5)$$

where  $sim_i$  is the predicted values in the test set,  $test_i$  is the actual value corresponding to it in the test set, and  $N$  is the number of data in the test set. The value of  $Err$  indicates the precision of the machine learning model, in which a trivial value presents a precise model.

It can be found that when the number of hidden layer nodes is set to 8,  $Err$  is the smallest and the training effect is the best as shown in Figure 7. Therefore, the hidden layer nodes are determined to be 8 in the model.

## 5 PREDICTION OF LLSI BASED ON MACHINE LEARNING

In order to study the degree of influence of eight variables on the LLSI, the average of each group of variables in the training set was taken as the benchmark value to obtain a set of benchmark values. Referring to the theory of control variables, the remaining benchmark values were kept constant. The corrugated steel web thickness of EOP ( $x_1$ ), the height of the main beam in the mid-span section ( $x_2$ ), the height of the main beam in the top-pier section ( $x_3$ ), the volume of steel in the single piece variable section area ( $x_4$ ), the single piece steel weight ( $x_5$ ), total steel weight ( $x_6$ ), cable force ( $x_7$ ), and maximum deflection of the main span ( $x_8$ ) were expanded by 20%, 40%, 60%, 80%, and 100% from the original ones, and the changes of LLSI were plotted in Figure 8. For example, in order to explore the influence of the change of corrugated steel web equivalent thickness  $x_1$  on LLSI,  $x_2 \sim x_8$  keep the base value unchanged and expand  $x_1$  by 20%, 40%, 60%, 80%, 100% on the

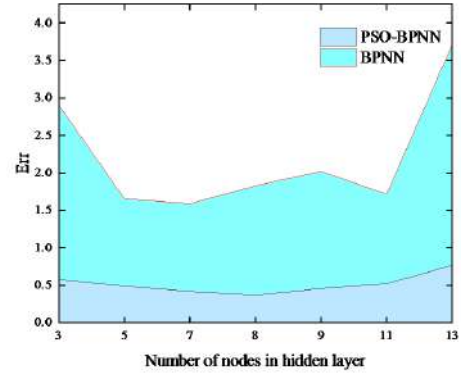


Figure 7: Training Effects of BPNN and PSO-BPNN

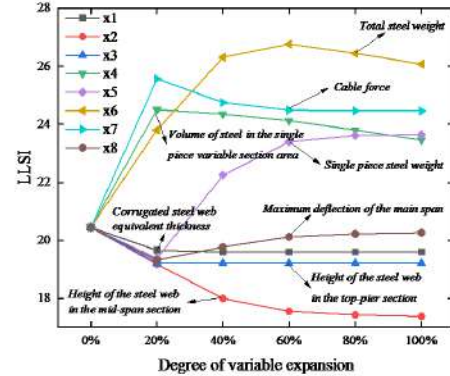


Figure 8: Prediction results by machine learning algorithms

basis of its base value, it is observed that the LLSI value decreases continuously with the expansion of  $x_1$  variable in Figure 8, which reflects the corrugated steel web equivalent thickness  $x_1$  has a positive correlation with LLSI. However, the effect of EOP on LLSI is small compared to that of the web height and weight. Obviously, the height of the CSW in the mid-span section ( $x_2$ ) is negatively correlated to LLSI. The total weight of steel ( $x_6$ ) has the greatest influence on LLSI, which is basically parabolic related to LLSI.

## 6 CONCLUSION

This paper introduces a novel index of Live Load Structural Index (LLSI) based on the engineering cost and overall stiffness of the corrugated steel web extradosed cable-stayed bridge. It is applied in the performance evaluation of corrugated steel web cable-stayed bridge, and the bridge is evaluated under the condition of changing the section size and the cable force. The following conclusions are drawn:

The influence of the total weight of steel and the overall stiffness of the bridge on LLSI varies proportionally, that is, their influence on the structural performance of the bridge is not constant. In the extradosed cable-stayed bridge with corrugated steel webs, the size of cable force has a great impact on the structural stiffness. The value of cable force should be reasonably set in consideration of the structural integrity and cost in practical engineering. The depth of mid-span section and the total weight of steel have the greatest influence on LLSI, with the former being negatively correlated to LLSI and the latter being parabolic correlated.

In this work, the stiffness and cable force's effects on LLSI are discussed. Other factors such as the span and the shape of corrugated web need to be further investigated in the future research.

## ACKNOWLEDGMENTS

The authors are grateful for the support of Science and Technology Project of Department of Transport of Shaanxi Province (Grant No. 20-40K), Elite Scholar Program of Northwest A&F University (Grant No. Z1010421003).

## REFERENCES

- [1] FUJINO, Y. and KAWAI, Y., 2016. Technical developments in structural engineering with emphasis on steel bridges in Japan. *Journal of JSCE*, 4(1), pp.211-226.
- [2] Zhang, L., Qiu, G. and Chen, Z., 2021. Structural health monitoring methods of cables in cable-stayed bridge: A review. *Measurement*, 168, p.108343.
- [3] Xiong Z., Hou X., Zhu H., Liu Y., Meng Y., Structural Performance and Cost Analysis of Multi-span Extradosed Cable-Stayed Bridge. *IABSE Congress Nanjing 2022*. 2022, 387-395, <https://doi.org/10.2749/nanjing.2022.0387>
- [4] Wang, S., Liu, Y., He, J., Xin, H. and Yao, H., 2019. Experimental study on cyclic behavior of composite beam with corrugated steel web considering different shear-span ratio. *Engineering Structures*, 180, pp.669-684.
- [5] Orcesi, A., Cremona, C. and Ta, B., 2018. Optimization of design and life-cycle management for steel-concrete composite bridges. *Structural Engineering International*, 28(2), pp.185-195.
- [6] Soliman, M. and Frangopol, D.M., 2015. Life-cycle cost evaluation of conventional and corrosion-resistant steel for bridges. *Journal of Bridge Engineering*, 20(1), p.06014005.
- [7] Xiong, Z., Li, J., Wang, S., Liu, Y. and Xin, H., 2018, February. Concrete filled tubular arch modified-VFT bridge and its LLSI analysis. In *2017 3rd International Forum on Energy, Environment Science and Materials (IFEESM 2017)* (pp. 1379-1382). Atlantis Press.
- [8] J. Wang, X. Xie, X. Cheng and Y. Wang, 2022. Improved density peaking algorithm for community detection based on graph representation learning, *Computer Systems Science and Engineering*, vol. 43, no.3, pp. 997–1008.
- [9] Ullah, F., Ullah S., Naeem M.R., Mostarda L., Rho S., and Cheng X., 2022. Cyber-Threat Detection System Using a Hybrid Approach of Transfer Learning and Multi-Model Image Representation. *Sensors* 22, no. 15, p.5883.
- [10] L. Yin, J. Feng, H. Xun, Z. Sun and X. Cheng, ,1 July-Sept. 2021. A Privacy-Preserving Federated Learning for Multiparty Data Sharing in Social IoTs, *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2706-2718.
- [11] Raheja, S., Kasturia, S., Cheng, X., & Kumar, M. 2021. Machine learning-based diffusion model for prediction of coronavirus-19 outbreak. *Neural computing & applications*, pp. 1–20.
- [12] Yue, Z.X., Ding, Y.L. and Zhao, H.W., 2021. Deep learning-based minute-scale digital prediction model of temperature-induced deflection of a cable-stayed bridge: case study. *Journal of Bridge Engineering*, 26(6), p.05021004.
- [13] Martins, A.M., Simões, L.M. and Negrão, J.H., 2020. Optimization of cable-stayed bridges: A literature survey. *Advances in Engineering Software*, 149, p.102829.
- [14] Hou, R. and Xia, Y., 2021. Review on the new development of vibration-based damage identification for civil engineering structures: 2010–2019. *Journal of Sound and Vibration*, 491, p.115741.
- [15] Najafzadeh, M., Barani, G.A. and Hessami-Kermani, M.R., 2015. Evaluation of GMDH networks for prediction of local scour depth at bridge abutments in coarse sediments with thinly armored beds. *Ocean Engineering*, 104, pp.387-396.
- [16] Xiong, Z., Li, J., Zhu, H., Liu, X. and Liang, Z., 2022. Ultimate Bending Strength Evaluation of MVFT Composite Girder by using Finite Element Method and Machine Learning Regressors. *Latin American Journal of Solids and Structures*, 19.
- [17] Cai, B., Lin, X., Fu, F. and Wang, L., 2022, October. Postfire residual capacity of steel fiber reinforced volcanic scoria concrete using PSO-BPNN machine learning. In *Structures* (Vol. 44, pp. 236-247). Elsevier.
- [18] Wang, Y. and Zhao, Y., 2022. Predicting bedrock depth under asphalt pavement through a data-driven method based on particle swarm optimization-back propagation neural network. *Construction and Building Materials*, 354, p.129165.

# An Adaptive Gradient Privacy-Preserving Algorithm for Federated XGBoost

Hongyi Cai  
Fuzhou University  
Minhou Xian, Fuzhou Shi, China  
hongyi.cai@qq.com

Jianping Cai  
Fuzhou University  
Minhou Xian, Fuzhou Shi, China

Lan Sun  
Fuzhou University  
Minhou Xian, Fuzhou Shi, China

## ABSTRACT

Federated learning (FL) is a novel machine learning framework in which machine learning models are built jointly by multiple parties. We investigate the privacy preservation of XGBoost, a gradient boosting decision tree (GBDT) model, in the context of FL. While recent work relies on cryptographic schemes to preserve the privacy of model gradients, these methods are computationally expensive. In this paper, we propose an adaptive gradient privacy-preserving algorithm based on differential privacy (DP), which is more computationally efficient. Our algorithm perturbs individual data by computing an adaptive gradient mean per sample and adding appropriate noise during XGBoost training, while still making the perturbed gradient data available. The training accuracy and communication efficiency of the model are guaranteed under the premise of satisfying the definition of DP. We show the proposed algorithm outperforms other DP methods in terms of prediction accuracy and approaches the lossless federated XGBoost model while being more efficient.

## CCS CONCEPTS

• Computing methodologies → Classification and regression trees; • Security and privacy → Privacy protections.

## KEYWORDS

federated learning, gradient boosting decision tree, differential privacy, security

### ACM Reference Format:

Hongyi Cai, Jianping Cai, and Lan Sun. 2023. An Adaptive Gradient Privacy-Preserving Algorithm for Federated XGBoost. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590051>

## 1 INTRODUCTION

In the era of big data, artificial intelligence technology enables people to extract valuable information from massive data and provide more efficient and accurate services. However, data security and user privacy issues have become pressing social concerns [10, 22].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590051>

Federated learning [17] is a collaborative machine learning framework where only a small amount of information is transmitted to build a model, ensuring that data remains in local areas. Current research primarily focuses on computationally complex deep learning models [3, 13], which often require high-performance hardware and have challenging-to-explain decision-making processes.

In contrast, gradient boosting decision tree (GBDT) [8] offers interpretability and efficiency during the training and reasoning processes for tabular data, without requiring high-performance hardware. We use XGBoost [4] as the efficient implementation of GBDT. In the federated environment, all parties collaboratively construct XGBoost by exchanging gradient information. However, there is still the risk of privacy leakage. The Attacker can restore the local data through reconstruction attack [15, 19] according to the intermediate parameters in the model training process.

To protect intermediate parameter privacy, researchers have relied on cryptography-based technologies like homomorphic encryption [5, 9] or multi-party computation [2, 7, 12]. However, these methods can be computationally expensive and slow. Differential privacy (DP) [6] as a mathematical-based privacy protection technology, with strict privacy definitions and high efficiency in calculation and communication. By adding random noise to data processing, DP can prevent attackers from inferring original data even if they access intermediate parameter information.

At present, studies have explored the use of DP for federated XGBoost training [14, 21, 23], including quantile sketch construction, leaf node split selection [11], and leaf node weight perturbation [16]. Additionally, the gradient in the XGBoost construction process also contains sensitive information. However, compared to cryptography methods such as homomorphic encryption that can directly encrypt a single element, DP often perturbs the statistical query of a set of data. It is challenging to add noise directly to a single data point.

To address this challenge, we propose an adaptive gradient privacy-preserving algorithm for federated XGBoost (AGPP-XGBoost) based on DP, which meets the definition of DP while retaining the availability of gradient data with minimal noise to reduce the model error.

## 2 PRELIMINARIES

### 2.1 Gradient Boosting Decision Tree / XGBoost

Gradient Boosting Decision Tree (GBDT) [8] is an ensemble learning algorithm. Each iteration process learns a new decision tree  $f(\mathbf{x})$  to minimize the objective function. Chen et al. [4] optimize the objective function with gradient of the loss function  $g_i = l'(y_i, \hat{y}_i^{t-1})$ ,

$h_i = l''(y_i, \hat{y}_i^{t-1})$  and regularization term  $\Omega(f)$ :

$$\mathcal{L}^t = \sum_{i=1}^n [l(y_i, \hat{y}_i^{t-1}) + g_i f_t(\mathbf{x}_i) + \frac{h_i}{2} f_t^2(\mathbf{x}_i)] + \sum_{k=1}^t \Omega(f_k) \quad (1)$$

In the training of XGBoost, it need choose the best split to determine the structure of tree and divide the sample set  $I$  into  $I_L$  and  $I_R$ , the gain that this scheme can provide is:

$$\mathcal{L}_{\text{split}} = \frac{1}{2} \left[ \frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (2)$$

where  $\lambda$  and  $\gamma$  are hyperparameters. In the federated scenario, XGBoost needs to exchange gradients between the server and the client to determine the best splitting scheme.

## 2.2 Differential Privacy

Differential Privacy (DP) [6] prevents data leakage by adding noise to queries, making it challenging to accurately identify individual data.

**2.2.1 ( $\epsilon, \delta$ )-Differential Privacy.** A randomized algorithm  $\mathcal{M}$  satisfies  $(\epsilon, \delta)$ -differential privacy [6] if for all  $S \in \text{Range}(\mathcal{M})$  and for any neighboring datasets  $D$  and  $D'$ :

$$\Pr[\mathcal{M}(D) = S] \leq e^\epsilon \Pr[\mathcal{M}(D') = S] + \delta \quad (3)$$

**2.2.2 Rényi Differential Privacy.** A randomized algorithm  $\mathcal{M}$  satisfies  $(\alpha, \bar{\epsilon})$ -Rényi differential privacy [18] if for all  $S \in \text{Range}(\mathcal{M})$  and for any neighboring datasets  $D$  and  $D'$ :

$$\text{Divergence}_\alpha(\mathcal{M}(D) \parallel \mathcal{M}(D')) \leq \bar{\epsilon}$$

$$\text{where } \text{Divergence}_\alpha(P \parallel Q) = \frac{1}{\alpha - 1} \log E_{x \sim Q} \left( \frac{P(x)}{Q(x)} \right)^\alpha \quad (4)$$

We can implement a Gaussian noise for a function  $f(x)$  that satisfies  $(\alpha, \bar{\epsilon})$ -Rényi differential privacy (RDP) as follows:

$$F(x) = f(x) + \mathcal{N}\left(\sigma^2\right), \quad \sigma^2 = \frac{\Delta f^2 \alpha}{2\bar{\epsilon}} \quad (5)$$

where  $\Delta f = \max_{x, x'} \|f(x) - f(x')\|$  denote the sensitivity of  $f(x)$ .

## 3 METHODOLOGY

### 3.1 Gradient Privacy-Preserving

We investigate using DP to protect gradient data privacy in federated XGBoost. Traditional DP methods add noise to statistical releases (e.g., sum, count, mean) [11] and adjust the noise based on the release's sensitivity, i.e., how much a single data point can impact the statistical results.

Let  $\{g_i, h_i\}$  denote the gradient information under privacy protection, where  $\Delta g$  and  $\Delta h$  represent the sensitivity of the gradient information for  $\{g_i, h_i\}$ , respectively. One of the most straightforward methods for applying DP to gradient information protection is to add appropriate noise to  $\{g_i, h_i\}$  individually:

$$\{g_i, h_i\} = \{g_i + \text{Noise}(\Delta g, \epsilon, \delta), h_i + \text{Noise}(\Delta h, \epsilon, \delta)\}$$

As shown in Figure 1(a), although this approach can satisfy the definition of DP, the noise generated by sensitivities  $\Delta g$  and  $\Delta h$  is of a similar order of magnitude to  $\{g_i, h_i\}$ . This can result in excessive noise being added that obscures the original data, which can undermine the usefulness of gradient information.

To reduce the amount of noise added to each sample's gradient, we drew inspiration from the k-anonymity [20] and the XGBoost quantile sketch [4] algorithm. Similar gradients were grouped, and the mean value of each group replaced all gradients in the group. Next, we added noise perturbation to each gradient mean value using DP.

Assuming that all gradient information is divided into  $k$  groups  $S = \{S_1, \dots, S_k\}$ , for the gradient elements in the  $j$ -th group, we calculate the gradient sum and gradient count values, and add noise to them:

$$\begin{aligned} \langle G_j \rangle &= \sum_{i \in S_j} g_i + \text{Noise}(\Delta g, \epsilon, \delta) \\ \langle H_j \rangle &= \sum_{i \in S_j} h_i + \text{Noise}(\Delta h, \epsilon, \delta) \\ \langle |S_j| \rangle &= |S_j| + \text{Noise}(1, \epsilon, \delta) \end{aligned}$$

Using the gradient mean after noise perturbation to replace the original gradient:

$$\{\langle g_i \rangle, \langle h_i \rangle\} = \{\langle G_j \rangle / \langle |S_j| \rangle, \langle H_j \rangle / \langle |S_j| \rangle\}, \quad i \in S_j$$

As shown in Figure 1(b), grouping the gradient means can greatly reduce the noise's impact on the gradient data. However, this substitution scheme is essentially a data generalization technique, and its effect on the original data depends on the quality of grouping. If gradients with significant differences are erroneously grouped, the gradient mean may not accurately represent the original gradient, resulting in a decrease in the model's performance.

Finding the optimal grouping scheme is an NP-Hard problem. Therefore, to balance the trade-off between noise disturbance and gradient generalization, we present an adaptive gradient privacy-preserving algorithm for federated XGBoost based on DP (AGPP-XGBoost). The algorithm has two stages, as shown in Figure 1(c):

- (1) Perform an adaptive weighted average of the gradient data of each sample within its neighborhood and the gradient data of other samples. The gradient weight depends on the distance measure between samples, and the sample gradient is generalized based on the result of the weighted average.
- (2) In the adaptive gradient mean calculation, noise is added to both the weighted gradient sum and the gradient weight sum, with their total privacy consumption calculated using DP's sequential composition.

### 3.2 Adaptive Gradient Mean

Before applying DP, we generalize the gradient by replacing original data with the gradient average in its neighborhood. This blurs individual feature information without losing data availability. Different from the simple mean calculation, we propose an adaptive gradient mean method that considers the non-uniform gradient distribution to preserve more statistical information from the original data. The calculation process is outlined in Algorithm 1.

For a given sample gradient  $g_i$ , we define its neighborhood using a width parameter  $r$  and include other sample gradients  $g_j$  within the neighborhood range. We measure the distance of each gradient  $g_j$  relative to  $g_i$  by computing the absolute difference for scalar

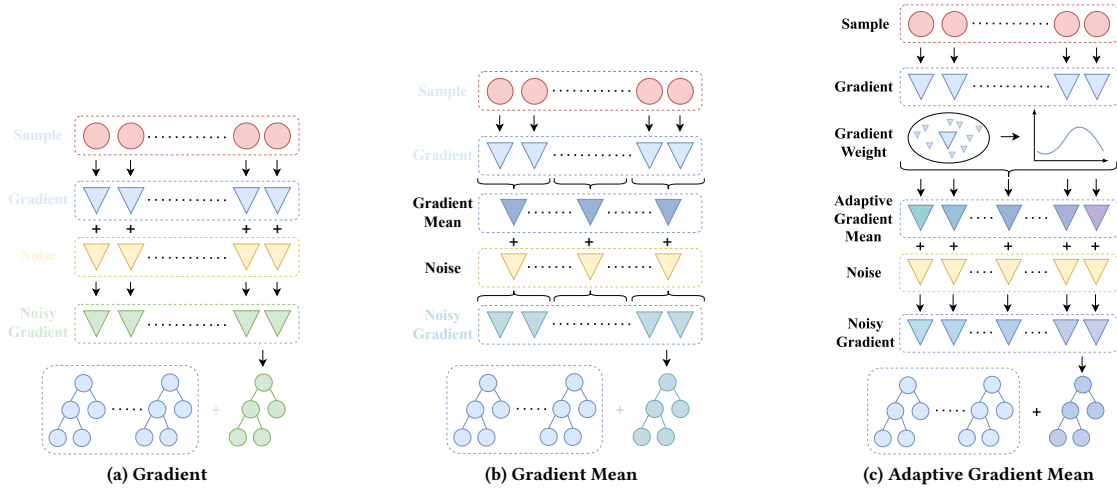


Figure 1: Three ways of adding noise to gradients

**Algorithm 1** Adaptive Gradient Mean**Input:** Original sample gradient  $\{g_i, h_i\}$ , neighborhood width  $r$ **Output:** Adaptive gradient mean of each sample  $\{\bar{g}_i, \bar{h}_i\}$ 

```

1: for each  $i \in [1, n]$  do
2:    $G \leftarrow 0, W_g \leftarrow 0$ 
3:   for each  $j \in \text{Neighborhood}(i, r)$  do
4:     Calculate  $\text{dist}(i, j)$  between  $g_i$  and  $g_j$ 
5:     Calculate  $\text{weight}(i, j)$  according to  $\text{dist}(i, j)$ 
6:      $G \leftarrow G + \text{weight}(i, j) * g_j$ 
7:      $W_g \leftarrow W_g + \text{weight}(i, j)$ 
8:   end for
9:   Calculate Adaptive Gradient Mean  $\bar{g}_i = \frac{G}{W_g}$ 
10:  Similarly, calculate the adaptive hessian mean  $\bar{h}_i = \frac{H}{W_h}$ 
11: end for

```

Subsequently, we calculate the weighted sum of the gradient of each sample and all gradient weights are accumulated:

$$G_i = \sum_j \text{weight}_g(i, j) * g_j$$

$$H_i = \sum_j \text{weight}_h(i, j) * h_j$$

$$W_i = \sum_j \text{weight}(i, j)$$

Then, we can calculate the adaptive gradient mean  $\{\bar{g}_i = \frac{G_i}{W_{g_i}}, \bar{h}_i = \frac{H_i}{W_{h_i}}\}$  of sample gradients  $\{g_i, h_i\}$ .

We utilize the adaptive gradient mean  $\{\bar{g}_i, \bar{h}_i\}$  to replace the original sample gradient  $\{g_i, h_i\}$  and blur the individual feature information while retaining the statistical information of the data.

gradients and the  $L_2$  norm for vector gradients:

$$\text{dist}_g(i, j) = \begin{cases} |g_i - g_j| & \text{for scalar} \\ \|\bar{g}_i - \bar{g}_j\|_2 & \text{for vector} \end{cases}$$

After obtaining the distances between each sample gradient and  $g_i$ , we utilize the Gaussian function to map the distance to a gradient weight. This approach ensures that the closer the gradient is to  $g_i$ , the greater its weight:

$$\text{weight}(i, j) = \text{Gauss}(\text{dist}(i, j)) = a \exp\left(-\frac{(\text{dist}(i, j) - b)^2}{2c^2}\right)$$

Where  $a$  represents the maximum weight that the gradient can have,  $b$  represents the central value of the Gaussian function and  $c$  is the standard deviation. Here, we set  $a = 1$ ,  $b = 0$  and determine the optimal value for the parameter  $c$  based on the loss function used in practice.

**3.3 Mean Perturbation**

After using the adaptive gradient mean, we noticed that the generalization of the sample gradient information improved. But for dense and high-frequency gradient information, the simple adaptive mean may not differ much from the original data, failing to hide individual characteristics.

We transformed the gradient data into means and added noise using DP to obtain perturbed means, addressing the issue of inability to conceal individual characteristics. This protects the sample's privacy by preventing attackers from obtaining private information. The perturbation process is detailed in Algorithm 2.

For the adaptive gradient mean calculation, we divide the gradient mean into the weighted gradient sum  $G$  and gradient weight sum  $W$ . To add Gaussian noise that satisfies the RDP [18], we compute the sensitivity  $\Delta G$  and  $\Delta W$ . The sensitivity is computed by value bounds using the derivative of the loss function. For an unbounded derivative function, clipping the gradient using a threshold value  $C$  can provide a sensitivity value [1].

**Algorithm 2** Mean Perturbation

---

**Input:** Original sample gradient  $\{g_i, h_i\}$ , RDP parameters  $(\alpha, \bar{\epsilon})$   
**Output:** Adaptive gradient mean after adding noise  $\{\langle \bar{g}_i \rangle, \langle \bar{h}_i \rangle\}$

- 1: **for** each  $i \in [1, n]$  **do**
- 2:   Calculate weighted gradient sum  $G$  and gradient weight sum  $W$  according to Algorithm 1
- 3:   Add Gaussian noise satisfying RDP definition to  $G$  and  $W$ :
- 4:    $\langle G \rangle \leftarrow G + \text{RDP\_GuassNoise}(\alpha, \bar{\epsilon}, \Delta G)$
- 5:    $\langle W \rangle \leftarrow W + \text{RDP\_GuassNoise}(\alpha, \bar{\epsilon}, \Delta W)$
- 6:   Calculate the adaptive gradient mean after perturbation  
 $\langle \bar{g}_i \rangle = \frac{\langle G \rangle}{\langle W_g \rangle}$
- 7:   Similarly, calculate the adaptive hessian mean after perturbation  
 $\langle \bar{h}_i \rangle = \frac{\langle H \rangle}{\langle W_h \rangle}$
- 8: **end for**

---

After computing the corresponding sensitivities, we select suitable parameters  $\alpha$  and  $\bar{\epsilon}$  and generate Gaussian noise according to the definition of RDP using Equation 5. We then add the noise to the original formula for the gradient mean to obtain the perturbed adaptive gradient mean  $\{\langle \bar{g}_i \rangle, \langle \bar{h}_i \rangle\} = \{\frac{\langle G_i \rangle}{\langle W_{gi} \rangle}, \frac{\langle H_i \rangle}{\langle W_{hi} \rangle}\}$ . Because the training process of XGBoost involves constructing multiple decision trees, using RDP enables us to obtain a tighter upper bound on the privacy budget and reduce total privacy consumption.

## 4 EXPERIMENT RESULTS

The following section will demonstrate the superiority of our proposed method in two key areas: utility and efficiency. In order to verify the performance of our algorithm, we use two public datasets from the Kaggle platform:

- Credit 1<sup>1</sup>: It involves the problem of classifying whether a user will experience severe financial problems. It consists of 150,000 user data, each containing 10 attributes and 1 label information.
- Credit 2<sup>2</sup>: It is a credit score dataset that relates to the task of predicting whether users will make timely payments. It consists of 30,000 user data, each containing 25 attributes and 1 label information.

In our experiment, we use 70% of each dataset for training and the remaining 30% for testing. The experimental environment is Intel(R) Core(TM) i7-12700H CPU @2.70GHz, 16GB memory, Windows 11 64-bit operating system.

### 4.1 Utility

This section conducts convergence experiments and analysis on the model based on noise protection to verify the utility of the model after adding noise. The proposed adaptive gradient privacy-preserving algorithm model (AGPP-XGBoost) is compared with the original XGBoost model (Plain-XGBoost), XGBoost model that directly adds noise to the gradient (Noisy-XGBoost), and XGBoost model that uses grouped gradient mean (Mean-XGBoost) by comparing the convergence of the loss curve.

Figure 2 shows that Noisy-XGBoost's loss does not converge due to excessive noise. While Mean-XGBoost can converge, the grouping uncertainty leads to instability and a significant difference from Plain-XGBoost. In contrast, AGPP-XGBoost generalizes the gradient by considering weights of neighboring gradients, resulting in stable loss convergence. Though accuracy is reduced by noise, final convergence loss is closer to Plain-XGBoost.

In order to verify the performance of the model, we compare the proposed method with other aforementioned methods and DP-XGBoost [11]. We use accuracy, area under ROC curve (AUC) and f1-score as evaluation metrics for the model.

Table 1 illustrates the performance comparison of various methods with the baseline given by Plain-XGBoost on the two datasets. The proposed method, AGPP-XGBoost, stands out in terms of the least performance loss among all other methods while providing privacy preservation. The experimental results on two datasets show that the proposed method performs well on not only ACC and AUC, but also on F1-score. Compared with DP-XGBoost, the proposed method reduces performance losses by 15.9% and 22.3% for ACC, and by 28% and 47% for AUC, while the one on F1-score is reduced by 41.9% and 45.1%, respectively.

AGPP-XGBoost is impacted by two key parameters: the width of the neighborhood and the privacy budget. We investigate their impact on model performance in this section. With a privacy budget of 2, Figure 4 shows that the model accuracy is unsatisfactory when the neighborhood width is less than or equal to 30, but improves as the width increases, stabilizing at 90 or more. With a neighborhood width of 100, Figure 5 shows that the model's performance is negatively impacted when the privacy budget is less than 0.5, but improves as the budget increases, converging at 1.5 or greater.

### 4.2 Efficiency

We also compared AGPP-XGBoost with SecureBoost, a method based on cryptography that uses homomorphic encryption to protect gradient privacy during the training process of federal XGBoost. However, homomorphic encryption is computationally expensive and results in a longer training process.

As shown in Figure 3, AGPP-XGBoost, which is based on DP, exhibits higher efficiency than SecureBoost. The training time of a single tree is shortened by nearly 75% compared to that of SecureBoost. Given the need for high communication efficiency, it is acceptable to trade-off a small error for efficient training.

## 5 CONCLUSION

We propose an adaptive gradient privacy-preserving algorithm for federated XGBoost. During the gradient transmission stage between the server and the client, our method calculates the adaptive gradient mean by combining the gradient distribution information in the gradient neighborhood of each sample, and adds noise that satisfies DP to protect the gradient privacy. We show the proposed algorithm outperforms general gradient protection methods and other DP-based methods on common public datasets, and its performance is similar to lossless protection and the original non-privacy protection model. In scenarios that require high communication efficiency, our method has a shorter training time than the cryptography-based method.

<sup>1</sup><https://www.kaggle.com/c/GiveMeSomeCredit/overview>

<sup>2</sup><https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset>

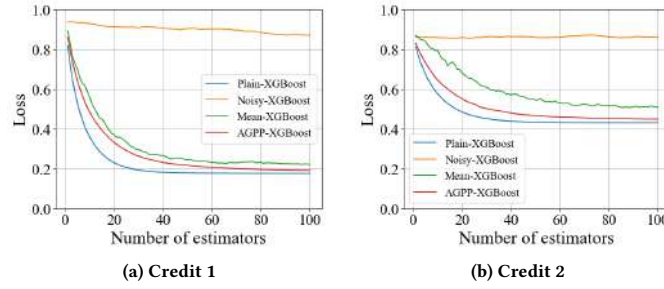


Figure 2: Loss convergence

Table 1: Performance of different approaches on public datasets

Model	Credit 1			Credit 2		
	ACC	F1-score	AUC	ACC	F1-score	AUC
Plain-XGBoost	93.80	28.49	86.65	81.95	46.15	78.01
Noisy-XGBoost	11.20 (-82.6)	12.11 (-16.4)	50.18 (-36.5)	34.63 (-47.3)	20.82 (-25.3)	52.96 (-25.1)
Mean-XGBoost	93.23 (-0.57)	19.04 (-9.45)	73.87 (-12.8)	79.72 (-2.23)	32.97 (-13.2)	69.69 (-8.32)
DP-XGBoost	93.36 (-0.44)	23.91 (-4.58)	82.84 (-3.81)	79.13 (-2.82)	36.17 (-9.98)	74.03 (-3.98)
AGPP-XGBoost	93.73 (-0.07)	26.57 (-1.92)	85.58 (-1.07)	81.32 (-0.63)	41.65 (-4.50)	76.13 (-1.88)

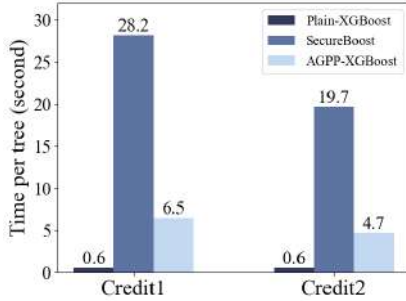


Figure 3: Tree building time

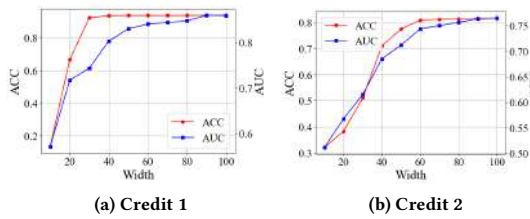


Figure 4: Performance on different Width

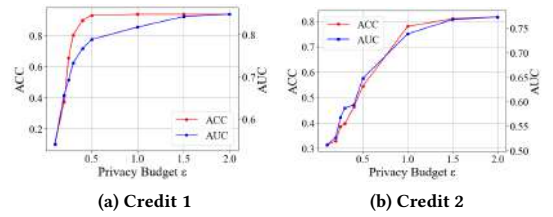


Figure 5: Performance on different Privacy Budget

## REFERENCES

- [1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
- [2] Mark Abspoel, Daniel Escudero, and Nikolaj Volgushev. 2020. Secure training of decision trees with continuous attributes. *Cryptology ePrint Archive* (2020).
- [3] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. 2019. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*. PMLR, 344–353.
- [4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- [5] Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. 2021. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems* 36, 6 (2021), 87–98.
- [6] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- [7] Wenjing Fang, Derun Zhao, Jin Tan, Chaochao Chen, Chaofan Yu, Li Wang, Lei Wang, Jun Zhou, and Benyu Zhang. 2021. Large-scale secure XGB for vertical federated learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 443–452.
- [8] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.
- [9] Fangcheng Fu, Yingxia Shao, Lele Yu, Jiawei Jiang, Huanran Xue, Yangyu Tao, and Bin Cui. 2021. Vf2boost: Very fast vertical federated gradient boosting for cross-enterprise learning. In *Proceedings of the 2021 International Conference on Management of Data*. 563–576.
- [10] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine* 38, 3 (2017), 50–57.
- [11] Nicolas Grislain and Joan Gonzalez. 2021. Dp-xgboost: Private machine learning at scale. *arXiv preprint arXiv:2110.12770* (2021).

- [12] Nhan Khanh Le, Yang Liu, Quang Minh Nguyen, Qingchen Liu, Fangzhou Liu, Quanwei Cai, and Sandra Hirche. 2021. Fedxgboost: Privacy-preserving xgboost for federated learning. *arXiv preprint arXiv:2106.10662* (2021).
- [13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* 2 (2020), 429–450.
- [14] Xiaochen Li, Yuke Hu, Weiran Liu, Hanwen Feng, Li Peng, Yuan Hong, Kui Ren, and Zhan Qin. 2022. OpBoost: a vertical federated tree boosting framework based on order-preserving desensitization. *arXiv preprint arXiv:2210.01318* (2022).
- [15] Lingjuan Lyu, Han Yu, and Qiang Yang. 2020. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133* (2020).
- [16] Samuel Maddock, Graham Cormode, Tianhao Wang, Carsten Maple, and Somesh Jha. 2022. Federated Boosted Decision Trees with Differential Privacy. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2249–2263.
- [17] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.
- [18] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 263–275.
- [19] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Membership inference attacks against adversarially robust deep learning models. In *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 50–56.
- [20] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.
- [21] Zhihua Tian, Rui Zhang, Xiaoyang Hou, Jian Liu, and Kui Ren. 2020. Federboost: Private federated learning for gbdt. *arXiv preprint arXiv:2011.02796* (2020).
- [22] Paul Voigt and Axel Von dem Bussche. 2017. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing* 10, 3152676 (2017), 10–5555.
- [23] Lingchen Zhao, Lihao Ni, Shengshan Hu, Yanliao Chen, Pan Zhou, Fu Xiao, and Libing Wu. 2018. Inprivate digging: Enabling tree-based distributed data mining with differential privacy. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2087–2095.

# Generate earthquake catalog using the VAE method

Zhangyu Wang

School of Earth and Space Sciences, University of Science  
and Technology of China, Hefei, Anhui, P. R. China

wangzyu@mail.ustc.edu.cn

Jie Zhang

School of Earth and Space Sciences, University of Science  
and Technology of China, Hefei, Anhui, P. R. China

jzhang25@ustc.edu.cn

## ABSTRACT

*The earthquake catalog is essential for seismic activity analysis and earthquake forecasting. Researchers would like to use a complete catalog for further study. In this study, we use a machine learning method to derive a double-variable model to learn the latent rules of catalogs and generate the synthetic ones from a historical catalog. In the first step, we obtain an individual cluster from the catalog by the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. Then we take the envelope of the magnitude-time curve of the clusters. In the end, we apply the Variational AutoEncoder (VAE) method to learn the inherent feature and produce the latent magnitude-time curves. We use the earthquakes in Southern California from 2016 January 1 to 2022 December 18 to train the VAE model. After training, the model can generate abundant magnitude-time curves and the result shows that the magnitude-time curves during this period can be divided into single-peak, double-peak, and treble-peak patterns. Furthermore, we can use this method to generate more clusters for swarm identification and analysis of regional seismic activity.*

## CCS CONCEPTS

• **General and reference** → Cross-computing tools and techniques; Performance.

## KEYWORDS

VAE, DBSCAN, catalog, earthquake sequence

### ACM Reference Format:

Zhangyu Wang and Jie Zhang. 2023. Generate earthquake catalog using the VAE method. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3590003.3590052>

## 1 INTRODUCTION

The sudden occurrence of a large earthquake may cause huge casualties and property losses. It is hoped that an effective earthquake prediction method could predict the occurrence of earthquakes to reduce loss in the disaster. Waveform data and catalogs are often used for seismic activity analysis and risk assessment. Waveform contains ground vibration information. By processing waveform

data, we can detect earthquakes and give earthquake early warnings. A catalog is a collection of events that record the parameters of earthquakes, including the original time, location, magnitude, depth of the focal point, etc. The earthquake catalog can be used for seismic hazard studies (Beauval et al., 2013), aftershock and mainshock prediction (Aceves et al., 1996), forecasting of volcano activities (Chen et al., 2015), etc. By analyzing the characteristics of original time, space distribution, and magnitude of seismic events in the seismic catalog, we can study the geological characteristics of seismic activity areas. The focal mechanism parameters of the seismic events in the earthquake catalog can be used to infer the stress changes. The seismic event information in the seismic catalog can be used to assess the earthquake risk of an area, like calculating the probability that a large earthquake may occur in the area in the next period of time. By updating the earthquake catalog in real-time, we can estimate the seismic activity in time and improve the accuracy of earthquake early warnings.

Most of the simulation of the catalog is based on physics. Davis developed a parametric model of earthquake behavior for generating synthetic earthquake catalogs (Davis et al., 1991). Earthquake simulators are computer programs that model long histories of earthquake occurrence and slip using various approximations of what is known about the physics of stress transfer due to fault slip and the rheological properties of faults (Tullis, 2012). Console considers the effect of minor earthquakes in redistributing stress and the interaction between earthquake sources (Console et al., 2015).

Machine learning has been used in various fields and brought about huge progress, such as image detection, and natural language processing. And Machine learning also has many applications in seismology, like earthquake discrimination (Li et al., 2018) and PhaseNet (Zhu & Beroza, 2019). The VAE method is widely used in signal and image processing to uncover the intrinsic structure of a large data set (Kingma & Welling, 2014). In this study, we propose a machine-learning method to generate the synthetic magnitude-time curves. This method only requires the historical catalog.

## 2 DATA

The Southern California Seismic Data Center (SCEDC) provides a complete data-sharing platform. From the platform, researchers can download raw waveforms recorded by the Southern California Seismic Network (SCSN), as well as high-quality earthquake catalog information. The catalog used in this study is from SCEDC. We use the earthquakes in Southern California from 2016 January 1 to 2022 December 18, as shown in Figure 1. The catalog contains 191, 469 events. During this period, the maximum magnitude is  $M_w 7.1$  and the minimum magnitude is  $M_L 0.0$ . There are 5 earthquakes with a magnitude of 5.5 or greater. In Figure 1, we can see that the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590052>

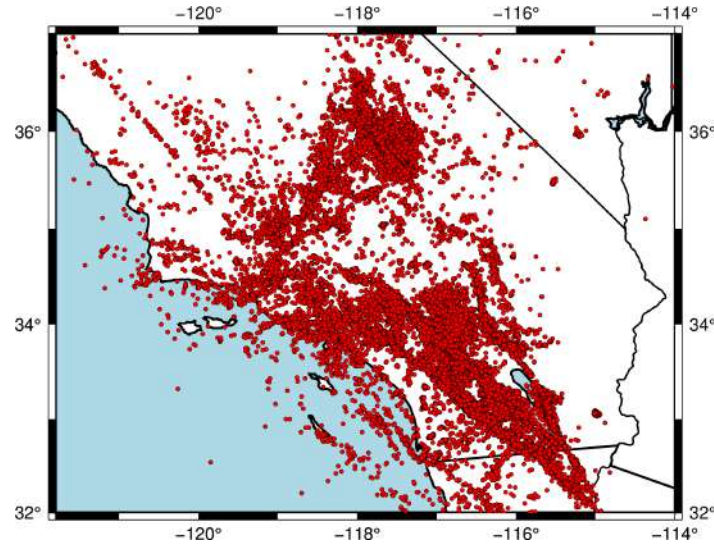


Figure 1: The distribution of the earthquakes from SCEDC. The red dots represent the location of earthquakes.

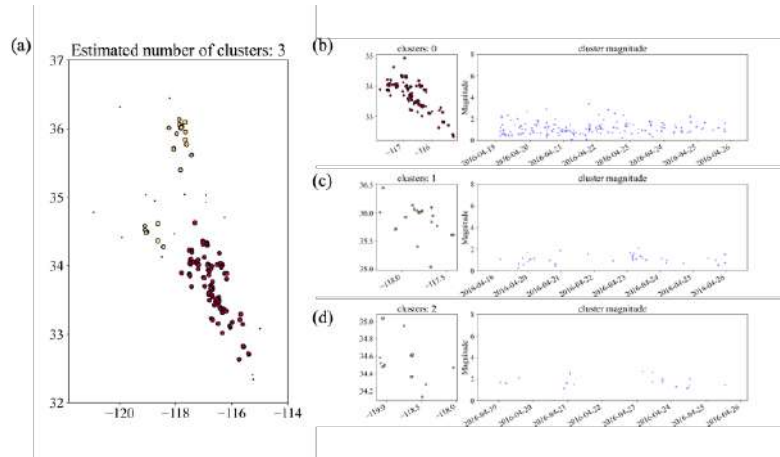


Figure 2: The cluster of a chosen catalog using the DBSCAN method. Fig (a) shows the distribution of different clusters. The colorful points represent the earthquakes of each cluster. And the black points represent the noise. Fig (b), Fig (c), and Fig (d) shows the distribution of earthquakes and the magnitude-time figure for cluster ‘0’, ‘1’, and ‘2’ separately.

earthquake distribution is relatively concentrated. With the catalog, we can seek specific earthquake sequences.

### 3 METHOD

#### 3.1 Extraction of seismic clusters

The catalog may contain quantities of earthquake sequences, like frequent aftershocks and repeated earthquakes. In order to facilitate research, we need to separate the different earthquake sequences from the catalog. The earthquake sequence can reflect the regional seismic activity to some degree. In this study, we use the DBSCAN algorithm (Ester et al., 1996) to obtain the separate earthquake sequence from the catalog simply. The DBSCAN mainly contains two key parameters, the adjacent radius  $R$ , and  $MinPt$ . The adjacent radius  $R$  means the minimum distance between two samples. If the

real distance between two events is less than the adjacent radius  $R$ , they will be classified into the same cluster. The  $MinPt$  means the minimum number of samples required to form a cluster. For earthquake clustering, we set the adjacent radius  $R$  to 500 km and  $MinPt$  to 10. The clustering result of a test is shown in Figure 2.

We apply the DBSCAN method to the entire catalog in SCEDC from 2016 to 2022. The sliding window is half of the month, 15 days. After the DBSCAN operation, we obtain the 227 separated clusters. Some clusters are shown in Figure 2. With the individual cluster, seismic activity can be analyzed.

For reconstructing the regional earthquake sequence, we first need to make the discrete sequence more consecutive. For earthquake sequences, we focus on the magnitude-time curve. It can reflect the evolution of earthquakes in time and space. We take the

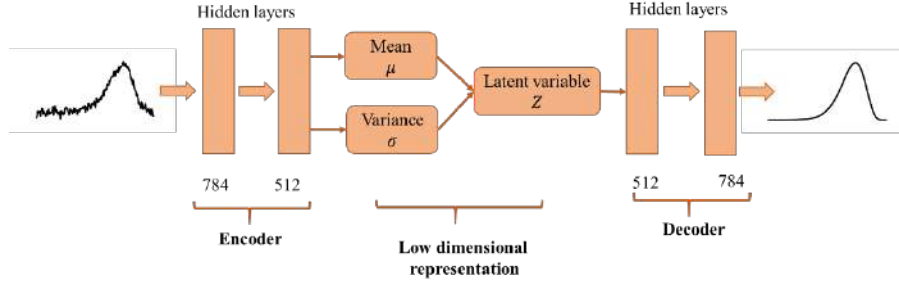


Figure 3: The network structure of the VAE model.

envelope of the magnitude-time curve by the Hilbert transform and 5-point smoothing algorithm. Then we can obtain a smooth and consecutive magnitude-time curve that still retains the shape feature.

### 3.2 The VAE model

The machine learning model can be divided into two types of models, discriminant model, and generative model. The discriminant models model conditional distributions, while the generative models model joint distributions. In the generative model, the joint probability is calculated, and then the conditional probability is calculated by the Bayesian formula. Therefore, the generative model can reflect more distribution information of the data itself, and its universality is wider. The VAE method has been widely used in seismology, like source spectra generation (Ma et al., 2022). In this study, we use the VAE method, one of the popular generative models, to achieve data reconstruction and generation.

The VAE model consists of an encoder, a low-dimensional representation, and a decoder. The encoder maps the original high-dimensional data to the low-dimensional feature space, which is usually smaller than the original data to compress dimensionality. The low-dimensional feature often becomes the latent representation. The decoder reconstructs the raw data based on compressed low-dimensional features. The VAE method is an unsupervised learning method and the model can be trained without labeled data. The VAE structure is shown in Figure 3. The encoder consists of two fully connected layers with 784 and 512 neurons separately. And the decoder consists of two fully connected layers with 512 and 784 neurons separately. The dimension of the input and output of the model is equal to achieve the data reconstruction. The reconstruction loss function of the model is defined as

$$\text{loss} = \|X_{\text{model}} - X_{\text{original}}\|_2 + \text{KL}[\mathcal{N}(\mu_2, \sigma_2) - \mathcal{N}(0, 1)]$$

The first item of the reconstruction loss calculates the Mean Square Error (MSE) loss between the output decoded by the decoder and the input. The second item calculates the Kullback-Leibler divergence. The KL divergence measures the difference between latent variables and normal distribution (Ma et al., 2022). We set 2 latent variables  $Z_1, Z_2$  to control the model. When the VAE model is well-trained, it will represent a real magnitude-time curve generator.

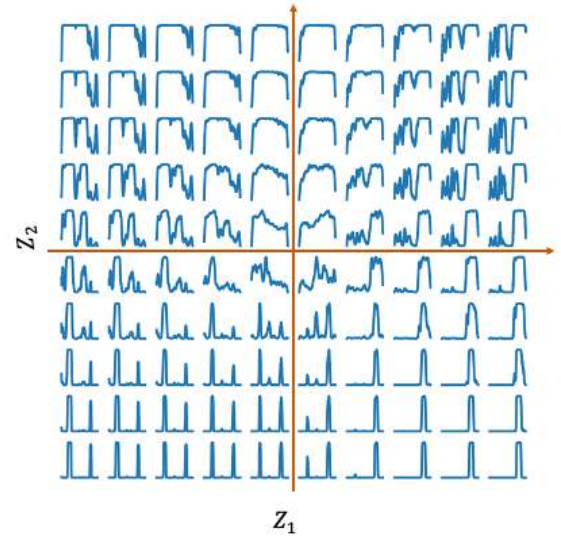


Figure 4: The envelopes of synthetic magnitude-time curves generated by the trained VAE model.

## 4 RESULT

The catalog from Southern California contains 191,469 events. After clustering by the DBSCAN algorithm, we can obtain 227 individual clusters. We use the 227 magnitude-time curves to train the double-variable VAE model. During the training procedure, the loss converges rapidly. The trained VAE model can reconstruct the magnitude-time curves. By giving the value of  $z_1$  and  $z_2$ , the model can generate a new magnitude-time curve.

To check the effect of the VAE model, we visualize the generated magnitude-time curves as shown in Figure 4. It shows 100 generated curves. When we choose different  $z_1$  and  $z_2$ , the generated magnitude-time curve changes according to a certain rule. This means the two latent variables can adjust the output of the model elaborately.

In Figure 4, we can see that the envelope of generated magnitude-time curves changes with an inherent rule. The magnitude-time curves show the different patterns, like single-peak, double-peak, and treble-peak patterns. Each pattern may correspond to different

seismogenic patterns and geological characteristics, which need further study. If we use more data to train the VAE model, more features may be found.

## 5 CONCLUSION

In this study, we use a double-variable VAE model to learn the latent rules of catalogs and generate the synthetic ones by historical catalog. The VAE method is a popular generative model which can make data reconstruction and generation. The workflow contains three steps. Firstly, we need to obtain the individual cluster from the catalog by the DBSCAN algorithm. Then we take the envelope of the magnitude-time curve of the clusters. Lastly, we apply the VAE method to learn the inherent feature and produce the latent magnitude-time curve. The earthquakes used in this study are in Southern California from 2016 January 1 to 2022 December 18. We use the envelopes of magnitude-time curves to train the VAE model. After training, the model can generate abundant magnitude-time curves with the inherent rule. We use the trained VAE model to produce 100 magnitude-time curves. The result shows that the magnitude-time curves in this period include single-peak, double-peak, and treble-peak patterns. These patterns may correspond to different seismogenic patterns and geological characteristics, which need further study. Moreover, we provide a perspective to discover the inherent feature of the earthquake sequences in one region. We can use the VAE method for dimension reduction of magnitude-time curves. In addition, we can use this method to generate more

clusters for swarm identification or analyze the regional seismic activity.

## REFERENCES

- [1] Beauval C, Yepes H, Palacios P, *et al.* An earthquake catalog for seismic hazard assessment in Ecuador[J]. *Bulletin of the Seismological Society of America*, 2013, 103(2A): 773-786.
- [2] Aceves R L, Park S K, Strauss D J. Statistical evaluation of the VAN method using the historic earthquake catalog in Greece[J]. *Geophysical research letters*, 1996, 23(11): 1425-1428.
- [3] Chen L, Chen X, Shao L. Method research of earthquake prediction and volcano prediction in Italy[J]. *International Journal of Geosciences*, 2015, 6(09): 963.
- [4] Davis S D, Frohlich C. Single-link cluster analysis, synthetic earthquake catalogues, and aftershock identification[J]. *Geophysical Journal International*, 1991, 104(2): 289-306.
- [5] Console R, Carluccio R, Papadimitriou E, *et al.* Synthetic earthquake catalogs simulating seismic activity in the Corinth Gulf, Greece, fault system[J]. *Journal of Geophysical Research: Solid Earth*, 2015, 120(1): 326-343.
- [6] Li Z, Meier M A, Hauksson E, *et al.* Machine learning seismic wave discrimination: Application to earthquake early warning[J]. *Geophysical Research Letters*, 2018, 45(10): 4773-4779.
- [7] Zhu W, Beroza G C. PhaseNet: a deep-neural-network-based seismic arrival-time picking method[J]. *Geophysical Journal International*, 2019, 216(1): 261-273.
- [8] Tullis T E. Preface to the focused issue on earthquake simulators[J]. *Seismological Research Letters*, 2012, 83(6): 957-958.
- [9] Ester M, Kriegel H P, Sander J, *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise[C]//kdd. 1996, 96(34): 226-231.
- [10] Ross Z E, Cochran E S. Evidence for latent crustal fluid injection transients in Southern California from long-duration earthquake swarms[J]. *Geophysical Research Letters*, 2021, 48(12): e2021GL092465.
- [11] Ma S, Li Z, Wang W. Machine learning of source spectra for large earthquakes[J]. *Geophysical Journal International*, 2022, 231(1): 692-702.
- [12] Kingma D P, Welling M. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*[J]. 2014.

# Robust Hypergraph-Augmented Graph Contrastive Learning for Graph Self-Supervised Learning

Zeming Wang

University of Chinese Academy of Sciences, Institute of  
Software Chinese Academy of Sciences  
Beijing, China  
wangzeming20@mails.ucas.ac.cn

Rui Wang\*

Science and Technology on Integrated Information System  
Laboratory, Institute of Software Chinese Academy of  
Sciences  
Beijing, China  
wangrui@iscas.ac.cn

Xiaoyang Li

Nankai University  
Tianjin, China  
xiaoyangli0903@163.com

Changwen Zheng

Institute of Software, Chinese Academy of Sciences  
Beijing, China  
changwen@iscas.ac.cn

## ABSTRACT

Graph contrastive learning has emerged as a promising method for self-supervised graph representation learning. The traditional framework conventionally imposes two graph views generated by leveraging graph data augmentations. Such an approach focuses on leading the model to learn discriminative information from graph local structures, which brings up an intrinsic issue that the model partially fails to obtain sufficient discriminative information contained by the graph global information. To this end, we propose a hypergraph-augmented view to empower the self-supervised graph representation learning model to better capture the global information from nodes and corresponding edges. In the further exploration of the graph contrastive learning, we discover a principal challenge undermining conventional contrastive methods: the *false negative sample problem*, i.e., specific negative samples actually belong to the same category of the anchor sample. To address this issue, we take the neighbors of nodes into consideration and propose the robust graph contrastive learning. In practice, we empirically observe that the proposed hypergraph-augmented view can further enhance the robustness of graph contrastive learning by adopting our framework. Based on these improvements, we propose a novel method called Robust Hypergraph-Augmented Graph Contrastive Learning (RH-GCL). We conduct various experiments in the settings of both transductive and inductive node classification. The results demonstrate that our method achieves the state-of-the-art (SOTA) performance on different datasets. Specifically, the accuracy of node classification on Cora dataset is 84.4%, which is 1.1% higher

than that of GRACE. We also perform the ablation study to verify the effectiveness of each part of our proposed method.

## CCS CONCEPTS

• **Computing methodologies** → **Unsupervised learning**; **Neural networks**; • **Computer systems organization** → **Neural networks**.

## KEYWORDS

graph contrastive learning, self-supervised, hypergraph, multi-view

## ACM Reference Format:

Zeming Wang, Xiaoyang Li, Rui Wang, and Changwen Zheng. 2023. Robust Hypergraph-Augmented Graph Contrastive Learning for Graph Self-Supervised Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590053>

## 1 INTRODUCTION

Recently, graph representation learning using Graph Neural Networks (GNN) has emerged as a promising approach for graph representation learning. However, these works usually establish existing GNN models in a supervised manner [12, 15, 30], requiring a large number of labeled samples for training. Although some attempts try to connect previous unsupervised objectives to GNN models [10, 16], these methods are heavily related to the preset graph proximity matrix.

To address the issues in the supervised manner, self-supervised learning has been introduced into the graph representation learning. Unlike supervised or semi-supervised learning, it can avoid the cost of labeling large-scale datasets. Self-supervised learning includes generative learning and contrastive learning (CL) methods. Self-supervised learning achieves impressive improvement in the fields of image representation learning [3, 13, 19, 20, 37], graph representation learning [6–8, 28, 36, 40], multi-modal learning [1, 18] and natural language processing [4, 14, 22]. Recently, self-supervised contrastive learning approaches, such as CMC [29], SimCLR [3], MoCo [11], and MetAug [21], have reached the competitive results compared to supervised SOTAs.

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590053>

Contrastive learning has recently become a mainstream paradigm for self-supervised learning methods in the field of computer vision, natural language processing, graph, and other domains. The goal of contrastive learning is embedding augmented views of the same sample close to each other while trying to push away embeddings from different samples. To achieve this, the similarity metric is introduced to determine the distance between embeddings. Representations of samples are learned through encoders to calculate the contrastive loss between samples.

Graph contrastive learning (GCL) applies the framework of CL to GNNs. [10] can be viewed as a traditional method based on local contrast, which reconstructs the adjacency information of nodes. [24] focuses on contrastive learning between the local and global graph structure and captures the structure information better. [39] designs four graph data augmentations and studies the influence of different combinations of graph augmentations on several datasets. [27] proposes AD-GCL to prevent GNNs from capturing redundant information during training by optimizing adversarial graph augmentation strategies used in GCL. A unified bilevel optimization framework is proposed by [38] to adaptively, automatically, and dynamically select data augmentations when performing graph contrastive learning on specific graph data. [34] puts perturbation into the GCL framework and does not require data augmentations.

In this paper, we propose a simple yet powerful framework for graph contrastive learning called RH-GCL, motivated by [42] and [17]. For traditional graph contrastive methods, the view of a simple graph focuses more on the local structure to learn the discriminant information of the local part. This brings up the issue that the model is lack of obtaining global information. To address this issue, we add a hypergraph-augmented view into the framework for graph contrastive learning, which generates the hypergraph from the original graph. The hypergraph view builds hyperedges connecting more nodes to capture the global information of nodes. Referring to [32], the embedding obtained from neighbor aggregation will concentrate toward the expectation of embedding belonging to the same class. Besides, we discover a principal challenge undermining conventional contrastive methods: the *false negative* sample problem, i.e., specific negative samples actually belong to the same category of the anchor sample. So we take the neighbors of nodes into consideration. Based on this motivation, we select samples found by k-NN search as positive samples and introduce a novel method called robust graph contrastive learning. Furthermore, we empirically observe that the proposed hypergraph-augmented view can further enhance the robustness of graph contrastive learning by adopting our framework.

Our major contributions are three-fold:

- Our paper is the first to add a hypergraph-augmented view into the framework for graph contrastive learning, which makes full use of the global structural information.
- We treat samples discovered by k-NN search as positive samples and propose a novel method called robust graph contrastive learning. The proposed method and contrastive loss can better guide the training process.
- We prove the validity of our proposed method on different datasets. Our method achieves the SOTA performance on transductive and inductive tasks.

## 2 PRELIMINARIES

### 2.1 Graph Neural Networks

Graph Neural Networks [15, 25, 30, 35] have become a promising method for graph representation learning in recent years. For graph  $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ ,  $\mathcal{V} = \{v_1, v_2, \dots, v_N\}$  denotes the node set, and  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$  represents the edge set. We denote the node feature matrix and the adjacency matrix as  $X \in \mathbb{R}^{N \times F}$  and  $A \in \{0, 1\}^{N \times N}$ , where  $x_i \in \mathbb{R}^F$  is the feature of  $v_i$ , and  $A_{ij} = 1$  iff  $(v_i, v_j) \in \mathcal{E}$ . As for node  $v_i \in \mathcal{V}$ ,  $h_i$  denotes the node representation, initialized as  $h_i^{(0)} = x_i$ . Node representations are updated by a GNN encoder  $f(\cdot)$ . Considering the  $k$ th layer of  $f(\cdot)$ , the propagation process can be described as:

$$a_i^{(k)} = \text{AGGREGATION}^{(k)} \left( \left\{ h_{i'}^{(k-1)} : i' \in \mathcal{N}(i) \right\} \right) \quad (1)$$

$$h_i^{(k)} = \text{COMBINE}^{(k)} \left( h_i^{(k-1)}, a_i^{(k)} \right) \quad (2)$$

where  $h_i^{(k)}$  is the representation of node  $v_i$  at the  $k$ th layer,  $\mathcal{N}(i)$  is the set containing neighbors of node  $v_i$ , and  $\text{AGGREGATION}^{(k)}(\cdot)$  and  $\text{COMBINE}^{(k)}(\cdot)$  are propagation functions of GNN. Then, the representation  $h_i^{(k)}$  is fed into a READOUT function:

$$f(X, A) = \text{READOUT} \left( \left\{ h_i^{(k)} : v_i \in \mathcal{V} \right\} \right) \quad (3)$$

READOUT function aggregates node features from the final layer of the GNN encoder to obtain the entire graph's representation. We aim at learning a GNN-based encoder  $f(X, A)$  to produce node features. We denote  $H = f(X, A)$  as the learned representations of nodes. Finally, a multi-layer perceptron (MLP) is applied for downstream tasks, such as node classification.

$$z_{\mathcal{G}} = \text{MLP}(f(\mathcal{G})) \quad (4)$$

### 2.2 Hypergraph

A hypergraph consists of nodes and hyperedges. Unlike the simple graph, a hyperedge in a hypergraph connects two or more nodes. A hypergraph can be denoted as  $\mathcal{G}' = \{\mathcal{V}', \mathcal{E}'\}$ , where  $\mathcal{V}'$  and  $\mathcal{E}'$  represent the node set and hyperedge set respectively. The feature matrix and the adjacency matrix can be denoted as  $X' \in \mathbb{R}^{N' \times F'}$  and  $A' \in \{0, 1\}^{N' \times N'}$ . The hypergraph  $\mathcal{G}'$  can also be defined by a  $|\mathcal{V}'| \times |\mathcal{E}'|$  incidence matrix  $H'$ , with entries defined as:

$$h'(v', e') = \begin{cases} 1, & \text{if } v' \in e' \\ 0, & \text{if } v' \notin e', \end{cases} \quad (5)$$

where  $v' \in \mathcal{V}'$  and  $e' \in \mathcal{E}'$ .

### 2.3 Graph Data Augmentation

In graph contrastive learning, different views of a graph are produced from the raw graph via data augmentations. [39] proposes four graph-level augmentations and [42] proposes two node-level augmentations. We apply two augmentations for node classification, Removing edges and Masking node attributes.

**Removing edges.** We randomly drop certain ratio of edges in the original graph. It is based on the basic prior that removing part of edges has no influence on the semantic meaning of the graph. We follow an i.i.d. uniform distribution to remove each edge.

**Masking node attributes.** We randomly mask certain ratio of dimensions in node features. Each dimension’s masking probability follows an i.i.d. uniform distribution. The basic prior of masking node attributes is that masking partial node features does not affect the model predictions much.

### 3 METHODS

#### 3.1 Hypergraph-Augmented View

In the traditional graph contrastive learning, there are two views generated from the raw graph. Different from the traditional framework, our proposed framework has three views, as shown in Figure 1. For graph  $\mathcal{G}$ ,  $\mathcal{G}_1$  and  $\mathcal{G}_2$  are two augmented views, the same as the traditional framework. We denote node features in two views as  $V_1 = f(X_1, A_1)$  and  $V_2 = f(X_2, A_2)$ , where  $X_*$  and  $A_*$  are the feature matrices and adjacency matrices of the views. To better capture the global information of nodes, we introduce a module called hypergraph-augmented view into our framework. Referring to [5], we build hyperedges in the hypergraph based on the distance between two nodes. We apply Euclidean distance to measure the distance between node  $v_i$  and  $v_j$ . Each hyperedge is generated by connecting one node and its  $K$  nearest neighbors. We set  $K$  as 2 in our experiments. Finally, we get the incidence matrix of the hypergraph and denote the augmented hypergraph view as  $\mathcal{G}_3$ . The node features of the hypergraph view can be denoted as  $V_3 = f(X_3, A_3)$ , where  $X_3$  and  $A_3$  are the feature matrix and adjacency matrix of  $\mathcal{G}_3$ .

After the data preprocessing, we get three views of the original graph  $\mathcal{G}$ . For node  $v$ ,  $v_1$ ,  $v_2$ , and  $v_3$  represent the augmented views of node  $v$  respectively. Samples of different views are augmented via data augmentations, including removing edges and masking node attributes. A GNN-based encoder extracts node-level representations  $h_1$ ,  $h_2$ , and  $h_3$  for three views respectively. Then, a projection head embeds  $h_1$ ,  $h_2$ , and  $h_3$  to  $z_1$ ,  $z_2$ , and  $z_3$  in the latent space, where the contrastive loss is calculated. In this phase, we apply a k-NN filter to select the positive samples of the node  $v$ , which will be discussed in Section 3.2. The GNN-based encoder and projection head are trained by minimizing the contrastive loss. Furthermore, we empirically observe that the proposed hypergraph-augmented view can further enhance the robustness of graph contrastive learning by adopting our framework.

#### 3.2 Robust Graph Contrastive Learning

For node  $v_i$  in the node set of  $N$  nodes, we denote the node embeddings in two different views as  $u_i$  and  $v_i$ . Figure 2 shows positive and negative pairs of node-level graph contrastive learning. For embedding  $u_i$ , the positive sample is the embedding in the other view, denoted as  $v_i$ . Negative samples include negative samples of intra-view and inter-view. Negative samples of intra-view are embeddings of different nodes in the same view, represented as  $u_j$ , where  $j \in [1, N]$  and  $j \neq i$ . Negative samples of inter-view are embeddings of different nodes in the other view, which can be defined as  $v_j$ , where  $j \in [1, N]$  and  $j \neq i$ . Cosine similarity is applied to measure the distance between embeddings. We denote the cosine similarity between  $u_i$  and  $v_i$  as  $\text{sim}(u_i, v_i)$ . Then, the contrastive loss can be calculated with cosine similarities of positive and negative pairs.

During the training of contrastive learning, the embedding obtained from neighbor aggregation will concentrate toward the expectation of embedding belonging to the same class. We also discover a principal challenge undermining conventional contrastive methods: the *false* negative sample problem, i.e., specific negative samples actually belong to the same category of the anchor sample. So we take the neighbors of nodes into consideration and propose a novel method called robust graph contrastive learning. Unlike the general scheme, we introduce a pool of positive samples into the contrastive loss. For embedding  $u_i$ , we use a k-NN filter to select the top-k embeddings of the highest similarity with  $u_i$ , which are the nodes in navy blue in Figure 2. The k-NN filter selects nodes via k-NN search. We denote the pool of positive samples for node  $v_i$  as  $P_i$ . After building the pool for node  $v_i$ , the contrastive loss for each positive pair  $(u_i, v_i)$  can be defined as:

$$\ell(u_i, v_i) = \log \frac{e^{\text{sim}(u_i, v_i)/\tau}}{\underbrace{e^{\text{sim}(u_i, v_i)/\tau}}_{\text{the positive pair}} + \underbrace{\sum_{k \neq i} e^{\text{sim}(u_i, v_k)/\tau}}_{\text{inter-view negative pairs}} + \underbrace{\sum_{k \neq i} e^{\text{sim}(u_i, u_k)/\tau}}_{\text{intra-view negative pairs}} + \underbrace{\sum_{p_k \in P_i} e^{\text{sim}(u_i, p_k)/\tau}}_{\text{positive pairs from } P_i}} \quad (6)$$

where  $p_k$  is the sample from  $P_i$  and  $\tau$  is the temperature parameter. For two views, we can define the overall loss as the average of all positive pairs:

$$\mathcal{L} = -\frac{1}{2N} \sum_{i=1}^N [\ell(u_i, v_i) + \ell(v_i, u_i)] \quad (7)$$

### 4 EXPERIMENTS

We compare our method with SOTAs in the settings of transductive and inductive node classification. For a comprehensive comparison, we use five widely-used datasets in our experiments. As shown in Table 1, we use two types of datasets: (1) citation networks including Cora, Citeseer, Pubmed, and DBLP [2, 26] for transductive learning and (2) biological protein-protein interaction (PPI) networks [43] for inductive learning on multiple graphs.

**Table 1: Introduction of datasets used in experiments.**

Dataset	Type	#Nodes	#Edges	#Features	#Classes
Cora	Transductive	2708	5429	1433	7
Citeseer	Transductive	3327	4732	3703	6
Pubmed	Transductive	19717	44338	500	3
DBLP	Transductive	17716	105734	1639	4
PPI	Inductive	56944	818716	50	121

#### 4.1 Experiment Settings

We adopt the linear evaluation pattern as in [42], and train each model in an unsupervised manner. Then, we use the learned embeddings to train and test a logistic regression classifier. The model is trained for twenty runs and we report the average results on each dataset. To compare the performance with SOTAs, we apply classification accuracy for transductive learning and micro-averaged F1-score for inductive learning.

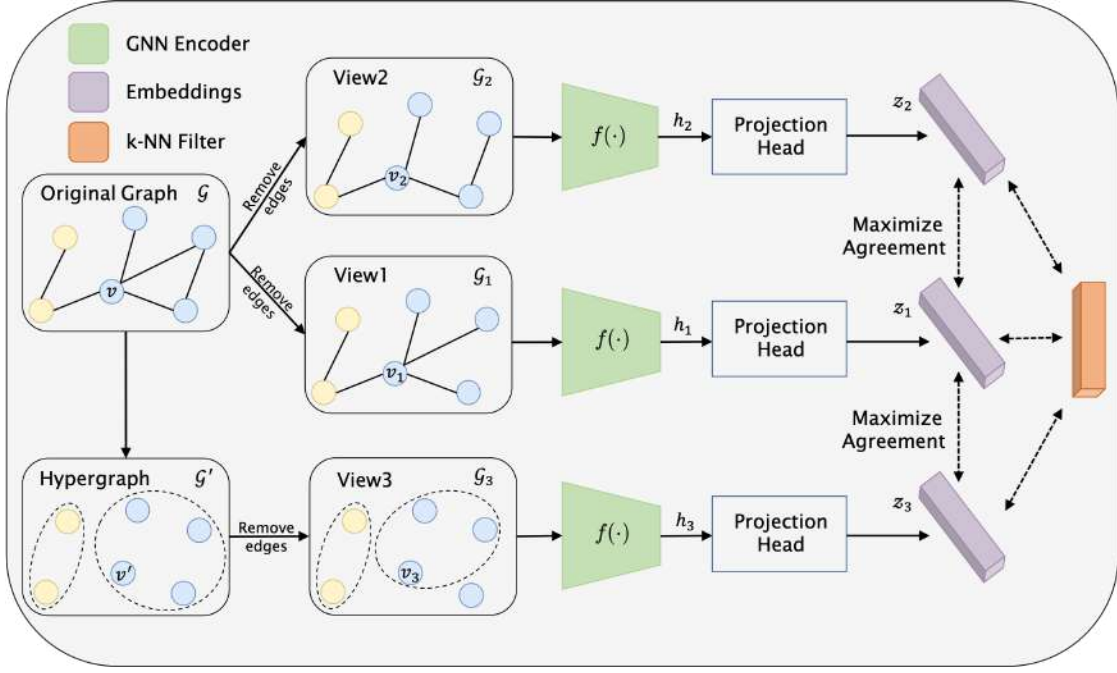


Figure 1: A framework of our method. For data augmentation, we take removing edges for example. The nodes in the same color belong to the same class.

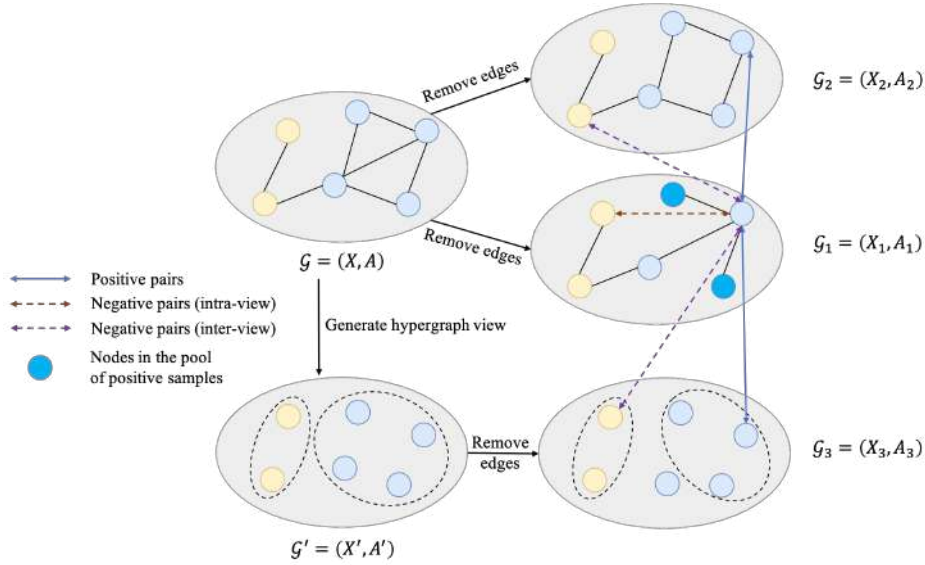


Figure 2: Positive and negative pairs of node-level graph contrastive learning framework. For data augmentation, we take removing edges for example. The nodes in the same color belong to the same class. The size of the pool of positive samples is 2.

## 4.2 Transductive Learning and Inductive Learning

In the setting of transductive learning, we compare our method with SOTAs on four datasets. We take three kinds of representation learning methods as baselines, including (1) traditional methods

DeepWalk [23] and node2vec [9], and (2) deep learning methods GAE, VGAE [16], and DGI [31], and (3) contrastive learning method GRACE [42]. We apply a logistic regression classifier on raw node features to compare the performance of methods except for DeepWalk. As for DeepWalk, embeddings concatenated with input node features are obtained to report the performance. Besides, we take

**Table 2: Comparing our method and SOTAs on node classification, accuracy for transductive tasks and micro-averaged F1-score for inductive tasks. The second column represents the training data for each method, where X,A,Y denote node features, the adjacency matrix, and labels respectively. Red numbers represent the best results of unsupervised methods on different datasets.**

Method	Training Data	Transductive				Inductive
		Cora	Citeseer	Pubmed	DBLP	PPI
Raw features	X	64.8	64.6	84.8	71.6	42.2
Node2vec	A	74.8	52.3	80.3	78.8	—
DeepWalk	A	75.7	50.5	80.5	75.9	—
DeepWalk + features	X,A	73.1	47.6	83.7	78.1	—
GAE	X,A	76.9	60.6	82.9	81.2	—
VGAE	X,A	78.9	61.2	83.0	81.7	—
GraphSAGE-mean	X,A	—	—	—	—	48.6
GraphSAGE-pool	X,A	—	—	—	—	50.2
DGI	X,A	82.6±0.4	68.8±0.7	86.0±0.1	83.2±0.1	63.8±0.2
GRACE	X,A	83.3±0.4	72.1±0.5	86.7±0.1	84.2±0.1	66.2±0.1
<b>Ours</b>	X,A	<b>84.4±0.1</b>	<b>72.6±0.1</b>	<b>87.0±0.1</b>	<b>84.4±0.1</b>	<b>66.2±0.2</b>
SGC	X,A,Y	80.6	69.1	84.8	81.7	—
GCN	X,A,Y	82.8	72.0	84.9	82.7	—
GaAN-mean	X,A,Y	—	—	—	—	96.9±0.2

**Table 3: Ablation study for two modules in our method, hypergraph view and k-NN filter. We report the accuracies of node classification on four datasets. Ours (-hyper) and Ours (-knn) denote the method without hypergraph view and k-NN filter respectively. The bold numbers represent the best results.**

Method	Cora	Citeseer	Pubmed	DBLP
Ours (-hyper)	83.8±0.7	72.2±0.5	86.0±0.3	84.1±0.3
Ours (-knn)	83.4±0.2	71.4±0.4	86.3±0.2	84.0±0.3
<b>Ours</b>	<b>84.4±0.1</b>	<b>72.6±0.1</b>	<b>87.0±0.1</b>	<b>84.4±0.1</b>

two supervised learning methods SGC [33] and GCN [15] for comparison. As shown in Table 2, our method reaches the SOTA performance on four datasets. The classification accuracy on Cora dataset is 84.4%, which is 1.1% higher than that of GRACE. On Citeseer dataset, the performance of our method is 0.5% higher than that of GRACE. Experimental results demonstrate our method’s superiority over benchmarks.

For inductive learning, baselines can be divided into three categories: (1) raw features, and (2) deep learning methods GraphSAGE [10] and DGI [31], and (3) contrastive learning method GRACE [42]. For a fair comparison, we only calculate negative samples for one anchor node as other nodes within the same graph, the same as [42]. Besides, we report the performance of the supervised learning method GaAN-mean [41]. We conduct an inductive learning experiment on PPI dataset. The performance of our method is competitive with GRACE and significantly better than other baselines, as shown in Table 2.

The results verify the effectiveness of hypergraph-augmented view and robust graph contrastive learning. The model can better capture the global information from nodes and corresponding edges by introducing the hypergraph-augmented view, which brings the

improvement of performance. Besides, the robust graph contrastive learning takes the neighbors of nodes into consideration and better guides the training process.

### 4.3 Ablation Study

We conduct experiments for the ablation study. *Ours (-hyper)* represents our method without the hypergraph-augmented view, and *Ours (-knn)* is our method removing the k-NN filter module. For the method without a hypergraph view, we only generate two views from the original graph. The method of removing the k-NN filter uses the traditional contrastive loss. As shown in Table 3, our method that jointly applies hypergraph view and k-NN filter outperforms *Ours (-hyper)* and *Ours (-knn)*. The results verify the effectiveness of our proposed method.

### 4.4 Hyperparameter Experiments

We study the performance of node classification with different options of the temperature  $\tau$  in the contrastive loss and further conduct exploration experiments on four benchmark datasets. Based on the experiments, we observe that for Pubmed and DBLP, RH-GCL with  $\tau = 0.7$  achieves the best performance, as shown in Figure 3. The result of Cora reaches the peak value when  $\tau = 0.4$ . Besides, our method with  $\tau = 0.9$  achieves the best performance for Citeseer.

We further explore the influence of layers of the GNN-based encoder on our method. The number of the GNN layer is sampled from the range of  $\{2, 3, 4, 5\}$ . As shown in Figure 4, the performance of our method reaches its best when the GNN layer number is 2 for all datasets. The model performance degenerates as the GNN layer number further increases. We conjecture the reasons behind such a phenomenon include: 1) the conventional over-smoothing issue of the GNN-based encoder, which can be proved by the general performance degradation on all datasets; 2) when the GNN-based encoder is over-deep, the features of negative pairs are supposed to be excessively homogeneous, which undermines the discriminativeness of the learned features, which can be proved by the difference of the performance degradation gap on the small-scale and large-scale datasets. Specifically, for small-scale datasets, such as Cora and Citeseer, a GNN-based encoder with 2 layers is enough to learn the representations of nodes, and the classification accuracy of our model decreases more significantly on small-scale datasets than the large-scale datasets when the GNN-based encoder is over-deep.

## 5 CONCLUSIONS

We propose a novel framework for graph contrastive learning called RH-GCL. Unlike the traditional framework, we introduce a hypergraph-augmented view into the framework to better capture the global information of nodes. Furthermore, we take the neighbors of nodes into consideration and propose a novel method called robust graph contrastive learning. Our method outperforms benchmark methods in the settings of transductive and inductive node classification. Specifically, the accuracy of node classification on Cora dataset is 84.4%, which is 1.1% higher than that of GRACE. We also perform the ablation study to verify the effectiveness of our proposed method. For future work, we plan to apply our method to graph-level downstream tasks. Another interesting direction is to study the impact of different encoders on our method.

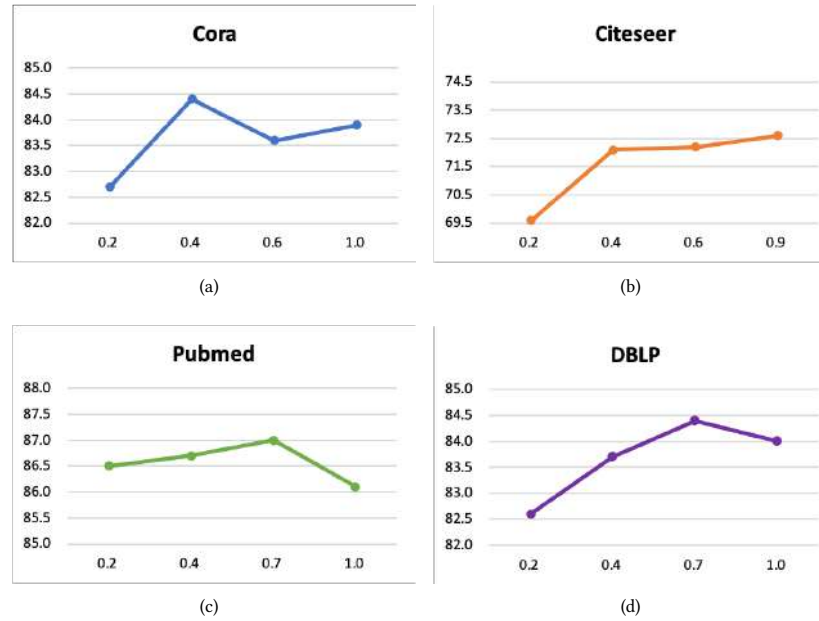


Figure 3: Performance of our method (node classification accuracy in transductive learning) with different choices of the temperature  $\tau$ . The x axis denotes the choice of different temperatures and y axis represents the classification accuracy.

	Cora	Citeseer	Pubmed	DBLP	
2	84.4	72.6	87.0	84.4	high
3	82.3	70.2	85.7	84.4	
5	75.3	58.1	84.7	83.4	low

Figure 4: Transductive node classification evaluation of different GNN layers. For one dataset, we apply the same settings.

## REFERENCES

- [1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*. PMLR, 1298–1312.
- [2] Aleksandar Bojchevski and Stephan Günnemann. 2017. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. *arXiv preprint arXiv:1707.03815* (2017).
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [4] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. Cert: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766* (2020).
- [5] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3558–3565.
- [6] Hang Gao, Jiangmeng Li, Wenwen Qiang, Lingyu Si, Fuchun Sun, and Changwen Zheng. 2022. Bootstrapping Informative Graph Augmentation via A Meta Learning Approach. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, Luc De Raedt (Ed.). ijcai.org, 3001–3007. <https://doi.org/10.24963/ijcai.2022/416>
- [7] Hang Gao, Jiangmeng Li, Wenwen Qiang, Lingyu Si, Bing Xu, Changwen Zheng, and Fuchun Sun. 2022. Robust Causal Graph Representation Learning against Confounding Effects. *CoRR* abs/2208.08584 (2022). <https://doi.org/10.48550/arXiv.2208.08584> arXiv:2208.08584
- [8] Hang Gao, Jiangmeng Li, Peng Qiao, and Changwen Zheng. 2022. Weight-Aware Graph Contrastive Learning. In *Artificial Neural Networks and Machine Learning–ICANN 2022: 31st International Conference on Artificial Neural Networks, Bristol, UK, September 6–9, 2022, Proceedings, Part II*. Springer, 719–730.
- [9] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 855–864.
- [10] Will Hamilton, Zhitaoying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).
- [11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.
- [12] Fenyun Hu, Yanqiao Zhu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Hierarchical graph convolutional networks for semi-supervised node classification. *arXiv preprint arXiv:1902.06667* (2019).
- [13] Xu Ji, Joao F Henriques, and Andrea Vedaldi. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9865–9874.
- [14] Yifan Jin, Jiangmeng Li, Zheng Lian, Chengbo Jiao, and Xiaohui Hu. 2022. Supporting Medical Relation Extraction via Causality-Pruned Semantic Dependency Forest. *arXiv preprint arXiv:2208.13472* (2022).
- [15] Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
- [16] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).
- [17] Namkyeong Lee, Junseok Lee, and Chanyoung Park. 2022. Augmentation-free self-supervised learning on graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 7372–7380.

- [18] Jiangmeng Li, Wenyi Mo, Wenwen Qiang, Bing Su, and Changwen Zheng. 2022. Supporting Vision-Language Model Inference with Causality-pruning Knowledge Prompt. *arXiv preprint arXiv:2205.11100* (2022).
- [19] Jiangmeng Li, Wenwen Qiang, Yanan Zhang, Wenyi Mo, Changwen Zheng, Bing Su, and Hui Xiong. 2022. MetaMask: Revisiting Dimensional Confounder for Self-Supervised Learning. *CoRR abs/2209.07902* (2022). <https://doi.org/10.48550/arXiv.2209.07902> arXiv:2209.07902
- [20] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, and Bing Su. 2022. RHMC: Modeling consistent information from deep multiple views via Regularized and Hybrid Multiview Coding. *Knowledge-Based Systems* 241 (2022), 108201.
- [21] Jiangmeng Li, Wenwen Qiang, Changwen Zheng, Bing Su, and Hui Xiong. 2022. Metaug: Contrastive learning via meta feature augmentation. In *International Conference on Machine Learning*. PMLR, 12964–12978.
- [22] Dongju Park and Chang Wook Ahn. 2019. Self-supervised contextual data augmentation for natural language processing. *Symmetry* 11, 11 (2019), 1393.
- [23] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 701–710.
- [24] Leonardo FR Ribeiro, Pedro HP Saverese, and Daniel R Figueiredo. 2017. struc2vec: Learning node representations from structural identity. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 385–394.
- [25] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*. Springer, 593–607.
- [26] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI magazine* 29, 3 (2008), 93–93.
- [27] Susheel Suresh, Pan Li, Cong Hao, and Jennifer Neville. 2021. Adversarial graph augmentation to improve graph contrastive learning. *Advances in Neural Information Processing Systems* 34 (2021), 15920–15933.
- [28] Hui Tang, Xun Liang, Yuhui Guo, Xiangping Zheng, and Bo Wu. 2022. Graph Fine-Grained Contrastive Representation Learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3478–3482.
- [29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 776–794.
- [30] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).
- [31] Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. 2019. Deep graph infomax. *ICLR (Poster)* 2, 3 (2019), 4.
- [32] Haonan Wang, Jieyu Zhang, Qi Zhu, and Wei Huang. 2022. Augmentation-free graph contrastive learning. *arXiv preprint arXiv:2204.04874* (2022).
- [33] Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*. PMLR, 6861–6871.
- [34] Jun Xia, Lirong Wu, Jintao Chen, Bozhen Hu, and Stan Z Li. 2022. Simgrace: A simple framework for graph contrastive learning without data augmentation. In *Proceedings of the ACM Web Conference 2022*. 1070–1079.
- [35] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
- [36] Longqi Yang, Liangliang Zhang, and Wenjing Yang. 2021. Graph adversarial self-supervised learning. *Advances in Neural Information Processing Systems* 34 (2021), 14887–14899.
- [37] Mang Ye, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. 2019. Unsupervised embedding learning via invariant and spreading instance feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6210–6219.
- [38] Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. 2021. Graph contrastive learning automated. In *International Conference on Machine Learning*. PMLR, 12121–12132.
- [39] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in neural information processing systems* 33 (2020), 5812–5823.
- [40] Jiaqi Zeng and Pengtao Xie. 2021. Contrastive self-supervised learning for graph classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 10824–10832.
- [41] Jiani Zhang, Xingjian Shi, Junyuan Xie, Hao Ma, Irwin King, and Dit-Yan Yeung. 2018. Gaan: Gated attention networks for learning on large and spatiotemporal graphs. *arXiv preprint arXiv:1803.07294* (2018).
- [42] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang. 2020. Deep graph contrastive representation learning. *arXiv preprint arXiv:2006.04131* (2020).
- [43] Marinka Zitnik and Jure Leskovec. 2017. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics* 33, 14 (2017), i190–i198.

# Quantum kernel subspace alignment for unsupervised domain adaptation

Xi He\*

Feiyu Du\*

xihe@sust.edu.cn

feiyu.du@sust.edu.cn

Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology

Xi'an, Shaanxi, China

## ABSTRACT

Domain adaptation (DA), the sub-realm of the transfer learning, attempts to deal with machine learning tasks on an unprocessed data domain with the different, but related labeled source domain. However, the classical DA can not efficiently deal with the cross-domain tasks in quantum mechanical scenarios. In this paper, the quantum kernel subspace alignment algorithm is proposed to achieve the procedure of DA by extracting the non-linear features with the quantum kernel method and aligning the two domains with the unitary evolution. The method presented in our work can be implemented on the universal quantum computer with the quantum basic linear algebra subroutines. Based on the algorithmic complexity analysis, the procedure of the QKSA can be implemented with at least quadratic quantum speedup compared with the classical DA algorithms.

## CCS CONCEPTS

• **Computing methodologies** → **Transfer learning; Machine learning; Kernel methods;**

## KEYWORDS

machine learning, quantum machine learning, kernel methods, domain adaptation

## ACM Reference Format:

Xi He and Feiyu Du. 2023. Quantum kernel subspace alignment for unsupervised domain adaptation. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3590003.3590054>

## 1 INTRODUCTION

Quantum machine learning (QML) is a interdisciplinary field, which combines the quantum computing (QC) and the machine learning (ML) [4]. It utilizes the properties of quantum entanglement and quantum superposition to realize the learning process on quantum

devices with quantum advantages [3]. On the one hand, the QML aims to accomplish machine learning tasks by QC techniques with quantum speed up compared with classical ML methods. On the other hand, the QML attempts to deal with tasks in the quantum physics scenario such as quantum many body problems, which can not be achieved efficiently by classical computing techniques. At present, the QML has developed to be a systematic research field. Specifically, the QML can be applied to the tasks of supervised learning such as classification [17, 20, 22, 26], linear regression [23, 25]; unsupervised learning such as dimensionality reduction [2, 5, 13, 19], clustering [1]; reinforcement learning [6, 7].

In the field of ML, the transfer learning (TL) aims to accomplish tasks in the target domain with the knowledge of the different but related source domain [16]. As an important branch of TL, the domain adaptation (DA) aligns the source domain with the target domain to achieve the procedure of TL [15]. In recent years, QC techniques are applied to the field of TL with quantum advantage. However, the quantum domain adaptation algorithms in ref. [10–12] utilizes linear transformations to achieve the alignment from the source domain to the target domain. Nonlinear features in the source and target domain can not be efficiently extracted to promote the performance of the domain transfer.

In this paper, we propose the quantum kernel subspace alignment algorithm (QKSA) for DA. The QKSA algorithm utilizes the kernel method to map the data of the source and target domain to the quantum kernel Hilbert feature space respectively. Subsequently, the dimensionality reduction such as the kernel principal component analysis in this paper is invoked to reduce the data dimension of the two domains in the high-dimensional feature space to the corresponding subspace. Then, the coordinate system of the source domain subspace is aligned to the target domain coordinate system by a unitary transformation. Finally, the classifier can be trained on the labelled source domain to predict the labels of the target domain data. Based on the algorithmic complexity analysis, the QKSA can be implemented with quantum speedup compared with the classical DA algorithms.

The contents of the manuscript are arranged as follows. In section 2, the basic settings of the DA are presented. Subsequently, the quantum kernel method is described in section 3. Then, the implementation of the QKSA is provided in section 4. Finally, the paper is concluded in section 5.

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590054>

## 2 DOMAIN ADAPTATION

Given a labelled source domain  $\mathcal{D}_s = \{x_i^{(s)}\}_{i=1}^{n_s} \in \mathbb{R}^D$  with labels  $\{y_i^{(s)}\}_{i=1}^{n_s} \in \mathcal{Y}_s$  and a target domain  $\mathcal{D}_t = \{x_j^{(t)}\}_{j=1}^{n_t} \in \mathbb{R}^D$ . Domain adaptation (DA) aims to predict the labels  $\{y_j^{(t)}\}_{j=1}^{n_t} \in \mathcal{Y}_t$  of  $\mathcal{D}_t$  with the learning process of  $\mathcal{D}_s$  based on the assumption that  $\mathcal{D}_s$  and  $\mathcal{D}_t$  specifically have the same feature and label space. The source domain data matrix  $X_s = (x_1^{(s)}, x_2^{(s)}, \dots, x_{n_s}^{(s)}) \in \mathbb{R}^{D \times n_s}$ . The target domain data matrix  $X_t = (x_1^{(t)}, x_2^{(t)}, \dots, x_{n_t}^{(t)}) \in \mathbb{R}^{D \times n_t}$ . In this paper, we discuss the most general type of DA, the unsupervised DA, which is initialized with a totally unlabelled  $\mathcal{D}_t$ . As presented in

$$\mathcal{D}_s : \{x_i^{(s)}, y_i^{(s)}\}_{i=1}^{n_s} \xrightarrow{\{x_j^{(t)}\}_{j=1}^{n_t}} f_s \xrightarrow{\{y_j^{(t)}\}_{j=1}^{n_t}} \{y_j^{(t)}\}_{j=1}^{n_t} \quad (1)$$

$$\downarrow \min \mathcal{L}_D \quad (2)$$

$$\mathcal{D}_t : \{x_i^{(t)}\}_{i=1}^{n_s} \xrightarrow{\quad} f_t \xrightarrow{\quad} \{y_j^{(t)}\}_{j=1}^{n_t}, \quad (3)$$

the labels of the target domain data can be predicted by DA techniques based on the knowledge of source domain, where  $f_s, f_t$  represent the feature extraction models of the source and target domain respectively;  $\mathcal{L}_D$  represents a measure of the divergence between the source and target domains.

## 3 QUANTUM KERNEL METHOD

Given the source and target domain data samples in the form of quantum states, they can be directly processed by the quantum devices. However, if the original data are classical vectors, they can be encoded as quantum states by parameterized quantum circuits (PQC) efficiently [24].

The source and target data  $X_s, X_t$  can be represented as the quantum states

$$|\psi_{X_s}\rangle = U_{enc}^{(s)}|0\rangle^{\log D} = \sum_{i=1}^{n_s} \sum_{m=1}^D x_{mi}^{(s)} |i\rangle|m\rangle = \sum_{i=1}^{n_s} |i\rangle|x_i^{(s)}\rangle, \quad (4)$$

$$|\psi_{X_t}\rangle = U_{enc}^{(t)}|0\rangle^{\log D} = \sum_{j=1}^{n_t} \sum_{m=1}^D x_{mj}^{(t)} |j\rangle|m\rangle = \sum_{j=1}^{n_t} |j\rangle|x_j^{(t)}\rangle, \quad (5)$$

respectively where  $U_{enc}^{(s)}, U_{enc}^{(t)}$  are the corresponding encoding quantum circuit;  $\sum_{m,i} |x_{mi}^{(s)}|^2 = \sum_{m,j} |x_{mj}^{(t)}|^2 = 1$  with the amplitude encoding.

In the realm of ML, the kernel method is one of the most efficient techniques with convincing theoretical foundation. Specifically, the utilization of quantum kernel methods is an alternative of the quantum feature map in the corresponding feature space, namely the reproducing kernel Hilbert space (RKHS) [9, 21]. Based on the quantum kernel, the original quantum data can be mapped to the quantum Hilbert space. Compared with the classical kernel, the data features which are particularly hard to be recognized classically can be efficiently dealt with by the quantum kernel. In addition, compared with the variational quantum circuit training, the kernel-based model can be trained smoothly with the guarantee of achieving the global minima. The kernel method utilizes the kernel map  $\phi : \mathcal{X} \rightarrow \mathcal{F}$  to obtain the corresponding source and target

domain data

$$\begin{aligned} |\phi(X_s)\rangle &= U_{PQC}^{(s)}|\psi_{X_s}\rangle = \sum_{i=1}^{n_s} \sum_{m=1}^D \phi(x_{mi}^{(s)}) |i\rangle|m\rangle = \sum_{i=1}^{n_s} |i\rangle|\phi(x_i^{(s)})\rangle \\ &= \sum_{m=1}^D |\tilde{\phi}(x_m^{(s)})\rangle|m\rangle, \end{aligned} \quad (6)$$

$$\begin{aligned} |\phi(X_t)\rangle &= U_{PQC}^{(t)}|\psi_{X_t}\rangle = \sum_{j=1}^{n_t} \sum_{m=1}^D \phi(x_{mj}^{(t)}) |j\rangle|m\rangle = \sum_{j=1}^{n_t} |j\rangle|\phi(x_j^{(t)})\rangle \\ &= \sum_{m=1}^D |\tilde{\phi}(x_m^{(t)})\rangle|m\rangle, \end{aligned} \quad (7)$$

respectively in the high-dimensional kernel space where  $\sum_{m,i} |\phi(x_{mi}^{(s)})|^2 = \sum_{m,j} |\phi(x_{mj}^{(t)})|^2 = 1$ . Subsequently, the source domain kernel matrix  $K_{ss}(x_i^{(s)}, x_j^{(s)}) = |\langle\phi(x_i^{(s)})|\phi(x_j^{(s)})\rangle|^2 = \text{Tr}(\rho_i^{(s)} \rho_j^{(s)})$  and the target domain kernel matrix  $K_{tt}(x_i^{(t)}, x_j^{(t)}) = |\langle\phi(x_i^{(t)})|\phi(x_j^{(t)})\rangle|^2 = \text{Tr}(\rho_i^{(t)} \rho_j^{(t)})$  where  $\rho_i^{(s)} = |\phi(X_s)\rangle\langle\phi(X_s)|, \rho_i^{(t)} = |\phi(X_t)\rangle\langle\phi(X_t)|$ .

Traditional QML algorithms can invoke the kernel methods directly with the basic independent and identically distributed (i.i.d) assumption. However, when the source and target domain are generated from different distributions, the distribution mismatch should be considered to achieve high accuracy cross-domain prediction. Inspired from TL, the quantum domain adaptation kernel between the source and target domain can be defined as

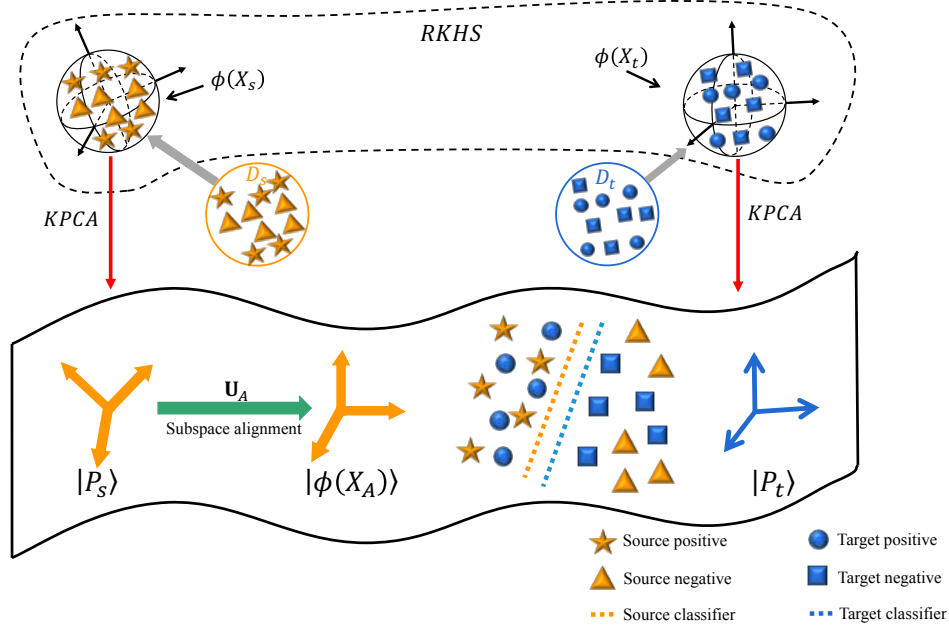
$$\begin{aligned} K_{st}(x_i^{(s)}, x_j^{(t)}) &= |\langle\phi(x_i^{(s)})|\phi(x_j^{(t)})\rangle|^2 = \text{Tr}(\rho_i^{(s)} \rho_j^{(t)}) \\ &= \left( \begin{array}{ccc} \langle\phi(x_1^{(s)})|\phi(x_1^{(t)})\rangle & \cdots & \langle\phi(x_1^{(s)})|\phi(x_{n_t}^{(t)})\rangle \\ \vdots & \ddots & \vdots \\ \langle\phi(x_{n_s}^{(s)})|\phi(x_1^{(t)})\rangle & \cdots & \langle\phi(x_{n_s}^{(s)})|\phi(x_{n_t}^{(t)})\rangle \end{array} \right) \in \mathbb{R}^{n_s \times n_t} \end{aligned} \quad (8)$$

to estimate the divergence between the two domains in the kernel space.

Specifically, the kernel estimation can be achieved by performing the kernel map  $U_k(x_i)$ , the swap test sequentially on the kernel-mapped source and target domain data. In the following, the quantum kernel-based domain adaptation can be implemented.

## 4 QUANTUM KERNEL SUBSPACE ALIGNMENT

The quantum kernel subspace alignment (QKSA) presented in this section is designed for achieving the transfer learning from  $\mathcal{D}_s$  to  $\mathcal{D}_t$  on the universal quantum devices. The schematic diagram of the QKSA algorithm is depicted in Fig 1. After performing the quantum kernel tricks on  $\mathcal{D}_s$  and  $\mathcal{D}_t$ , the quantum kernel data are projected to the same subspace resulting in the corresponding coordinate systems. Because of the divergence between the two domains, the two coordinates point to different directions respectively. Thus, the unitary evolution  $U_A$  is designed for aligning the source domain coordinate system to the target domain. Specifically, given the kernel matrix estimated from the source and target domain data,



**Figure 1: The schematic diagram of the QKSA.** The given source domain data  $\mathcal{D}_s$  and target domain data  $\mathcal{D}_t$  are firstly mapped to the RKHS resulting in  $\phi(X_s)$  and  $\phi(X_t)$  respectively. Subsequently, we utilize the principal component analysis operations to project  $\phi(X_s), \phi(X_t)$  to the corresponding subspace. After the quantum embedding, the source domain kernel state  $\rho_s$  is aligned to the target domain kernel state  $\rho_t$  by performing the domain adaptation  $U_A$ . Finally, the source classifier can be transferred to the target domain to achieve the labels  $Y_t$ .

the corresponding quantum state  $\rho_{K_Q} = \text{tr}_i\{|\phi(X_Q)\rangle\langle\phi(X_Q)|\}$  for  $Q \in \{s, t\}$  can be obtained trivially.

For the specified domain  $Q$ , the quantum phase estimation (QPE) [14]

$$\mathbf{U}_{\text{PE}}(\rho_{K_Q}) = (\mathbf{QFT}^\dagger \otimes \mathbf{I}) \left( \sum_{\tau=0}^{T-1} |\tau\rangle\langle\tau| \otimes e^{i\rho_{K_Q} \tau \Delta t} \right) (\mathbf{H}^{\otimes \log(2n_Q)} \otimes \mathbf{I}) \quad (9)$$

is performed on  $\rho_{K_Q}$  with  $O(\Delta t^2/\epsilon)$  copies resulting in the quantum state

$$|\psi_1\rangle = \sum_{i=1}^{n_Q} |i\rangle |\lambda_i^{(Q)}\rangle |\alpha_i^{(Q)}\rangle \quad (10)$$

where  $\mathbf{QFT}^\dagger$  represents the inverse quantum Fourier transform; the slice time  $\Delta t = t/l$  for some large  $l$ . Therefore, the coordinate system of the kernel domain data can be obtained as

$$|P_Q\rangle = \sum_{i=1}^d \sum_{p=1}^D \alpha_{pi}^{(Q)} |i\rangle |p\rangle = \sum_{i=1}^d |i\rangle |\alpha_i^{(Q)}\rangle \quad (11)$$

by sampling the  $d$  largest eigenvalues  $\lambda_i^{(Q)}$  and the corresponding eigenvectors  $|\alpha_i^{(Q)}\rangle$ . The quantum state

$$\rho_Q = \text{tr}_i\{|P_Q\rangle\langle P_Q|\} \quad (12)$$

can be subsequently obtained by tracing out the  $|i\rangle$  register.

Inspired from the tricks in ref. [8, 18], the quantum state  $|A\rangle$  which represents the alignment matrix  $A$  can be obtained as

$$\begin{aligned} |\psi_0\rangle &= \sum_{i=1}^D \sum_{j=1}^d |i\rangle^{I_1} |j\rangle^{I_2} |0\rangle^B |0\rangle^{C_1} \\ &\xrightarrow{\mathbf{U}_P(P_s, P_t)} |\psi_1^{(s)}\rangle = \sum_{i=1}^D \sum_{j=1}^d |i\rangle^{I_1} |j\rangle^{I_2} \\ &\quad \otimes \left[ \frac{1}{\sqrt{2}} (|0\rangle |\alpha_i^{(s)}\rangle + |1\rangle |\alpha_j^{(t)}\rangle) \right]^{BC_1} \\ &\xrightarrow{\mathbf{U}_{\text{PE}, C - R_y(2 \arcsin(\langle \tilde{\phi}(x_i^{(s)}) | A_{s_j} \rangle))}} |\psi_2^{(s)}\rangle \\ &\quad \otimes \left( \sqrt{1 - \langle \alpha_i^s | \alpha_j^t \rangle^2} |0\rangle^R + \langle \alpha_i^s | \alpha_j^t \rangle |1\rangle^R \right) \\ &\xrightarrow{\text{Uncompute \& measurement}} |A\rangle \\ &= \sum_{i=1}^D \sum_{j=1}^d \langle \alpha_i^s | \alpha_j^t \rangle |j\rangle \end{aligned} \quad (13)$$

where  $I_1, I_2, B, C_1, C_2$  are the quantum registers;  $\mathbf{U}_S(P_Q)|i\rangle|0\rangle = |i\rangle|\alpha_Q\rangle$ ;  $\mathbf{U}_P(P_s, P_t) = I^{I_2} \otimes |0\rangle\langle 0|^B \otimes \mathbf{U}_S^{I_1 C_1}(P_s) + I^{I_1} \otimes |1\rangle\langle 1|^B \otimes \mathbf{U}_S^{I_2 C_1}(P_t)$ . For simplicity, the procedure of computing the transformation matrix state  $|A\rangle$  is specifically represented by  $|A\rangle = \mathbf{U}_A(P_s, P_t)$ , where  $U_A$  represents the whole process of eq. (13). Therefore, the total runtime of equation

(13) is in the scale  $O(\text{poly}(D))\epsilon^{-1}\|P_s\|_F\|P_t\|_F/\|P_s^T P_t\|_F$  represents the Frobenius norm;  $\epsilon$  is the error coefficient. Similarly, the subspace data  $|\phi(\hat{X}_Q)\rangle$  can be obtained by performing  $U_A(P_Q, \phi(X_Q))$  on  $|\psi_0\rangle$  in the computational complexity  $O(\text{poly}(D))\epsilon^{-1}\|P_Q\|_F\|\phi(X_Q)\|_F/\|P_Q^T \phi(X_Q)\|_F$ .

The domain transfer map  $U_A$  is performed to align  $|\phi(\hat{X}_s)\rangle$  to  $|\hat{X}_t\rangle$ . According to the kernel theory, this operation can be reduced to align the coordinate system  $|P_s\rangle$  to  $|P_t\rangle$  by  $|\phi(X_A)\rangle = U_A(A, \phi(\hat{X}_s))$  in corresponding run time  $O(\text{poly}(D))\epsilon^{-1}\|A\|_F\|\phi(\hat{X}_s)\|_F/\|A^T \phi(\hat{X}_s)\|_F$ .

After the procedure of DA, the quantum classifier is trained on the source aligned data  $\{|\phi(X_A)\rangle, Y_s\}$ . The trained classifier is then performed on the target data  $\{|\phi(\hat{X}_t)\rangle\}$  to predict the corresponding labels  $Y_t = \{y_j^{(t)}\}_{j=1}^{n_t}$ . Concretely, the global classifier such as the quantum support vector machine (qSVM) can be invoked to achieve the label prediction with logarithmic resources in the scale and dimension of the given data. In addition, the quantum  $k$ -nearest neighbors algorithm can be utilized to predict the target labels with quadratic speedup compared to the corresponding classical classifier.

## 5 CONCLUSION AND DISCUSSION

In this paper, the QKSA algorithm, based on the quantum kernel method and the unitary evolution, is proposed to achieve the procedure of DA with quadratic speedup compared to the classical algorithms. The QKSA method utilizes the quantum kernel to extract the nonlinear features of the data, and aligns the source domain to the target domain to realize the DA. The procedure of the domain alignment can be implemented in time  $O(\sqrt{D})$ . Thus, the complexity analysis shows that the QKSA can achieve quadratic speedup compared to the classical process.

However, some problems need further exploration. The implementation of the QKSA requires fully coherent evolution and high-depth quantum circuits. The unitary evolution for aligning the two domains should be decomposed into considerable amount of one and two-qubit quantum gates. The depth of the PQC of the QKSA increases rapidly resulting in the difficulty of hardware implementation. Thus, how to implement the procedure of kernel subspace alignment on the noisy intermediate-scale quantum devices is an interesting research topic in the future.

## ACKNOWLEDGMENTS

This paper is supported by Natural Science Basic Research Program of Shaanxi (Program No. 2022JQ-018).

## REFERENCES

- [1] Esma Aïmeur, Gilles Brassard, and Sébastien Gambs. 2013. Quantum speed-up for unsupervised learning. *Machine Learning* 90, 2 (2013), 261–287.
- [2] Mohammad H Amin, Evgeny Andriyash, Jason Rolfe, Bohdan Kulchytksyy, and Roger Melko. 2018. Quantum boltzmann machine. *Physical Review X* 8, 2 (2018), 021050.
- [3] Frank Arute, Kunal Arya, Ryan Babbush, Dave Bacon, Joseph C Bardin, Rami Barends, Rupak Biswas, Sergio Boixo, Fernando GSL Brandao, David A Buell, et al. 2019. Quantum supremacy using a programmable superconducting processor. *Nature* 574, 7779 (2019), 505–510.
- [4] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. 2017. Quantum machine learning. *Nature* 549, 7671 (2017), 195–202.
- [5] Iris Cong and Luming Duan. 2016. Quantum discriminant analysis for dimensionality reduction and classification. *New Journal of Physics* 18, 7 (2016), 073011.
- [6] Daoyi Dong, Chunlin Chen, Hanxiong Li, and Tzyh-Jong Tarn. 2008. Quantum reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 38, 5 (2008), 1207–1220.
- [7] Vedran Dunjko, Jacob M Taylor, and Hans J Briegel. 2017. Advances in quantum reinforcement learning. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 282–287.
- [8] Aram W Harrow, Avinandan Hassidim, and Seth Lloyd. 2009. Quantum algorithm for linear systems of equations. *Physical review letters* 103, 15 (2009), 150502.
- [9] Vojtěch Havlíček, Antonio D Córcoles, Kristan Temme, Aram W Harrow, Abhinav Kandala, Jerry M Chow, and Jay M Gambetta. 2019. Supervised learning with quantum-enhanced feature spaces. *Nature* 567, 7747 (2019), 209–212.
- [10] Xi He. 2020. Quantum correlation alignment for unsupervised domain adaptation. *Physical Review A* 102, 3 (2020), 032410.
- [11] Xi He. 2020. Quantum subspace alignment for domain adaptation. *Physical Review A* 102, 6 (2020), 062403.
- [12] Xi He, Feiyu Du, Mingyuan Xue, Xiaogang Du, Tao Lei, and AK Nandi. 2023. Quantum classifiers for domain adaptation. *Quantum Information Processing* 22, 2 (2023), 105.
- [13] Xi He, Li Sun, Chufan Lyu, and Xiaoting Wang. 2019. Quantum locally linear embedding. *arXiv preprint arXiv:1910.07854* (2019).
- [14] Michael A. Nielsen and Isaac L. Chuang. 2011. *Quantum Computation and Quantum Information: 10th Anniversary Edition*. Cambridge University Press. <https://www.amazon.com/Quantum-Computation-Information-10th-Anniversary/dp/1107002176?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1107002176>
- [15] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.
- [16] Lorient Y Pratt. 1993. Discriminability-based transfer between neural networks. In *Advances in neural information processing systems*. 204–211.
- [17] Patrick Rebentrost, Masoud Mohseni, and Seth Lloyd. 2014. Quantum support vector machine for big data classification. *Physical review letters* 113, 13 (2014), 130503.
- [18] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. 2018. Quantum singular-value decomposition of nonsparsely low-rank matrices. *Physical review A* 97, 1 (2018), 012327.
- [19] Jonathan Romero, Jonathan P Olson, and Alan Aspuru-Guzik. 2017. Quantum autoencoders for efficient compression of quantum data. *Quantum Science and Technology* 2, 4 (2017), 045001.
- [20] Maria Schuld, Alex Bocharov, Krysta M Svore, and Nathan Wiebe. 2020. Circuit-centric quantum classifiers. *Physical Review A* 101, 3 (2020), 032308.
- [21] Maria Schuld and Nathan Killoran. 2019. Quantum machine learning in feature hilbert spaces. *Physical review letters* 122, 4 (2019), 040504.
- [22] Maria Schuld and Francesco Petruccione. 2018. Quantum ensembles of quantum classifiers. *Scientific reports* 8, 1 (2018), 1–12.
- [23] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. 2016. Prediction by linear regression on a quantum computer. *Physical Review A* 94, 2 (2016), 022342.
- [24] Xiaoming Sun, Guojing Tian, Shuai Yang, Pei Yuan, and Shengyu Zhang. 2023. Asymptotically optimal circuit depth for quantum state preparation and general unitary synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* (2023).
- [25] Nathan Wiebe, Daniel Braun, and Seth Lloyd. 2012. Quantum algorithm for data fitting. *Physical review letters* 109, 5 (2012), 050505.
- [26] Nathan Wiebe, Ashish Kapoor, and Krysta M Svore. 2018. Quantum nearest-neighbor algorithms for machine learning. *Quantum Information and Computation* 15 (2018).

# Cross-Modal Audio-Text Retrieval via Sequential Feature Augmentation

Fuhu Song  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
songfh20@mails.jlu.edu.cn

Jifeng Hu\*  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
hujf21@mails.jlu.edu.cn

Che Wang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
wch675413020@163.com

Jiao Huang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
huangjiao20@mails.jlu.edu.cn

Haowen Zhang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
2803702356@qq.com

Yi Wang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
wangyi2220@mails.jlu.edu.cn

## ABSTRACT

The goal of cross-modal audio-text retrieval is to retrieve the target audio clips (textual descriptions), which should be relevant to a given textual (audial) query. It is a challenging task because it necessitates learning comprehensive feature representations for two different modalities and unifying them into a common embedding space. However, most existing cross-modal audio-text retrieval approaches do not explicitly learn the sequential representation in audio features. Moreover, their method of directly employing a fully connected neural network to transform the different modalities into a common space is detrimental to sequential features. In this paper, we introduce a sequential feature augmentation framework based on reinforcement learning and feature fusion to enhance the sequential feature for cross-modal features. First, we adopt reinforcement learning to explore effective sequential features in audial and textual features. Then, a recurrent fusion module is applied as a feature enhancement component to project heterogeneous features into a common space. Extensive experiments are conducted on two prevalent datasets: the AudioCaps and the Clotho. The results demonstrate that our method gains a significant improvement over previous state-of-the-art methods.

## CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**; **Top-k retrieval in databases**; • **Theory of computation** → *Reinforcement learning*.

## KEYWORDS

cross-modal task, audio-text retrieval, reinforcement learning

\*Corresponding author: Jifeng Hu (hujf21@mails.jlu.edu.cn)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590056>

## ACM Reference Format:

Fuhu Song, Jifeng Hu, Che Wang, Jiao Huang, Haowen Zhang, and Yi Wang. 2023. Cross-Modal Audio-Text Retrieval via Sequential Feature Augmentation. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590056>

## 1 INTRODUCTION

Cross-modal audio-text retrieval consists of two individual tasks: text-to-audio and audio-to-text retrieval. Text-to-audio retrieval aims to retrieve relevant audio clips in an audio database with a natural language query and vice versa. With the growth of enormous amounts of user-generated multimedia data in the Internet era, the demand for efficiently retrieving ever-growing multimedia databases has never decayed. While other cross-modal retrieval tasks, such as image-text retrieval, have drawn active attention over a decade [2, 27], audio-text retrieval has only attracted growing attention in the past three years, and its performance still needs to improve. Therefore, it is vital to develop a practical algorithm for cross-modal audio-text retrieval to meet the rising need in reality.

The mainstream approach to cross-modal audio-text retrieval task [7, 18, 22, 23] uses pre-trained deep neural networks to encode modal inputs. The two different encoded modal inputs are then projected into a common space to measure cross-modal similarities. The semantically similar heterogeneous embeddings are pulled close by applying various metric learning objectives, while the dissimilar ones are pushed away in the common space. However, there are two significant problems that remain unsettled. **Problem I:** How to learn the sequential features from two modal inputs? Previous studies usually choose to extract audio features with a pre-trained convolutional neural network [22, 23], which is not able to discover the sequential features. **Problem II:** How to retain the sequential features while unifying them into the common space? Most existing studies simply apply fully connected layers [7, 22, 23] to transform encoded embeddings into the common space, which is not powerful enough to retain the sequential features.

In view of this, we introduce a Sequential Feature Augmentation framework (SFA), which can learn the sequential features from multi-modal embeddings and enhance the sequential features to a common space. Notably, we explore the potential sequential features in modal embeddings by utilizing the trial-and-error characteristic of reinforcement learning [29]. Ineffective sequential features

receive a penalty, such as a lower reward, while effective sequential features gain a higher reward. With this mechanism, the potential effective sequential features will be extracted. Besides, by utilizing recurrent feature fusion, a more profound sequential feature is integrated with the former extracted modal feature, which can improve the representation of sequential features in the common space. Furthermore, the results of our experiments demonstrate that, for the majority of tasks, our method achieves new state-of-the-art performance. In summary, the contributions of our work can be divided into three parts:

- We introduce a reinforcement learning algorithm for cross-modal audio-text retrieval. Our approach extends existing models with a reinforcement learning module that aims to explore potential sequential features from audial and textual embeddings.
- We introduce a recurrent fusion module to learn deeper sequential features from cross-modal embeddings and retain the sequential features in the common embedding space.
- We carry out extensive experiments to confirm the effectiveness of our SFA framework. In comparison with the state-of-the-art methods on AudioCaps and Clotho datasets, SFA outperforms them in text-to-audio retrieval task by a relative 13.4% and 2.4% on R@1 metric, and audio-to-text retrieval task by 28.1% and 11.5%, respectively.

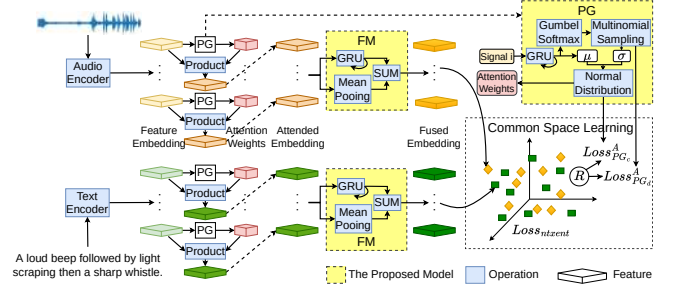
## 2 RELATED WORKS

### 2.1 Cross-modal Audio-text Retrieval

Early works [3, 10, 28] on audio-text retrieval are limited to class labels, where the text queries are separate words rather than natural language descriptions. Language-based cross-modal audio-text retrieval has gotten much attention recently due to the advent of audio captioning datasets [9, 16]. Current research generally follows two directions: (a) utilizing methods from other retrieval tasks: [24] employs two video-text retrieval frameworks to obtain the first result in audio-text retrieval. [18] further applies another video retrieval model based on [24]. [7] proposes a framework based on the latest transformers with contrastive learning; (b) solving the potential problems in audio-text retrieval: [22] incorporates different sequence aggregation methods for learning audio-text alignment. [23] investigates the effect of various metric learning objectives. [31] studies the influence of negative sampling for contrastive learning in audio-text retrieval. However, most of the above methods do not explicitly learn the sequential feature from modal input, hence failing to represent the sequential feature in the common space. Therefore, for cross-modal audio-text retrieval, these methods lead to sub-optimal performance.

### 2.2 Reinforcement Learning

With the rapid development of reinforcement learning (RL) methods, attempts to utilize RLs for cross-modal tasks have also been made. [20] applies a policy gradient algorithm to boost image captioning performance. [26] employs policy gradient methods to treat the learning between image and text as a bidirectional game. [32] proposes a discrete-continuous policy gradient algorithm to learn a



**Figure 1: Overview of the Sequential Feature Augmentation Framework for Cross-modal Audio-text Retrieval. “PG” is the reinforcement learning module. “FM” represents the recurrent fusion module.**

joint image-text representation via the attention mechanism. However, as far as we know, it is the first time that reinforcement learning has been explicitly discussed and utilized in audio-text retrieval, especially for the exploration of sequential features.

### 2.3 Feature Fusion

Feature fusion is often implemented via simple operations, such as addition or concatenation, and it has been an omnipresent part of modern multi-modal network architectures. For example, [15] uses feature fusion to exploit diverse features for text-to-video retrieval. [21] studies recurrent residual fusion to generate a power feature representation in image-text retrieval. [30] initially employs an attentional feature fusion to enable their model to handle variable-length inputs in audio-text retrieval. Inspired by these, we follow up on feature fusion, an essential yet mostly unexplored area for audio-text retrieval. Particularly, we introduce a much simplified feature fusion module that only involves averaging and summation to map heterogeneous representations into a common space.

## 3 METHODOLOGY

In this section, we provide a comprehensive explanation of the SFA framework. Fig. 1 illustrates the overview of the SFA framework. A pre-trained audio encoder first processes the audio clips to get feature embedding. The feature embedding is then evaluated by the discrete-continuous policy gradient (PG) algorithm to produce attention weights, which are subsequently utilized to explore sequential representations comprehensively. Furthermore, the attended feature embedding is fed into the recurrent fusion module (FM) to enhance the sequential features and gather audial representations into a more effective common space. Equivalently, the caption feature embedding is attended by the attention weights produced by the reinforcement learning module and then fused to learn a more textual representation into the joint embedding space. The final audial and textual embedding are then aligned by the normalized temperature-scaled cross entropy (NT-Xent) loss [4], the discrete PG loss, and the continuous PG loss in the common space to reduce the gap between the cross-modal embeddings.

### 3.1 Audio Relation Reasoning

**3.1.1 Audio Encoder.** We utilize pre-trained audio neural networks (PANNs) [19] that have been trained on AudioSet [13] to encode audio inputs, as PANNs have demonstrated strong performance across various audio-related tasks. Specifically, given an audio clip, we employ a pre-trained ResNet-38 module in PANN to encode each audio frame and take the output of the last convolutional block with 1024-dimension as frame features. The audio feature can be represented as a collection of feature vectors:  $\mathcal{A} = \{A^1, A^2, \dots, A^T\}$ , where  $A^i, i = 1, \dots, T$  denotes the feature vector of the  $i$ -th frame.

**3.1.2 Reinforcement Learning Module.** To explore the potential and deeper sequential relations among audio frames, we adopt the discrete-continuous policy gradient algorithm to generate attention weights. Discrete-continuous policy gradient algorithm considers the action space is formed by multiple normal distributions, each with different mean  $\mu$  and standard deviation  $\sigma$  value. The  $\mu$  value is a discrete action that is sampled from a pre-defined action space. The  $\sigma$  value is a continuous value that is outputted by a neural model. Attention weight is a continuous action sampled from the normal distribution formed by  $\mu$  and  $\sigma$ .

Specifically, the generation of attention weights is modeled as Markov Decision Process (MDP). The state space is composed of the input audio features and the generated attention weights. The action space is the range of attention weight. The transition function is implemented by a gated recurrent unit (GRU) [5]. First, a discrete action is sampled from a pre-defined  $\mathbb{N}$  action category via multinomial sampling. Mathematically,

$$\begin{aligned} h^t &= GRU_{mdp}(A^t, h^{t-1}), t = 1, \dots, T, \\ a^t &= h^t * W_{\mu}^t, \\ a_g^t &= \text{Gumbel} - \text{softmax}(a^t), \\ a_s^t &\sim \text{Multinomial}(a_g^t), \end{aligned} \quad (1)$$

where  $A^t$  is the  $t$ -th audio feature after PANN encoding.  $W_{\mu}^t \in \mathbb{R}_{s \times n}$  are the learnable parameter weights in the neural model.  $s$  represents the size of the audio feature vector.  $a_g^t$  is the action probability.  $a_s^t$  is the action sampled from multinomial distribution.

Then  $\mu$  is generated from the sampled  $a_s^t$  after the logistic activation, and  $std$  is generated from hidden state  $h^t$ .

$$\begin{aligned} \mu^t &= \text{Logistic}\left(\frac{a_s^t}{\mathbb{N}}\right), \\ std^t &= h^t * W_{std}^t \end{aligned} \quad (2)$$

where  $W_{std}^t \in \mathbb{R}_{s \times 1}$  are the parameter weights to be learned in the neural model.

The attention weights  $Att^t$  can be sampled from a normal distribution. It can be formulated as:

$$\begin{aligned} \text{Sample} &\sim N(\mu^t, \sigma^t), \\ Att^t &= \text{Sigmoid}(\text{Sample}). \end{aligned} \quad (3)$$

The reward comprises two parts: an online evaluation of the different feature embeddings using recall at rank  $k$  ( $R@k$ ) and an instance-wise Average Precision (AP) among a batch of feature embeddings with each feature embedding as a query. To be more concrete, we calculate the sum of  $R@k$  ( $k \in \{1, 5, 10\}$ ) and AP as

the reward:

$$\begin{aligned} \mathcal{R} &= Rsum + AP, \\ Rsum &= R@1 + R@5 + R@10. \end{aligned} \quad (4)$$

Then, the training of the PG algorithm is guided by this reward to generate attention weights to learn the sequential relations in audio features.

**3.1.3 Recurrent Fusion Module.** We use the attention weights generated by the reinforcement learning module to adjust the encoded audio feature via dot product for a more effective audial representation. Meanwhile, we do an averaging fusion operation to transform the adjusted audio embedding into the common space. Then to further learn sequential contexts in the audio feature, the adjusted audio feature is inputted to the GRU. At last, the fused audio embedding is a summation of the average of the adjusted feature and the GRU output.

Aforementioned, the encoded feature is denoted as  $\mathcal{A} = \{A^1, A^2, \dots, A^T\}$ , and the generated attention weights for encoded audio embedding are  $ATT^A = \{Att^1, Att^2, \dots, Att^T\}$ , we can unroll above procedure as follows:

$$\begin{aligned} A_A &= \mathcal{A} * ATT^A, \\ h_g^t &= GRU_{gr}^A(A_A^t, h_g^{t-1}), t = 1, \dots, T, \\ A_F &= h_g^T + \left(\sum_{t=1}^T A_A^t\right)/T, \end{aligned} \quad (5)$$

where  $A_A$  is the adjusted audio feature.  $A_F$  is the fused audio feature.

### 3.2 Text Relation Reasoning

Text relation reasoning is similar to audio relation reasoning. Given a caption of length  $N$ , a “<CLS>” token is added at the beginning of it, then the pre-trained BERT [8] module is used to encode each word of the given caption. The output caption feature is a sequence of 768-dimensional word embeddings. The caption feature is represented as  $C = \{C^1, C^2, \dots, C^N\}$ . Similarly, we use the reinforcement learning module to learn attention weights  $ATT^T$  from  $C$ . Then the text feature adjustment and feature fusion are conducted in sequence: (1) dot product is operated between the encoded feature  $C$  and attention weights  $ATT^T$  to get a more sequential representation  $C_A$ , (2) GRU is used on  $C_A$  to learn deeper sequential relations, and (3)  $C_A$  and the output of GRU are mapped into the same common space with previously introduced fused audio feature via feature fusion operation.

### 3.3 Loss Function

To fulfill the cross-modal audio-text retrieval task, we apply the PG losses and the NT-Xent loss, which is demonstrated in [23] to outperform the triplet-based losses, to train the model. The goal of the NT-Xent loss is to increase the similarity score between the positive audio-text pairs while decreasing the negative pairs within a batch. The PG losses are to maximize the long-term reward, which means maximizing  $Rsum$  and  $AP$ . Meanwhile,  $Rsum$  and  $AP$  are related to the similarity score of the positive audio-text pairs within a batch. In other words, the PG losses are also beneficial to maximize the similarity of the positive audio-text pairs. The final

loss function can be expressed as follows:

$$\mathcal{L} = \text{Loss}_{\text{ntxent}} + \text{Loss}_{\text{PG}_c}^A + \text{Loss}_{\text{PG}_d}^A + \text{Loss}_{\text{PG}_c}^T + \text{Loss}_{\text{PG}_d}^T, \quad (6)$$

where  $\text{Loss}_{\text{ntxent}}$  is the NT-Xent loss.  $\text{Loss}_{\text{PG}_c}^A$  and  $\text{Loss}_{\text{PG}_c}^T$  are continuous PG loss for audio and text.  $\text{Loss}_{\text{PG}_d}^A$  and  $\text{Loss}_{\text{PG}_d}^T$  are discrete PG loss for audio and text.

**3.3.1 NT-Xent Loss.** The NT-Xent loss is described as follows:

$$\text{Loss}_{\text{ntxent}} = -\frac{1}{B} \left( \sum_{i=1}^B \log \frac{\exp(S(A^i, C^i)/\tau)}{\sum_{j=1}^B \exp(S(A^i, C^j)/\tau)} + \sum_{i=1}^B \log \frac{\exp(S(A^i, C^i)/\tau)}{\sum_{j=1}^B \exp(S(A^j, C^i)/\tau)} \right), \quad (7)$$

where  $B$  is the batch size.  $\tau$  is a temperature hyper-parameter.  $S(\cdot)$  is the cosine similarity function.

**3.3.2 Discrete PG Loss.** The optimization objective of the policy gradient algorithm to maximize the long-term reward can be described as follows:

$$\nabla_{\theta} J(\theta) = E_{\tau \sim \pi_{\theta}(\tau)} \left[ \left( \sum_{t=0}^T \nabla_{\theta} \log \pi_{\theta}(a_t | s_t) \right) \left( \sum_{t=0}^T r(s_t, a_t) \right) \right]. \quad (8)$$

The Monte Carlo method [29], which can be viewed as  $\infty$ -step TD, is used to approximate the cumulative reward so that we can approximate  $\sum_{t=0}^T r(s_t, a_t)$  in equation (8) to  $\sum_{t=0}^T \mathcal{R}$ . Meanwhile, we can derive  $\log \pi_{\theta}(a_t | s_t) = \log \text{prob}_a^t = \log a_g^t(a_s^t)$  from equation (1). Therefore, we can present the discrete PG loss as follows:

$$\text{Loss}_{\text{PG}_d} = - \sum_{b=1}^B \left[ \left( \sum_{t=0}^T \nabla_{\theta} \log \text{prob}_a^t \right) \left( \sum_{t=1}^T \mathcal{R} - \text{baseline}_b \right) \right], \quad (9)$$

$$\text{baseline}_b = \frac{1}{B-1} \sum_{j \neq b} \mathcal{R}_j, \quad (10)$$

where  $\text{baseline}_b$  denotes the baseline of  $b$ -th sample to enhance the stability of PG training, which is an average of all the rewards within a batch.  $B$  is the batch size.  $\mathcal{R}_j$  is the reward of  $j$ -th sample.

**3.3.3 Continuous PG Loss.** Similarly, we can derive the continuous PG loss as follows:

$$\text{Loss}_{\text{PG}_c} = - \sum_{b=1}^B \left[ \left( \sum_{t=0}^T \log(f(\text{Att}^t; \mu^t, \sigma^t)) \right) \left( \sum_{t=0}^T \mathcal{R} - \text{baseline}_b \right) \right], \quad (11)$$

where  $\log(f(\text{Att}^t; \mu^t, \sigma^t))$  is the log probabilities of the normal distribution mentioned in equation (3), and we can attain the closed form of the log probability  $-\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum (\text{Att}^t - \mu^t)^2$ . With these two kinds of losses, the distance of relevant audial and textual embeddings is reduced, and the distance of irrelevant is increased in the common space.

## 4 EXPERIMENTS AND ANALYSIS

In this section, to evaluate our SFA framework's effectiveness, we carry out experiments on both text-to-audio and audio-to-text retrieval tasks on two widely-used datasets. Our goal is to answer the following three research questions:

**RQ1:** Is the proposed SFA framework effective, i.e., whether the sequential feature can be learned?

**RQ2:** How does each module of the SFA framework affect the performance of the framework?

**RQ3:** How do crucial hyper-parameters make a difference to the performance?

### 4.1 Experimental Setup

**4.1.1 Datasets.** We perform our experiments on AudioCaps [16] and Clotho [9] datasets. AudioCaps is sourced from AudioSet with more than 50k audio clips, which are 10 seconds long. Each audio corresponds to one human-annotated text description for the training split but five text descriptions for the validation and testing split. We use the same training, validation, and testing split as [23], consisting of 49,274 audios, 494 audios, and 957 audios, respectively. The latest ClothoV2, in which the audio duration is between 15 and 30 seconds, sourced from Freesound [12], is used in experiments. Unlike AudioCaps, each audio in ClothoV2 matches five human-annotated text descriptions for all splits. We also follow the splits of [23] for the Clotho dataset, which includes 3,839 audios for training, 1,045 audios for validation, and 1,045 audios for testing.

**4.1.2 Implementation Details.** Our model is built with PyTorch toolkit [25]. Audio features are represented as Log-Mel spectrogram sampled from original audio samples at 32 kHz, with a Hanning window of 1024 points, a hop length of 320, and 64 mel bins. The hidden size of the GRU models and common space dimension are both 1024. The temperature parameter  $\tau$  in NT-Xent loss is 0.07, which is empirically better. We use the Adam optimizer [17] with a learning rate of  $10^{-4}$ , which is decayed by 0.1 every 20 epochs. For all experiments, the model is trained using a batch size of 32 for 50 epochs. We use ResNet38 and CNN14 as the audio encoder for AudioCaps and Clotho datasets, respectively.

**4.1.3 Evaluation Metrics.** As standard metrics in information retrieval tasks, we choose recall at rank  $k$  ( $R@k$ , higher indicates better) and median rank (MedR, lower indicates better) to evaluate our results.  $R@k$  is defined as the percentage of matched targets among top- $k$  retrieved results. The MedR calculates the median rank position of the correct items in the retrieved ranked list. We present the mean and standard deviation of  $R@1$ ,  $R@5$ ,  $R@10$  and MedR as the final result based on three different training seed runs.

**4.1.4 Baseline Methods.** We compare our SFA method with five state-of-the-art cross-modal audio-text retrieval baselines, including MoEE [24], CE [24], MMT [18], CNN14+NetRVLAD [22], and ASE [23]. MoEE and CE are re-evaluated on the latest ClothoV2 dataset for a fair comparison. Among them, the former three are the models adopted from video-text retrieval, and the latter two are the state-of-the-art methods that utilize the CNN structure pre-trained module and fully connected layers to project multi-modal embeddings into common space. Comparing with these methods can validate the superiority of our method, which enhances sequential feature learning.

### 4.2 Overall Results and Analysis (RQ1)

Table 1 presents the results of our method compared to the current state-of-the-art methods on the AudioCaps and Clotho datasets.

**Table 1: Audio-Text Retrieval Results with Different Methods on AudioCaps and Clotho Datasets.**

Model	Text-to-Audio				Audio-to-Text			
	R@1↑	R@5↑	R@10↑	MedR↓	R@1↑	R@5↑	R@10↑	MedR↓
AudioCaps								
MoEE	22.5±0.3	54.4±0.6	69.5±0.9	5.0±0.0	25.1±0.8	57.5±1.4	72.9±1.2	4.0±0.0
CE	23.1±0.8	55.1±0.9	70.7±0.7	4.7±0.6	25.1±0.9	57.1±1.0	73.2±1.0	4.0±0.0
MMT	36.1±3.3	<b>72.0±2.9</b>	<b>84.5±2.0</b>	2.3±0.6	39.6±0.2	76.8±0.9	86.7±1.8	2.0±0.0
CNN14+NetRVLAD	29.3±0.3	65.2±0.5	79.3±1.0	3.0±0.0	33.3±0.5	67.6±0.5	80.6±0.8	3.0±0.0
ASE	33.9±0.4	69.7±0.2	82.6±0.3	-	39.4±1.0	72.0±1.0	83.9±0.6	-
<b>SFA (Ours)</b>	<b>47.6±0.5</b>	<b>69.8±0.6</b>	<b>78.3±0.4</b>	<b>2.0±0.0</b>	<b>67.7±0.5</b>	<b>86.3±0.7</b>	<b>91.0±0.9</b>	<b>1.0±0.0</b>
Clotho								
MoEE	8.5±0.1	26.5±0.1	38.2±0.9	19.3±0.6	9.7±0.4	27.0±0.1	38.7±0.6	17.3±0.6
CE	9.0±0.4	26.8±0.2	38.6±0.6	18.0±0.0	9.4±0.9	27.2±1.5	39.6±1.5	17.0±1.0
MMT	6.5±0.6	21.6±0.7	32.8±2.1	23.0±2.6	6.3±0.5	22.8±1.7	33.3±2.2	22.3±1.5
CNN14+NetRVLAD	13.1±0.2	33.1±0.6	45.1±0.2	<b>13.0±0.0</b>	13.0±0.2	32.9±0.7	45.4±0.8	13.0±0.0
ASE	14.4±0.4	<b>36.6±0.2</b>	<b>49.9±0.2</b>	-	16.2±0.7	37.5±0.9	50.2±0.7	-
<b>SFA (Ours)</b>	<b>16.8±0.2</b>	34.1±0.2	43.6±0.1	16.0±0.0	<b>29.6±0.4</b>	<b>49.1±0.3</b>	<b>58.7±0.6</b>	<b>6.0±0.0</b>

As shown in the table, our SFA method has gained a significant improvement in the audio-to-text task. While for the text-to-audio task, even the performance increment is not much more dramatic than audio-to-text. We can also conclude that it is comparable to the SFA method and other approaches because of the significant improvement in R@1 metric. To be more accurate, our R@1 metric increases by 28.1% and 11.5% in audio-to-text and text-to-audio retrieval tasks on AudioCaps, respectively. Simultaneously, the increment is 13.4% and 2.4% on Clotho. As discussed in [18], the Clotho task is more challenging due to the much more diversity of captions for each audio clip. Therefore, the 2.4% and 13.4% improvement on R@1 metric in text-to-audio and audio-to-text retrieval over the Clotho benchmark indicates that the SFA method can learn the sequential features better.

Meanwhile, we notice that the SFA method performs worse than the current state-of-the-art method in the text-to-audio retrieval task on R@5, R@10 and MedR metrics. This might be attributed because the BERT pre-trained module used to encode text learns sequential features well, so the reinforcement learning algorithm might learn an overfitting sequential textual feature, which causes a negative impact. However, for the audio-to-text retrieval task, the current popular PANN modules do not learn the sequential feature in audio. Thus our reinforcement learning module can make up for this and achieves the new state-of-the-art.

### 4.3 Ablation Studies

**4.3.1 Effect of Reinforcement Learning Module (RQ2).** We design our experiments in two cases: whether to add the reinforcement learning module with the recurrent fusion module (FM, PG+FM) and whether to add the reinforcement learning module without the recurrent fusion module (FC, PG+FC). In the latter case, we replace the missing fusion module with a fully connected layer to project different modal features into a common space. From Table 2, we can observe that with the reinforcement learning module, the performance of audio-to-text retrieval is significantly improved in both cases, which proves the feasibility of using reinforcement

learning to learn sequential features. However, the R@10 metric slightly drops under the text-to-audio retrieval task, which might be the conflict between the sequential feature learned by BERT and our reinforcement learning module, as previously analyzed.

**4.3.2 Effect of Recurrent Fusion Module (RQ2).** We also perform our experiments in two cases: whether to add the recurrent fusion module with the reinforcement learning module (PG+FM, PG+FC) and whether to add the recurrent fusion module without the reinforcement learning module (FC, FM). Table 2 shows that with the recurrent fusion module, the performance of both audio-to-text and text-to-audio retrieval tasks get improved, as the recurrent fusion module can increase sequential feature retention when transforming into the common space. The reason FM performs better with PG might be explained because FM does better on the sequential feature PG learns first, where PG plays the same role as data pre-processing.

**4.3.3 Effect of Recurrent Neural Network (RQ2).** We then conduct ablation studies on the recurrent neural network in the reinforcement learning and the recurrent fusion module. The results reflect that GRU is better than LSTM [14], and both outperform Elman RNN [11]. It is well-known that RNN suffers from gradient exploding and vanishing problems [1], which lead to hard convergence. Also, our experiments on RNN get the best module in the last one or two epochs. It demonstrates the existence of analysis problems, which lead to the expected RNN result. The reason why GRU outperforms LSTM might be interpreted because GRU can manage the information stream from preceding activation as studied in [6], which naturally suits the state transition in reinforcement learning.

**4.3.4 Effect of Discrete Action Space (RQ3).**  $\mu$  plays a vital role in attention weight generation. It is the size of the pre-defined action space that affects the range of  $\mu$  in the normal distribution. We carry out ablation studies of the action space size with three different values, and the result indicates that  $N = 500$  results in the best overall performance. It is likely that 500 approximates the maximum value of the token number in sentences and the feature

**Table 2: Ablation Studies on Clotho Dataset.**

Methods	Text-to-Audio			Audio-to-Text		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
FC	8.0±0.2	24.9±0.3	37.1±0.1	10.2±0.3	28.2±0.7	39.8±0.4
FM	10.1±0.2	31.1±0.3	44.3±0.1	12.1±0.8	32.5±1.4	45.8±1.6
PG+FC	9.2±0.1	23.6±0.5	33.0±0.4	15.1±0.6	34.6±0.2	45.2±0.6
PG+FM (Ours)	16.8±0.2	34.1±0.2	43.6±0.1	29.6±0.4	49.1±0.3	58.7±0.6
RNN	4.8±0.7	14.5±0.5	22.7±0.9	6.6±0.5	20.0±2.1	30.2±2.8
LSTM	9.3±0.4	23.3±0.2	33.6±0.6	14.1±0.5	31.8±0.1	42.0±0.3
GRU (Ours)	16.8±0.2	34.1±0.2	43.6±0.1	29.6±0.4	49.1±0.3	58.7±0.6
N = 400	16.5±0.4	33.8±0.2	43.8±0.2	30.8±0.8	48.8±0.1	58.0±1.1
N = 500 (Ours)	16.8±0.2	34.1±0.2	43.6±0.1	29.6±0.4	49.1±0.3	58.7±0.6
N = 600	16.1±0.4	33.4±0.3	43.1±0.6	28.5±0.5	48.2±0.7	58.0±0.3

number in audios of one batch. Hence, 500 is powerful enough to distinguish between the audio signals and caption items.

## 5 CONCLUSIONS

In this paper, we investigate the subject of sequential feature augmentation for cross-modal audio-text retrieval task since the sequential feature is a significant part of both audio and text. However, most existing approaches ignore its learning and underestimate its importance for representing a robust audial-textual common embedding space. Based on these two issues above, we introduce an SFA framework that uses reinforcement learning to explore the sequential features and a recurrent fusion module to enhance the sequential features to bridge the gap between audial and textual features in a common space. In addition, experimental performance beats the benchmark of the audio-to-text retrieval task on two widely-used datasets and is competitive in the text-to-audio retrieval task. This work can show a promising future of reinforcement learning and feature fusion in cross-modal audio-text retrieval.

## REFERENCES

- [1] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.
- [2] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text Retrieval: A Survey on Recent Research and Development. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 5410–5417.
- [3] Gal Chechik, Eugene Ie, Martin Rehn, Samy Bengio, and Dick Lyon. 2008. Large-scale content-based audio retrieval from text queries. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 105–112.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International conference on machine learning*, Vol. 119. PMLR, 1597–1607.
- [5] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).
- [6] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR abs/1412.3555* (2014).
- [7] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2022. Audio Retrieval with WavText5K and CLAP Training. *arXiv:2209.14275*
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
- [9] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an Audio Captioning Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 736–740.
- [10] Benjamin Elizalde, Shuayb Zarar, and Bhiksha Raj. 2019. Cross Modal Audio Search and Retrieval with Joint Embeddings Based on Text and Audio. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4095–4099.
- [11] Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science* 14, 2 (1990), 179–211.
- [12] Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*. 411–412.
- [13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.
- [14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
- [15] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. 2022. Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval. In *European Conference on Computer Vision*. Springer, 444–461.
- [16] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 119–132.
- [17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
- [18] A. Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio Retrieval with Natural Language Queries: A Benchmark Study. *IEEE Transactions on Multimedia*, 1–1.
- [19] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.
- [20] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017. Improved Image Captioning via Policy Gradient Optimization of SPIDER. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [21] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE international conference on computer vision*. 4107–4116.
- [22] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. 2022. Audio-Text Retrieval in Context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4793–4797.
- [23] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang. 2022. On Metric Learning for Audio-Text Cross-Modal Retrieval. *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)* (2022), 4142–4146.
- [24] Andreea-Maria Oncescu, A. Sophia Koepke, João F. Henriques, Zeynep Akata, and Samuel Albanie. 2021. Audio Retrieval with Natural Language Queries. In *INTERSPEECH*. 2411–2415.
- [25] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

- [26] Jinwei Qi and Yuxin Peng. 2018. Cross-modal Bidirectional Translation via Reinforcement Learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, 2630–2636.
- [27] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*. 251–260.
- [28] Malcolm Slaney. 2002. Semantic-audio retrieval. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 4. IEEE, IV–4108.
- [29] R.S. Sutton and A.G. Barto. 1998. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks* 9, 5 (1998), 1054–1054.
- [30] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. [arXiv:2211.06687](https://arxiv.org/abs/2211.06687)
- [31] Huang Xie, Okko Räsänen, and Tuomas Virtanen. 2022. On Negative Sampling for Contrastive Audio-Text Retrieval. [arXiv:2211.04070](https://arxiv.org/abs/2211.04070)
- [32] Shiyang Yan, Li Yu, and Yuan Xie. 2021. Discrete-continuous action space policy gradient-based attention for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8096–8105.

# Health monitoring system for elderly people based on Raspberry Pi

Qingsong Peng

Center of Bioinformatics, Shanghai Technical Institute of Electronics & Information, Shanghai, China  
peng.qingsong@qq.com

## ABSTRACT

We present a comprehensive overview of the application of Raspberry Pi in the field of health monitoring for elderly people with disabilities. Firstly we discuss the advantages of using artificial intelligence technology for health monitoring of elderly people, and the significance of using information technology devices to achieve health monitoring for the elderly, while keeping the cost of the devices low. And then we examine the development of Raspberry Pi and its advantages for health monitoring of elderly people with disabilities, such as its low cost, portability, and ease of use. After that we outline the methods of collecting data for health monitoring of elderly people, such as using sensors to measure heart rate, oxygen levels, and blood pressure, and integrating these sensors into a single device. We also discuss the implementation of a Raspberry Pi-based health monitoring system for elderly people, and the ways in which health data can be utilized to optimize the performance of the system. The work provides useful insights for those who are interested in using Raspberry Pi for health monitoring applications for elderly people with disabilities.

## KEYWORDS

CCS CONCEPTS • Information systems Information systems applications Mobile information processing systems Additional Keywords and Phrases: Raspberry Pi, Health Monitoring System, Deep Learning

### ACM Reference Format:

Qingsong Peng. 2023. Health monitoring system for elderly people based on Raspberry Pi. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3590003.3590057>

## 1 INTRODUCTION

The use of artificial intelligence technology for health monitoring of elderly people is of great significance. Elderly people's health needs to be achieved through information technology devices, and the total cost of these devices should not be too high. With the development of artificial intelligence technology, more and more intelligent health monitoring systems are being developed to meet the needs of elderly people. Hossain developed a contactless syndromic

surveillance platform FluSense that aims to expand the current paradigm of influenza-like illness (ILI) surveillance<sup>[1]</sup>. Dong proposed a novel approach that tries to connect together these sparse observations using a model of how individuals interact with each other and how social interactions happen in terms of a sequence of proximity interactions<sup>[2]</sup>. Osthus performed a controlled experiment, taking into account data backfill, to improve clarity on the benefits and limitations of augmenting an already good flu forecasting model with internet-based nowcasts<sup>[3]</sup>. Rakhmatulin presented the method for the quick and anonymous alcoholism diagnosis by neural networks<sup>[4]</sup>. Seco explored two signal analysis techniques, Matching Pursuit (MP) and Fast Fourier Transform (FFT), for differentiation between two states, eyes open (EO) and eyes closed (EC), through the detection of EEG alpha activity obtained from seven scalp regions, using a portable EEG device<sup>[5]</sup>.

These systems can monitor the elderly's physical and mental health, detect potential health problems in advance, and provide timely medical assistance. In addition, these systems can also provide personalized health advice and reminders to the elderly, helping them to maintain a healthy lifestyle. Moreover, by using artificial intelligence technology, the cost of health monitoring systems can be greatly reduced. As a result, the use of artificial intelligence technology for health monitoring of elderly people can help to reduce the cost of health monitoring systems while providing effective and reliable monitoring services. Part one gives the introduction to the background, part two and three are the characteristic of Raspberry Pi, and the necessity of monitor the health data respectively. Part four gives the design and the conclusion is in Part five.

## 2 THE PROPERTY OF RASPBERRY PI

Raspberry Pi is a low-cost, single-board computer developed in UK. It was first released in 2012 and has since become one of the most popular single-board computers in the world. The Raspberry Pi is powered by an ARM processor and runs the Linux operating system. It is equipped with a range of ports, including HDMI, USB, Ethernet, and audio jacks. The device is also equipped with a variety of sensors, such as temperature, humidity, and light sensors.

The Raspberry Pi is widely used in a variety of applications, such as home automation, robotics, and health monitoring. It is also used in educational projects, as it is easy to use and cost is low. Moreover, the Raspberry Pi is highly portable and can be used in a variety of environments, such as in hospitals, homes, and in remote locations. The Raspberry Pi is also highly customizable and can be used to create custom applications. Furthermore, it is highly energy efficient, making it suitable for applications that require low power consumption. It is a low-cost, highly versatile device that is suitable for a variety of applications, including health monitoring for elderly people.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590057>

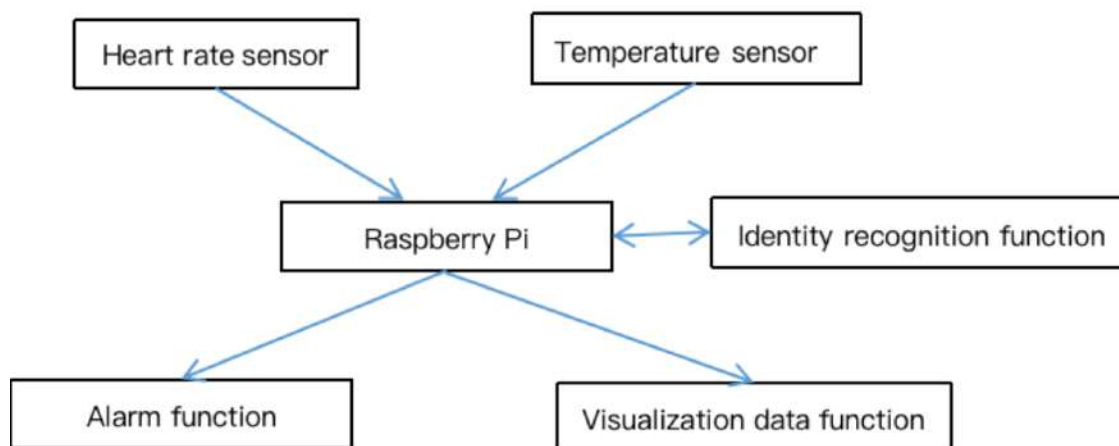


Figure 1: Health monitoring system for elderly people based on Raspberry Pi.

### 3 THE NECESSARY STEP TO COLLECT HEALTH DATA OF ELDERLY PEOPLE

Health monitoring of elderly people requires the collection of data from various sources. To do this, a variety of sensors must be used to collect data such as heart rate, oxygen levels, and blood pressure. These sensors must be integrated into a single device in order to collect the necessary data. For example, the Raspberry Pi can be used to integrate multiple sensors into a single device. And it is also equipped with a range of sensors, such as temperature, humidity, and light sensors. Furthermore, the Raspberry Pi is highly energy efficient, making it suitable for applications that require low power consumption. As a result, the Raspberry Pi can be used to collect data from a variety of sources and integrate them into a single device. This makes it an ideal device for health monitoring of elderly people.

### 4 THE IMPLEMENTATION OF A HEALTH MONITORING SYSTEM FOR ELDERLY PEOPLE

The implementation of a health monitoring system for elderly people based on Raspberry Pi involves the integration of multiple sensors into a single device. It is also equipped with a range of sensors, such as temperature, humidity, and light sensors. These sensors can be used to collect data such as heart rate, oxygen levels, and blood pressure. Data collected from these sensors can then be stored in a database and analyzed using machine learning algorithms. This analysis can be used to detect potential health issues and provide personalized health advice to the elderly. In addition, the data can also be used to create health-related reports and insights. By utilizing health data in this way, the health monitoring system can provide more accurate and reliable results. Furthermore, the use of health data can also help to reduce the cost of health monitoring systems, as the data can be used to optimize the performance of the system.

The use of Raspberry Pi devices for health monitoring of elderly people can greatly reduce the risk of illness among the elderly. By

using Raspberry Pi devices, elderly people can be monitored in real time, allowing health issues to be detected in advance and providing timely medical assistance. In addition, the use of Raspberry Pi devices can also enable personalized health advice and reminders to be given to the elderly, helping them to maintain a healthy lifestyle. Furthermore, the low cost of Raspberry Pi devices makes them suitable for use in a variety of settings, including in hospitals, homes, and remote locations.

#### 4.1 Wireless module design

Wireless single-chip microcontrollers can be designed to provide a cost-effective solution for embedded applications. They are typically used in a wide range of applications such as industrial, automotive, medical, and consumer electronics.

Wireless single-chip microcontrollers are based on a single integrated circuit (IC) that contains all the necessary components for operation. The components include a processor, memory, communication interface, and other peripherals. The design of a wireless single-chip microcontroller is based on a modular approach. The processor and other components are organized in a single IC package. The processor is typically connected to the other components through an on-chip bus. The bus is used to communicate data and instructions between the processor and other components. The processor is also connected to the memory, which stores the program code and data. The communication interface is connected to the processor and is used to communicate with other devices. The other peripherals are also connected to the processor and are used for specific tasks.

The design of a wireless single-chip microcontroller includes several steps. First, the processor and other components are selected and integrated into a single IC package. The processor is then connected to the other components through the on-chip bus. The communication interface is also connected to the processor. The memory is then programmed with the program code and data. Finally, the other peripherals are connected to the processor and configured for the specific tasks. The design must also be optimized for power consumption and performance. Once the design is

complete, the microcontroller is tested to ensure that it meets the application requirements.

## 4.2 The design of heart rate sensor module

The design of heart rate sensor module includes sensor selection, circuit design, software design and so on. Sensor selection is the basis, and the appropriate sensor should be selected according to the actual application requirements. For example, if the heart rate sensor module is used for medical purposes, then a highly sensitive and accurate sensor should be selected. If the heart rate sensor module is used for sports purposes, then a sensor with a wide range of measurement should be selected.

Circuit design is the core of the heart rate sensor module, and the appropriate circuit should be designed according to the characteristics of the sensor to meet the actual application requirements. The circuit design should consider the signal conditioning, signal amplification, signal transmission, signal processing, power supply and other aspects. For example, the signal conditioning should be designed to filter out noise and interference signals, and the signal amplification should be designed to ensure the signal accuracy and sensitivity.

Software design is the key to realizing sensor data acquisition and processing, and the appropriate program should be designed according to the requirements of the application system. The software design should consider the data acquisition, data processing, data storage, data analysis, data visualization and other aspects. For example, the data acquisition should be designed to acquire the sensor data in real time, and the data processing should be designed to process the acquired data accurately and efficiently.

In summary, the design of heart rate sensor module requires the selection of appropriate sensors, the design of appropriate circuits, and the design of appropriate programs. The design should consider the actual application requirements, and the design should be optimized to ensure the accuracy and reliability of the sensor module.

## 4.3 The design of temperature sensor module

The design of the temperature sensor module of Raspberry Pi mainly includes temperature sensor, analog-to-digital converter (ADC), processor, controller, and output module. The temperature sensor is used to measure the temperature, the analog-to-digital converter (ADC) is used to convert the temperature signal into digital signal, the processor is used to process the temperature data, the controller is used to control the working of the temperature sensor, and the output module is used to output the processed data to the external device.

The temperature sensor module of Raspberry Pi is usually composed of a temperature sensor, an analog-to-digital converter (ADC), a processor, a controller, and an output module. The temperature sensor is responsible for measuring the temperature of the environment, and the analog-to-digital converter (ADC) is used to convert the analog temperature signal into digital signal. The processor is responsible for processing the digital temperature data, and the controller is used to control the working of the temperature sensor. The output module is responsible for sending out the processed temperature data to the external device.

In order to ensure the accuracy of the temperature sensor module, the temperature sensor should be calibrated before use. The calibration process includes setting the temperature range, adjusting the offset, and setting the temperature coefficient. After the calibration is completed, the temperature sensor module can be used to measure the temperature of the environment accurately. The design of the temperature sensor module of Raspberry Pi is composed of temperature sensor, analog-to-digital converter (ADC), processor, controller, and output module. The temperature sensor module can be used to measure the temperature of the environment accurately, and the measured temperature data can be used for various purposes.

## 4.4 Identity recognition function design

The design of a recognition model for timely judging whether the physical state of the elderly is suddenly abnormal should take into account heart rate, body temperature, and other information. The model should be able to detect any sudden changes in the physical state of the elderly, such as a sudden increase in heart rate or a sudden drop in body temperature.

The recognition model should include sensors that measure heart rate, body temperature, and other information. The sensors should be able to detect any sudden changes in the physical state of the elderly. The model should also include a processor that processes the data collected by the sensors and compares it with a predefined threshold. If the data exceeds the threshold, the processor should trigger an alert. In addition to the sensors and processor, the recognition model should also include a communication module that allows the model to send alerts to the elderly's family and caregivers. The communication module should be able to send alerts via a variety of channels, such as text messages, emails, and phone calls. The recognition model should also be able to store data for later analysis. This data can be used to track the elderly's physical state over time and identify any potential trends. The design of a recognition model for timely judging whether the physical state of the elderly is suddenly abnormal should take into account heart rate, body temperature, and other information. The model should include sensors, a processor, a communication module, data storage, and a user interface.

## 4.5 Visualization data function design

The visualization display model for displaying the physical health information of the elderly in real time in the way of a cockpit in the computer should support multi-condition compound query. The model should be able to display various physical health information of the elderly in real time, such as heart rate, body temperature, blood pressure, oxygen saturation, and other vital signs. It should include sensors that measure the physical health information of the elderly. The sensors should be able to detect any sudden changes in the physical state of the elderly. The model should also include a processor that processes the data collected by the sensors and compares it with a predefined threshold. If the data exceeds the threshold, the processor should trigger an alert.

In addition to the sensors and processor, the visualization display model should also include a communication module that allows the model to send alerts to the elderly's family and caregivers. The

communication module should be able to send alerts via a variety of channels, such as text messages, emails, and phone calls. It should also be able to store data for later analysis. This data can be used to track the elderly's physical state over time and identify any potential trends. And it should also include a graphical user interface that allows users to view the physical health information of the elderly in real time in the form of a cockpit. The user interface should be easy to use and should include features such as data visualization, multi-condition compound query, and alert notifications.

In conclusion, the visualization display model for displaying the physical health information of the elderly in real time in the way of a cockpit on the computer should support multi-condition compound query. The model should include sensors, a processor, a communication module, data storage, and a graphical user interface.

#### 4.6 Alarm function design

The alarm module for providing multiple alarm modes for the results submitted by the recognition module should be able to detect any sudden changes in the physical state of the elderly. The alarm module should be able to monitor the results submitted by the recognition module and trigger an alert if the results exceed a predefined threshold.

The alarm module should include sensors that measure the physical health information of the elderly. The sensors should be able to detect any sudden changes in the physical state of the elderly. The alarm module should also include a processor that processes the data collected by the sensors and compares it with a predefined threshold. If the data exceeds the threshold, the processor should trigger an alert. In addition to the sensors and processor, the alarm module should also include a communication module that allows the model to send alerts to the elderly's family and caregivers. The communication module should be able to send alerts via a variety of channels, such as text messages, emails, and phone calls.

The alarm module should also include multiple alarm modes that can be triggered in different situations. For example, the alarm module should be able to trigger a low-level alert when the results

submitted by the recognition module exceed a certain threshold, or a high-level alert when the results exceed a higher threshold. The alarm module should also be able to trigger an alert when the results submitted by the recognition module remain unchanged for a certain period of time. The alarm module for providing multiple alarm modes for the results submitted by the recognition module should include sensors, a processor, a communication module, and multiple alarm modes.

With the development of technology, new modules should be added in time, new functions should be developed and promoted in time. Sensors such as blood oxygen and plantar pressure should also be increased to provide a more complete health monitoring system without increasing too much cost.

## 5 CONCLUSION

The use of health data collected from Raspberry Pi devices can help to optimize the performance of the system, further reducing the cost of health monitoring systems. In conclusion, the use of Raspberry Pi devices for health monitoring of elderly people can help to reduce the risk of illness among the elderly, while also providing cost-effective and reliable monitoring services.

## REFERENCES

- [1] Forsad Al Hossain, Andrew A. Lover, George A. Corey, Nicholas G. Reich, and Tauhidur Rahman. 2020. FluSense: A Contactless Syndromic Surveillance Platform for Influenza-Like Illness in Hospital Waiting Areas. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 1, Article 1 (March 2020), 28 pages. <https://doi.org/10.1145/3381014>.
- [2] Wen Dong, Tong Guan, Bruno Lepri, and Chunming Qiao. 2019. PocketCare: Tracking the Flu with Mobile Phones using Partial Observations of Proximity and Symptoms. *arXiv preprint arXiv:1905.02607* (2019).
- [3] Dave Osthus, Ashlynn R Daughton, and Reid Priedhorsky. 2019. Even a good influenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited. *PLoS computational biology* 15, 2 (2019), e1006599.
- [4] Rakhmatulin, I. (2020). Deep learning and machine learning for EEG signal processing on the example of recognizing the disease of alcoholism. <https://doi.org/10.1101/2021.06.02.21258251>.
- [5] Seco, G., *et al.* (2019). EEG alpha rhythm detection on a portable device. *Biomedical Signal Processing and Control*, 52, 97-102. <https://doi.org/10.1016/j.bspc.2019.03.014>.

# Multiple Frequency Bands Temporal State Representation for Deep Reinforcement Learning

Che Wang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
wch675413020@163.com

Jifeng Hu\*  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
hujf21@mails.jlu.edu.cn

Fuhu Song  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
songfh20@mails.jlu.edu.cn

Jiao Huang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
huangjiao20@mails.jlu.edu.cn

Zixuan Yang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
yangzx9920@mails.jlu.edu.cn

Yusen Wang  
Jilin University  
Chaoyang Qu, Changchun Shi, China  
1092718626@qq.com

## ABSTRACT

Deep reinforcement learning has achieved significant success in solving sequential decision-making tasks. Excellent models usually require the input of valid state signals during training, which is challenging to encode temporal state features for the deep reinforcement learning model. To address this issue, recent methods attempt to encode multi-step sequential state signals so as to obtain more comprehensive observational information. However, these methods usually have a lower performance on complex continuous control tasks because mapping the state sequence into a low-dimensional embedding causes blurring of the immediate state features. In this paper, we propose a multiple frequency bands temporal state representation learning framework. The temporal state signals are decomposed into discrete state signals of various frequency bands by Discrete Fourier Transform (DFT). Then, feature signals filtered out different high-frequency bands are generated. Meanwhile, the mask generator evaluates the weights of signals of various frequency bands and encodes high-quality representations for agent training. Our intuition is that temporal state representations considering multiple frequency bands have high fidelity and stability. We conduct experiments tasks and verify that our method has obvious advantages over the baseline in complex continuous control tasks such as Walker and Crawler.

## CCS CONCEPTS

• **Theory of computation** → **Reinforcement learning**; • **Computing methodologies** → **Knowledge representation and reasoning**.

\*Corresponding author: Jifeng Hu (hujf21@mails.jlu.edu.cn)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590058>

## KEYWORDS

deep reinforcement learning, representation learning, Fourier transform, continuous control task.

### ACM Reference Format:

Che Wang, Jifeng Hu, Fuhu Song, Jiao Huang, Zixuan Yang, and Yusen Wang. 2023. Multiple Frequency Bands Temporal State Representation for Deep Reinforcement Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590058>

## 1 INTRODUCTION

Deep Reinforcement Learning (RL) has achieved significant success among a wide range domains [6, 26, 30], including robotics [23, 32], video games [11], and medical applications. [31]. Deep RL aims to get the optimal policy to maximize the accumulated rewards during the interaction with the environment. State representation learning plays a significant role in the training process of the data-driven policy model. For example, COBERL uses LSTM-transformer architecture to learn better representations for RL without the need of hand-engineered data augmentations [3]. JueWu-MC captures underlying relations between action and representation by action-aware representation learning [17]. OpenAI enabled the agent to generate masks from the frontal vision cone and line of sight, thus reducing the training difficulty in hide-and-seek tasks [2]. In summary, excellent models usually have strong representational capabilities of state features.

Researchers have shown some algorithms fail to converge using Markov Decision Process [19], where policy becomes brittle because a small change of state can cause a large change in the agent's action. To extract generalized state representations, the state vectors of each time step are saved in a sequence for output action. There are two main methods to extract representation from temporal state sequences, i.e., recurrent networks [12, 15, 22] and attention-based models [21, 24]. Recently, researchers tried to combine the two methods for training, which is helpful for the model to understand the complex sequence feature [3, 28].

Although existing methods have successfully extracted temporal state representation, there are still two remaining challenges. **Challenge I:** How to retain accurate immediate state features in complex continuous control tasks? When the action space is continuous and high-dimensional, selecting some actions requires accurate latest state features. However, the aggregation of state sequences blurs the

features of the latest state. **Challenge II:** How to improve sampling efficiency in sequence representation learning? The combination of data-driven deep learning modules leads to increased samples required for model training. Poor sample efficiency is the key problem in the application of deep reinforcement learning[8, 18, 20]. Based on the above challenges, we utilize a non-data-driven approach known as the Fourier transform to extract both real-time features and trend features of the state simultaneously.

Fourier transform is a method that can process discrete time series signals[14], which transforms time-domain signals into frequency-domain for further analysis. It is often used in the physical analysis, mathematical calculation, image processing, and other fields[4, 5, 29]. Intuitively, the state sequences are filtered out of various high-frequency band signals by Fourier transform so that the latest state variables are biased to different values due to the influence of the state sequences.

In view of this, we propose a Multiple Frequency Bands Temporal State Representation Learning(MBTS) framework, which can automatically select the frequency bands according to state sequences so as to determine how much the latest state is affected by the past states. Specifically, we transform the discrete state signals from the time domain to the frequency domain and filter the signals out with multiple low-frequency-band filters, which generate multiple feature sequences with different bands. Meanwhile, we design the weight allocation method of frequency bands, which allocates the agent's attention to various bands when the agent constantly interacts with the environment. Furthermore, we conduct experiments in four continuous control tasks to verify the effectiveness of our method. By automatically selecting relevant frequency bands, our method can improve the efficiency and accuracy of state representation learning, which is demonstrated by the promising experimental results. To clearly illustrate the focus of our research, we divide the contributions of this article into three parts:

- A practical framework is proposed for temporal state representation learning. Under this framework, agents' performance is significantly improved in complex continuous tasks.
- We have designed a mechanism for the weight distribution of multiple state features, which effectively adjusts various state variables to suitable frequency bands.
- We verify the influence of temporal state sequence on existing methods. The performances of other methods decrease with the increasing sequence length. The final results verify that our algorithm can effectively solve this problem.

## 2 BACKGROUND

### 2.1 Deep Reinforcement Learning

Deep Reinforcement Learning(RL) can be regarded as a finite-horizon Markov decision process defined by a tuple  $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \rho_0, \gamma \rangle$ , where  $\mathcal{S}$  is the set of states and  $\mathcal{A}$  is the set of actions,  $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$  is the dynamic function representing the transition probability from the current state to the next state by executing the action,  $\rho_0$  is initial state distribution, and  $\gamma$  is the discount factor. The reward function  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  outputs a supervisory signal based on the current state  $s_t$  and action  $a_t$  [27]. The problem requires the agent to learn the policy  $\pi_\theta : \mathcal{S} \rightarrow \mathcal{A}$  that maximizes

the expected cumulative return by a rollout in the environment. The objective function is defined as:

$$J(\theta) = \mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)} \left[ \sum_t \gamma^t r(s_t, a_t) \right] \quad (1)$$

The Action-value function is used to estimate the expected return of a state or state-action pair, which is defined as  $Q(s, a) = \mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s, a_0 = a]$ . DRL methods usually learn parameterized policy  $\pi_\theta$  and value function  $Q_\psi(s, a)$  with mini-batch samples. We adjust the parameters of the value function to make the action value  $Q_\psi(s, a)$  close to return  $y = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$ . Its parameters  $\psi$  are updated so that minimizing the mse loss:

$$L^Q(\psi) = \mathbb{E}_{(s,a,s',r) \sim D} [y - Q_\psi(s, a)]^2 \quad (2)$$

With the development of DRL research, there are more and more methods for updating policy network parameters  $\theta$ . Proximal Policy Optimization(PPO) is a general algorithm that is easy to implement and has better sample complexity[25]. PPO updates the policy network cautiously, which controls the probability ratio  $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$  within a certain range. The main objective is the following:

$$L(\theta) = \mathbb{E}_t [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t)] \quad (3)$$

In the continuous control task, the previous reinforcement learning algorithms, such as PPO have achieved significant success. Our study builds on these algorithms and attempts to improve the model's sampling efficiency and the agent's score.

### 2.2 Fourier Transform

Fourier transform is a fundamental algorithm in digital signal processing. Fourier principle holds that any continuously measured time series or signal can be represented as an infinite superposition of sine and cosine signals with different frequencies. Fourier transform converts the original intractable time-domain signal into an easy-to-analyze frequency-domain signal that some tools can process for analyzing frequency-domain signals. The formula for the Fourier transform is:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \quad (4)$$

where  $w$  is the frequency of the test signal. The Fourier transform algorithm computes the convolution between the test signal and the original signal to obtain the amplitude and frequency of the test signal. The test signals are sine and cosine signals of different frequencies, and their results are stored in the plural. According to Euler's formula:

$$e^{-j\omega t} = \cos(\omega t) - j \sin(\omega t) \quad (5)$$

It is necessary to obtain its corresponding time domain signal as the final result after the signal is analyzed and processed in the frequency domain. Inverse Fourier transform usually be used to convert these frequency domain signals into time domain signals:

$$f(t) = \int_{-\infty}^{\infty} F(\omega)e^{-j\omega t} d\omega \quad (6)$$

In addition, many methods have been developed on the Fourier transform. Discrete Fourier transform(DFT) was used to process

discretely sampled signals, and Fast Fourier transform(FFT) sped up the calculation process in practical engineering tasks.

### 3 METHODOLOGY

In this section, we introduce the Multi-Frequency Temporal State Representation Learning (MBTS) architecture, which is a framework for temporal state feature extraction based on multi-scale frequency signals. First, we introduce the generating multi-frequency state feature method based on a fast Fourier transform. Then, In this section, we introduce the Multi-Frequency Temporal State Representation Learning (MBTS) architecture, which is a framework for temporal state feature extraction based on multi-scale frequency signals. First, we introduce the generating multi-frequency state feature method based on a fast Fourier transform. Then, we adopt the Multi-frequency masked mechanism to assign weight to state features over different frequency bands.

#### 3.1 Multi-frequency State Feature Extraction

During model training, the agent constantly interacts with the environment to sample a sequence of state-action pairs, referred to as a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_T, a_T\}$ . The agent can not only obtain the immediate state  $s_t$  to accurately describe the agents' features in the current time step  $t$  but also the temporal features through the state sequence  $\hat{S} = \{s_0, s_1, \dots, s_t\}$  to describe the changing trends of the state variables. For example, the mean value of all position variables in the sequence can help the agent to predict and explore the unreached area. In other words, we endow the agent can extract global features and local features from temporal state sequences like humans. Further, the powerful feature extraction ability improves the sampling efficiency of the agent in deep reinforcement learning.

To handle the above challenges, we study the influence of state feature signals by Fourier analysis. The state sequence is a discrete signal in the time domain. We can decompose it into multi-frequency signals in the frequency domain, where: 1) The low-frequency signals represent the overall trend of the temporal state signal. 2) High-frequency signals contain the details of immediate state features. The curves of the state variables are sharpened or passivated to filter out signals of different frequency bands. We propose the Multi-frequency State Feature Extraction method based on the above ideas.

In practice, the last  $N$  samples are saved as a sequence  $\bar{S}_{t,N} = \{s_{t-N+1}, \dots, s_{t-1}, s_t\}$  to get discrete state signals in the time domain, which were transformed into the frequency-domain signal  $\bar{X}_N$  was calculated by Discrete Fourier Transform(DFT). To extract the trend feature of the state curve, high-frequency signals larger than the threshold  $\lambda$  are filtered. The different  $\lambda$  enable the feature vectors of multiple frequency bands to be generated. Based on this idea, the changed frequency domain state signal is defined by:

$$x_k^\lambda = \begin{cases} \sum_{n=0}^{N-1} s_{t-n} e^{-i2\pi kn/N} & k \leq \lambda \\ 0 & \text{other} \end{cases}, \quad (7)$$

where  $k$  is the index of the samples of the frequency domain signal.  $x_k$  represents the matching degree of the test signal with a frequency of  $2\pi kn/N$  and the original signal.  $\lambda$  is the threshold that limits the range of high-frequency signals.  $x_k$  will be equal

to 0 when the value of  $k$  is higher than  $\lambda$ . Otherwise,  $x_k$  is the convolution of the state signals  $\bar{S}_{t,N}$  and series  $e^{-i2\pi kn/N}$  which can be decomposed into sine and cosine testing signals:

$$e^{-i2\pi kn/N} = \cos(2\pi kn/N) - i \sin(2\pi kn/N). \quad (8)$$

Finally, the discrete state signals are converted from the frequency domain to the time domain by inverse Fourier transform:

$$s_t^\lambda = \frac{1}{N} \sum_{k=0}^{N-1} x_k e^{i2\pi kn/N}. \quad (9)$$

We have obtained a synthetic signal in the frequency bands between 0 and  $2\pi\lambda/N$  by the above method. We can change the value of  $\lambda$  to output state representation of different scales.  $s_t$  is the original state signal, equivalent to  $s_t^\lambda$  where  $\lambda = N$ . The  $\lambda$  is smaller, the feature curve is smoother, and the historical state significantly influences the state variable.

#### 3.2 Multi-frequency Masked Mechanism

In this section, we design a masking mechanism to select appropriate frequency band features for each variable in the state vector. To tackle the challenge of learning good representation, a mask generator net is trained to assign weights to the state features with various  $\lambda_i$ . The original state  $s_t$  passes through LSTM and mask net consisting of many Fully Connected(FC) layers. Meanwhile, we generate  $M$  (4 in the experiment) discrete state signals with different values of  $\lambda$  and obtain the state of the last step in their sequence respectively  $\{s_t^{\lambda_0}, s_t^{\lambda_1}, s_t^{\lambda_2}, \dots, s_t^{\lambda_{M-1}}\}$  by the method in Section 3.1. The feature mask  $\omega \in \mathbb{R}^{M \times W}$  is calculated by the mask net  $M$  vectors of shape  $1 \times W$  and concatenates them together to generate  $\mathbf{h} \in \mathbb{R}^{M \times W}$ .  $W$  is the size of the state vector. The  $i$ -th column vector  $\omega_i$  of the feature mask  $\omega$  is defined as:

$$\omega_i = \text{softmax}(h_i^{\lambda_0}, h_i^{\lambda_1}, \dots, h_i^{\lambda_{M-1}}). \quad (10)$$

We obtain embeddings of the multiple frequency bands state by a module  $f(\cdot)$  composed of multiple FC networks. Each input vector of different frequency bands corresponds to an FC network. The state representation  $\mathbf{z}_t$  is extracted according to multiple frequency bands states  $s_t^{\lambda_0}, s_t^{\lambda_1}, \dots, s_t^{\lambda_{M-1}}$  and feature map  $\omega$ :

$$\mathbf{z}_t = \omega \odot f(s_t^{\lambda_0}, s_t^{\lambda_1}, \dots, s_t^{\lambda_{M-1}}). \quad (11)$$

The feature mask should focus on only one of the multiple  $\lambda$  state signals instead of evenly distributing the weights. Evenly distributed weights cause all the multiple-state features to be blurry. Therefore, we add the entropy of the mask generation network to the loss function to reduce uncertainty. The loss function of the model follows:

$$L_{total}(\theta) = L_{policy} + \eta_0 L_{critic} - \eta_1 \sum_{i=0}^{W-1} \sum_{j=0}^{M-1} \log \omega_i^{\lambda_j}. \quad (12)$$

Intuitively, the attention mechanism can replace the mask generator net. However, the truth is that the training of the policy model converges more slowly and requires more samples, which conflicts with the improvement of sampling efficiency.

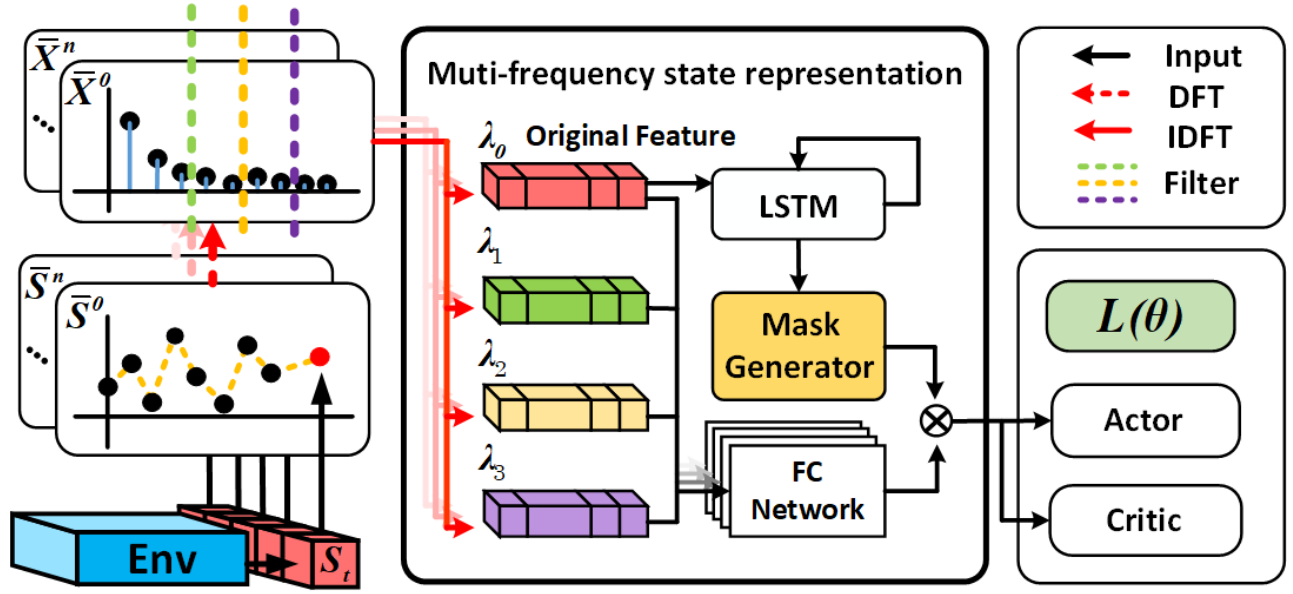


Figure 1: The overall framework of Multi-frequency State Representation Learning contains (a)Multi-frequency State Feature Extraction and (b)Multi-frequency Masked Mechanism. (a) process state sequences as discrete signals with different  $\lambda$  thresholds representing frequency ranges. We output all the states  $s_t^\lambda$  of all the ends of the sequences. (b) need to learn how to output the signals that determine which frequency different state variables should be selected at the objective function  $L_{total}(\theta)$

## 4 EXPERIMENT

In this chapter, we first introduce the environment settings and baselines. The MBTS effectiveness is verified in continuous control tasks through experimental analysis. Overall, we answer the following three questions:

- **Q1:** Can MBTS improve agents' performance for reinforcement learning?
- **Q2:** Does MBTS have advantages over other temporal state encoders?
- **Q3:** How does the sequence length affect temporal state encoders?

### 4.1 Experiment Setting

**4.1.1 Environment Setup.** The Unity Machine Learning Agents Toolkit (ML-Agents) is an open-source project that simulates an environment for training intelligent agents[1, 9, 10]. We evaluate the complexity level of tasks based on the size of the state and action space and test the performance of algorithms on these tasks. We conduct experiments in the following four tasks:

- **3DBall.** A balance-ball task, where the agent balances the ball on its head. It has 8 observed variables and 2 continuous actions.
- **Worm.** A worm with a head and 3 body segments moves its body toward the goal direction. It has 64 observed variables and 9 continuous actions.
- **Crawler.** A creature with 4 arms and 4 forearms moves its body toward the goal direction. It has 172 observed variables and 20 continuous actions.

- **Walker.** A humanoid agent with 26 degrees of freedom moves its body toward the goal direction without falling. It has 243 observed variables and 39 continuous actions.

**4.1.2 Baselines.** To demonstrate the effectiveness of multi-frequency temporal state representations in complex continuous control tasks, we selected baselines for analysis in the ML-Agents tasks. **PPO**[25] is a optimization algorithm that learns with the goal of a surrogate objective. **SAC**[7] secures diversity of samples by constructing policies with high entropy while learning with the same goal of maximizing expected rewards. PPO and SAC have been proven to have good results in continuous action space tasks. **DDPG**[16] is an algorithm that has strengths in solving tasks with continuous action space. **R2D2**[13] alleviates the problems of representational drift and recurrent state staleness that can suffer when using RNNs for off-policy RL through burn-in and stored state strategies.

### 4.2 Results and Analysis (Q1)

To prove that our method can significantly improve agents' performance, we compare the performance of five different algorithms in two continuous control tasks. Figure 2 presents the score curves of various methods in complex continuous control tasks. As Figure 2(a) shows, the scores of agent increase significantly and converge at about 3 million steps. Figure (b) shows that the curve of MBTS has declined at about 9 million steps. We find that the walker who once mastered standing skills fell again when the mask generator changed the state variables' frequency band. However, the walker quickly learns how to stand and walk after the frequency bands are

fixed. Finally, our method exceeds the results of other baselines at about 14 million steps.

In particular, R2D2 is the only algorithm that uses the recurrent network to encode state sequences, which performed significantly worse in the crawler task. The agent could learn how to stand up and move in a particular goal-relevant direction. However, agent behavior has become more cautious than other algorithms, which means the agent’s joints change less and move more slowly. We wanted to know how to handle sequential state sequences that would lead to this result. Therefore, we analyze various approaches to processing sequential states in the ablation experiment. It is worth noting that we used the replay buffer technique to improve sample collection efficiency for both PPO and SAC. We found in the relatively challenging tasks we conducted that SAC was not more effective than PPO. We believe this is related to the fact that PPO has been improved to be an efficient off-policy algorithm.

### 4.3 Ablation Study (Q2)

To analyze the impact of various temporal state coding methods on model training, we combine temporal encoding modules with PPO for ablation study. We extend the experiment to continuous control tasks with multiple difficulty levels. The methods for comparison are as follows: 1) **None**: The primary method hasn’t any encoder. 2) **MBTS**: The PPO algorithm applies Multiple Frequency Bands Temporal State Representation. 3) **LSTM**: The LSTM is used to process the state sequence, and its results mainly represent the performance of the recurrent network in continuous control tasks. 4) **Attention**: The attention mechanism is used to assign weights to the states at different time steps.

Figure 3 shows the results for different state sequence encoders. LSTM has the best results when faced with simple continuous control tasks (Figure 3(a) and Figure 3(b)). However, its performance degrades by increasing the observation and action space size (Figure 3(c) and Figure 3(d)). Attention isn’t easy to learn good state representations because the model cannot remember the order information of the state sequences. Figures 3 (c) and (d) show that our method always converges faster to a better policy. MBTS extracts the temporal state representation to improve the sampling efficiency in complex control tasks.

### 4.4 Parameter Sensitivity (Q3)

Table 1 shows the performance of various state sequence encoders for state sequences of different lengths. The results show that the LSTM and Attention Mechanism (ATTEN) performances decrease when the sequence length increases to 256. By rendering the environment, we found that the end of the agent’s manipulator swings less and more often when the agent receives a longer sequence of states. We believe that aggregating temporal state sequences weaken immediate state features. However, according to the state variables, our method automatically selects an appropriate frequency band, which means that model balances the proportion of temporal and immediate features. As shown in Table 1, with increasing the length of state sequences, the state signals in multiple frequency bands have less noise, but LSTM and ATTEN aggregate more temporal state information. Therefore, MBTS gets the best

score when the length is 256, while other methods’ performance is worse.

## 5 CONCLUSION

In this paper, we propose Multi-Frequency State Representation Learning, an efficient method to extract representations from state sequences in complex continuous control tasks. In order to extract temporal state features while maintaining high sampling efficiency, we designed a Multi-frequency State Feature Extraction based on the Fourier transform. Multi-frequency Masked Mechanism enables the model to adjust the frequency of various input signals to the appropriate frequency band. Intuitively, our method enables the agent to consider whether or not to choose state features influenced by historical signals and finally helps the agent to learn the optimal policy quickly. Experimental results demonstrate that our method is effective when the control task is complex. In conclusion, we provide a multiple frequency bands temporal state representation learning method, and future work can develop new methods with Fourier analysis for deep reinforcement learning.

## REFERENCES

- [1] Laura Almon-Manzano, Rafael Pastor-Vargas, and José Manuel Cuadra Troncoso. 2022. Deep Reinforcement Learning in Agents’ Training: Unity ML-Agents. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Springer, 391–400.
- [2] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. 2019. Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528* (2019).
- [3] Andrea Banino, Adrià Puidomenech Badia, Jacob Walker, Tim Scholtes, Jovana Mitrovic, and Charles Blundell. 2021. Coberl: Contrastive bert for reinforcement learning. *arXiv preprint arXiv:2107.05431* (2021).
- [4] Anuja Bhargava and Atul Bansal. 2021. Fruits and vegetables quality evaluation using computer vision: A review. *Journal of King Saud University-Computer and Information Sciences* 33, 3 (2021), 243–257.
- [5] Chuan-Zhi Dong and F Necati Catbas. 2021. A review of computer vision-based structural health monitoring at local and global levels. *Structural Health Monitoring* 20, 2 (2021), 692–743.
- [6] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, Joelle Pineau, et al. 2018. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning* 11, 3-4 (2018), 219–354.
- [7] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR, 1861–1870.
- [8] Zhiyu Huang, Jingda Wu, and Chen Lv. 2022. Efficient deep reinforcement learning with imitative expert priors for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [9] Phan Thanh Hung, Mac Duy Dan Truong, and Phan Duy Hung. 2022. Tuning Proximal Policy Optimization Algorithm in Maze Solving with ML-Agents. In *International Conference on Advances in Computing and Data Sciences*. Springer, 248–262.
- [10] Arthur Juliani, Vincent-Pierre Berges, Ervin Teng, Andrew Cohen, Jonathan Harper, Chris Elion, Chris Goy, Yuan Gao, Hunter Henry, Marwan Mattar, and Danny Lange. 2020. Unity: A general platform for intelligent agents. *arXiv preprint arXiv:1809.02627* (2020).
- [11] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. 2019. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374* (2019).
- [12] Steven Kapturovski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. 2018. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*.
- [13] Steven Kapturovski, Georg Ostrovski, John Quan, Remi Munos, and Will Dabney. 2018. Recurrent experience replay in distributed reinforcement learning. In *International conference on learning representations*.
- [14] Thomas William Körner. 2022. *Fourier analysis*. Cambridge university press.
- [15] Xujun Li, Lihong Li, Jianfeng Gao, Xiaodong He, Jianshu Chen, Li Deng, and Ji He. 2015. Recurrent reinforcement learning: a hybrid approach. *arXiv preprint arXiv:1509.03044* (2015).

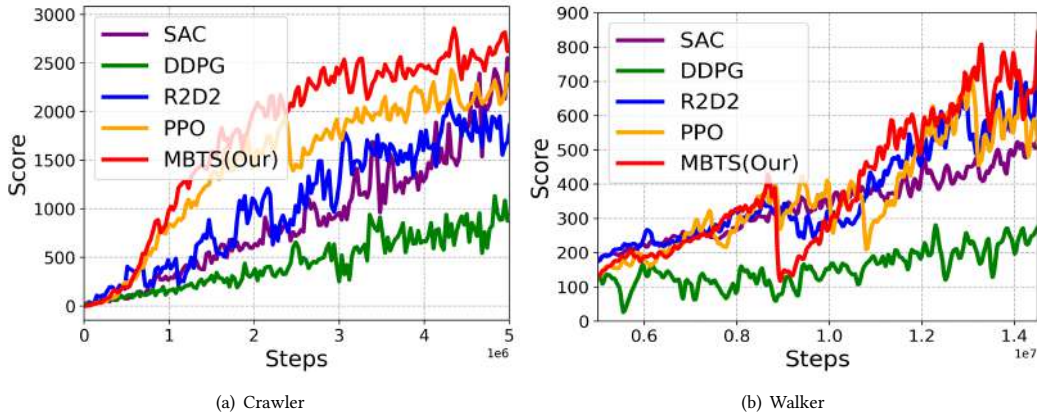


Figure 2: Experimental results of many RL methods in complex continuous control tasks.

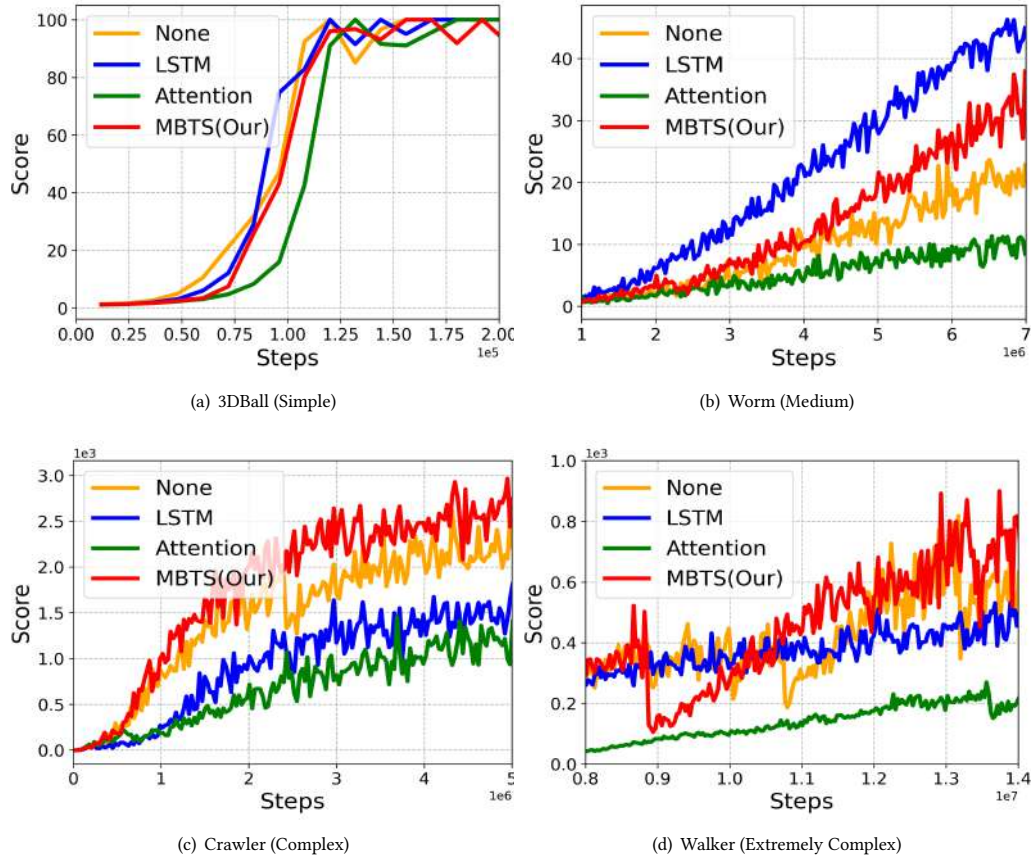


Figure 3: Results of various methods for encoding temporal state sequences

- [16] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).
- [17] Zichuan Lin, Junyou Li, Jianing Shi, Deheng Ye, Qiang Fu, and Wei Yang. 2021. Juewu-mc: Playing minecraft with sample-efficient hierarchical reinforcement learning. *arXiv preprint arXiv:2112.04907* (2021).

- [18] Haochen Liu, Zhiyu Huang, Jingda Wu, and Chen Lv. 2022. Improved deep reinforcement learning with expert demonstrations for urban autonomous driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 921–928.
- [19] David W Lu. 2017. Agent inspired trading using recurrent reinforcement learning and lstm neural networks. *arXiv preprint arXiv:1707.07338* (2017).

**Table 1: Impact of state sequence length on different methods.**

Length	Crawler			Walker		
	MBTS	LSTM	ATTEN	MBTS	LSTM	ATTEN
16	1304 $\pm$ 99.54	1442 $\pm$ 59.04	<b>1052<math>\pm</math>32.01</b>	703 $\pm$ 43.71	594 $\pm$ 23.19	<b>216<math>\pm</math>20.12</b>
32	1542 $\pm$ 107.96	1271 $\pm$ 48.99	1011 $\pm$ 47.64	732 $\pm$ 39.05	<b>633<math>\pm</math>27.32</b>	172 $\pm$ 17.47
64	1932 $\pm$ 133.32	1241 $\pm$ 69.50	926 $\pm$ 40.70	806 $\pm$ 43.32	563 $\pm$ 24.04	87 $\pm$ 12.63
128	2377 $\pm$ 164.35	<b>1525<math>\pm</math>78.44</b>	522 $\pm$ 23.06	774 $\pm$ 60.32	387 $\pm$ 22.71	70 $\pm$ 13.77
256	<b>2417<math>\pm</math>173.74</b>	1308 $\pm$ 63.84	504 $\pm$ 20.55	<b>823<math>\pm</math>67.87</b>	231 $\pm$ 19.04	84 $\pm$ 9.31

- [20] Vincent Mai, Kaustubh Mani, and Liam Paull. 2022. Sample efficient deep reinforcement learning via uncertainty estimation. *arXiv preprint arXiv:2201.01666* (2022).
- [21] Anthony Manchin, Ehsan Abbasnejad, and Anton van den Hengel. 2019. Reinforcement learning with attention that works: A self-supervised approach. In *International conference on neural information processing*. Springer, 223–230.
- [22] Ananya Paul and Sulata Mitra. 2021. Management of traffic signals using deep reinforcement learning in bidirectional recurrent neural network in ITS. In *2021 5th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*. 60–64.
- [23] Antonin Raffin, Jens Kober, and Freek Stulp. 2022. Smooth exploration for robotic reinforcement learning. In *Conference on Robot Learning*. PMLR, 1634–1644.
- [24] Yongming Rao, Jiwen Lu, and Jie Zhou. 2017. Attention-aware deep reinforcement learning for video face recognition. In *Proceedings of the IEEE international conference on computer vision*. 3931–3940.
- [25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).
- [26] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
- [27] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.
- [28] Adam R Villaflor, Zhe Huang, Swapnil Pande, John M Dolan, and Jeff Schneider. 2022. Addressing optimism bias in sequence modeling for reinforcement learning. In *International Conference on Machine Learning*. PMLR, 22270–22283.
- [29] Aichen Wang, Wen Zhang, and Xinhua Wei. 2019. A review on weed detection using ground-based machine vision and image processing techniques. *Computers and electronics in agriculture* 158 (2019), 226–240.
- [30] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincai Huang, Xin Xu, Bin Dai, and Qiguang Miao. 2022. Deep reinforcement learning: a survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [31] S Kevin Zhou, Hoang Ngan Le, Khoa Luu, Hien V Nguyen, and Nicholas Ayache. 2021. Deep reinforcement learning in medical imaging: A literature review. *Medical image analysis* 73 (2021), 102193.
- [32] Kai Zhu and Tao Zhang. 2021. Deep reinforcement learning based mobile robot navigation: A review. *Tsinghua Science and Technology* 26, 5 (2021), 674–691.

# Federated Learning-Based Intrusion Detection Method for Smart Grid

Bin Dongmei

Electric Power Research Institute of  
Guangxi Power Grid Co., Ltd  
bin\_dm.sy@gx.csg.cn

Li Xin

Electric Power Research Institute of  
Guangxi Power Grid Co., Ltd  
Li\_xin.sy@gx.csg.cn

Yang Chunyan

Electric Power Research Institute of  
Guangxi Power Grid Co., Ltd  
Yang\_Cy.sy@gx.csg.cn

Han Songming

Electric Power Research Institute of  
Guangxi Power Grid Co., Ltd  
songming\_h@gx.csg.cn

Ling Ying

Electric Power Research Institute of  
Guangxi Power Grid Co., Ltd  
Ling\_y.sy@gx.csg.cn

## ABSTRACT

Power systems have revealed serious security problems in the process of gradual opening, and intrusion detection as an important security defense measure can detect potential intrusions in a timely manner. In the big data environment of electric power, there are information silos between different electric power data owners, and in order to obtain intrusion detection models with better performance, traditional methods need to fuse data from all parties, which often brings difficulties in information security and data privacy protection. In this paper, we propose a distributed intrusion detection framework based on federated learning and apply it to network traffic data analysis. The framework aims to ensure the information security of each local power data while establishing a collection of decentralized data and completing the joint training of models from multiple data sources. The experimental results show that the scheme achieves 98.1% accuracy on the simulated data set, which is better than other commonly used intrusion detection algorithms. In addition, the method well ensures the security and privacy of data because the data are not interoperable among each participant under the federated learning mechanism.

## CCS CONCEPTS

• **Security and privacy** → Intrusion/anomaly detection and malware mitigation; Intrusion detection systems; Artificial immune systems.

## KEYWORDS

smart grid, intrusion detection, federal learning, neural network

## ACM Reference Format:

Bin Dongmei, Li Xin, Yang Chunyan, Han Songming, and Ling Ying. 2023. Federated Learning-Based Intrusion Detection Method for Smart Grid. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590060>

(CACML 2023), March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590060>

## 1 INTRODUCTION

In the context of Industry 4.0, with the increasing degree of informatization and business needs, the traditional power grid system is gradually transformed into a more open smart grid system. In 2015, Ukraine's power infrastructure was attacked by malicious code, which led to a long and large-scale power outage and seriously affected the normal operation of society. normal functioning of society. Similar cyber-attacks on power grid systems include the Stuxnet virus that attacked Iran's nuclear power plants, the VPN-Filter virus that infected data acquisition and monitoring systems (SCADA), and the Snake attack that disrupted the internal network of Enel, the European power company. The internal traffic of the power grid industrial control system is mainly the information of the devices at the edge of the network collected by the intranet, and the system lacks in-depth analysis of network packets and corresponding security protection means such as intrusion detection and traffic monitoring. Therefore, to enhance the detection mechanism of network intrusion in the power grid industrial control system, in order to ensure the efficient and safe operation of the power grid system.

Intrusion detection technology is a technique to obtain and analyze data in the network and determine whether it is a legitimate access or network intrusion. Network intrusions in power grid industrial control systems mainly include: attacks against industrial network protocol vulnerabilities; intruders entering the intranet through the external network to form intrusions, such as unauthorized access; and attacks against wireless transmission, which is widely used in the environment. Intrusion detection technology can identify unauthorized operations from within the system and malicious intrusions and probes from outside, so that security personnel can take timely countermeasures, thus the deployment of intrusion detection systems within the power terminal network units has become an effective means to ensure the security of smart grid systems.

Many experts have already conducted in-depth research on intrusion detection in power networks and proposed a series of detection algorithms. However, in order to train a better model, these methods require massive amounts of data for training to complete modeling in order to improve intrusion detection accuracy and avoid missing

real intrusions due to the generation of large amounts of false alarm information. However, a single power organization often has difficulty training a model with good performance due to insufficient sample data, and thus needs to combine data from multiple parties to achieve joint training of the model. In reality, for data privacy and information security reasons, data from different power agencies are often not shared with each other. Therefore, how to improve the accuracy of intrusion detection models while securing data privacy is an important concern for the current power grid system by using the dispersed massive power data.

Federated learning, as a privacy-preserved learning approach, can build a shared model by passing only the encrypted intermediate parameters in the training process without guaranteeing the data out of the local area, and can achieve practical requirements after multiple rounds of iteration and communication. As an important infrastructure related to the people's livelihood, the data generated by the power grid system is characterized by high confidentiality and sensitivity, and there is a high risk of data leakage in the process of long-distance transmission. Therefore, this paper proposes an intrusion detection method based on federated learning, aiming to solve the problems that the training of high-precision models requires massive data and the data cannot be shared due to privacy and security issues.

## 2 RELATED WORK

In recent years, there are some researches on intrusion detection of smart grid. Andresini et al. [1] proposed a novel deep learning approach to provide an efficient way for computer networks to analyze network traffic in order to distinguish malicious activities. Liu et al. [2] in 2017 designed a detection model using binary logistic regression to obtain critical network data from the routing layer and used it to analyze the power sensor behavior with an accuracy of 96%. Siniosoglou et al. [3] proposed an intrusion detection system for smart grids exploiting self-encoders and adversarial generative networks, and demonstrated the effectiveness of the system through experiments. Mendonça et al. [4] presented a tree-convolutional neural network layering algorithm and a soft root-symbol activation function method, which can reduce the training time for generating models and detect DDoS attacks. Wu et al. [5] proposed a process state transition intrusion detection algorithm for industrial control systems that implements a two-stage anomaly detection.

With data security issues becoming more and more important to the public, federation learning as a promising tool to address data silos and data privacy has attracted wide attention from experts and scholars in various industries and applied to the field of intrusion detection. In 2020, Liu et al. [6] proposed an efficient communication federation learning method for industrial IoT anomaly detection, which firstly improves model generalization using a federation learning framework, and secondly proposes a convolutional neural network combined with a model of long and short-term memory for anomaly detection, and reduces the communication overhead by 50% based on accurate anomaly detection through a gradient compression mechanism. 2021 Wang et al. [7] proposed a hierarchical federation learning-based anomaly detection for industrial IoT, which uses federation learning techniques to build a generic monitoring model and then deep reinforcement learning

algorithms to train the local model, achieving high throughput, low latency, and high accuracy.

## 3 SYSTEM MODEL

### 3.1 Problem Description

Smart grid is a heterogeneous network interconnected by multiple devices and technologies, as shown in Figure 1, which can be divided into three parts according to coverage and communication rate, Wide Area Network (WAN), Neighborhood Area Network (NAN) and Home Area Network (HAN) [8]. Among them, the HAN includes power users and related information, while the NAN contains various types of power devices, and the WAN realizes data communication between various types of devices.

Two-way communication between smart grid devices is usually done by wired or wireless means. Due to the poor security of the power equipment in the HAN, it is likely to be subject to DOS attacks that interfere with its normal operation. Probe attacks may also scan the equipment or HAN for vulnerabilities, search system configuration or network topology. The data collected within the NAN is mainly uploaded by HAN, which is characterized by low computing and storage capacity, and thus cannot effectively resist intrusion, and is more vulnerable to power lifting (U2R) attacks. The data transmitted in the WAN is highly sensitive and relates to the normal operation of the entire power grid. If attacked, it may lead to the grid system being paralyzed, thus becoming the target of remote to local (R2L) attacks.

### 3.2 Network Structure

An intrusion detection problem can be defined a classification problem where a classification model is trained by supervised learning and then the trained model is used to complete predictions on unknown data. Neural networks have nonlinear adaptive information processing capability to obtain the identification of unknown anomalous behavior by analyzing a large number of training samples. In traffic detection, neural networks can be used to analyze real-time data and detect anomalous behavior by establishing a mapping relationship between specific traffic patterns and the system security state through a learning/training process. Therefore, in order to effectively detect cyber-attacks against smart grid, a convolutional neural network model GCNN with the gating mechanism is developed. The structure is shown as Figure 2. The GCNN model has 10 layers, including one input layer, three convolutional layers with gating mechanism, three Dropout layers, one maxpooling layer, one fully connected layer and one softmax layer.

The detailed model as follows.

(1) Input layer: the pre-processed raw intrusion data is fed into the convolutional neural network for extracting high-level feature.

(2) Convolutional layers: layers 2, 4 and 6 are all convolutional layers. The concept of threshold convolution is used in the convolution layer, and there are two parts in the threshold convolution: one part is the activation value of the convolution, i.e., B; the other part is the direct linear to get the convolution, i.e., A. The two parts of A and B are multiplied to get the corresponding convolution value. It has been demonstrated in the literature [9] that smaller convolutional kernels give better local feature and classification performance, so in terms of the design of convolutional kernel size,

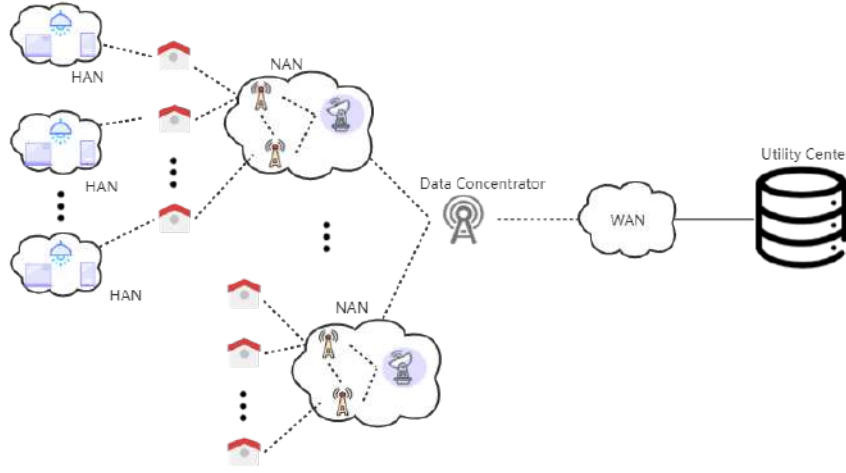


Figure 1: Smart grid application scenarios

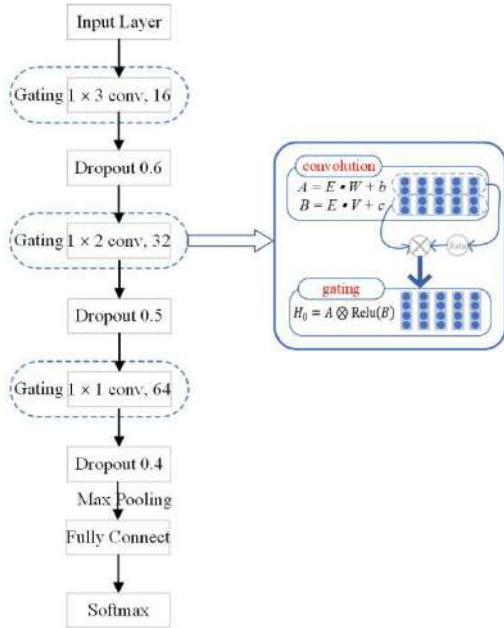


Figure 2: Architecture of the gated convolutional network

GCNN adopts a small convolutional kernel strategy with  $1 \times 3$ ,  $1 \times 2$  and  $1 \times 1$  kernels, and the number of convolutional kernels are 16, 32 and 64, respectively. In addition to small convolutional kernels can cluster the learned features, which to some extent alleviates the impact of convolutional. The model performance is affected by the redundancy.

(3) Dropout layer: because traditional convolutional neural network models are prone to overfitting during training, dropout is used in layers 3, 5, and 7 of the GCNN model to mitigate the impact of overfitting on model performance.

(4) MaxPooling layer: considering the introduction of pooling layer can reduce the calculation amount, the 8th layer of GCNN model is the Max-Pooling layer, and the stride is set to 2, which makes the number of parameters halved.

(5) Fully connected layer: this layer is responsible for mapping the learned distributed features to the sample tag space.

(6) Softmax layer: GCNN uses Softmax regression, a standard classifier for multi-classification or binary classification problems, as a binary classifier.

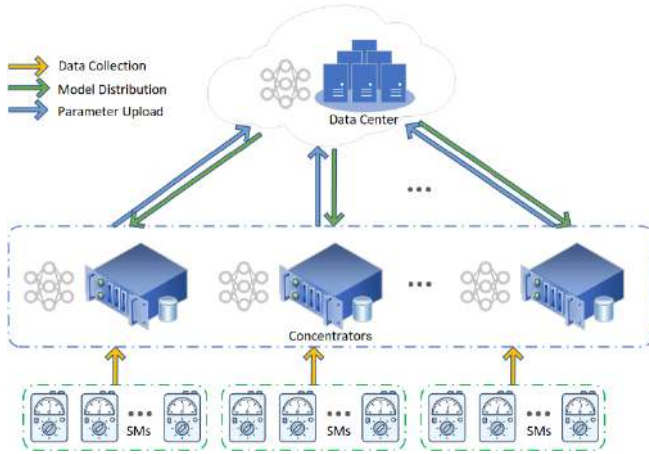
### 3.3 Learning Architecture

A natural server-client federation learning architecture is used to implement the algorithms involved. As shown in Figure 3, the data center is responsible for coordinating the centralizers for model training and global modeling. Federated learning consists of a federation server and multiple participants, each of which holds its own local data, and the data of each participant does not leave the participant's local area during the process. The federated model is trained for multiple rounds until it meets the requirements, such as reaching a certain test accuracy or exceeding a limit on the number of training rounds.

In this paper, each concentrator is involved in model training as a client. The set of clients is  $C = \{C_k | k = 1, 2, \dots, N\}$ , the data owned by the client  $C_k$  is denoted as  $D_k$ , and the data set owned by all clients constitutes the total data set as  $D$ . Before the client can train the local model, it must first download the global model on the server side  $m$  and the initial weight parameters  $w$ . Then, the local data is used for model training. Let  $F_k(w)$  be the objective function of the client  $C_k$ .

$$Fk(w) = \frac{1}{|Dk|} \sum_{i \in Dk} f_i(w) \quad (1)$$

where  $f_i(w)$  is the corresponding loss function. The loss function of  $D_k$  the loss function of all the data in the client and divide it by the amount of data  $|D_k|$  to get the average loss function of the client.



**Figure 3: Federation learning architecture under consideration**

In this paper, we use the aforementioned GCNN as the training model and cross-entropy as the loss function for anomaly detection of traffic data.

$$H(y, y') = - \sum_{i=1}^n y_i \log(y'_i) \quad (2)$$

where  $y, y'$  are the predicted and true values, respectively, and  $n$  is the classification category.

The model accuracy can be continuously improved and better performance can be obtained by solving the minimum value of the loss function through the optimizer's continuous optimization search. The local client model weight update formula is as follows, where  $w_k^t$  denotes the model weights of client  $C_k$  at round  $t$  of training, and  $\eta$  is the learning rate.

$$w_k^t = w_k^{t-1} - \eta \nabla F_k(w) \quad (3)$$

After the local client finishes training, it uploads  $w_k^t$ . After the server completes the weight aggregation, the global weights are downloaded again to update the local model and the next round of iterative training is performed until the global model training is completed.

The server side controls the entire model training process. Firstly, the server side coordinates each client to confirm the model to be trained and completes the system initialization settings. In addition, to ensure the clients' privacy, the training data are only stored locally in the clients, and each client only uploads the updated weights  $w$ . The server side need to compute the whole model parameters  $w$  and complete the aggregation of the whole model parameters. At present, there are various methods for aggregation of model parameters, and in this paper, the aggregation method FedAvg proposed in [10] is used, and its calculation formula is

$$w^{t+1} = \sum_{k=1}^N p_k w_k^t \quad (4)$$

where  $p_k$  denotes the client  $C_k$  weight in the overall model training, generally  $\frac{|D_k|}{|D|}$ .

In summary, the proposed Federated Learning-based Intrusion Detection Algorithm called FL-GCNN as follows.

---

**Algorithm 1: FL-GCNN algorithm**

---

**Input:** clients set  $C$ , data features vector  $x_i$ , communication number  $R$ .

**Output:** The comprehensive intrusion detection model

```

1 Function Server()
2 For  $C_k$  in  $C$  Do
3    $get(w_k^t)$ 
4    $w^{t+1} \leftarrow \sum_{k=1}^N p_k w_k^t$ 
5    $send(w^{t+1})$ 
6 Function Client $_k$ ()
7 For  $t = 1$  to  $R$  Do
8    $get(w^t)$ 
9    $GCNN.update(w^t)$ 
10   $y' \leftarrow GCNN(x_i)$ 
11   $Loss \leftarrow cross\_entropy(y', y)$ 
12   $backward(Loss)$ 
13   $w_k^t \leftarrow w_k^t - \eta \nabla F_k(w)$ 
14   $send(w_k^t)$ 

```

---

## 4 EXPERIMENTAL EVALUATION

### 4.1 Implementation

**Data selection.** In recent years, open communication protocols such as Industrial Ethernet and TCP/IP have been introduced into a new generation of power systems, the system platform tends to be open and standardized, and the connection with external networks becomes closer and more frequent, which makes the inherent vulnerabilities and attack surface of power systems increasing, and the security attacks faced by the Internet are introduced into the power system. In this paper, the NSL-KDD dataset is selected to simulate the intrusion traffic in the smart grid.

There are 22 attacks in the training dataset, and the testing dataset includes not only the attack types in the training set, but also 17 attacks that are not found in the training set, which can be further categorized into four classes of attacks, as shown in Figure 4. Each traffic data consists of 41 features and a category label, where the features can be classified into two types of continuous features and discrete features by value type, and into basic features, traffic features and content features by meaning [11]. The complete dataset contains a larger number of normal connection records

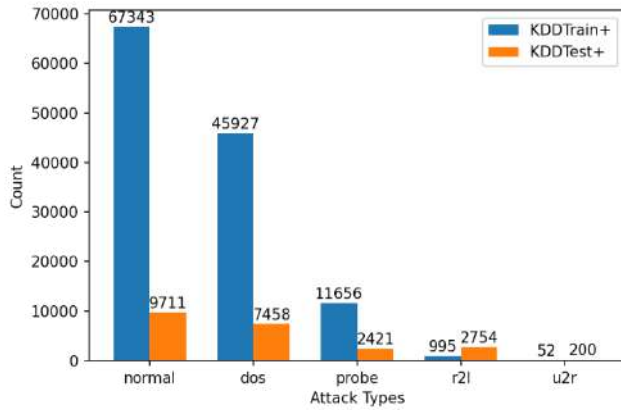


Figure 4: NSL-KDD Dataset Distribution Statistic

compared to attacks, and the number of different kinds of attack samples in the training set is very unbalanced.

Data preprocessing. Considering that U2R and R2L attacks cause more damage to the grid, this paper uses the SMOTE algorithm [12] to oversample these two kinds of connection records and increase their data proportion, in addition, 20% of the connection records are randomly removed from the normal and DOS connection records to make the number of the four kinds of intrusions more balanced and obtain a new training dataset.

Considering that only numerical dimensions can be processed by convolutional neural networks, the character-based data dimensions are transformed into numerical values by a unique thermal coding technique. Then the Min-Max standard normalization algorithm is used to map the data between 0 and 1 without destroying the original mapping relationship:

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Where the minimum value of sample data  $x$  is denoted as  $x_{\min}$  and the maximum value of sample data  $x$  is denoted as  $x_{\max}$ . The normalization helps to eliminate the singular data, speed up the gradient descent of the neural network, and improve the classification detection accuracy.

Testing environment. The experimental computer is configured with 16G RAM, R5-5600H processor, and Ubuntu operating system, based on the TensorFlow Federated open-source federation learning framework for federation intrusion detection simulation experiments.

## 4.2 Performance Indicators

The typical indexes of accuracy, precision, recall rate and F1 score were used in this experiment. The index accuracy and precision refers to the proportion of data correctly judged by the algorithm and samples correctly predicted as attacks among all samples predicted as attacks respectively. The index recall rate refers to the proportion of samples correctly predicted as attacks in the attack sample set. The index F1 score is a comprehensive index, which takes into account the impact of accuracy and recall rate, and can provide a more comprehensive evaluation of the performance of a

single classifier. The higher the F1 score, the higher the quality of intrusion detection.

## 4.3 Experiment Scheme

To verify the model performance proposed in this paper, the Singh et al. [13] proposed model (FL-CNN) as the baseline model for comparison, which is a federal learning model that directly applies the convolutional neural network to the intrusion detection scenario, and can effectively respond to the improvement in the convergence speed and accuracy of model. Meanwhile, in order to verify the performance of the proposed method on the premise of data privacy protection, we also compare the classification performance of the centralized learning model and the federated learning model on the NSL-KDD. The centralized model uses the GCNN model proposed in the previous paper for experiments, and the experimental goal is to detect all the abnormal traffic records, i.e., to perform binary classification on the dataset.

During the training process, keeping the parameter settings consistent, the optimizer selects the adaptive moment estimation algorithm Adam, the activation function uses Relu, the number of training rounds is set to 100, the batch size is set to 32, and the learning rate is set to 0.001. In addition, in federal learning model, the training data are randomly divided into 10 copies, and it is assumed that 10 clients participate in training at the same time, and the formula (4) for parameter aggregation.

## 4.4 Analysis of results

To demonstrate that the proposed algorithm can improve the intrusion detection accuracy in the federal learning environment, the effect of CNN with gating mechanism is compared with the standard CNN with same number of convolutional layers for intrusion detection. In addition, this experiment also compares the differences in intrusion detection accuracy and training time between centralized learning and federated learning.

Figure 5 shows the comparison of the accuracy of the two federal learning models and the centralized learning model. (a) and (b) show the training results on the training set and the test set, respectively. It can be seen that the intrusion detection model can obtain better classification accuracy in the centralized training mode. When the number of training rounds reaches 100, the classification accuracy of the model can reach 99.2%. Compared with the centralized training mode, the federated learning mode, in which each client trains the model locally and only uploads the model parameters to the server, can effectively protect data privacy by keeping the data out of the local area during the process. Compared with the FL-CNN model, the FL-GCNN can achieve higher classification accuracy and better final results with the same number of training rounds, indicating that the introduction of the gating mechanism in the convolutional layer can improve the performance of the intrusion detection model.

Figure 6 shows the comparison of the overall training loss of the two federal learning models and the centralized learning model, (a) and (b) are the results on the training and test sets, respectively. From the figure, we can see that the centralized learning converges faster than the federal learning, but the final loss is not much different, which verifies the feasibility of the federal learning. The loss of

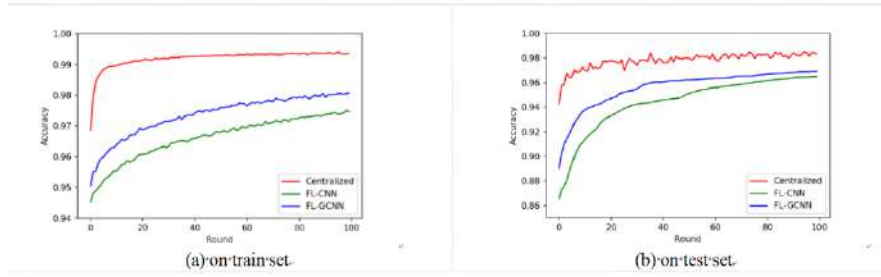


Figure 5: Comparison of accuracy

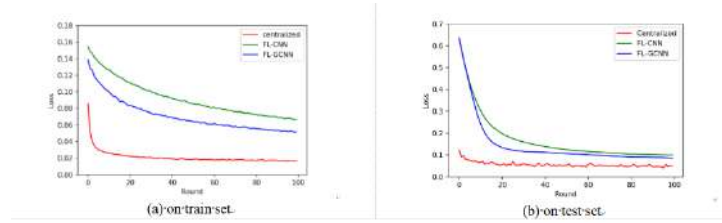


Figure 6: Comparison of loss

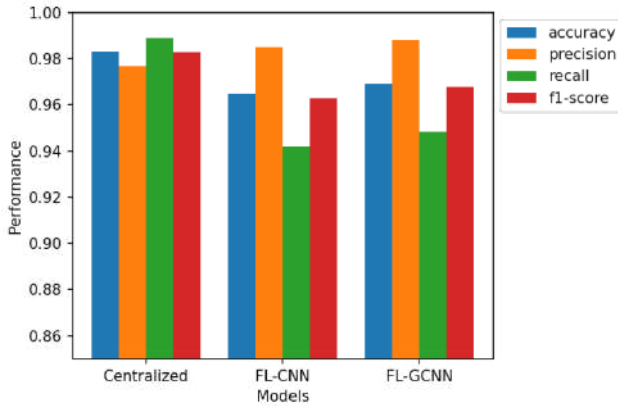


Figure 7: Comparison of Evaluation Metrics In Test Set

FL-GCNN on the test set is close to the minimum when the model is trained up to the 20th round, while the loss value of FL-CNN only tends to be constant at the 60th round of training, and FL-GCNN always outperforms FL-CNN when the number of training rounds is the same, indicating that the gating mechanism can improve the convergence speed of the model.

The experimental results of the three models on the test set are shown in Figure 7. It can be seen that the proposed model in the paper outperforms the baseline model in all evaluations. In addition, the performance of the proposed federal model has a small degradation compared to the centralized learning model, but it is within the acceptable range. That is, federal learning is able to obtain the desired accuracy while ensuring data privacy. Therefore, considering its high performance in intrusion detection as well as

protecting the privacy of data resources, the proposed model has some implications for smart grid intrusion detection.

## 5 CONCLUSION

In this paper, we propose a federation intrusion detection model based on gated convolutional neural networks for solving the problem of cyber attacks and power data protection in smart grids. First, we propose a smart grid intrusion detection framework using a federated learning mechanism. We also design a convolutional neural network model with a gating mechanism, which can extract higher-level, more abstract features by stacking CNNs. In addition, the linear gating unit not only effectively reduces gradient dispersion, but also retains the ability to be nonlinear, making model convergence and training easier and enabling efficient detection and classification of network attacks on smart grids. Finally, extensive experiments on the NSL-KDD dataset demonstrate the effectiveness of the proposed FL-GCNN framework, which provides a new idea for privacy-preserving intrusion detection in smart grids.

## ACKNOWLEDGMENTS

This paper is supported by project: Research and Application Project of Network Mutual Trust Support System for New Electricity System Business Subjects (Project No.: GXKJXM20210257), Guangxi Power Grid Company Science and Technology Project Funding.

## REFERENCES

- [1] Andresini, Giuseppina, Annalisa Appice and Donato Malerba, 2021. Near-est cluster-based intrusion detection through convolutional neural networks. *Knowledge-Based Systems* 216:106798 doi:<https://doi.org/10.1016/j.knsys.2021.106798>.
- [2] Yufei Liu and Dechang Pi. 2017. A Novel Kernel SVM Algorithm with Game Theory for Network Intrusion Detection. *KSII Transactions on Internet and Information Systems*, 11, 8, (2017), 4043-4060. DOI: 10.3837/tiis.2017.08.016.

- [3] Siniosoglou, Ilias, Panagiotis Radoglou-Grammatikis, Georgios Efstathopoulos, Panagiotis Fouliras and Panagiotis Sarigiannidis, 2021. A Unified Deep Learning Anomaly Detection and Classification Approach for Smart Grid Environments. *IEEE Transactions on Network and Service Management* 18(2):1137-1151 doi:10.1109/TNSM.2021.3078381.
- [4] Mendonça, Robson V., Arthur A. M. Teodoro, Renata L. Rosa, Muhammad Saadi, Dick Carrillo Melgarejo, Pedro H. J. Nardelli and Demóstenes Z. Rodríguez, 2021. Intrusion Detection System Based on Fast Hierarchical Deep Convolutional Neural Network. *IEEE Access* 9:61024-61034 doi:10.1109/ACCESS.2021.3074664.
- [5] Wu, Kehe, Zuge Chen and Wei Li, 2018. A Novel Intrusion Detection Model for a Massive Network Using Convolutional Neural Networks. *IEEE Access* 6:50850-50859 doi:10.1109/ACCESS.2018.2868993.
- [6] Liu, Yi, Neeraj Kumar, Zehui Xiong, Wei Yang Bryan Lim, Jiawen Kang & Dusit Niyato, 2020. Communication-Efficient Federated Learning for Anomaly Detection in Industrial Internet of Things.
- [7] Abdel-Basset, Mohamed, Nour Moustafa, Hossam Hawash, Imran Razzak, Karam M. Sallam and Osama M. Elkomy, 2022. Federated Intrusion Detection in Blockchain-Based Smart Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems* 23(3):2523-2537 doi:10.1109/TITS.2021.3119968.
- [8] Murat, Kuzlu, Pipattanasomporn Manisa and Rahman Saifur, 2014. Communication network requirements for major smart grid applications in HAN, NAN and WAN. *Computer Networks* 67:74-88 doi:https://doi.org/10.1016/j.comnet.2014.03.029.
- [9] Simonyan, Karen and Andrew Zisserman, 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv e-prints:arXiv:1409.1556.
- [10] Brendan McMahan, H., Eider Moore, Daniel Ramage, Seth Hampson and Blaise Agüera y Arcas, 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv e-prints:arXiv:1602.05629.
- [11] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani, 2009. A detailed analysis of the KDD CUP 99 data set.
- [12] Han, Hui, Wen-Yuan Wang and Bing-Huan Mao, Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In: Huang, D.-S., X.-P. Zhang & G.-B. Huang (eds) *Advances in Intelligent Computing*, Berlin, Heidelberg, 2005// 2005. Springer Berlin Heidelberg, p 878-887.
- [13] Singh, Praneet, Jishnu Jaykumar P, Akhil Pankaj and Reshmi Mitra, 2021. Edge-Detect: Edge-Centric Network Intrusion Detection using Deep Neural Network.

# Deep Learning AD Detection Model based on a Two-Layer Ensemble Module with Data Augmentation and Contrastive Learning

Weicheng Wang

Radley College

weichengwang2005@outlook.com

## ABSTRACT

Abstract—Alzheimer’s Disease (AD) is a long-term disease that gradually decreases cognitive functioning, such as thinking, memory and behavior. In 2015, 29.8 million AD cases were recorded and 1.9 million AD-related deaths were reported worldwide. Early detection and intervention are critical for such a deadly and costly disease. I present to tackle the detection of AD and its severity using a deep-learning architecture that consists of a two-stage ensemble system with contrastive learning and data augmentation. I evaluated it on the ADReSS Challenge’s dataset, which is subject-independent and balanced in terms of age and gender. When compared against a one-stage ensemble baseline approach, my two-stage ensemble system was able to achieve better results, with a F1-score of 95.7% in the AD classification task, and an RMSE score of 5.432. Moreover, I found that the data augmentation can effectively improve the robustness of the AD detection performance, particularly when there are sensor noises in the training and test data. Besides data augmentation, I also explored whether contrastive loss can further boost the robustness, and the results showed that contrastive learning might not be necessary when we have data augmentation.

## CCS CONCEPTS

• **Applied computing** → Life and medical sciences; Bioinformatics.

## KEYWORDS

Alzheimer’s Disease, Two-Layer Ensemble Module, Data Augmentation

### ACM Reference Format:

Weicheng Wang. 2023. Deep Learning AD Detection Model based on a Two-Layer Ensemble Module with Data Augmentation and Contrastive Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590061>

## 1 INTRODUCTION

Alzheimer’s dementia is a progressive neurodegenerative disease that causes cognitive and physical impairment. This in turn affects

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590061>

their quality of life, prognosis, and social relationships. Consequently, it has imposed increased health risks and a significant financial burden on patients, caregivers, families, and healthcare institutions. The number of people with dementia worldwide in 2020 was estimated at 55 million and reaching 139 million in 2050 [1]. According to the World Health Organization, the global economic burden is currently nearly a trillion dollars which amounts to 1.1% of the global GDP [2], with 63% of people with dementia living in low- and middle-income countries [3].

Currently, Alzheimer’s detection methods are classified into two different categories, invasive and non-invasive. Though accurate, invasive methods require obtaining data from patients through procedures that are mostly unsafe and uncomfortable for the patients, such as lumbar puncture and blood extraction. On the other hand, non-invasive tests are harmless and more convenient for diagnosis, such as neuroimaging and behavior analysis [4]. The traditional methods of AD diagnosis are generally subpar since the misdiagnosis rate can be as high as one in five [5], even for the most experienced clinicians. An early diagnosis can assist immensely in better management of their healthcare needs. Such that an accurate clinical diagnosis of AD as distinguishing AD from other causes of dementia in life is critical for prognosis, treatment and research.

There have been existing works in the field that uses machine learning techniques for AD diagnosis. Some of these studies have focused on using Magnetic Resonance Image (MRI) scan or cognatic test results such as the Mini Mental-State Exam (MMSE). However, MRI scanning involves a lot of costs and time while also being invasive, thus, we aim to develop a cost-effective, reliable and non-invasive approach to detect the onset of AD and estimate its severity using MMSE scores.

In this work, I propose multiple strategies to improve the robustness of deep learning models. There are two stages to my approach, two-layer ensemble, and data augmentation in addition to contrastive learning.

**Two-layer ensemble:** The first layer of the ensemble is performed across all five folds within each sub-model of distinctive features, and the second layer of the ensemble leverages the output from all three sub-models to improve the overall model’s performances.

**Data Augmentation and Contrastive Learning:** On top of the two-layer ensemble, data augmentation was performed by injecting noise into the dataset to improve the generalization of the model. then, the contrastive loss is calculated as the distance between the augmented and original data in order to boost the accuracy of the predictions.

Our contributions can be summarized in two folds: 1) the two-layer model ensemble, and 2) the data augmentation and contrastive loss.

## 2 RELATED WORKS

There are currently a few AD detection studies that have produced their own methodology and results.

[6] proposed a shared task on the recognition of Alzheimer’s Dementia called the ADReSS challenge, which includes pre-processed and balanced dataset consisting of speech recordings and transcripts. The challenge includes two tasks: the Alzheimer’s speech classification task and the neuropsychological score regression task. The ADReSS challenge serves as a standardized platform for future fair model comparisons.

Based on [6, 7] compares two approaches on AD detection: using knowledge based handcrafted features and finetuning with BERT based classification models. Through a 10-fold cross validation, the paper shows that BERT outperforms domain knowledge-based models (SVM, random forest, naive bayes) on multiple metrics.

[8] aims to tackle the ADReSS challenge with a multi-modal ensemble system that leverages acoustic, cognitive, and linguistic features. This paper built three models based on the aforementioned three types of information and employed an ensemble module in the end. It achieves the state-of-the-art results on both AD classification task and MMSE score regression task.

[9] proposes a multi-model system on AD classification task for the ADReSS challenge. It processed the acoustic and textual information respectively and combined the results in the end, achieving 81.25% accuracy on the test dataset.

[10] aims to avoid the suffering of extracting reliable features when the acoustic quality is poor. It employed time alignment information and confidence scores from the automatic speech recognition (ASR) system to identify audio segments of good quality. With the identified segments used, the paper shows that the F-measure improves on the Dementia task.

[11] is about using x-vectors to characterize speech signals while BERT models for transcriptions, and also evaluating features from silence segments of the ADReSS challenge dataset, the results indicate that the fusion of both models provides the best results, indicating the individual models contain complementary information.

[12] compares different NLP techniques on the ADReSS challenge, including Support Vector Machines (SVMs), Gradient Boosting Decision Trees (GBDT), and Conditional Random Fields (CRFs) and Transformer based models. The top performing models are a SVM model with TF-IDF as input and DistilBERT, yielding test score of 0.81-0.82 and a RMSE of 4.58 for AD classification and RSME tasks, respectively.

[13] also aims to tackle the ADReSS challenge. The paper proposes a model that obtains unimodal decisions from two LSTMs on different modalities of text and audio, and a gating mechanism is used to fuse the two outputs for the final prediction. They achieved 0.792 accuracies and a RMSE of 4.54 on test set of AD classification and regression experiments, respectively.

[14] compares AD detection using different acoustic and linguistic features and classifiers. The features explored in this paper are the result using ComParE, X-vector, linguistics, TFIDF and BERT. The models compared here are LDA, SVM and AT-LSTM. They found out that linguistic features outperform acoustic features.

Also, automatically extracted transcripts have comparable performance with manual ones, suggesting automatic detection of AD is viable. features can achieve high accuracy and sensitivity.

[15] focuses on AD detection with non-invasive data of speech and eye movements. They show that eye tracking data is predictive of AD. And using eye tracking data along with speech data outperforms using unimodality, indicating they are complementary to each other. experiments on classification using both speech and eye movement data, it was able to show that eye tracking data is predictive of AD.

## 3 METHODOLOGY

### 3.1 Feature Selection

The ADReSS Challenge Dataset [6] contains 1,955 speech segments from 78 non-AD subjects and 2,122 speech segments from 78 AD subject, elicited through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam [16]. The average number of speech segments produced by each participant was 24.86 (standard deviation  $sd = 12.84$ ). A system called CHAT coding [17] is used to generate the transcripts based on the above speech information, which is annotated with disfluencies, pauses, utterances and more complex events. The dataset contained both fully enhanced audio, and normalized audio chunks.

However, not every subject has a complete set of fully enhanced audio, normalized audio chunks, transcription data and Mini-Mental State Examination (MMSE). As none of the aforementioned data types had the same number of patients, the data involving some of the patients were removed with the implementation of contrastive learning to the system.

I set my focus on the ADReSS challenge dataset from INTER-SPEECH 2020. The baseline paper introduced a dataset containing over four thousand speech segments and transcripts from 78 AD and 78 non-AD participants, elicited through the Cookie Theft picture from the Boston Diagnostic Aphasia Exam and annotated using the CHAT coding system. Subsequently, we conducted literature reviews on various related works, which aims to provide a state-of-the-art approach to AD detection or MMSE prediction.

The ADReSS Challenge Dataset extracts cognitive and acoustic features using three different strategies, Disfluency, Acoustic and Interventions.

**Disfluency:** The disfluency features are extracted from the transcripts, such as word rate, intervention rate and different kinds of pauses and utterances. These are normalized by the respective audio lengths and scaled thereafter.

**Acoustic:** Also known as ComParE, the features include energy, spectral mel-frequency cepstral coefficients (MFCC) and low-level descriptors (LLDs). The extraction process was performed on the speech segments using the openSMILE toolkit.

**Interventions:** The Intervention features include one hot encoded sequence of speaker of both the participant (PAR) and the investigator (INV) extracted from the transcripts, which indicates the loss of train of thoughts.

The first three figures in 2 illustrates the architecture of the disfluency, acoustic and intervention models respectively. The disfluency model is a multi-layer perceptron (MLP) and the acoustic model is an MLP with a single hidden layer, while the interventions model

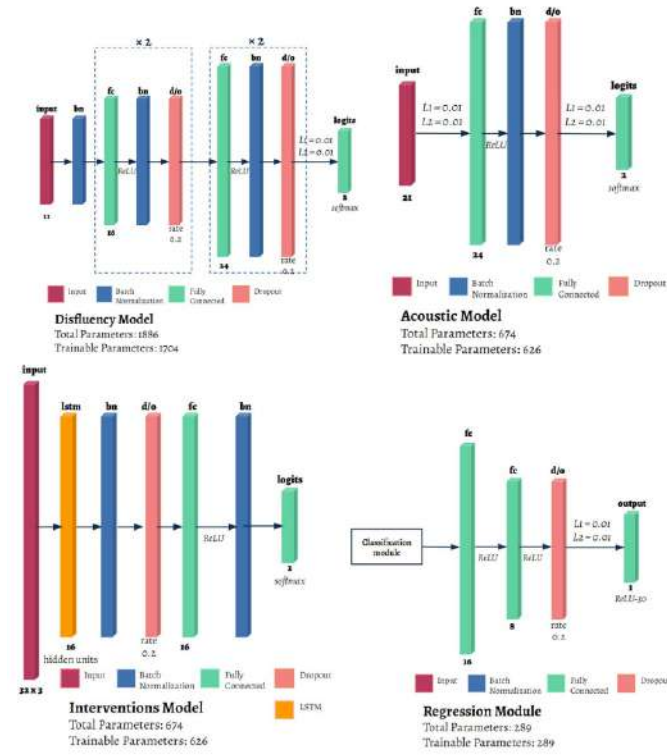


Figure 1: Acoustic, Disfluency and Intervention Model Architecture, and Regression Module

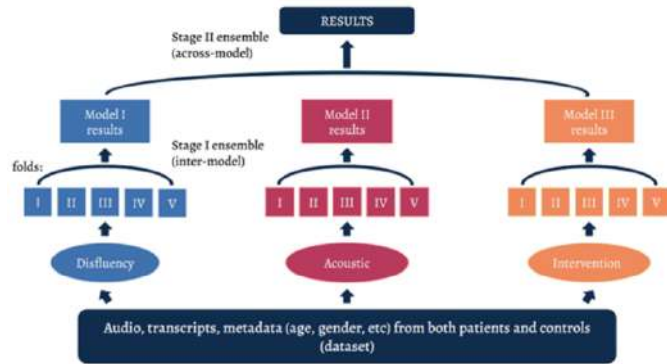


Figure 2: Two-Stage Ensemble System

uses a recurrent architecture. The fourth figure in 2 is the regression module, which replaced the previous three models' output layer during MMSE regression task.

### 3.2 Feature Selection

Figure 2 illustrates the overall architecture of our two-stage ensemble model. The first-layer of ensemble happens within each individual model: acoustic, pause and intervention, where for each model, we train different folds of sub-models, and ensemble the results for each sub-model to get a better result. The second-layer

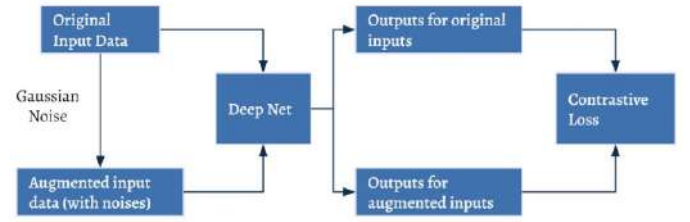


Figure 3: Contrastive Learning

of ensemble happens across different models, i.e., we ensemble the detection results based on the three different types of models to further boost the accuracy of the final results.

The reason of why we introduced this two-layer ensemble is that we want to improve the model accuracy without significantly increasing our training dataset.

### 3.3 Data Augmentation

To further improve the robustness of our AD detection network, I introduced data augmentation on top of the two-stage ensemble system as shown in Figure 2. A more robust model will make the deep learning-based AD detection algorithm more practically useful since it is inevitable that the real test data is subjected to different noises. For instance, the audios from real patients might not be as clean as our training data. Data Augmentation is a very powerful method to handle such potential discrepancy between real testing data and our training data.

In this work, we tried one type of data augmentation by injecting noise into our training data to mimic the issues with real testing data. We created random values drawn from a Gaussian distribution and added them to the original dataset so that our training data includes "contaminated" data with noise.

### 3.4 Contrastive Learning

On top of data augmentation, contrastive learning [18] is a machine learning technique to improve the robustness of machine learning models by adding extra losses to encourage the models to learn the inherent embedding of the problems instead of spurious correlations between the input and output. Thus, I introduce contrastive learning in my study.

I implemented the contrastive loss as shown in 3. First, I generated some augmented data by adding Gaussian noises on to the original input data. Then both the original data and the augmented input data are served as inputs to the deep net model which yields two sets of outputs: the outputs for the original inputs and those for the augmented ones. The contrastive loss is defined to minimize the discrepancy between the two sets of outputs, i.e., we would like to encourage the deep learning model to generate robust outputs even if the inputs were contaminated by potential noises.

## 4 RESULTS

Experiments were performed to complete two separate tasks: AD classification task, which uses a binary-classification model to distinguish between AD and non-AD patients, and MMSE prediction

**Table 1: AD Regression MMSE RMSE Across All Folds for each Individual Model**

	Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
<b>Train</b>	Compare	5.2098675	5.6776607	5.5008616	4.5624375	5.491896
	Pause	7.933954	7.082717	7.010587	7.595532	13.60928
	Intervention	8.746698	11.297575	5.2404222	8.148167	7.0735
<b>Validation</b>	Compare	4.657672	9.715276	17.862017	4.636089	4.9021597
	Pause	4.470912	4.02136	5.0452147	5.523724	6.257209
	Intervention	6.8988004	8.359658	7.2196684	5.915.67526472	6.3817735

**Table 2: AD Classification Accuracies Across All Folds for each Individual Model**

	Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
<b>Train</b>	Compare	0.882	0.847	0.871	0.859	0.847
	Pause	0.953	0.906	0.788	0.824	0.929
	Intervention	0.847	0.812	0.882	0.871	0.976
<b>Validation</b>	Compare	0.955	0.909	0.909	0.909	0.773
	Pause	0.955	0.864	0.682	0.682	1.000
	Intervention	0.682	0.591	0.682	0.773	0.773

**Table 3: AD Classification Comparison w/Baseline and Luz et. al**

Model	Accuracy	Precision	Recall	F1-Score
Luz. et al [6]	0.77	0.77	0.77	0.76
One-layer (2nd layer) ensemble (Baseline)	0.773	0.917	0.733	0.815
Two-layer ensemble (Mine)	0.955	1	0.917	0.957

task, which generates a regression model to predict the severity of AD in both groups based on the MMSE scores. The results for the classification tasks were recorded using a combination of accuracy, precision, recall and F1-score (the harmonic mean of the precision and recall).

#### 4.1 Sub-model performance

I first present the learning performance for each individual model in Table 1 and Table 2. We can see that for both the regression and classification problems, the three individual models can effectively learn the AD detection task. For instance, for AD classification, the compare model can achieve 89.1% accuracy on average, while Pause 83.7%, and Intervention 70%.

#### 4.2 Performance with the two-layer ensemble

I also implemented the two-layer ensemble approach on the dataset, and compared the results with the reference model in [6] and the algorithm with only one-layer ensemble (i.e., we only ensemble across different models).

The results for AD classification are shown in Table 3. We can see that compared to both baseline approaches, our two-layer ensemble approach can effectively improve the AD classification accuracy: the baselines' accuracy is around 77%, while my approach can achieve an accuracy as high as 95.5% and a recall as high as 91.7%.

Such a significant boosting in terms of both accuracy and recall is extremely meaningful for AD patients because it means potential patients can get early diagnosis signal.

The results for MMSE Regression are shown in Table 4. The individual features achieved competitive performance in comparison with the baseline results, as both Pause and Intervention yielded lower RMSE scores. The overall RMSE score of the two-layer ensemble also outperforms the baseline ensemble, with a RMSE that is 30.5% lower in validation. This improvement in performance reflects my system's accuracy in predicting MMSE scores.

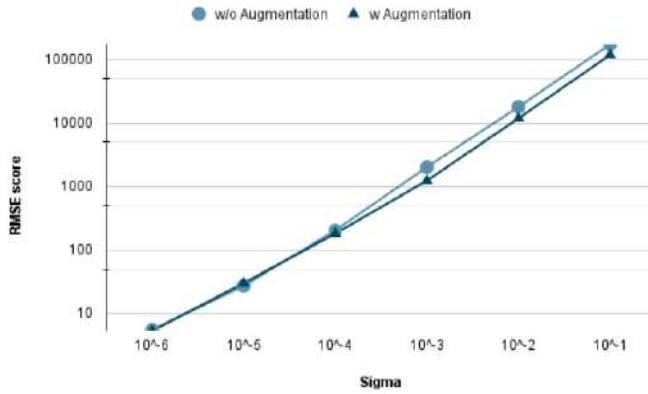
#### 4.3 Performance with the two-layer ensemble

As we all know, real patient data is inevitably contaminated by sensor noises. For instance, the audio signals can include some background noises, and the intervention signals might include some unnecessary breaks / interventions. Hence, to further improve the robustness of the AD detection algorithm, I introduced data augmentation. To be more specific, I used Gaussian noise to mimic the potential sensor noises.

In this subsection, I implemented Data Augmentation, and compared the results with and without data augmentation, as shown in Figure 4. The x-axis is the magnitude of Gaussian noise I introduced in training or validation dataset, and y-axis is the RMSE score. We can see that when there is no data augmentation during training, if

**Table 4: AD Regression MMSE RMSE Comparison w/Baseline**

Model	Compare	Pause	Intervention	Ensemble
One-layer (2nd layer) ensemble (Baseline)	6.2385836	7.798862	10.282262	7.819
Two-layer ensemble (Mine)	8.315	7.009	7.111	5.432

**Figure 4: Augmented data AD Regression RMSE score compared with Non-Augmented data**

we introduce some Gaussian noise in the validation data, the model performance will degrade significantly, which means the model is not able to generalize or not robust to sensor noise. But if I used data augmentation during training, we can see that the performance on the noisy validation dataset can be improved effectively, which means that data augmentation is very effective to make the model more robust / less sensitive to sensor noises.

#### 4.4 Performance with data augmentation and contrastive loss

Table 5 and 6 were obtained from validation results of each fold for each feature with data augmentation turned on during training. The noise level, or sigma selected is 0.1.

## 5 CONCLUSION

In this work, I have proposed to construct a robust AD detection algorithm based on deep learning with two-layer ensembling, data augmentation and contrastive loss learning. I used three different models: Acoustic, Pause, and Intervention. The two-layer ensemble structure not only ensemble across different model types, but also ensemble across different folds of the data within each model type. Beyond side, I further introduced data augmentation with Gaussian noises and contrastive loss learning to boost the robustness of the algorithm. I compared the performance of the proposed algorithm with other baseline algorithms on the ADRess challenge. The results showed that two-layer ensembling outperforms one-layer ensembling in both AD classification and MMSE regression tasks. I also verified the robustness of the proposed algorithm, and found that data augmentation and contrastive learning improves the robustness and generalization of the models.

Further work can be aimed at expanding the dataset, which would improve the robustness and generalization of the model, this can be done by extracting other acoustic features from the audio files, such as emobase and eGeMAPS, or combine the cognitive data with eye-tracking data. Other directions include the use of an automatic speech recognition system to extract reliable features, or implementing large deep nets to further improve the model accuracy.

**Table 5: AD Regression MMSE RMSE w/Data Augmentation and Contrastive Learning across all Folds for each Individual Model**

	Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
<b>Train</b>	Compare	6.6421485	6.3208413	5.7627554	6.6360235	7.8343773	6.6392292
	Pause	6.7397037	6.437809	7.002092	8.234497	7.7982354	6.0353895
	Intervention	6.786	6.307374	7.0152397	8.861489	8.16959	6.1899487

**Table 6: AD Classification w/Data Augmentation and Contrastive Learning across all Folds for each Individual Model**

	Feature	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
<b>Train</b>	Compare	0.3636364	0.3636364	0.5454546	0.6363636	0.4545455	0.4727282
	Pause	0.3636364	0.6363636	0.6363636	0.7727273	0.4545455	2.8636088
	Intervention	0.6363636	0.3636364	0.6363636	0.6363636	0.5454546	0.56363636

## ACKNOWLEDGMENTS

Words cannot express my appreciation to my instructor Dr. Sun for her invaluable patience and helpful feedback. I also would like to extend my sincere thanks to my parents who supported me throughout the entire research process.

## REFERENCES

- [1] Alzheimer's Disease International. World Alzheimer Report 2020.
- [2] 2016 alzheimer's disease facts and figures. *Alzheimer's Dementia*, 12(4):459–509, 2016.
- [3] World Health Organisation. The top 10 causes of death, 12 2020.
- [4] Juan Manuel Fernández Montenegro, Barbara Villarini, Anastassia Angelopoulou, Epaminondas Kapetanios, Jose Garcia-Rodriguez, and Vasileios Argyriou. A survey of alzheimer's disease early diagnosis methods for cognitive assessment. *Sensors*, 20(24):7292, 2020.
- [5] Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the clinical diagnosis of alzheimer disease at national institute on aging alzheimer disease centers, 2005–2010. *Journal of neuropathology and experimental neurology*, 71(4):266–273, 2012.
- [6] Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. *arXiv preprint arXiv:2004.06833*, 2020.
- [7] Aparna Balagopalan, Benjamin Eyre, Frank Rudzicz, and Jekaterina Novikova. To bert or not to bert: Comparing speech and language-based approaches for alzheimer's disease detection. *arXiv preprint arXiv:2008.01551*, 2020.
- [8] Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. *arXiv preprint arXiv:2009.00700*, 2020.
- [9] Anna Pompili, Thomas Rolland, and Alberto Abad. The inesc-id multi-modal system for the adress 2020 challenge. *arXiv preprint arXiv:2005.14646*, 2020.
- [10] Yilin Pan, Bahman Mirheidari, Markus Reuber, Annalena Venneri, Daniel Blackburn, and Heidi Christensen. Improving detection of alzheimer's disease using automatic speech recognition to identify high-quality segments for more robust feature extraction. *Proc. Interspeech 2020*, pages 4961–4965, 2020.
- [11] Raghavendra Pappagari, Jaejin Cho, Laureano Moro-Velazquez, and Najim Dehak. Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. *Proc. Interspeech 2020*, pages 2177–2181, 2020.
- [12] Thomas Searle, Zina Ibrahim, and Richard Dobson. Comparing natural language processing techniques for alzheimer's dementia prediction in spontaneous speech. *arXiv preprint arXiv:2006.07358*, 2020.
- [13] Morteza Rohanian, Julian Hough, and Matthew Purver. Multi-modal fusion with gating using audio, lexical and disfluency features for alzheimer's dementia recognition from spontaneous speech. *arXiv preprint arXiv:2106.09668*, 2021.
- [14] Jinchao Li, Jianwei Yu, Zi Ye, Simon Wong, Manwai Mak, Brian Mak, Xunying Liu, and Helen Meng. A comparative study of acoustic and linguistic features classification for alzheimer's disease detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6423–6427. IEEE, 2021.
- [15] Oswald Barral, Hyeju Jang, Sally Newton-Mason, Sheetal Shajan, Thomas Soroski, Giuseppe Carenini, Cristina Conati, and Thalia Field. Non-invasive classification of alzheimer's disease using eye tracking and language. In *Machine Learning for Healthcare Conference*, pages 813–841. PMLR, 2020.
- [16] Carole Roth. *Boston Diagnostic Aphasia Examination*, pages 428–430. Springer New York, New York, NY, 2011.
- [17] Brian MacWhinney. *The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database*, 2000.
- [18] James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. *Advances in Neural Information Processing Systems*, 26, 2013.

# RhySpeech: A Deployable Rhythmic Text-to-Speech Based on Feed-Forward Transformer for Reading Disabilities

Yixuan Lin  
YK Pao School  
sailerneeter@qq.com

## ABSTRACT

Dyslexia was first proposed in 1877, but this century-old problem still troubles many people today [1]. Dyslexia is marked by difficulty in reading despite having normal or superior conditions in their environment and intellectual ability, is curable using multi-sensory learning, which involves providing audio stimulus, sometimes generated from expressive text-to-speech. However, such generated audio lacks rhythmic features, marked by inadequate insertion of pauses. In response to such technological difficulty, this paper proposes RhySpeech, which models rhythm using feed-forward transformer neural networks and an LRV (Latent Rhythm Vector). The LRV receives input from the pitch, energy, and duration features encoded using a Transformers network along with the numeric encoding of the previous 16 phonemes, which together build a strong sense of context for the pause prediction. This LRV is trained to generate adequate lengths and positions of pa uses, allowing the synthesized audio to have more accurate pausing

## CCS CONCEPTS

• **Applied computing** → Life and medical sciences; Bioinformatics.

## KEYWORDS

Text-to-Speech, Deployable, Rhythmic, Transformer, Latent Vector

### ACM Reference Format:

Yixuan Lin. 2023. RhySpeech: A Deployable Rhythmic Text-to-Speech Based on Feed-Forward Transformer for Reading Disabilities. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3590003.3590062>

## 1 INTRODUCTION

Human speech consists of sentences, units of speech representing relatively complete and independent meaning; which can be broken down into words, which have their definition, and humans tend to pause between words. By definition, a text-to-speech process is a digital process that takes text as input, then, through multiple steps, converts the text into sound waves (which resemble human speech) which are outputted. Sometimes, the term "end-to-end" is

added as a modifier to emphasize that the system is responsible for the entire process of text-to-speech.

Modern end-to-end text-to-speech systems consist of three steps (which text goes through to become audio) as shown in figure: text analysis, acoustic modeling, and the vocoder, which will be explained below in figure 1.

Text analysis involves converting input text (to-be-read) into a format that the acoustic model can accept as input. Usually, this step involves the conversion of numbers and symbols, which exists in some texts such as textbooks and science books, into their word form (which is what words will be spoken if a person is to read that number or symbol out loud). However, the spelling of a word doesn't carry a one-to-one relationship about how the word should be pronounced. For example, the "h" grapheme in "hill" is pronounced with the "h" sound, but the same "h" grapheme in "honor" is not pronounced. To solve this issue, a grapheme-to-phoneme (G2P) conversion is carried out. Either a G2P dictionary or a G2P conversion model is used for the process.

The acoustic model is the part of text-to-speech that converts phonemes to acoustic features, or indicators of how text is to be pronounced, complete with details such as volume and sound frequency. Variance encoders, which are popular approaches to capturing key features of pronouncing phonemes, are situated in this step. Variance encoders operate by extracting a specific feature of audio from the training set that varies with the phoneme involved (such as the volume of speech which can be extracted from the amplitude of the sound wave) and training the feature of the sound wave in association to the corresponding phoneme. Variance encoders are usually trained using machine learning, where their parameters will adjust according to training data. An alternative approach to modeling core characteristics of audio is GAN, which is short for generative adversarial networks. A GAN involves a generator trying to generate audio as close to a human voice as possible, while the discriminator tries to tell apart the generator's artificial audio from a real human's speech. Through training, the generator and discriminator both improve in abilities, thus generating increasingly realistic speech audio.

The vocoder is the final part of text-to-speech that converts acoustic features to sound waves, what humans perceive as speech. Modern vocoders tend to use neural networks as the approach to speech synthesis.

Recent studies on text-to-speech can be divided into autoregressive and non-autoregressive approaches. In the autoregressive approach, speech generation models utilize computational results and byproducts as a basis for future computation, therefore requiring the model to construct the final speech audio linearly, from the start to the end of the text for which the speech needs to be generated. The inability to perform generative computations in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590062>



Figure 1: A pipeline adumbrating the general text-to-speech process.

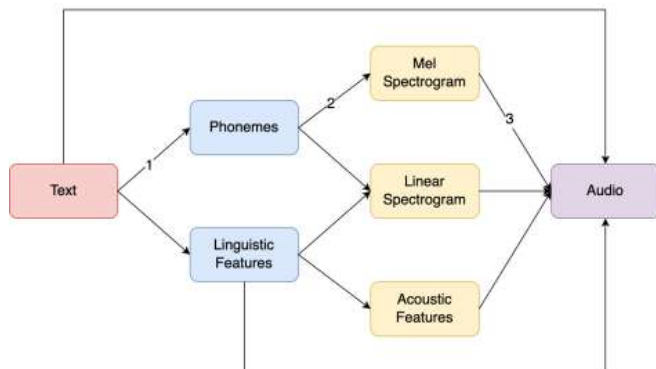


Figure 2: Variations in the process from text to speech [2]. Note that the pathway from the red to blue boxes is the phoneme encoding process, from the blue to the yellow boxes is the acoustic model, and from the blue to the purple box is the vocoder. The research follows a pathway labeled 1, 2, and 3 (from text to phonemes, to meet spectrogram, and finally to audio).

parallel impairs its processing speed, which is a significant letdown for text-to-speech applications where responses must be generated in very short times. In an attempt to reduce the computation time, the non-autoregressive approach does not rely on previous computations to make future computations, instead using other methods to ensure the quality of the generated speech, allowing parallel construction of speech audio, thus decreasing the generative latency of speech. The high potential speeds that non-autoregressive models process makes it the center of attention for researchers, believing that the non-autoregressive approaches have large potential for development. Due to the advantage of fast processing speeds, this research also uses the non-autoregressive approach.

Also, approaches to the text-to-speech task can be classified according to intermediate steps of speech synthesis. As shown in Figure 2, each approach selectively passes through steps of acoustic and linguistic feature analysis. This research follows a pathway from text to phonemes, mel spectrogram, and finally to audio.

Upon the problem of speech synthesis systems having inadequate rhythmic prediction capabilities under a small model size, this research proposes RhySpeech, a non-auto-regressive text-to-speech system based on Fastspeech2. RhySpeech specially features a pause predictor based on the transformer architecture of the audio’s pitch and energy values. The training of the system is enhanced by RPSigmoid, a trainable activation function. The contributions of this research can be summarized as two main points:

1. Pause prediction via transformer architecture’s input
2. RPSigmoid, a randomly initialized and parametric activation function based on the sigmoidal curve in figure 2.

## 2 METHOD

The overall process for text-to-speech in this research will be introduced in the section. Figure 3 shows the process of the inference stage when text is converted to speech, and Figure 4 shows the process of training.

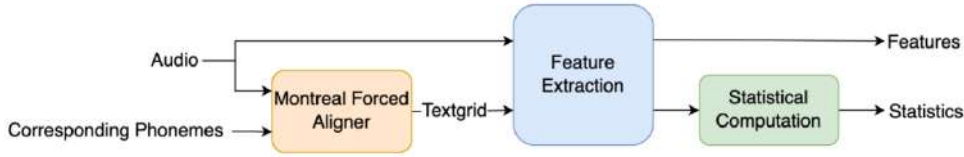
### 2.1 Input Text Processing

The processing of inputted text involves breaking it down into individual phonemes. For this research, this is done using cmudict, which creates a long dictionary of words and their respective phoneme sequence, which is an ordered set of smallest units of distinct sound. Each phoneme sequence is described with a space inserted between each phoneme, and the vowels are marked with numbers to indicate the amount of stress that should be applied to them: 0 means there should be no stress applied, 1 means there should be primary (strongest) stress applied, and 2 means there should be secondary (next-strongest) stress applied. Note that the phonemes words are broken up into can be spelled differently from the original word since different spellings can have the same sound in English, thus making them all belong to the same phoneme. The same word, if labeled with 1 and 2 beside them, can be pronounced differently, such as presents primary stress is on the first “e” if the word is a noun, and on the second “e” if the word is a verb.

The processing of inputted text involves breaking it down into individual phonemes. For this research, this is done using cmudict, which creates a long dictionary of words and their respective phoneme sequence, which is an ordered set of smallest units of distinct sound. Each phoneme sequence is described with a space inserted between each phoneme, and the vowels are marked with numbers to indicate the amount of stress that should be applied to them: 0 means there should be no stress applied, 1 means there should be primary (strongest) stress applied, and 2 means there should be secondary (next-strongest) stress applied. Note that the phonemes words are broken up into can be spelled differently from the original word since different spellings can have the same sound in English, thus making them all belong to the same phoneme. The same word, if labeled with 1 and 2 beside them, can be pronounced differently, such as presents primary stress is on the first ‘e’ if the word is a noun, and on the second ‘e’ if the word is a verb.

Next, the phonemes are encoded into vectors for ease of storage because the storage of character sequences as characteristics of words is memory intensive and unable to be handled by neural networks, which only accept numerical inputs.

The last step of the preprocessing of data is to reduce the dimensionality of the vector. Aiming to use the available computational resource wisely, the vectors are run through an encoder algorithm, outputting more space-efficient vectors.



**Figure 3:** This diagram shows how the audio samples are preprocessed to obtain the features which are used as the ground truth by the acoustic model. Corresponding phonemes are also inputted to assist with the process of ground-truth-extraction.

**Table 1:** A selection of words from cmudict. The words are broken up into phoneme sequences. Screenshot from cmudict, extracted from <https://raw.githubusercontent.com/cmuphinx/cmudict/master/cmudict.dict>

Word (grapheme)	Phoneme (individual sounds that make up the word)
Ablate	AH2 B L EY1 T
Ablation	AH2 B L EY1 SH AH0 N
Ablaze	AH0 B L EY1 Z
Able-Bodied	EY1 B AH0 L B AA1 D IY0 D
Abled	EY1 B AH0 L D
Abler	EY1 B AH0 L ER0
Abler(2)	EY B L ER0

## 2.2 Audio Preprocessing

The audio preprocessing stage obtains the ground truth values for the later acoustic model by preparing and extracting useful information from the speech recording, which are the attributes of pitch, duration, and energy, along with the mel spectrogram. A diagram for the audio preprocessing process is included in Figure 5.

The inputted audio recorded from an English speaker is fed into the system as .wav files, where the sound wave is saved as floating point numbers, indicating the height of the wave at that instant.

To accommodate readers with color vision differences, figures should still be usable when printed in grayscale. Refer to elements of the figure with non-color terms, for example “indicated as squares” instead of “indicated in blue”. Use different patterns in bar charts, different line patterns in graphs, and different shapes in plots to distinguish groups of elements and reinforce color differences.

**2.2.1 Montreal Forced Aligner.** The audio waveform along with the corresponding phonemes (produced by the input text processing) are passed into the Montreal Forced Aligner (MFA) [3], which marks the duration of each phoneme by referring to the audio waveform.

The MFA is an open-source model pretrained on the Speech dataset (with approximately 1000 hours of English speech). Pre-training follows a regime of 40 iterations of monophonic training, which by considering phonemes one at a time in alignment, then 35 iterations of triphone training, which considers one phoneme before and after the to-be-aligned phoneme. The monophonic training phrase aims to offer the model general alignment capabilities, and the triphone training fine-tunes the model by requiring it to

consider the context. Realignment is performed on 20 (out of 40) iterations during monophonic training and 15 (out of 35) during triphone training, allowing the model to improve upon itself.

On a closer scale, the MFA model is trained with the acoustic feature of mel-frequency cepstral coefficients (MFCCs), which represent the intensity of different meet frequencies (frequencies with equal perceptual distance to humans, which puts it roughly on a log scale). The MFCCs are then put through Cepstral Mean and Variance Normalization (CMVN) to obtain a zero-median and unit-standard-deviation distribution to improve robustness.

**2.2.2 Feature Extraction.** The audio information is also passed along with the textgrid into the feature extraction, which finds certain characteristics of the audio inputted.

The first component of the extraction process is the mel spectrogram constructor. A mel spectrogram is an image that can represent the features of the audio. Its x-axis denotes the time of the audio, its y-axis is the log of the human-perceive frequency, and the color of a given pixel is rendered based on the intensity of that specific frequency at that specific timing.

To perform this construction, Short-Term Fourier Transform (STFT) is first applied to the audio recordings. The transformation process involves breaking the audio recording into short frames of around 20 to 50 milliseconds. The process only works through these short frames because human speech, on such a small scale, is relatively regular and predictable. Short-Term Fourier Transform utilizes this regularity to break the complex sound wave in each frame into compositions of sinusoidal waves at varying frequencies and varying amplitudes (intensities). Algebraically, this is represented as:

$$fk = \frac{1}{2\pi} \int dx f(x) e^{-ikx}$$

where  $f$  is the result for  $x$ 's Fourier Transform and  $k$  is the wavenumber.

To tackle the human pitch-determining inaccuracy, where humans typically perceive sounds higher than 1000 Hz lower than their actual pitch, a mel spectrogram adjusts for this hearing subjectivity, making it reflect what humans actually perceive.

The pitch is computed using the dio tool from Pyworld, which receives the .wav audio file as input and output the fundamental frequency (the loudest pitch at a specific time, which is the one that listeners hear and interpret) of audio at a specific time.

The energy is computed alongside the mel spectrogram when the Short-Term Fourier Transform represents the audio in an instant as differing intensities of differing frequencies. The energy captures this intensity.

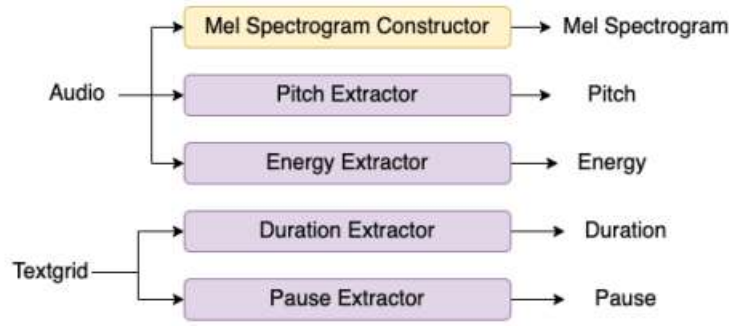


Figure 4: This diagram shows the extraction process used to create the ground truth predictions from the textgrid and the audio.

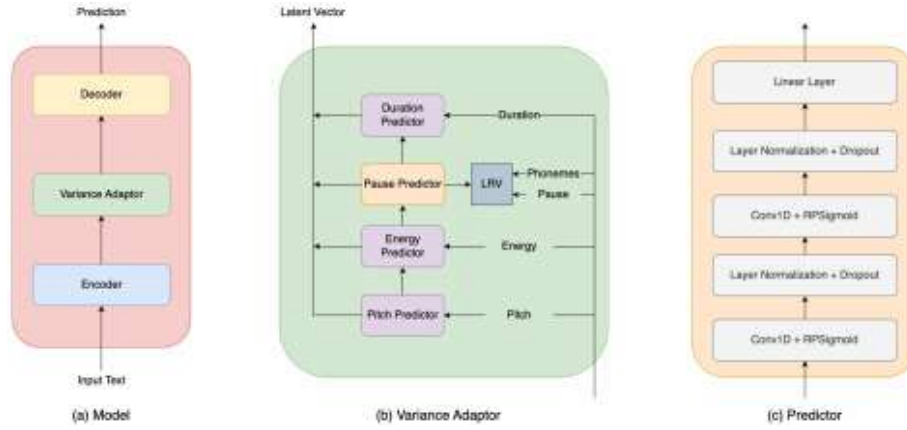


Figure 5: This diagram shows the acoustic model and its components used in this research

The duration, along with the usage of pauses in human speech, is extracted from the textgrid. Since the textgrid holds the start and end time of each phoneme and silence, such time values can directly be used as the duration and pause values. Specifically, if the timings on the textgrid correspond to an uttered phoneme, that timing length is labeled to the duration of the corresponding phoneme; if the timings correspond to silence, that timing length is labeled to the pause after the phoneme before it. If there is no silence after a phoneme, the pause information for that phoneme is labeled as 0. The last phoneme's pausing is also labeled as 0 since there is no need to save silence after the last phoneme because that silence doesn't improve audio quality but increases the file size in figure 5.

**2.2.3 Statistical Computation.** The values for pitch and energy are standardized using the StandardScalar() of scikit-learn [4]. First, the mean and standard deviation (measuring how spread-out the data is, relative to the center of it) are computed using built-in functions of the StandardScalar. Note that the standard deviation can be computed using the equation.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

where  $\sigma$  represents the standard deviation of the data,  $x_i$  is the  $i$ th value in the data,  $\mu$  is the mean of the data, and  $N$  is the number of elements in the data.

Then, the values of pitch and energy are transformed (using specific operations while not distorting the relative data) into a normal distribution: a median of 0, and a standard deviation of 1. This is achieved through a formula of

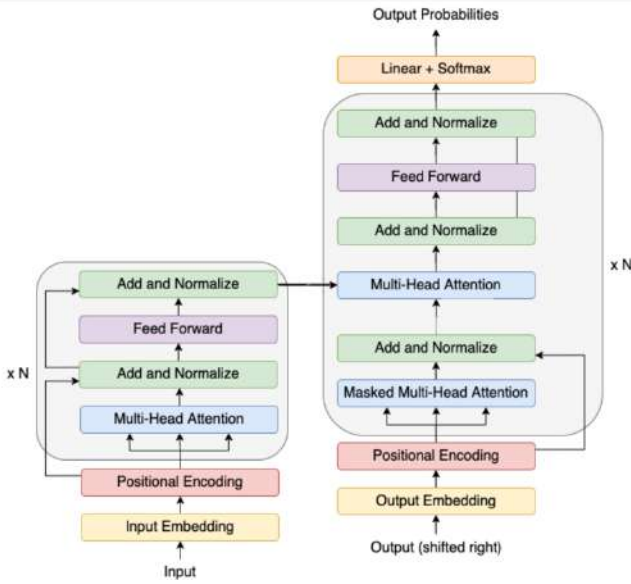
$$z = \frac{x - u}{s}$$

where  $z$  is the output of the standard scalar,  $x$  is the input,  $u$  is the mean, and  $s$  is the standard deviation. This allows for pitch and energy to be compared "equally", so that the influence of the exact value of pitch and energy, confounding variables, are removed.

Additionally, the minimum and maximum values for pitch and energy are computed. This would serve the purpose of establishing bounds in later stages of the research in functions such as rendering mel spectrograms.

## 2.3 Acoustic Model

Aiming to introduce abilities of expression and naturalness to the generated audio, this research project holds the ability for the audio-generation system to learn duration, energy, and pitch by building on Fastspeech 2; and additionally allowing the system to learn the rhythm. The framework for the acoustic model is shown in Figure 5



**Figure 6: This diagram shows the structure of the transformer neural network**

**2.3.1 Transformer Architecture.** The acoustic model uses a Transformer neural network architecture [5], as shown in Figure 6. The usage of a Transformer network allows the acoustic model to predict of features in multiple levels of focus, helping the vocoder access more detailed and accurate information about the acoustic features of the audio to be generated.

The Transformer structure includes an encoder, a variance adaptor (the name comes from its ability to learn the differing nuances in speech audio), and a decoder. The encoder is responsible for converting information into vector form useful for the future. The variance adaptor learns the variance present in speech. The decoder converts information from the encoded vectors back into normal values

Note that there could be multiple layers of encoder and decoder components to increase the model’s capabilities. This means multiple degrees of content can be reframed into vectors, allowing the variance adaptor to receive better quality information.

The encoder and decoder share a similar architecture, which is the Transformer architecture, as shown in Figure 6.

To start with, the vectors entering the Transformer neural network are embedded, or converted, into vectors. This is because neural networks can only deal with groups of numbers (namely, vectors).

Afterward, positional encoding allows each element in the group of data to have a unique real number to indicate its position in the dataset. This allows for position to be effectively represented and considered in later steps, which is useful for time-dependent data analysis. The positional encoding process uses the formula of

$$PE \parallel \text{mod}_2(n) = 0 : \sin\left(\frac{n}{1000^{\frac{2}{d}}}\right)$$

$$PE \parallel \text{mod}_2(n) = 1 : \cos\left(\frac{n-1}{1000^{\frac{2}{d}}}\right)$$

where PE represents the result of positional encoding,  $n$  is the index of the to-be-positionally-encoded data, and  $d$  is the dimensions in the vector.

Such a method of positional encoding brings the benefit of not belittling other encoded data (if using the index to encode, the index might get much larger than other non-positional-encoding values, therefore making the model only focus on positional encoding values). This method also attributes similar weighting to the position regardless of data vector size, so that the difference of one element apart sets the positional encoding values to a similar difference.

Afterward, a multi-head attention operation is applied to the values passed along, regardless of coming from the “input” or “output” channel. Attention networks, in general, have the formula of

$$y = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)$$

where  $y$  is the output of the attention network,  $Q$ ,  $K$ , and  $V$  are vectors from the data, and  $d_k$  is the number of dimensions of the data vector  $K$ . The division by  $\sqrt{d_k}$  prevents dimensionality from changing the variance in the data, removing dimensionality as a confounding variable in the analysis. In multi-head attention,  $Q, K, V$  come from different places.

For the “output” channel, values undergo a masking process before this multi-head attention process. Iterating through the data, for every element, those elements later than it will be hidden (or masked), so that during the later attention mechanism, any later elements cannot be referenced (which is the case in real-world applications since results generated earlier do not know what results will be generated later).

After the multi-head attention, the original value of the positional encoding is combined with the result of the multi-head attention to allow the positional information to be retained in the data. Then, batch normalization (represented as “normalization” in the diagram) is performed to put the data into a zero-mean, unit-standard-deviation state, which enhances stability in the neural network.

For the “input” channel, the value then passes through a feed-forward fully-connected neural network, and that is added to results before the fully-connected network, and finally normalized. This result represents the encoding from the “input” channel. Such a process can be performed multiple times, and the number of times it is repeated is equal to the number of encoder layers.

The result (of one or more repeats) is received by the “output” channel and joined at a multi-head attention block from the previous data in the “output” channel. Again, the adding of values and batch normalization is performed, then another fully connected neural network, another adding and batch normalization is performed. Such a process starting from the masked multi-head attention to the adding and batch normalization is performed for some repetitions, as represented by the layers of the decoder.

Finally, the result of the “output” channel is computed over a single fully-connected layer of a neural network, and softmax is applied to it. Softmax is an activation function that ensures the sum of all neuron outputs in the softmax layer of the neural network

adds up to one. In other words, this output of one is split up between different neurons based on the size of their input to the softmax function. Since probabilities add up to one, softmax makes an effective activation function when predicting output probabilities, which is the result to be outputted.

**2.3.2 Rhythm Prediction.** The main innovative point of this research is its ability to effectively guide the acoustic model to generate rhythmic speech. The embedding of pitch and energy is passed to the rhythm predictor, along with the 16 letters prior to the phoneme to predict pausing for better context (if the phoneme doesn't have 16 letters before it, -1 is used as a placeholder), to check whether the previous embedding is suggestive of appropriate pausing in the given context, and if not, such error is back-propagated on the loss function. Through training, the acoustic model will eventually generate audio with appropriate pausing.

This step is done after the pitch and energy prediction because pausing length is highly dependent on what the text is trying to express, which can be modelled to some extent by the pitch and energy predictors. This step is done before the duration prediction because the embedding of duration changes the dimensionality of the embedding vector, which wouldn't match the phoneme-by-phoneme requirement for the pause predictor.

**2.3.3 Variance Predictor.** The variance predictor receives a latent vector, which is passed through conv1d (one-dimensional convoluted neural network), RPSigmoid, layer normalization [6], dropout layer [7], conv1d, RPSigmoid, layer normalization, and dropout layer in this order and setup is used. Note that RPSigmoid is a new and trainable activation function. RPSigmoid takes the formula of

$$y = \frac{abx}{(1 + b|x|^c)^{\frac{1}{c}}} + dx$$

where a,b,c,d are trainable parameters randomized at the start of the training process.

The usage of convolutional neural networks ensures a low count of parameters while achieving decent quality in synthesized audio. The usage of layer normalization brings the benefits of training stability and faster convergence by ensuring different layers treat data that is approximately within the same magnitude. The usage of a dropout layer prevents the model from overfitting by changing some neurons' output to zero, introducing variation, and forcing the model to learn the patterns despite such perturbations (instead of just memorizing the correspondences of the previous layer's output and the correct answer).

There are two types of math equations: the *numbered display math equation* and the *un-numbered display math equation*. Below are examples of both.

## 2.4 Vocoder

The acoustic model can directly output the mel spectrogram, but multi-sensory learning requires the use of audio instead of mel spectrogram to stimulate dyslexic learners. To facilitate the transformation from mel spectrograms to audible waveforms, a vocoder is used. This research uses the HiFi-GAN (short for High-Fidelity Generative Adversarial Network) vocoder [8], the current state-of-art approach. HiFi-GAN is a pretrained model, pretrained using one generator rebuilding audio from mel spectrograms and two types

of discriminators operating on multi-period and multi-scale bases. Each multi-period discriminators only accepts input from certain discrete time points (multiples of a certain prime number), thus having each discriminator focus on different parts of the audio. There are three individual multi-scale discriminators, each focusing on a different scale: the original data, data that has undergone 2x pooling, and data that has undergone 4x pooling. The usage of multi-scale discriminators introduces the ability to distinguish continuity to the discriminator group since the multi-period discriminators each look at discrete data points.

The HiFi-GAN model is pretrained on LJSpeech, which contains around 24 hours of a single English speaker. HiFi-GAN is built to improve using three losses: GAN (as performed by the aforementioned discriminators), mel-spectrogram (which reconstructs the mel-spectrogram from the produced audio), and feature-matching (a comparison between the generated and ground truth audio samples). The weighting of these losses in the final training process is 1 to 2 to 45 respectively.

## 3 EXPERIMENTAL SETUP

The experiment to verify the effectiveness of the aforementioned method will be described below. First, the usage of the dataset will be described, which is followed by how the aforementioned method is carried out, and finally a summary of how the quality of the text-to-speech process in this research is evaluated.

### 3.1 Dataset

This experiment tests the model using the LJSpeech dataset. This dataset consists of around 24 hours of speech performed by one female English speaker. The speech consists of the reading of seven non-fiction books therefore the tone is more informative and less dramatic. The 24 hours of speech is broken up into 13100 audio clips, each within 10 seconds of length. This dataset is around 2.6 GB large.

This dataset is selected because it is one of the most commonly-used datasets in text-to-speech modeling. Moreover, the reasonable size of the dataset (if compared to other datasets such as LibriSpeech) makes it a viable option since the computational capacity is limited.

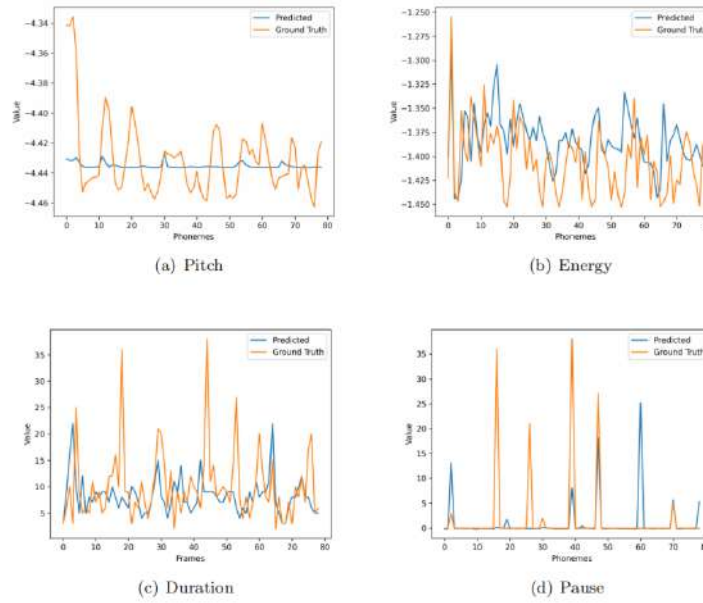
### 3.2 Training on the Dataset

A NVIDIA Tesla T4 GPU equipped with 15 GB of memory was used for this experiment. The proposed method uses a batch size of 16 along with an epsilon of  $10^{-9}$ , an optimizer of Adam, betas of 0.9 and 0.8, 2000 steps for warm-up, and trained for a total of 10,000 steps.

### 3.3 Evaluating the System

For this text-to-speech system, which features high expressiveness in pausing, the following evaluative metrics will be used.

Mean Opinion Scale (MOS): the subjective evaluation of listeners upon the quality of the model- synthesized audio. Twelve listeners (six female and six male) with fluent English capabilities are asked to listen to the generated audio. Each listener is asked to give a score of 1 (the worst, where the audio is unintelligible) to 5 (the audio sounds perfect) to each piece of audio they hear. To ensure fairness,



**Figure 7: The comparison between the ground truth and the predicted values for pitch, energy, duration, and pause features on unseen data.**

each listener will listen to audio generated from both FastSpeech2 (the baseline) and RhySpeech without knowing the audio-model correspondences (unbiased judgment).

Pause Prediction Effectiveness Statistics: accuracy (the fraction of all entities that are judged correctly), precision (the fraction of those judged positive (paused) that are supposed to be positive), recall (the fraction of those that should be positive (paused) that are judged positive), and F1 values that reflect the effectiveness of the prediction of pausing.

Loss-by-EPOCH: this is specifically a comparison to show whether the activation function (its purpose is to aid faster convergence) is effective in helping the model converge.

Model Parameter Count: this is a metric used in reference with the MOS (quality of prediction) to determine whether the acoustic model of this research is deployable, which would require a comparatively small parameter count.

## 4 RESULTS

After introducing the methods of this research, results will be given to justify the validity of the aforementioned design. The results section will begin by presenting the general outputs of this study. Afterward, changes in synthesized vocal quality after the addition of the pause predictor will be presented. Finally, the usage of the RPSigmoid activation function will be justified.

### 4.1 General Results

The RhySpeech model was able to achieve a MOS value of 3.87, which means that the audio quality is decent. To offer a context of comparison, the MOS value of the baseline is only 3.73, as shown in

**Table 3: The comparison of parameter count (fewer means requires less storage space, therefore is better) between the baseline (FastSpeech2) and RhySpeech.**

Model Name	Parameter Count
FastSpeech2 (baseline)	43.8M
RhySpeech	41.4M

Table 2. At first glance, it is already evident of RhySpeech’s superior audio quality.

To visualize the high quality of generated audio, line graphs demonstrating values of vocal characteristics (pitch, energy, duration, and pause) will be presented for the ground truth and the generated results of RhySpeech in Figure 7. Note that RhySpeech has not trained on these pieces of audio in advance.

The similarity between the ground truth and the generation of RhySpeech shows that RhySpeech is strong at modeling speech characteristics even in unseen scenarios as shown in Figure 7.

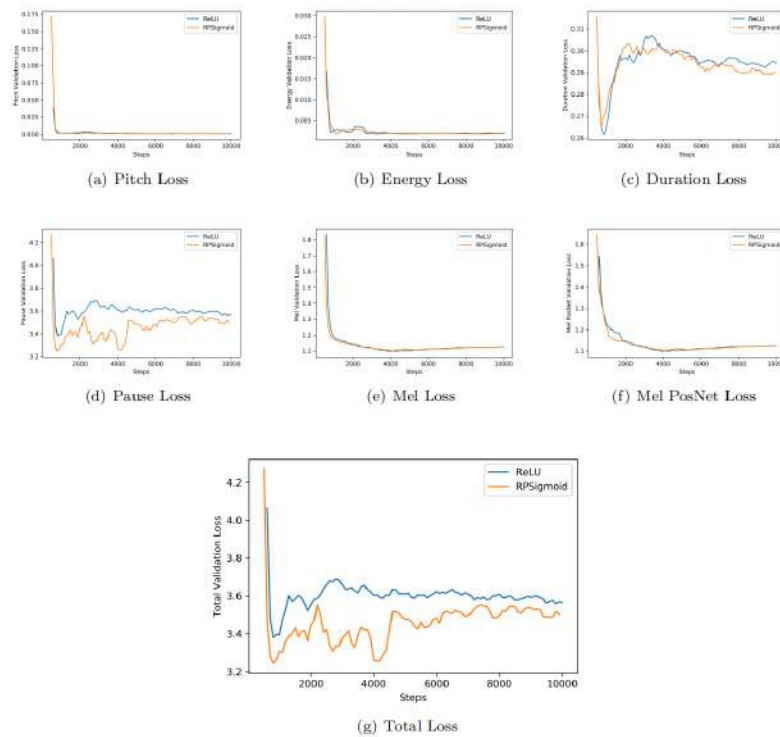
Since the scenario requires a relatively lightweight model for implementation, the parameter count is also presented in Table 3. From the smaller model size of RhySpeech, it is shown that RhySpeech is comparatively small, and therefore more usable in resource-restricted scenarios.

### 4.2 Pause Predictor Results

To prove that pausing has been effectively captured using the pause predictor, statistical comparisons between the baseline (FastSpeech2) and RhySpeech regarding whether a pause has been predicted will be made in Table 4. The comparison is done based on

**Table 4: The statistical quantities that reflect the effectiveness of Fastspeech2 and RhySpeech at predicting the existence of pausing when no punctuation information is available. Note that Fastspeech2 is unable to predict pausing when not given punctuation so no pausing has been predicted, leading to the presence of NaN values. The accuracy for RhySpeech is lower than Fastspeech2 due to the sparseness of pausing (less than 7 percent of the data used to generate the statistics are supposed to be where pausing is present).**

Quantity	Baseline (Fastspeech2)	RhySpeech (proposed method)
Accuracy	0.957	0.948
Precision	NaN	0.679
Recall	0	0.430
F1	NaN	0.527



**Figure 8: The loss-per-epoch curves for various parts of the text-to-speech system and the total loss comparing ReLU and RPSigmoid**

the LJSpeech validation set (which neither model has been trained on before).

### 4.3 RPSigmoid Results

To justify RPSigmoid’s effectiveness as an activation function in this model, its effects of convergence are compared with the state-of-art activation function in text-to-speech fields: ReLU. In Figure 8.

### 4.4 Discussion

In conjunction with the successes in this research, there are also potential directions to research into in the future. Firstly, an effort can be made to enhance stability of the quality of generated audio, ensuring that when this text-to-speech system is deployed, people

with dyslexia can enjoy a reliable auditory stimulus. Additionally, an approach of taking pausing into the embedding could be attempted, in order to make the pausing prediction directly impact the audio, creating even more natural rhythmic patterns in synthesized speech. Lastly, a user-friendly system can be developed on top of the theoretical basis introduced in this paper, allowing for dyslexic people to access auditory stimulus more conveniently

## 5 CONCLUSION

This paper proposes a component to model rhythmic information in the human speech by predicting pausing under the Transformer architecture based on pitch and energy embedding. The pausing prediction generates a loss which is used to improve the acoustic

model to produce more accurate pausing patterns. To achieve desirable results under a smaller parameter count, this pause predictor is coupled with RPSigmoid (Randomized and Parameterized Sigmoid) as its activation function. Such a method has resulted in an increase in MOS and pausing recall compared to FastSpeech2, while enjoying the convenience of less parameters than FastSpeech 2.

## REFERENCES

- [1] Philip Kirby. Dyslexia debated, then and now: A historical perspective on the dyslexia debate. *Oxford Review of Education*, 46(4):472–486, 2020.
- [2] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.
- [3] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kald. In *Interspeech*, volume 2017, pages 498–502, 2017.
- [4] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.
- [7] Nitish Srivastava. Improving neural networks with dropout. *University of Toronto*, 182(566):7, 2013.
- [8] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

# Research on Natural Scene Vehicle Nameplate Text Detection Based on Improved DBNet

Du Yucheng, Dong Jinsong

Research Institute of Highway Ministry of Transport, Beijing 100088, China

## ABSTRACT

Vehicle nameplate information as the main content of vehicle test, it is an important guarantee for the test quality of automobile testing institutions, and an important basis for the transportation authorities to determine the consistency of vehicle parameter configuration. Aiming at the problems of diverse text distribution, variable scale and complex background in vehicle nameplate detection, this paper proposes a dense connection and feature enhancement based on differentiable Binarization (DBNet) semantic segmentation algorithm. This algorithm uses the Dense Atrous Spatial Pyramid Pooling (DASPP) module to establish the connection between multiple dilated convolutions, capture dense sampling point pixels, and improve the utilization of high-level feature information. Secondly, the Feature Pyramid Enhancement Module (FPEM) is used to enhance the expression ability of the multi-layer feature information output from the backbone network, and the Feature Fusion Module (FFM) is used to fuse the feature information of different scales output from the FPEM, which improves the complementary ability between the features of each layer and obtains more comprehensive feature map information. Finally, the output of the DASPP and the FFM are concatenated to get the final segmentation results. The experimental results show that the improved algorithm can effectively locate the nameplate text area in the complex background. The detection accuracy on the self-defined datasets reaches 90.4 %, which is 2.6 % higher than the original algorithm DBNet.

## CCS CONCEPTS

• Computing methodologies; • Machine learning; • Machine learning approaches; • Neural networks;

## KEYWORDS

Vehicle nameplate, Vehicle test, DBNet, Dense Atrous Spatial Pyramid Pooling, Feature Pyramid Enhancement Module, Feature Fusion Module

### ACM Reference Format:

Du Yucheng, Dong Jinsong. 2023. Research on Natural Scene Vehicle Nameplate Text Detection Based on Improved DBNet. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590064>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org). CACML 2023, March 17–19, 2023, Shanghai, China

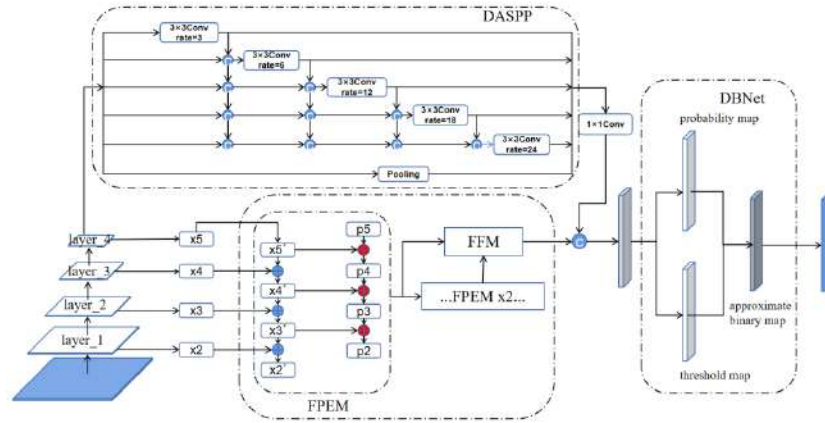
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-9944-9/23/03...\$15.00 <https://doi.org/10.1145/3590003.3590064>

## 1 INTRODUCTION

Parameter consistency is an important part of vehicle test. The automobile production enterprise is the declaration subject of the vehicle model, and is responsible for submitting the design parameters and technical documents of the vehicle model; the testing institution is the main body of the implementation of testing activities, which checks and tests the sample vehicles provided by the enterprise; as the technical support unit of road transport vehicle management, the Highway Research Institute of the Ministry of Transport carries out technical review according to standards and regulations. Among them, test is one of the key links in vehicle management. In this link, enterprises need to provide key technical parameters, sample vehicles and product inspection plans to the testing institutions, the testing institutions need to check the consistency of parameters with the sample vehicles and record them to the inspection report at the same time, so as to provide evidence materials for the technical support unit.

The testing institutions should check the basic information, structure layout, vehicle configuration, safety protection and other parameter configuration of the sample vehicle. Among them, the vehicle nameplate is the main information source for verification, including the whole car nameplate, engine nameplate, chassis nameplate, interchangeability nameplate, etc., recording the key data of the vehicle model, vehicle identification number, chassis model, engine model and vehicle performance design parameters. The traditional method is to manually copy the vehicle nameplate data, or check the consistency between the vehicle nameplate and the vehicle declaration information. It is a simple and repetitive but important work, which relies heavily on manual identification, and the accuracy of manual identification in complex scenes is limited. Therefore, the study of automatic detection and recognition of vehicle nameplate text can achieve fast and efficient entry of nameplate information.

The traditional Optical Character Recognition (OCR) technology has been able to accurately identify characters from document scans, but the text font size, color, density, alignment direction, and contrast in natural scenes are very different [1]. The vehicle nameplate text is one of the natural scene text images, which is more complex than the character background in the document scan. The characters are often submerged by the cluttered background and cannot be extracted. The material of the commercial vehicle nameplate itself is mostly metal, and there may be interference reflections in the nameplate image. In addition to the harsh environment, the metal nameplate image often has degradation conditions such as stains, blurs, and scratches [2]. In addition, the nameplate images taken in natural scenes cannot guarantee a fixed shooting angle and horizontal direction, which brings many difficulties to the automatic recognition of nameplates, and the difficulty is much higher than that of traditional optical character recognition. Therefore, it



**Figure 1: Vehicle nameplate detection network structure based on dense connection and feature enhancement**

is necessary to adopt a recognition method with strong environmental adaptability and anti-interference. Deep learning is a new development of artificial neural network model. This method of automatic learning to obtain features [3] is especially suitable for text image detection in complex background.

The most important task of vehicle nameplate text detection is to locate the position of the text in the nameplate image. In the absence of open source datasets, this paper manually collects and labels vehicle nameplate image datasets in natural scenes. Aiming at the problem of poor detection effect and omission of text detection in multi-scale dense text, this paper proposes a semantic segmentation algorithm with dense connection and feature enhancement based on DBNet scene text detection network. A large number of experiments are carried out on the self-built datasets. The results show that the accuracy of the proposed method is greatly improved compared with the original DBNet algorithm.

#### Network algorithm

### 1.1 Network algorithm structure

In this paper, a semantic segmentation algorithm based on dense connection and feature enhancement is designed in text detection network DBNet. The network structure is shown in Figure 1. It mainly includes three parts: The first part is the basic backbone network. ResNet18 is used as the backbone network to extract the output of  $x_2$  to  $x_5$  size features from layer \_ 1 to layer \_ 4. The second part is used the DASPP module to establish the connection between multiple dilated convolutions. The output of the previous dilated convolution and the input features are concatenated as the input of the next dilated convolution, which enhances the semantic association between local information. the FPEM module is used to enhance the expression ability of the multi-layer feature information output from the backbone network, and FFM module is used to fuse the feature information of different scales output from the FPEM, which improves the complementary ability between the features of each layer and obtains more comprehensive feature map information. Finally, the output of DASPP and FFM are concatenated and convoluted. The third part is the differential binarization

module. Through the final feature map prediction, a probability map and a threshold map are obtained, and then the approximate binary map is obtained by the differential binarization formula. Finally, the final detection result is obtained by post-processing.

### 1.2 DASPP network structure

The DASPP module proposed in this paper expands the receptive field of the model while ensuring the resolution of the feature map, which is conducive to extracting multi-scale and long-distance context information. Its design concept is inspired by the semantic segmentation algorithm DenseASPP [4]. The DenseASPP semantic segmentation algorithm combines the spatial pyramid pooling module (ASPP) [5] and the dense convolution connection idea in the DenseNet [6] network. In this paper, the DASPP module is added after the backbone network to make the network have larger receptive fields and denser sampling points. In addition, with the increase of the dilated convolution expansion rate, the effectiveness is attenuated. The global average pooling is added to the DASPP module to extract the global features, and the global features and local features are fused to improve the segmentation accuracy.

**1.2.1 Dilated Convolutions.** Compared with other public datasets in natural scenes, the text on the vehicle nameplate belongs to a small target object, and the length distribution of the bounding box of the vehicle nameplate text area is shown in Figure 2. Note that the number of text characters is not fixed, and the length of the image bounding box varies greatly. If the receptive field of the text detection model is not enough to cover the area, and the text area is interrupted, the prediction model will output a shorter text box and cannot detect the complete long text. In addition, in some photos taken in darker environments, the etched words of metal nameplates have the problems of low contrast and difficult separation. In summary, the accurate extraction of features requires the model to have a larger receptive field.

In order to expand the receptive field of the model, the common method is to perform pooling or downsampling operation on multi-dimensional feature maps. However, the direct pooling operation



Figure 2: Vehicle nameplate label and detection effect

is too simple and crude. As the network deepens, the resolution of the feature map will gradually decrease, and the information of the image on the spatial scale will be easily lost. The Dilated Convolutions [7] is introduced into the semantic segmentation algorithm. The basic principle is to superimpose different factors of expansion rate around the standard convolution to expand the convolution. Due to the different setting of expansion rate, the receptive field will obtain a larger range of feature information, while ensuring that the resolution of the feature map does not change.

Assuming that the expansion rate is  $d$  and the size of the convolution kernel is  $K$ , the value of 0 is inserted between every two pixels of the convolution kernel in the traditional convolution operation.  $y[i]$  represents the information of the  $i$ th pixel of the output feature matrix, and  $x[i]$  represents the information of the  $i$ th pixel of the input feature matrix. The operation process is:

$$y[i] = \sum_{k=1}^k x[i + d \cdot k] \cdot w[k] \quad (1)$$

$w[k]$  denotes the  $k$ th parameter of the convolution kernel of size  $K$ . The receptive field  $R_{K,d}$  is:

$$R_{K,d} = (d - 1) \times (K - 1) + K \quad (2)$$

**1.2.2 DenseNet.** In the segmentation of vehicle nameplates, there are text characters of different scales, and the proportion of text is too large and too small. Although the dilated convolution can effectively increase the receptive field, the dilated convolution inserts a value of 0 when expanding the convolution kernel. As can be seen in Figure 3, the information of some pixel positions may not be calculated from beginning to end. If the sampling is not dense, it

will be prone to gridding effect, which will cause the problem of detail loss. In addition, the ASPP module proposes to use multiple dilated convolutions with different dilation rates in parallel to fuse multi-scale features, but the receptive field and resolution obtained are still not dense enough. Therefore, this paper sets a reasonable expansion rate and cascades multiple dilated convolution layers. The output of each dilated convolution layer is used as the input to the subsequent dilated convolution layer, so that the output features cover the multi-scale range and the coverage is more dense. At the same time, it also avoids the loss of detailed features caused by dilated convolution with excessive expansion rate.

DenseNet is proposed to solve the problem of gradient disappearance in deep convolutional neural networks. Similar to the idea of ResNet network, it is to establish a cross-connection between shallow networks and deep networks, but the implementation method is different from the residual structure of direct stacking and cross-connection in ResNet network. As shown in Figure 4, DenseNet concatenates the output of all previous layers in the depth direction as input to the next layer and repeats the feature extraction. This dense convolution connection not only reduces the problem of gradient disappearance to a certain extent, but also effectively utilizes each feature layer. The input and output relationship expressions of ResNet residual module and Dense Block module are as follows:

$$x_i = H_l(x_{l-1}) + x_{l-1} \quad (3)$$

$$x_l = H([x_0, x_1, \dots, x_{l-1}]) \quad (4)$$

In formula (3),  $x_l$  is the output characteristic of the  $l$  layer,  $x_{l-1}$  is the output characteristic of the previous layer, and  $H_l$  is the nonlinear transformation. In formula (4),  $[x_0, x_1, \dots, x_{l-1}]$  represent the concatenating of features from the original input to the  $x_{l-1}$

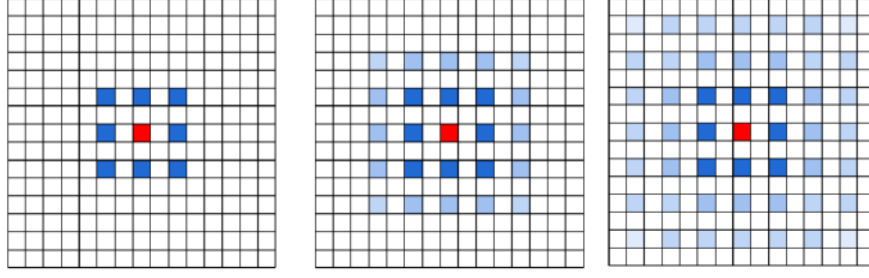


Figure 3: Dilated convolution gridding effect

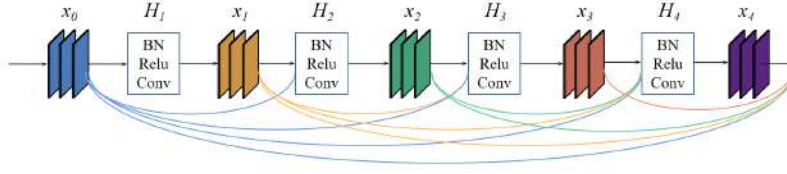


Figure 4: Dense convolution connection of DenseNet network

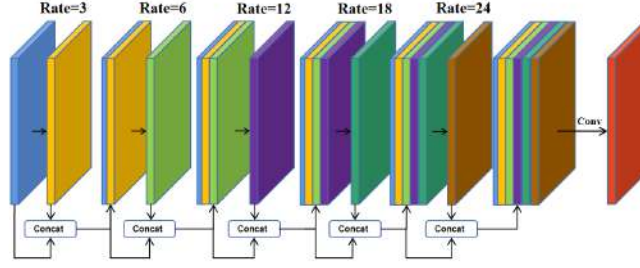


Figure 5: DASPP module details

layer.  $H_l$  represents a series of nonlinear transformations, which are composed of BN + ReLU +  $1 \times 1$  Conv + BN + ReLU +  $3 \times 3$  Conv.

**1.2.3 DASPP module.** The DASPP module expands the receptive field and covers denser feature pixels by repeatedly concatenating dilated convolutions of different dilation rates in a densely connected manner. The structure is composed of multiple dilated convolutions with dense connections of different dilation rates and global average pooling. The dilation rates of 3, 6, 12, 18 and 24 are used to fuse the extracted multi-scale information, and  $1 \times 1$  convolution is used to reduce the number of channels of the fused features. Firstly, the feature map  $x_5$  from the backbone network is input into the DASPP module. After the dilated convolution with the dilation rate of 3, the output feature map and the feature map  $x_5$  are concatenated in the channel dimension, as the input of the dilated convolution with the dilation rate of 6 in the first layer. Then the output feature map of the second layer is concatenated with the input feature map of the second layer, as the input of the dilated convolution with the dilation rate of 12 in the third layer. So repeatedly, the DASPP module will build a larger receptive field

and more dense feature pixel sampling, the specific structure is shown in Figure 5.

Stacking multiple convolutional layers can obtain a larger receptive field. Assuming that the stacked two convolution kernels are convolutional layers of  $K_1$  and  $K_2$ , the size of the new receptive field is:

$$K = K_1 + K_2 - 1 \quad (5)$$

Combined with the definition of the receptive field of the dilated convolution in formula (2), the receptive field in the DASPP module can be calculated as:

$$R_{\max} = R_{3,3} + R_{3,6} + R_{3,12} + R_{3,18} + R_{3,24} - 4 = 127 \quad (6)$$

It can be seen that the receptive field of each pixel on the feature map is 127. As the dilation rate increases layer by layer, the convolution operation of each layer can make full use of all previous feature layers, making the acquisition of feature pixels more dense and greatly reducing the loss of feature map information. In addition, in the case of large dilation rate, the effective weight of convolution will become smaller. The method of parallel global average pooling layer in DASPP module can capture multi-scale

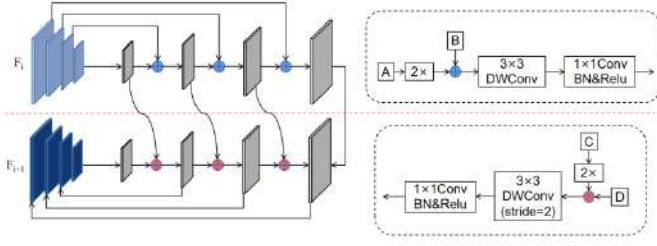


Figure 6: FPPEM module details

local dense information and global context information regardless of the size of input features.

### 1.3 Feature pyramid fusion and enhancement module

The Feature Pyramid Network (FPN) adopts the feature pyramid model. Although the FPN module can improve the target detection accuracy of different scales in the network, the feature fusion only uses upsampling and vector splicing technology, which can not fully integrate the low-level and high-level semantic information [8]. It will also lead to large network computing overhead and can not guarantee the real-time performance of the network. In the Up-Scale stage, the top features output from the backbone network are added to the corresponding feature elements of the next layer after 2 times upsampling, and then through 3×3 deformable convolution and 1×1 convolution, the fused features are obtained. The above operations are performed layer by layer until the bottom features of the feature pyramid at this stage, and the top-bottom feature fusion is realized. In the Down-Scale stage, the bottom features output from the Up-Scale stage are added to the corresponding feature elements of the upper layer after 2 times upsampling, and then through 3×3 deformable convolution with stride 2 and 1×1 convolution, the fused features are obtained. The above operations are performed layer by layer until the top features of the feature pyramid at this stage, and the bottom-top feature fusion is realized.

In addition, the FPPEM module has the following advantages: FPPEM is a cascade module. With the increase of the number of cascades, the feature maps of different scales can be fused more fully and the receptive field of the features can be increased. This paper uses deformable convolution instead of conventional convolution to construct the connection part of FPPEM module, which can greatly reduce the network computing overhead.

For semantic segmentation, high-level semantics facilitate the extraction of pixel categories, and low-level semantics facilitate the extraction of edge and texture information. Each type of semantic information has an irreplaceable role. The FFM fuses the different levels of features generated by the FPPEM cascade. The schematic diagram of the FFM module is shown in Figure 7. In order to enhance the feature expression ability of different scale texts, the feature maps of the same scale are merged by the method of element addition, and then the feature maps of different scales are upsampled to make the feature maps have the same size. and the final feature map is output by splicing. Compared with the method of directly

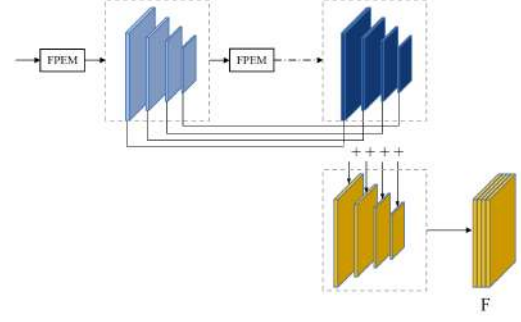


Figure 7: FFM module details

upsampling the feature maps of different scales and then cascading them all, the FFM module can effectively reduce the number of feature channels, thereby accelerating the prediction speed and obtaining better segmentation results.

### 1.4 Differential binarization

The third part is the detection head. The detection head needs to separate the pixels in the core area of the text from the background in the predicted image P, which is equivalent to clustering the pixels in the image. The simplest clustering is to set a fixed threshold, and divide each pixel according to the threshold  $t$ . The two classification method of fixed threshold is :

$$B_{i,j} = \begin{cases} 1, & P_{i,j} \geq t \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

The setting of fixed threshold depends on artificial experience and is not suitable for scene images with complex background. Therefore, it is considered to design the threshold as a learnable parameter to add to the optimization iteration of the neural network, the decision threshold of each pixel will be adaptive [9]. The differential binarization equation is shown in formula (8). a, b, c represent the values of the  $(i,j)$  position points on the approximate binary map B, the probability map P, and the threshold map T. The back propagation from the gradient is shown in formula (9), in the case of cross entropy loss, the calculation of positive sample  $l_+$  and negative sample  $l_-$  can be expressed as formula (10) and formula (11), and the partial derivative of input  $x$  is shown as formula (12) and formula (13). It can be seen that  $k$  is the gradient gain factor, and the gradient has a large gain range for false prediction, and the habitual setting  $k = 50$ . Therefore, the differentiable binarization with adaptive threshold not only helps to separate the text area from the background area, but also separates the similar instances.

$$B_{ij} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \quad (8)$$

$$f(x) = \frac{1}{1 + e^{-kx}} \quad (9)$$

$$l_+ = -\log \frac{1}{1 + e^{-kx}} \quad (10)$$

$$l_- = -\log \left( 1 - \frac{1}{1 + e^{-kx}} \right) \quad (11)$$

$$\frac{\partial l_+}{\partial x} = -kf(x)e^{-kx} \quad (12)$$

$$\frac{\partial L}{\partial x} = kf(x) \quad (13)$$

## 2 EXPERIMENTS AND RESULTS ANALYSIS

### 2.1 Datasets

Self-defined datasets: The collection of vehicle nameplates in complex environments includes more than 500 text pictures from different angles. The materials of the nameplates are mostly metal materials, and a few are flexible materials. At the same time, data enhancement operations are taken on the self-defined datasets, including random rotation, random perspective transformation, adding blur, adding noise, random cropping, random modification of tone, brightness, contrast, etc. During the training process, each text image will randomly use data augmentation operations according to a fixed probability (0.5). The text box area is recorded with four coordinate points of the rectangle.

### 2.2 Experimental environment and evaluation indicators

The experimental platform is Windows10 (64bit) operating system, R9-5900HX CPU, 32G memory, NVIDIA GeForce GTX 3070 GPU, 8GB video memory, CUDA version 11.3, CUDNN version 8.2, using PyTorch framework to build network model. In this paper, The experiment is loaded with 100 epochs pre-trained on the ICDAR2015 datasets and 500 epochs fine-tuned training on self-defined datasets. The BatchSize is set to 8, the initial learning rate is set to 0.001, and the gradient is updated using the Adam optimizer. In order to improve the generalization ability of the training model, random data enhanced rotation, random cropping and random flip are used to augment the training image. Finally, the image is adjusted to 640 pixel  $\times$  640 pixel and sent to the network for training.

In order to evaluate the improved model, Precision, Recall and F1-Score commonly used in text detection tasks are taken as evaluation indicators. True Positive represents the number of actual text but judged as non-text (TP), False Positive represents the number of actual non-text but judged as text (FP), False Negative represents the number of actual text but judged as non-text (FN), such as formulas (13), (14), (15):

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (16)$$

### 2.3 Custom Datasets Results

In order to more intuitively reflect the feasibility of the model, Figure 8 shows the comparison between the proposed method and the basic algorithm on the self-defined datasets. The first column is the original image, the second column is the segmentation result of DBNet, and the third column is the segmentation result of this paper. DBNet has achieved good detection results in public datasets, but it has not achieved the same advantages in the vehicle nameplate scene data set, because the vehicle nameplate scene datasets are mostly small targets and long texts. The combination of the two

networks can better predict the image and make it more accurate in the detection process.

### 2.4 Ablation experiment

In order to analyze the effectiveness of different improved methods, under the same experimental environment and parameter configuration, the improved DASPP module, FPEM + FFM module and the combination of the two methods are trained and tested on the self-defined datasets, and the performance is compared with the original network. The experimental results are shown in Table 1.

Line 1 is the baseline, representing the replication results of the DBNet algorithm.

Line 2 is to add the DASPP module, the P, R and F1 indicators are all improved compared with the original network, which verifies that the dilated convolution can increase the local receptive field and the dense connection can improve the semantic interaction between the branches.

Line 3 is to add the FPEM + FFM module. Although the accuracy is reduced, the ability of text detection is improved. The R indicator is increased from 87.5 % to 88.8 %, and the F1 indicator is increased from 87.6 % to 88.2 %. This method can improve the expression ability of different scale features in the backbone network and effectively enhance the text region features.

Line 4 is the combination of DASPP and FPEM + FFM modules, which will increase the P and R indicators by 2.6 % and 0.9 % respectively. It can not only better enhance the detection of small targets and long text, but also balance false detection and missed detection. Compared with the baseline results, the F1 indicator is the best.

Figure 9 shows the effectiveness of each module intuitively. Among them, the first column is the original map, the second column is the segmentation map of the original algorithm, the third column is the segmentation map of the DASPP module, the fourth column is the segmentation map of the FPEM + FFM module, and the last column is the final segmentation map of both DASPP and FPEM + FFM. From the visualization results, although DBNet can find text instances in the map, the detected text boxes are not fine enough, and there are problems such as discontinuous segmentation edges and incorrect segmentation of adjacent text. The text localization boxes generated by the improved model is more accurate, with less false detection and missed detection. It also achieves better detection results on multi-scale text detection, and can correctly separate and detect the closely arranged text lines. It basically realizes the accurate positioning of vehicle nameplate detection, and a more complete detection area is conducive to text recognition.

## 3 CONCLUSION

Due to the harsh working environment of vehicle detection, the vehicle nameplate is affected by illumination changes, complex backgrounds, and noise interference. Characters are often submerged by complex backgrounds and cannot be extracted. The nameplate image taken in the natural scene cannot guarantee better clarity, fixed shooting angle and horizontal nameplate image direction. This paper proposes a DBNet semantic segmentation algorithm based on dense connection and feature enhancement. The DASPP module is used to extract multi-scale information of specific resolution features, and capture dense sampling point pixels to obtain

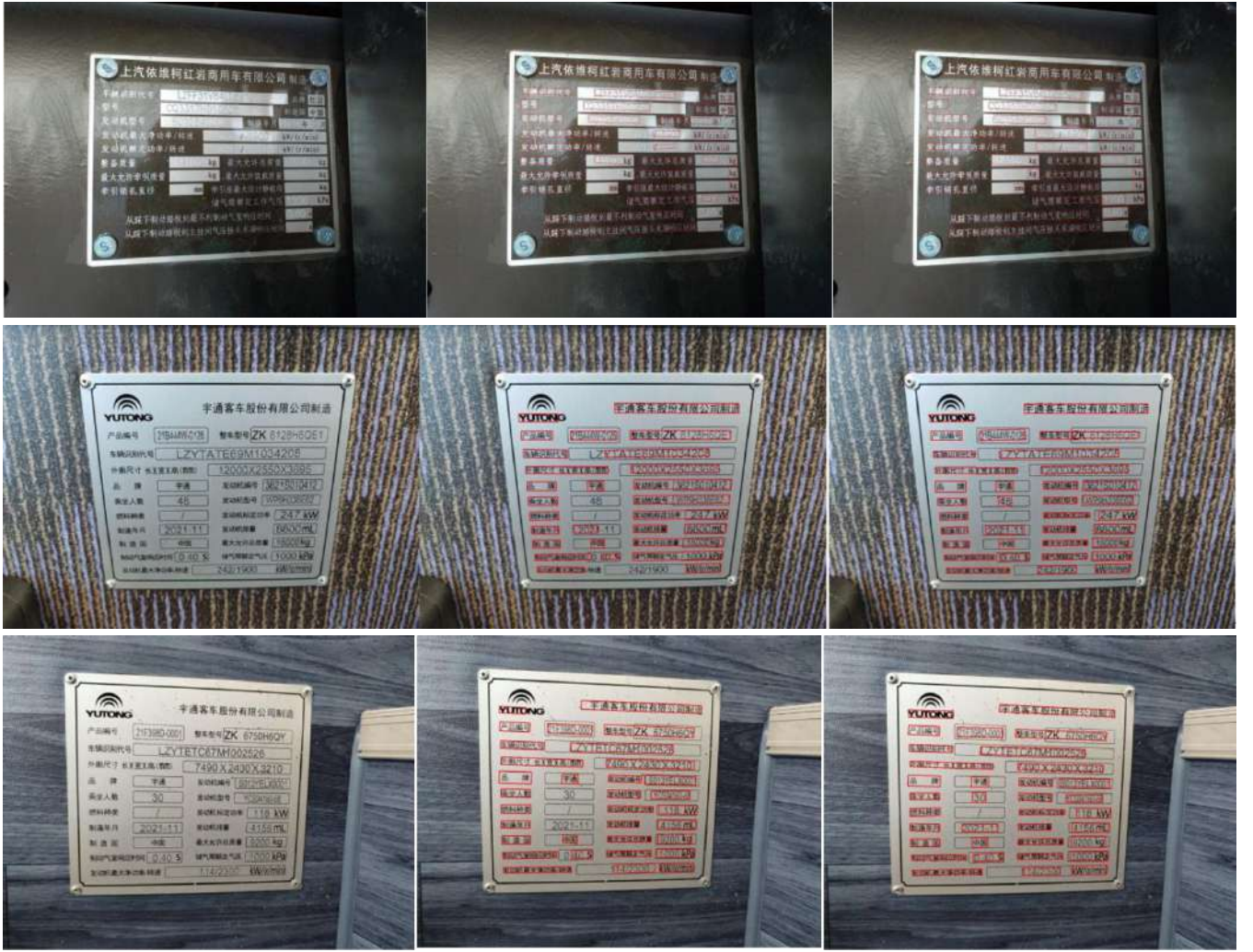


Figure 8: Visualization results of custom validation set

Table 1: Ablation experimental results on custom datasets

Backbone	DASPP	FPEM+FFM	Custom Dataset		
			$P$	$R$	$F1$
ResNet18	×	×	87.8	87.5	87.6
ResNet18	√	×	88.5	87.8	88.1
ResNet18	×	√	87.5	88.8	88.2
ResNet18	√	√	90.4	88.4	89.4

more comprehensive local information. At the same time, the multi-layer features of the backbone network are reused, and the features are enhanced to improve the representation ability of the features. Finally, the features obtained by the two branches are fused. The experimental results show that this algorithm effectively solves the problem of segmentation edge discontinuity and false segmentation, and achieves ideal segmentation results. Compared with

the original DBNet algorithm, the method in this paper has a certain improvement in the evaluation indicators, which proves the effectiveness of the method. The follow-up work will research on the structured representation of target text, model lightweight and text recognition, and further improve the detection accuracy and detection speed of the model.

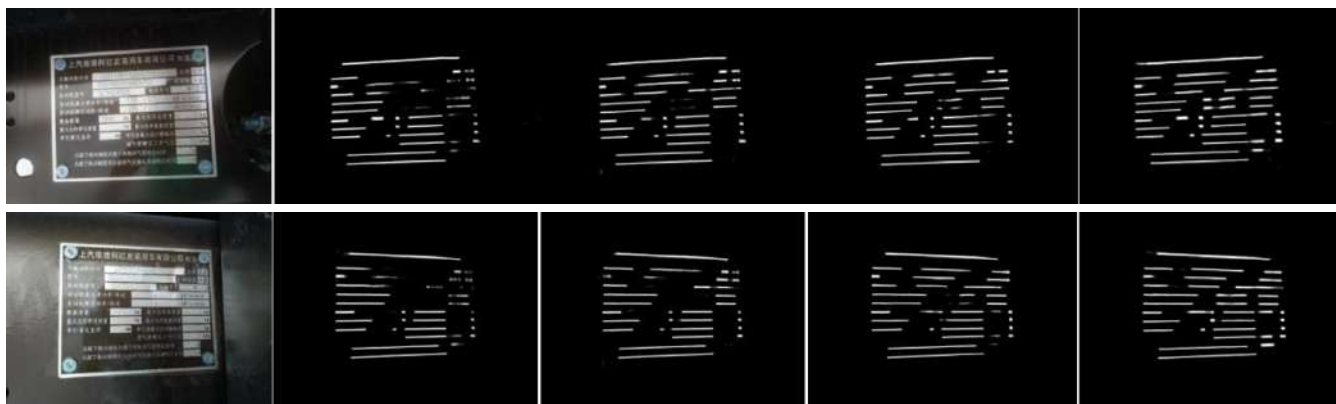


Figure 9: the segmentation image visualization results of each module

## REFERENCES

- [1] Wang R, Sang N, Ding D, *et al*. Text Detection in Natural Scene Image: A Survey [J]. ACTA AUTOMATICA SINICA. 2018,44(12):2113-2141.
- [2] Chen X, Chen X, Yuan J, *et al*. Electricity equipment nameplate recognition based on deep learning[J]. Journal of Guangxi University( Nat Sci Ed). 2018,43(06):2216-2226.
- [3] Gong F, Liu F, Li J, *et al*. Scene Text Detection and Recognition Based on Deep Learning[J]. Computer Systems & Applications. 2021,30(08):179-185.
- [4] M, Yu K, Zhang C, *et al*. Denseaspp for semantic segmentation in street scenes[C].2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.3684-3692.
- [5] L C, Papandreou G, Kokkinos I, *et al*. Semantic image segmentation with deep convolutional nets and fully connected crfs [J]. Computer Science,2014,(4):357-361.
- [6] G, Liu Z, Van Der Maaten L, *et al*.Densely connected convolutional networks[C].2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).2017.2261-2269.
- [7] F, Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[C].IEEEInternational Conference on Learning Representations, 2016.
- [8] W H, XIE E Z, LI X,*et al*. PAN++:Towards Efficient and Accurate End-to-End Spotting of Arbitrarily-Shaped Text[J].IEEE Transactions on Pattern Analysis Machine Intelligence,2021.
- [9] M, Wan Z, Yao C, *et al*. Real-time scene text detection with differentiable binarization [C]Proceedings of the AAAI Conference on Artificial Intelligence.2020, 34(07): 11474-11481.

# Performance evaluation of agricultural logistics enterprises based on GA algorithm

Yebin

ChengDu Neusoft University, ChengDu  
SiCuan, 611844, yebin@nsu.edu.cn

## CCS CONCEPTS

• **Applied computing** → Electronic commerce; E-commerce infrastructure.

## KEYWORDS

Agriculture, logistics enterprises, performance evaluation, GA, BP neural network

## ACM Reference Format:

Yebin. 2023. Performance evaluation of agricultural logistics enterprises based on GA algorithm. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590068>

## 1 INTRODUCTION

With the rapid development of agricultural logistics, the market demand for agricultural products logistics continues to grow. However, how to better improve the efficiency of enterprise logistics is still a major problem in rural agricultural development. Therefore, the performance evaluation of agricultural product logistics enterprises can clearly understand the development situation faced by enterprises, and at the same time, can also find problems in enterprise management. <sup>[1]</sup>Based on the purpose of improving the economic efficiency of enterprises, this paper mainly designs the evaluation index system from the perspective of financial efficiency, and uses the GA neural network algorithm to evaluate the performance of agricultural cold chain logistics enterprises more scientifically.

## 2 INDEX SYSTEM OF PERFORMANCE EVALUATION

### 2.1 Basic principles for establishing evaluation index system

In the process of selecting indicators, in order to ensure the scientificity of indicators and lay a good foundation for later prediction, it is mainly based on the following four principles: objectiveness, representativeness, quantifiability, and the combination of static indicators and dynamic indicators

The so-called objective evaluation index system should be aimed at improving the core competitiveness of the enterprise and the social and economic benefits of the company. The index can play a decisive role in this purpose; Representativeness means that the selected indicators should reflect all aspects of logistics more comprehensively; <sup>[2]</sup> Quantifiable means that the index data can be obtained and quantified as input layer data in order to facilitate the later application in the artificial neural network; The principle of combining static and dynamic means to consider the short-term development goals of enterprises and the long-term development trend of agricultural cold chain logistics enterprises at the same time, because the operation mode of agricultural cold chain logistics enterprises is not a static process, but a dynamic, long-term and complex process from the production to distribution of fresh agricultural products.

### 2.2 Screening of evaluation indicators

In the process of research, this paper consulted the relevant indicators of performance evaluation, mainly on agricultural products logistics, including the author Lv Qingqing in his paper "Research on the performance evaluation of agricultural products cold-chain logistics enterprises" <sup>[3]</sup>, and the author Sheng Zhonghua in his paper "Research on the performance evaluation of agricultural products cold-chain logistics enterprises" <sup>[4]</sup>, as well as the author Fu Yanmei's paper, Relevant indicators in its "Research on Performance Evaluation of Cold Chain Logistics Enterprises" <sup>[5]</sup>. At the same time, relevant industrial policies and specifications were consulted, such as the China Logistics Statistical Yearbook issued by the China Federation of Logistics and Procurement in 2019, the Comprehensive Work Plan for Energy Conservation and Emission Reduction in the 13th Five-Year Plan issued by government departments, and the Logistics Industry Adjustment and Revitalization Plan.

Based on the principle of scientific indicator selection, as well as the full interpretation of policies and documents, this paper initially proposes a set of three-level evaluation indicator system to examine the comprehensive performance of agricultural cold chain logistics enterprises with five dimensions.

The final indicator system is shown in Table 1

### 2.3 Data standardization

The evaluation sample objects selected in this paper are from Guotai An Database and <http://www.cninfo.com.cn>. The specific relevant information is mainly obtained from the "Enterprise Annual Report" publicly disclosed by logistics enterprises, as well as the query and collation of relevant data. By analyzing the data of listed agricultural products logistics enterprises in recent years, and on the basis of constructing the performance evaluation system of agricultural

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590068>

**Table 1: Evaluation index system and quantification method**

Index name	Specific description	Quantification method
connection	Customer growth rate C11	Main customer sales/annual total sales
	Customer ratio C12	Sales revenue of major customers/Total turnover
	Customer Dependency C13	Annual sales of top 5 customers
Cold chain operation	Utilization rate of refrigerated truck C21	Transportation expenses/sales expenses
	Utilization rate of cold chain equipment C22	Annual depreciation rate of equipment
	Cold chain production ratio C23	Production number/total number
Green and environmental protection	Whether there is environmental certification C31	Yes is 2, No is 1, Not applicable is 0
	Whether it is a key pollutant discharging unit C32	Yes is 2, No is 1, Not applicable is 0
Scientific and technological development	R&D investment rate C41	R&D investment/operating income
	Proportion of technical personnel C42	R&D personnel/total personnel * 100%
	Proportion of highly educated personnel C43	Number of undergraduate students/total number * 100%
finance	Operating profit margin C51	Net profit/total revenue
	Net profit margin C52	Net profit/total revenue

**Table 2: Original data of evaluation indicators**

	A	B	C	D	E	F	G	H
	21.66	1.25	2.18	-1.42	15.45	1.11	-2.97	-0.95
C12	52.94	74.61	2.64	5.63	24.85	22.44	24.19	17.16
C13	0.00	0.00	3.02	0.95	0.00	0.00	0.11	0.00
C21	9.50	9.70	9.50	8.71	9.50	9.50	9.50	9.00
C22	35.48	62.79	5.40	71.69	55.07	58.35	45.68	76.76
C23	16.13	2.33	0.00	12.71	25.21	15.13	7.65	7.93
C31	0.00	0.00	2.00	2.00	2.00	1.00	2.00	1.00
C32	0.00	0.00	2.00	2.00	2.00	2.00	2.00	1.00
C41	0.00	10.47	4.00	4.59	0.01	11.98	32.76	3.84
C42	12.90	11.63	4.16	4.05	20.46	10.73	43.95	4.61
C43	72.69	44.20	63.90	59.43	30.70	46.96	53.36	49.09
C51	6.04	6.59	3.98	2.08	0.13	0.50	-150.12	5.96
C52	4.91	5.40	3.00	1.43	0.34	0.49	-174.36	5.31

products logistics enterprises, the quantitative three-level indicators will be used to evaluate agricultural products logistics enterprises.

According to the actual values of the three indicators of the five dimensions previously determined, the original data of the three indicators of the eight companies (A-H) in 2019 are shown in Table 2 below:

C11 to C52 in the above table refer to the corresponding indicators in Table 1. This paper only presents the data of eight companies in 2019. In the actual calculation process, due to the need of sample training, the data of more than one listed company from 2016 to 2021 is used.

As the actual performance value of the output layer, the index is weighted by the entropy weight method, and then the performance score is calculated according to the weight and each index value, that is, the actual performance value.

It is set that there are  $n$  and  $m$  evaluation indicators, and  $X_{ij}$  means the  $j$ th indicator of the  $i$ th company ( $i=1,2,3,\dots,n$ ;  $j=1,2,3,\dots,m$ ). The calculation process is as follows:

Calculate the characteristic proportion of the  $i$ th enterprise for the  $j$ th index:

$$p'_{ij} = \frac{X'_{ij}}{\sum_{i=1}^n p_{ij}} \quad (1)$$

Entropy value of index  $j$ :

$$e_i = -\frac{1}{\ln n} \sum_{i=1}^n p_{ij} \ln(p_{ij}), 0 \leq e_j \leq 1 \quad (2)$$

Calculate the coefficient of difference:

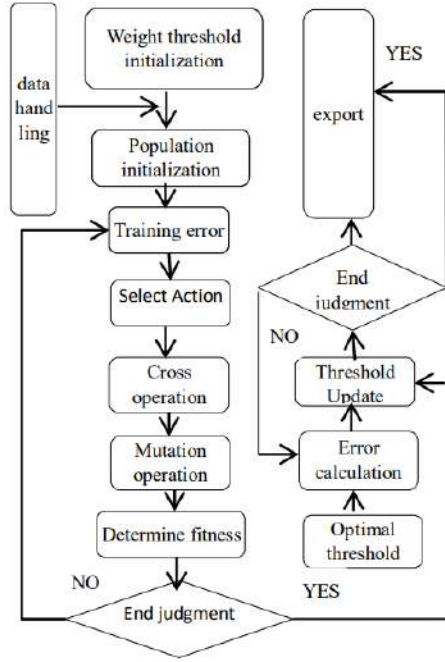
$$g_j = 1 - e_j \quad (3)$$

Weight value  $W_j$  calculation:

$$w_j = \frac{g_j}{\sum_{j=1}^m g_j} \quad (j = 1, 2, \dots, n) \quad (4)$$

Calculation of performance score  $S$ :

$$S = \sum_{j=1}^m w_j * p_{ij} \quad (5)$$



**Figure 1: Flow chart of optimized BP neural network algorithm**

In the genetic neural network model, the index value of each indicator is used as the input layer value, and the performance score calculated by the above formula is used as the output layer value. Before training, the data will be normalized to make the corresponding values between [0,1], so as to improve the training efficiency.

### 3 EVALUATION MODEL ALGORITHM DESIGN

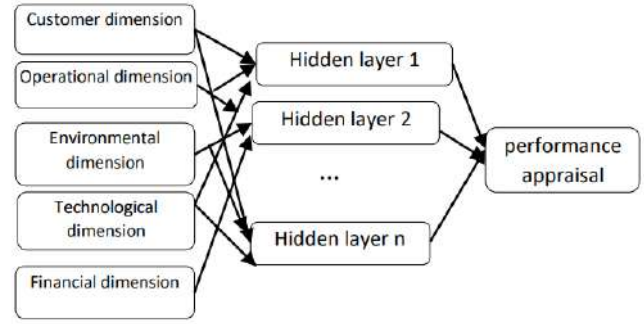
In this paper, genetic algorithm and BP artificial neural network algorithm are organically integrated, and the original BP artificial neural network algorithm is optimized by genetic algorithm. The flow chart is as follows:

In the figure above, the implementation process of genetic algorithm optimization and original BP neural network is included. The key points of the whole algorithm optimization mainly include the establishment of population, the calculation of individual fitness, the control of selectivity, mutation and cross calculation. [6] This paper focuses on the error calculation process, selection process, mutation calculation process, relevant methods and key calculation processes in the algorithm, including the following: [7]

Calculation of training error value, the error value of BP neural network training determines the value of individual fitness. [8] Therefore, it is very important. The calculation formula of fitness value is as follows:

$$F = k \left( \sum_{i=1}^n asb(y_i - o_i) \right) \quad (6)$$

In this formula,  $k$  represents the coefficient, while  $y_i$  indicates the last expected value of the  $i$ th node. Use  $o_i$  represents the actual



**Figure 2: Topological structure of neural network**

operation value of the  $i$ th node, and the number of output layers is  $n$ .

Select calculation process, in this paper, roulette algorithm is used as the selection operation method, which is based on the fitness ratio, and the probability value of the selection of the  $i$ th individual,  $p_i$  Calculation method:

$$f_i = \frac{k}{F_i} \quad (7)$$

$$p_i = \frac{f_i}{\sum_{j=1}^N f_j} \quad (8)$$

In the above two formulas,  $f_i$  represents the fitness of individual  $i$ , with  $k$  as the department;  $R$  The number of individuals in the population is represented by  $N$ . Variation calculation process, the formula is as follows:

$$a_{ij} = \begin{cases} a_{ij} + (a_{ij} - a_{max}) * f(g) & r > 0.5 \\ a_{ij} + (a_{min} - a_{ij}) * f(g) & r \leq 0.5 \end{cases} \quad (9)$$

$$f(g) = r_2 \left( 1 - \frac{g}{G_{max}} \right)^2 \quad (10)$$

In the above formula,  $a_{ij}$  mainly represents the variation value of the  $j$ th gene of an individual,  $a_{max}$  represents gene  $a_{ij}$  The maximum value of  $ij$ ; The minimum value is  $a_{min}$ ;  $F(g)$  is a random number,  $g$  is used to identify the number of iterations;  $G_{max}$  represents its limit evolution times; The value range of  $r$  is [0,1].

This paper adopts a three-layer topology of input layer, hidden layer and output layer. The input layer data is the evaluation index value after standardization, and the output value is the performance value calculated. Its basic structure is as follows:

Where, the final output is "performance value", and the empirical calculation formula of hidden number is: [9]  $L = \frac{3\sqrt{mn}}{2}$  (11)

In the formula,  $n$  and  $m$  represent the output quantity and the input quantity respectively;  $L$  is the number of hidden layers. Of course, according to different designs, corresponding adjustments can be made according to the requirements of accuracy and convergence. [9] The main algorithm process is as follows:

**Algorithm 1** Iterative Algorithm

---

```

1. Input evaluation training data;
2. Standardize data processing;
3. Construct artificial neural network;
4. Initialize the genetic algorithm;
5. maxge=70;
6. sizepo=20; % Set the size of the population;
7. cross=[0.3]; % Set the probability of crossing;
8. mutation=[0.3]; % Set the probability of variation;
9. Set the number of nodes;
10. Initialize the population;
11. Calculate chromosome fitness;
12. Random initialization of population;
13. individuals=Select (individuals, sizepo)% to select;
14.individuals.chrom=Cross(cross,lenchrom,individuals.chrom,sizepo,bound);
% Carry out cross operation;
15.individuals.chrom=Mutation(pmutation,lenchrom,individuals.chrom,sizepo,maxge,bound);
% Perform mutation operation;
16.Functionret=mutation(pmutation,l
enchrom,chrom,sizepop,num,maxgen,bound);% The optimal
threshold and weight are given to the network;
17. Calculate individual fitness.
```

---

#### 4 COMPARATIVE ANALYSIS OF ALGORITHM RESULTS

Through the empirical formula for calculating the number of hidden layers in this paper, you can try to run the program with different hidden layers to find the optimal number of hidden layers. The comparison of validation data is shown in the following table:

After comparison, when the number of hidden layers is 5, the error is relatively small. The accuracy of evaluation will vary with different methods and different number of samples. Compare BP neural network with this algorithm and cart regression tree. The accuracy data is shown in Table 4

Through comparison, it can be seen that when the number of training samples is small, the results of each method are not very large, while when the number of samples is relatively large, the algorithm used in this paper performs better.

#### 5 CONCLUSION

By comparing with other commonly used methods and combining with the analysis of the algorithm results in this paper, it can be seen that the BP artificial neural network algorithm based on genetic algorithm improved in this paper can well evaluate the performance of cold chain logistics companies in the field of agricultural products, and the prediction results are relatively accurate. Therefore, in the process of practical application, through prediction and evaluation in advance, when the evaluation result is found to be unsatisfactory, it can give early warning to relevant enterprises and urge them to adjust their business strategies in time, which also fully reflects that the algorithm in this paper can play a positive role in the production and operation of agricultural cold chain logistics enterprises.

At the same time, the indicator system for performance evaluation of agricultural products cold-chain logistics enterprises designed in this paper is based on other literature or experience. In the actual production and operation, agricultural products cold-chain logistics enterprises may encounter other unpredictable external influences, which will also have a great impact on the evaluation results. Therefore, in the subsequent research, other influencing factors can be considered as much as possible, and the input sample data of the algorithm can be adjusted, Make the results more objective.

In addition, the sample data of this study may be insufficient, including a small number and the limitations of the company size. In terms of quantity, this paper uses at most 300 sets of sample data. The limitation is that these data come from the number of listed companies for many years, which may only reflect the situation of these large and medium-sized enterprises. However, the evaluation of small and medium-sized enterprises is still insufficient. Therefore, in the later application process, we will try to collect more and broader enterprise samples, so as to make the results more scientific and effective.

#### ACKNOWLEDGMENTS

This work was supported by “Research Center for the Coordinated Development of Education, Economy and Society in Chengdu-Chongqing Region”, “Sichuan E-Commerce and Modern Logistics Research Center”, Grant no: CYJXF22032, DSWL-22.

**Table 3: Error comparison of different hidden layers**

Number of hidden layers	3	4	5	6	7
maximum error	0.141	0.095	0.071	0.089	0.112
Training steps	655	481	285	370	461

**Table 4: Comparison of prediction accuracy of different algorithms**

Number of samples	Cart regression tree	BP neural network	Algorithm in this paper
300	91.58%	93.41%	95.23%
200	90.73%	92.56%	94.60%
100	92.77%	91.65%	93.26%

## REFERENCES

- [1] Vola Federico, Benedetto Vera, Vainieri Milena, Nuti Sabina. The Italian interregional performance evaluation system[J]. Research in Health Services & Regions, 2022, 1(1).
- [2] Jeon Kwang Sub. A Study on Project Goals and Performance Evaluation System in Urban Regeneration Project Evaluation System[J]. Korea Real Estate Academy, 2019, 79.
- [3] Lv Qingqing. Research on performance evaluation of agricultural cold chain logistics enterprises [J]. China Collective Economy. 2019 (02): 131-132
- [4] Sheng Zhonghua. Research on performance evaluation of agricultural cold chain logistics enterprises [D]. Shandong: Shandong University of Technology, 2018
- [5] Fu Yanmei. Research on performance evaluation of cold chain logistics enterprises [D].
- [6] Sasaki Aiichiro. Effectiveness of Artificial Neural Networks for Solving Inverse Problems in Magnetic Field-Based Localization[J]. Sensors, 2022, 22(6).
- [7] Kommula Bapayya Naidu, Kota Venkata Reddy. An Effective Sustainable Control of Brushless DC Motor using Firefly Algorithm – Artificial Neural Network based FOPID Controller[J]. Sustainable Energy Technologies and Assessments, 2022, 52(PB).
- [8] Tang Jun, Zhao Bo, Li Wenxing. Prediction of wear state of disc shaped milling cutter based on genetic algorithm and BP neural network [J]. Journal of Henan University of Technology: Natural Science Edition, 2017, (5): 66-7
- [9] Bi Yunfan, Zhang Jian, Xu Xiaohui, Sun Wenhui, Zhang Zhisheng. Electric short-term load forecasting model based on gradient lifting decision tree [J]. Journal of Qingdao University: Engineering Technology Edition, 2018, 33 (3): 70-75

# SSGAR: A Genetic-based Routing Solution for Aeronautical Networks aided by Software Defined Satellite Network

Kaixuan Sun  
University of Science and Technology  
of China  
Hefei, Anhui, China  
skx96@mail.ustc.edu.cn

Ke Wu  
University of Science and Technology  
of China  
Hefei, Anhui, China  
wkwkwk@mail.ustc.edu.cn

Wenke Yuan  
University of Science and Technology  
of China  
Hefei, Anhui, China  
ywk4432@mail.ustc.edu.cn

Guangyuan Wei  
University of Science and Technology  
of China  
Hefei, Anhui, China  
wgy1997@mail.ustc.edu.cn

Huasen He\*  
University of Science and Technology  
of China  
Hefei, Anhui, China  
hehuasen@ustc.edu.cn

## ABSTRACT

In the next generation network, both the satellite network layer and aeronautical network layer will play significant roles, leading the world into the era of global interconnectivity. However, the large-scale and high-mobility characteristics of aircraft networks greatly challenge the application of traditional routing algorithms. Therefore, this paper aims to solve this challenge by exploiting a Software Defined Satellite Network (Sat-SDN) to facilitate the routing in aeronautical networks. By centrally controlling aeronautical routing through satellites, the computation and communication overhead for aeronautical networks are relieved, since frequent packet flooding and broadcasting for synchronizing the rapidly-fluctuating topology of aeronautical networks can be avoided. To extend the aeronautical networking and transmission mechanism to a global scale, a multi-domain extension mechanism is proposed, while the concept of dynamic inter-domain telescope nodes is introduced to greatly simplify the network topology. A Sat-SDN aided Genetic-based Aeronautical Routing (SSGAR) algorithm is further designed to solve the problem of huge routing calculation space and long convergence time in large-scale multi-node network scenarios. Moreover, experiments and simulations are conducted using real aircraft data, which demonstrate that our proposed SSGAR algorithm can effectively reduce communication costs and improve transmission quality compared to existing solutions.

This work was supported by the National Key R&D Program of China (Grant No. 2022YFB3902800), by the National Natural Science Foundations of China (Grant No. 62173315, 62101525, 62201543 and 62021001), by the Youth Innovation Promotion Association CAS (Grant No. 2020450), by the Strategic Priority Research Program of CAS (Grant No. XDC07020200), by the Fundamental Research Funds for the Central Universities, and by the China Environment for Network Innovations (CENI).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590069>

## CCS CONCEPTS

• Networks → Routing protocols.

## KEYWORDS

Software Defined Satellite Network, Aeronautical Networks, Routing Algorithm

## ACM Reference Format:

Kaixuan Sun, Ke Wu, Wenke Yuan, Guangyuan Wei, and Huasen He\*. 2023. SSGAR: A Genetic-based Routing Solution for Aeronautical Networks aided by Software Defined Satellite Network. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590069>

## 1 INTRODUCTION

With the continuous breakthrough of communication technology, we have entered the era of the fifth generation (5G) network [11], which has greatly changed various aspects of our lives. However, the limited coverage and high deployment cost of 5G network have hindered their development [6]. Therefore, researchers have shifted their focus to the space-air-ground integrated network, attempting to use the satellite network and the aeronautical networks to construct a low-cost, full-coverage network and propose the architecture of the sixth generation (6G) network system. Especially in recent years, the development of millimeter-wave technology and high-gain antennas has made long-distance communication between aircraft possible [5]. The aeronautical network layer, which is one of the most promising parts of the next generation networks, has attracted lots of attention of researchers. Currently, there are thousands of aircraft flying in the sky, and by deploying high-gain antennas and related routing equipment, the forming aeronautical networks can achieve multiple functions. Firstly, unlike satellite networks, aeronautical network effectively alleviates the deployment cost problem. It does not require to launch specific satellites and equipment to provide network services. The existing civil aircraft scale is large enough to form the network. Secondly, due to the low flight altitude of the aeronautical network, it can provide faster network services. In addition, the aeronautical network has extensive coverage, thus it can not only meet the demand for in-flight

internet service, but also can provide seamless coverage to remote mountain villages, sea-going vessels, etc., which is not limited by terrestrial buildings and other factors.

However, the implementation of the aeronautical networks still faces many difficulties, including short survival cycles of aircraft communication links, frequent updates of topology between aircraft, high deployment cost of base stations, limited signal reception distance, and difficulty in predicting the flight trajectory of aircraft [14].

Currently, research on dynamic network routing mainly falls into three categories: proactive routing algorithms, reactive routing algorithms, and geographic-based routing algorithms. As one of the most classic dynamic routing algorithms, proactive routing such as Optimized Link State Routing (OLSR) [7], often uses network slicing model, which requires each node to obtain real-time connection status information of the entire network at the beginning of each time slot to support routing decisions. However, this type of routing algorithm requires high communication bandwidth, which often severely affects normal data flow transmission, especially in scenarios with a large number of nodes. Therefore, reactive routing, like the Ad hoc On-demand Distance Vector (AODV) [10], proposes a new mechanism, in which nodes do not actively obtain the full network topology. They only broadcast a route-finding packet to the destination when the transmission demand arrives which to a certain degree avoids unnecessary communication overhead. However, this mechanism is only suitable for low-traffic network scenarios, where only few nodes are active. And the improvement effect is limited and may introduce additional path lookup latency. Geographic-based routing, such as Greedy Perimeter Stateless Routing (GPSR) [8], uses a greedy forwarding strategy to forward data packets to the nearest neighbor to the destination. Although this method solves the problem of routing broadcast packets to some extent, it also brings a high risk of transmission failure. Therefore, algorithms belonging to this kind always need to introduce additional path recovery measures, which in turn affect forwarding efficiency sometimes.

In summary, traditional dynamic routing algorithms are often suitable for small-scale or low-mobility systems and cannot adapt to the large-scale and high-mobility scenarios of aeronautical networks, which requires the study of new routing algorithms to solve this problem. The greatest challenge in aeronautical networking is how to synchronize the state between aircraft [12]. Introducing satellite network layer is a promising solution for addressing the challenge of real-time information perception for large-scale aeronautical networks [13]. The satellite network has a broader coverage range, such as Geosynchronous Earth Orbit (GEO). Once utilizing the satellite network to centrally control aircraft, the widely distributed aircraft nodes can be effectively managed. At the same time, the aircraft layer is only responsible for data transmission, without the need to perceive real-time information of other nodes. Then researchers can fully leverage the regularity and predictability of flight schedules to realize the network design and route calculation.

Software Defined Networking (SDN) [1] adopts the strategy of separating the control plane and data plane. The controller is responsible for formulating network policies and publishing forwarding rules, using the OpenFlow protocol to send control information to switches and routers in the data plane. The control information will

modify the forwarding or routing tables of the data plane, which effectively achieves network control and management. Switches and routers just need forward data packets following the received forwarding rules without the need to perceive the network state.

There have been many studies on the application of SDN in satellite networks, among which [2] proposed an SDN architecture based on the combination of high-orbit satellites and terrestrial networks. However, such a space-ground architecture often suffers from long-distance transmission delays. Similarly, in [3], the authors proposed an SDN solution for future inter-satellite networks by combining the characteristics of high, medium, and low orbit satellite structures. That is, separating the data plane and control plane of the satellite network, and letting the high-orbit control plane perform system calculations and resource management operations, while other satellite nodes only need to perform simple hardware configuration and data forwarding. In [9], the authors envisioned the application of SDN in the next-generation network architecture by combining the space-air-ground integrated fusion structure.

Therefore, based on the ideas of using software-defined satellite networks (Sat-SDN) to assist in routing and transmission in the aeronautical networks, this paper designs the following system structure: the terrestrial node acts as the user plane, responsible for generating data requests, while the aeronautical networks act as the data forwarding plane who forward data packet by the received control message and look for the target reachable path after receiving the request from the user plane. The satellite network serves as the SDN control plane, which is responsible for state collection and route calculation. This architecture aims to reduce the status synchronization overhead in large-scale and time-varying aeronautical networks. Due to the centralized management of the satellite control plane, the network status can be obtained in real-time and network resources can be flexibly deployed and allocated. To adapt to the requirement for quick routing decision-making by the control plane in large-scale and complex scenarios, this paper combines genetic algorithms [4] and uses the proposed telescope node to further simplify the network topology, meeting the requirements for fast routing decision-making and high-quality path updates in time-varying scenarios.

Compared with existing works, the main contributions of this paper can be summarized as follows:

- 1) We design a Sat-SDN assisted aeronautical network architecture, combining the multi-layer satellite network coverage capability with the high-quality and low-latency services of the aeronautical networks to provide a new approach for global device interconnection and communication.
- 2) Based on a reasonable domain partition mechanism, combined with the concept of dynamic inter-domain telescope nodes, we design an efficient and flexible inter-domain routing algorithm. By introducing genetic principles, the challenge of large routing calculation state space and difficult convergence is well resolved in large-scale multi-node network scenarios.
- 3) Using real aircraft data for experimental simulation, we verify the superior performance of the designed Sat-SDN Genetic-based Aeronautical Routing (SSGAR) algorithm, over other benchmark algorithms in various aspects.

The rest of this paper is organized as follows: Section II introduces the system model, Section III describes the inter-domain routing design, Section IV designs experiments and presents performance evaluation, and finally, Section V summarizes the paper and key contributions.

## 2 SYSTEM MODEL

### 2.1 System Structure

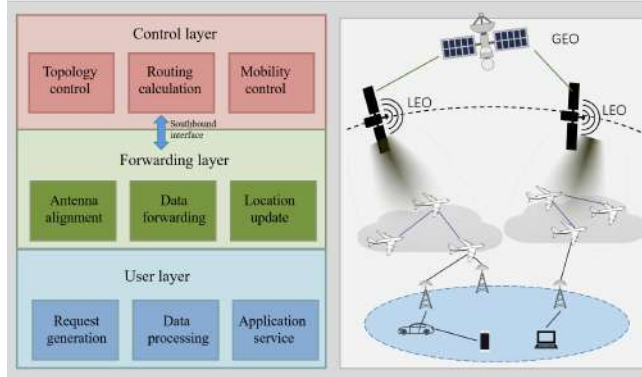


Figure 1: System Structure.

The Sat-SDN assisted aeronautical network architecture, as shown in Fig. 1, consists of three layers: the satellite network layer, aeronautical network layer and terrestrial layer. The satellite network layer is divided into two layers. The upper layer has SDN controllers on GEO satellites, who are responsible for managing and configuring the satellite network. While the lower layer is composed of low Earth Orbit (LEO) satellites that handle communication with the aircraft within their coverage. In the proposed architecture, each LEO satellite communicates with the aircraft within its coverage and reports their status and communication requests to the GEO satellites for synchronization. Once the GEO satellites receive communication requests from aircraft and calculate the route path, they use OpenFlow to distribute routing information to the aircraft nodes. The aircraft nodes then perform data forwarding based on the issued flow table. The status information of the aircraft is synchronized to its satellite node according to its flight reservation information and updated promptly when the flight trajectory changes. While the terrestrial layer is the customer of aeronautical networks, that generates data demand and searches for reachable aircraft nodes to submit the transmission task.

### 2.2 Domain Model and Scope Setting

Considering the fact that thousands of aircraft fly simultaneously in the existing aeronautical networks with a wide distribution range and a huge scale, single domain management is no longer applicable. Therefore, a domain partitioning mechanism, as shown in Fig. 2, is proposed as a multi-domain solution. We propose a partitioning strategy based on longitude and latitude to divide the aircraft into multiple domains. The telescope nodes are nodes that could directly connect to other domains, while ordinary nodes need to communicate with nodes in other domains through telescope nodes.

When constructing inter-domain routes in GEO, we ignore the ordinary connecting nodes within the domain and thus could simplify the graph by only keeping the connection among telescope nodes, which greatly reduces the computational complexity in large-scale scenarios.

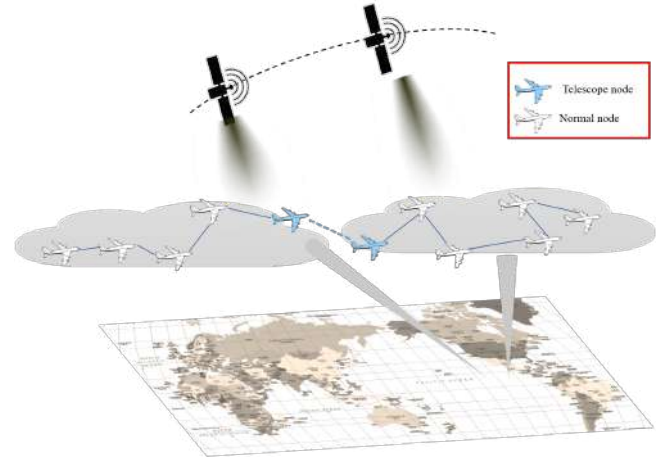


Figure 2: Domain Model and Telescope Node.

### 2.3 Communication Settings

The aeronautical networks are mobile ad-hoc networks, in which each aircraft can communicate directly with neighboring aircraft within a fixed distance threshold. They use millimeter wave technology to achieve high link capacity and meet the network transmission requirements of Air-to-Air (A2A) and Air-to-Ground (A2G) communication. Communication between aircraft and satellites can be achieved through microwave links. And since only a small amount of control information needs to be transmitted, the bandwidth requirement for the link is relatively small. The terrestrial network mainly consists of base stations and user devices. In cases where the terrestrial network cannot provide the network service, communication can be recovered through reachable aircraft and satellite network layers to connect users to far places. LEO satellites need to periodically communicate via optical links to update the belonging aircraft's status information. And all aircraft's latest status information is stored in the corresponding GEO SDN controller after being synchronized with other GEOs to construct a global information topology.

## 3 INTER-DOMAIN ROUTING POLICY

Intra-domain routing algorithms such as shortest path routing are relatively mature, and have been widely used to compute reachable paths in small-scale networks. Therefore this paper mainly optimizes long-distance multi-domain transmission scenarios. As for intra-domain routing, we choose the OLSR [7].

### 3.1 Telescope Graph Forming and Genetic Setting

As mentioned earlier, the large-scale characteristic greatly challenges the deployment of the routing algorithms. Thus, we first propose a topology simplification scheme based on our proposed

**Algorithm 1:** Genetic algorithm of SSGAR.

---

**Input:** The initial population of the selected paths  $POP$ , the maximum number of iterations  $loop_{max}$ , Maximum number of the population  $M$ , the mutation rate  $\epsilon$ .

**Output:** The best routing path  $P_{best}$  after several iterations.

```

1 for  $loop < loop_{max}$  do
2    $loop = loop + 1$ ;
3   for  $i (path) \in POP$  do
4     //Select shared gene pool;
5      $Pool_i \leftarrow []$ ;
6     for  $j (path) \in POP$  do
7       if  $i \neq j$  then
8          $Pool_i += Gen_{i,j}$ ;
9       end
10    end
11  end
12  //crossover and optimization;
13  random pick several  $k, pool_k$  pairs;
14  //for each  $k, pool_k$ ;
15  if  $random() > \epsilon$  then
16     $New_{path} \leftarrow Crossover(k, pool_k)$ ;
17  end
18  else
19     $New_{path} \leftarrow Mutations(k)$ ;
20  end
21  if  $New_{path} \notin POP$  then
22     $POP += New_{path}$ ;
23  end
24  if  $len(POP) > M$  then
25    Screen the best  $M$  individuals;
26  end
27 end
28  $P_{best} = \arg \min POP.hop$ ;
29 return  $P_{best}$ ;

```

---

telescope nodes. We extend the concept of neighbors, which include telescope pairs that can be connected to each other through ordinary nodes. By establishing a set of generalized neighbors for telescope nodes, we further remove ordinary nodes and build a connectivity graph for telescope nodes based on the preset generalized neighbor relationships. This will greatly simplify the network topology and reduces the cost of following route calculation. It is worth mentioning that, for source and destination nodes that not belong to the telescope node set, we temporarily change their attributes to telescope nodes when calculating their routing paths.

Fig. 3 shows two paths that start from the source of the route, pass through multiple telescope nodes, and finally reach the destination. Such paths are defined here as chromosomes. The connected inter-domain telescope pairs in the chromosome are considered genetic operators containing genetic information. In the subsequent chromosome crossover stage, just these genetic operators will combine and change, leading to the generation of offspring individuals.

During system initialization, a set of chromosomes, called a population, is randomly selected using the broadcast priority search tree algorithm. In order to avoid the curse of dimensionality caused

by the huge state search space, we limit the number of times each node can be visited. Even though we have used telescope nodes to simplify the network topology, the search space of reachable paths in actual aeronautical network scenarios is still extremely large. Therefore, it is necessary to limit the number that node can be visited when obtaining reachable paths to the destination, to feed the requirement that the system needs to quickly converge and make route decisions. However, this limitation may inevitably lead to the fact the selected path population sometimes does not contain the optimal paths. Therefore, in the following section, we design a genetic-based iteration and selection strategy to keep combining the high quality path genetic operators in the population and finally form the optimal routing path.

### 3.2 Evolution-based Inter-domain Routing Algorithm

In Fig. 3, an aircraft located in domain 1 ( $DM_1$ ) has two paths to connect to the user in  $DM_5$ . One path passes through the  $DM_1$ 's telescope node, then  $DM_2$ ,  $DM_3$ ,  $DM_4$ , and finally connects to  $DM_5$ . The other path passes through the  $DM_1$ 's telescope node, then  $DM_2$ ,  $DM_3$ ,  $DM_7$ , and finally connects to  $DM_5$ . Note that  $TN_{1,out}$  and  $TN'_{1,out}$  are different telescope nodes in the same domain.

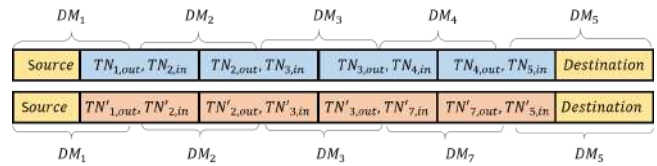


Figure 3: Initial Chromosome Pair.

The following crossover procedure of the two chromosomes is illustrated in Fig. 3 and Fig. 4. As shown in Fig. 3, these two paths have two common genetic operators, i.e., telescope pair of  $(DM_{1,out}, DM_{2,in})$ ,  $(DM_{2,out}, DM_{3,in})$ , even though the two paths may have different telescope nodes in the same genetic operator. The following crossover result of the two chromosomes is illustrated in Fig. 4 which takes the genetic operator  $(DM_{1,out}, DM_{2,in})$  as an example. This crossover procedure produces two new offspring.

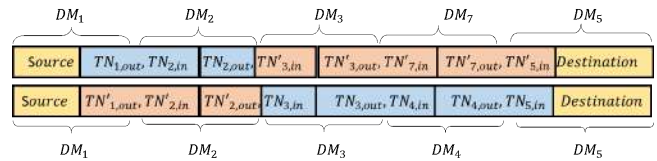


Figure 4: Hybridized Chromosome Pair.

To be specific, the GEO randomly initializes several paths containing multiple domains from the start domain to the destination as the initial population when getting the communication request. In order to find the optimal path, the GEO need to keep performing the crossover procedure of the population. During each crossover stage, one path will pick up random members in the selected population that has the same genetic operator with it, after which, the number of the population could be further increased. And following a stage of high-quality offspring selection will be organized. Some

individuals who have high quality paths in the population will be retained, while the rest will be discarded.

In addition to the crossover procedure, we also defined the mutation rate to represent the condition that a chromosome may discard part of its segments with a certain probability and choose a randomly connected path to the destination (or source) as a replacement. The specific genetic algorithm of SSGAR executed on GEO can be seen in Algorithm 1:

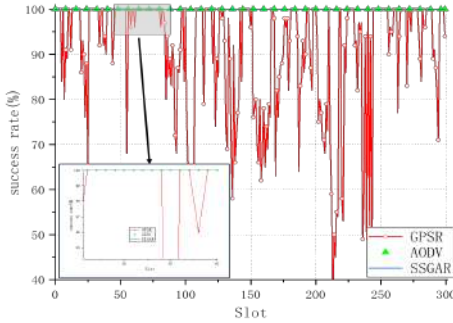


Figure 5: Forward Success Rate of SSGAR, AODV and GPSR.

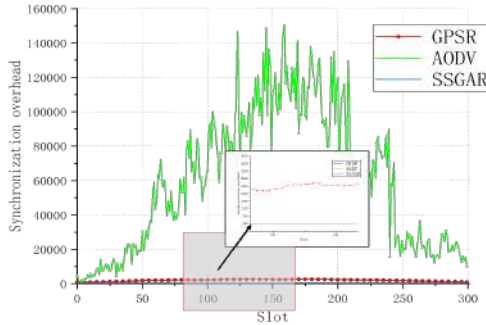


Figure 6: Synchronization Overhead of SSGAR, AODV and GPSR.

#### 4 PERFORMANCE EVALUATION

This paper selects real aircraft data to verify the performance of our proposed SSGAR in actual scenarios. We tracked the two-hour position trajectories of 408 aircraft in the central region of the United States as our data set, and further divided this region into four domains based on geographic location. We also set a GEO right above the middle region as the controlling and routing center. Data requests are randomly generated by users on the ground, and transmitted to the nearest aircraft. The comparison algorithm uses the GPSR [8] and AODV [10] mentioned above as the benchmark algorithm.

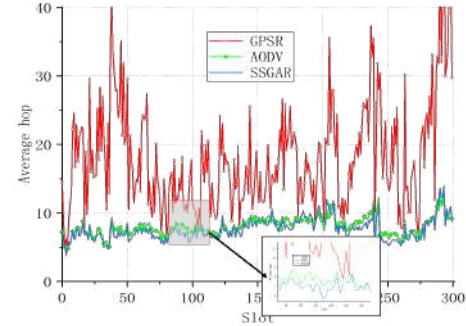


Figure 7: Average Hop of SSGAR, AODV and GPSR.

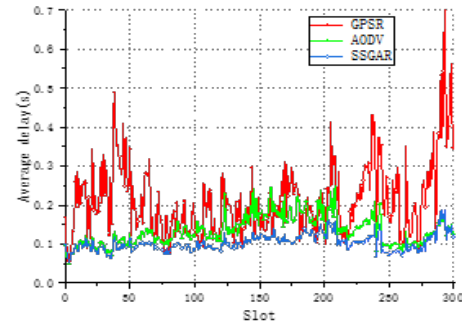


Figure 8: Average Delay of SSGAR, AODV and GPSR.

As shown in Fig. 5, we compare the forwarding success rate of GPSR, AODV and SSGAR algorithms in the aeronautical networks. As GPSR lacks global state information and can only perform greedy forwarding, the packets are often transmitted to blank areas or local optimal places, resulting in forwarding failure. In contrast, our proposed SSGAR algorithm can still achieve nearly 100% forwarding success rate while retaining the simplified inter-domain telescope graph.

Fig. 6 compares the difference in routing learning overhead between SSGAR and other benchmark algorithms. Due to the fact that AODV needs to broadcast learning packet to find the reachable path to the destination, it will produce a large communication overhead. While GPSR only needs to interact regularly with neighbors to obtain their location information for route calculation and data forwarding. As for SSGAR, nodes only need to perform simple data interaction with upper-level satellites to complete data forwarding, which greatly reduces the communication overhead.

In Fig. 7, we compare the average number of routing hops of SSGAR, GPSR and AODV during the simulation. It can be concluded that SSGAR has fewer average hops in complex environments than GPSR and AODV. This is because SSGAR simplifies the network topology and uses multiple iterations for optimal path selection, instead of relying on existing path information like AODV. Meanwhile, GPSR only uses location information for greedy forwarding,

often requiring more time and hops for route recovery when forwarding fails.

As shown in Fig. 8, the performance of GPSR, AODV and SSGAR in terms of communication delay is compared. The figure shows that SSGAR achieves the best performance compared to benchmark algorithms. This is because SSGAR does not need to perform frequent path recovery like GPSR or use a large number of broadcast packets to search for a path like AODV, which consumes too many bandwidth resources and increases total transmission time due to the queuing delay.

In summary, compared to benchmark routing algorithms, the proposed SSGAR routing algorithm can effectively reduce synchronization overhead, improve transmission success rate, and reduce transmission latency.

## 5 CONCLUSION

This paper proposes a satellite SDN-based solution to cope with the routing challenges in the next-generation integrated aeronautical networks. To reduce the synchronization overhead of large-scale nodes, we design an inter-domain telescope solution to simplify the inter-domain network topology, and propose an inter-domain routing strategy, named SSGAR, based on genetic algorithm. Simulation experiments demonstrate that, compared with benchmark routing algorithms, SSGAR can effectively reduce communication costs and improve transmission quality.

## REFERENCES

- [1] Kamal Benzekki, Abdeslam El Fergougui, and Abdelbaki Elbelrhiti Elalaoui. 2016. Software-defined networking (SDN): a survey. *Security and communication networks* 9, 18 (2016), 5803–5833.
- [2] Lionel Bertaux, Samir Medjah, Pascal Berthou, Slim Abdellatif, Akram Hakiri, Patrick Gelard, Fabrice Planchou, and Marc Bruyere. 2015. Software defined networking and virtualization for broadband satellite networks. *IEEE Communications Magazine* 53, 3 (2015), 54–60.
- [3] Luca Boero, Roberto Bruschi, Franco Davoli, Mario Marchese, and Fabio Patrone. 2018. Satellite networking integration in the 5G ecosystem: Research trends and open challenges. *IEEE Network* 32, 5 (2018), 9–15.
- [4] Jingjing Cui, Halil Yetgin, Dong Liu, Jiankang Zhang, Soon Xin Ng, and Lajos Hanzo. 2021. Twin-component near-Pareto routing optimization for AANETs in the north-Atlantic region relying on real flight statistics. *IEEE Open Journal of Vehicular Technology* 2 (2021), 346–364.
- [5] Xiaojing Huang, J Andrew Zhang, Ren Ping Liu, Y Jay Guo, and Lajos Hanzo. 2019. Airplane-aided integrated networking for 6G wireless: Will it work? *IEEE Vehicular Technology Magazine* 14, 3 (2019), 84–91.
- [6] Marianna Ivashina, Artem Vilenskiy, Hsi-Tseng Chou, Joachim Oberhammer, and M. Ng Mou Kehn. 2021. Antenna Technologies for Beyond-5G Wireless Communication: Challenges and Opportunities. In *2021 International Symposium on Antennas and Propagation (ISAP)*. 1–2.
- [7] Philippe Jacquet, Paul Muhlethaler, Thomas Clausen, Anis Laouiti, Amir Qayyum, and Laurent Viennot. 2001. Optimized link state routing protocol for ad hoc networks. In *IEEE International Multi Topic Conference, (INMIC)*. 62–68.
- [8] Brad Karp and Hsiang-Tsung Kung. 2000. GPSR: Greedy perimeter stateless routing for wireless networks. In *Annual international conference on Mobile computing and networking*. 243–254.
- [9] Jiajia Liu, Yongpeng Shi, Zubair Md Fadlullah, and Nei Kato. 2018. Space-air-ground integrated network: A survey. *IEEE Communications Surveys & Tutorials* 20, 4 (2018), 2714–2741.
- [10] Elizabeth M Royer and Charles E Perkins. 1999. Ad hoc on demand distance vector routing. In *IEEE Workshop on Mobile Computing Systems and Applications*, Vol. 2. 90–100.
- [11] Shen Wang, Rihui Li, Yongjian Han, and Ming Yao. 2021. Opportunities and Challenges of Antenna Design for Future 5G Mobile Terminals. In *2021 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)*. 115–116.
- [12] Jian Yang, Kaixuan Sun, Huasen He, Xiaofeng Jiang, and Shuangwu Chen. 2022. Dynamic virtual topology aided networking and routing for aeronautical ad-hoc networks. *IEEE Transactions on Communications* 70, 7 (2022), 4702–4716.
- [13] Michael J Zernic, Lawrence C Freudinger, and A Terry Morris. 2001. Aeronautical Satellite Assisted Process for Information Exchange through Network Technology (Aero-SAPIENT) Project: The Initial Trials. In *2001 IEEE Aerospace Conference Proceedings*, Vol. 3. 3–1375.
- [14] Jiankang Zhang, Taihai Chen, Shida Zhong, Jingjing Wang, Wenbo Zhang, Xin Zuo, Robert G Maunder, and Lajos Hanzo. 2019. Aeronautical Ad Hoc networking for the Internet-above-the-clouds. *Proc. IEEE* 107, 5 (2019), 868–911.

# Mathematical models of colony population dynamics and hive placement

Zixuan Zhang  
Shenzhen College of International  
Education  
zxx.zhang@outlook.com

Dongyi He  
The Stony Brook School  
donyi.he@sbs.org

Hanwen Zhang  
Georgetown Preparatory School  
hzhang@gprep.org

## ABSTRACT

Animal pollinators have been supporting the lives of human beings on Earth. Bee pollinators are the biggest contributors to the pollination of crops, providing humans with food. Given such circumstances, this paper investigates the population of honeybee colonies and the processes of bee pollination. We constructed Honeybee Colony Population Model (BCPM) to predict the population of a honeybee colony over time. We first outlined the life cycle of a honeybee, including eggs, larval stage, pupal stage, and adult bee stage. Within the adult bee stage, bees transition back and forth between foragers and hive bees depending on the number of available resources and the workload of nursing tasks. By listing out factors that affect the population in each stage, we established equations representing the rate of change in each of the stages of a honeybee's life cycle, as well as an equation describing the change in resource storage. We also evaluated the death rate and the resources in each month of the year and calculated each group's typical maximum, minimum, and mean population in a honeybee colony: 3 years after the establishment of the colony, the total adult population follows a seasonal change with recurring patterns each year, giving a maximum of 100862 bees and a minimum of 35676 bees. The annual average population is found to be 64877 bees. We then conducted a sensitivity analysis on BCPM and found that the initial number of bee hives and the initial amount of available resources have the most significant impact on the population of the colony. We also observed an unusual pattern in the cross-analysis of the two factors and constructed Simplified Colony Collapse Disorder Model (SCCDM) to predict whether a colony will collapse using only one equation. In response to estimate the number of hives needed to support the pollination of a specific land area, we constructed Hive Deployment Model (HDM). We first divided the land into 20 nodes and then found the most appropriate locations to place the hives. After establishing the equations for movements between nodes per day per forager group, we developed an iterating algorithm to find the number of hives needed to pollinate crops on 20 acres of land. We collected data for 9 typical bee-pollinated plants and found the number of hives needed for each type of plant based on

the algorithm, with blueberries being the most demanding, requiring 83 hives, whilst apples and roses only required 2 hives at the other end of the spectrum. Then, we established a sensitivity analysis to ensure the stability of the model by changing two arbitrary parameters. Finally, we discussed the potential advantages and disadvantages of our model. We have also created a non-technical blog that summarizes our investigation, presenting our results in a simplified way

## CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation; Model development and analysis; Modeling methodologies.

## KEYWORDS

Colony population dynamics, hive placement, mathematical method

### ACM Reference Format:

Zixuan Zhang, Dongyi He, and Hanwen Zhang. 2023. Mathematical models of colony population dynamics and hive placement. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3590003.3590070>

## 1 INTRODUCTION

Primarily consisting of the queen, drones, and worker bees, one beehive can contain tens of thousands of bees. Bees demonstrate impressive communicative abilities for a group of this size, with pheromones acting as an important social glue that binds the colony together. In particular, the Queen mandibular pheromone (QMP) released by the queen bee regulates the social behavior of the hive, including swarming and mating, insofar as to facilitate the inhibition of ovary development in worker bees [1]. Other communication tools, such as bee dances, are also a testament to bee colonies' complexity.

These intelligent creatures pollinate one-third of the human diet, feeding billions of people all over the world [2]. Nowadays, more than 90 crops are dependent on bee pollination, including apples, squash, broccoli, and many more. In addition to humans, bee pollination also helps feed 80% of all the birds in the US. Bees' value also translates into economic profits. In the US alone, bees contribute \$20 billion to the agricultural sector every year, forming an indispensable component of the global economy.

However, the status quo of bees is nothing short of grim—the population of bees peaked in 1947 and has been declining since; nearly one in every ten wild bee species face extinction. Every year, two out of three beekeepers in the US lose 40% of their bee colonies due to compounded reasons, including diseases, climate

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China  
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590070>

change, and habitat destruction. A study published in Science found that around 50% of the 109 bee species investigated had been lost, and the remaining species see a decline in their pollination ability. Therefore, it is vital that we, the direct beneficiaries of bees, develop a sound understanding of the pollination capabilities and population trends of bee colonies.

We are asked to construct a mathematical model (Model 1) that estimates the population over time of one honeybee colony. Because bees go through three developmental stages — egg, larva, and pupa — we should take into consideration the time each stage takes to determine the rate at which a bee hive produces new bees. We should also take into consideration the initial amount of resources that the bees can access so that there is a starting point for future projections.

We should conduct sensitivity analysis on our Model 1 and thus determine the change of which initial variables lead to the most significant change in the average population when the colony is in a stable stage.

We should construct another mathematical model (Model 2) which estimates the minimum number of hives needed to support pollination in a 20-acre parcel of land. Various factors, such as pollination efficiency and changes in the number of flowers, should be taken into account, and data from Model 1, such as bee population, can be reused in this model. Then, we should test our Model 2 on some plant species to yield results.

Based on the data we acquired in the previous questions, we should put together a straightforward blog that beekeepers or farmers could use as a potential guide. In the blog, we should point out the factors that have the most influence over the bee population and give our estimate of the approximate hive numbers needed to cover the 20-acre land.

## 2 ASSUMPTIONS AND JUSTIFICATIONS

Assumption 1: All eggs are hatched after 3 days they are laid. The larval stage lasts for 6 days, and the pupal stage lasts for 10 days.

Justification 1: Typically, an egg laid by the queen will be hatched in 3 days, the larval stage lasts for approximately 5.5-6.5 days, and the pupal stage lasts for approximately 7.5-14.5 days. For the purpose of constructing a model, we simplify this premise by taking averages of the duration in the larval and the pupal stage, which will hardly have an impact on our results since the bee population is large.

Assumption 2: 50% of foragers die every day if no resource is available.

Justification 2: To simulate the effect of an exponential decrease in foragers, we set a rate of decline in population each day when the storage of resources falls below 0. We do not expect an immediate collapse of the colony, nor do we want to trivialize the impact of this shortage. Therefore, we set the rate of death to 50% to imitate a temperate decline in the population but preserve the colony's chances of recovering.

Assumption 3: The 20-acre farmland has a rectangular shape consisting of 20 smaller 1-acre squares arranged in a  $4 \times 5$  pattern.

Justification 3: Since we do not have any information about the shape of the farmland, we assume that it is close to a typical rectangular farmland. For the purpose of our model, we divided the

**Table 1: Variables Used in BCPM**

Symbol	Description
$t$	Time from start of model
$n_H$	Number of hive bees
$n_E$	Number of eggs
$n_B$	Number of larva and pupa
$n_F$	Number of foragers
$R$	Mass of available resources
$c$	Rate of resource collection
$\theta$	Rate of resource consumption
$k$	Rate of forager death
$r$	Rate of transition to foraging
$l$	Maximum egg laying rate
$\phi$	Maximum eclosion rate
$L(t)$	Egg laid per day
$E(t)$	Daily eclosion number
$F(t)$	Net forager transition rate

farmland into 20 nodes, each a 1-acre square, so that it is easier to calculate the distance between farmlands. The model can be altered slightly so that it can be applied to other shapes of farmlands since the process of HDM is the same.

Assumption 4: For each forager, 25 visits are made to each acre every time. Besides, foragers are assigned in groups of 200 and operate together.

Justification 4: The number of visits and foragers is so huge that it will strongly affect our model and undermine the importance of other factors in our model due to time constraints. Considering the visits and forager behaviors as a group not only helps us solve the problem of time constraints but also has little impact on the final data we get from the model because we are simply expecting all the groups to operate at an average efficiency. The fluctuations of individuals will not influence the bees' overall behavior.

## 3 HONEYBEE COLONY POPULATION MODEL (BCPM)

### 3.1 Variables

The table of variables used in this model are shown in Table 1.

### 3.2 Typical Honeybee Lifecycle

To establish a model for the population of a honeybee colony, we first draw an outline of the life cycle of a typical honeybee. The queen lays eggs, which will be taken care of by the worker bees (hive bees), and the eggs are hatched within 3 days, followed by the larval stage which lasts for 6 days, and the pupal stage which lasts for 7-14 days as shown in Table 1. The bees then enter the adult stage [9].

After the bees fully mature, they become worker bees, responsible for nursing tasks in the hives [9]. However, they can transition into foragers, focusing on resource collection and pollination. The reverse transition can also occur when there is more need for nursing than foraging [10].

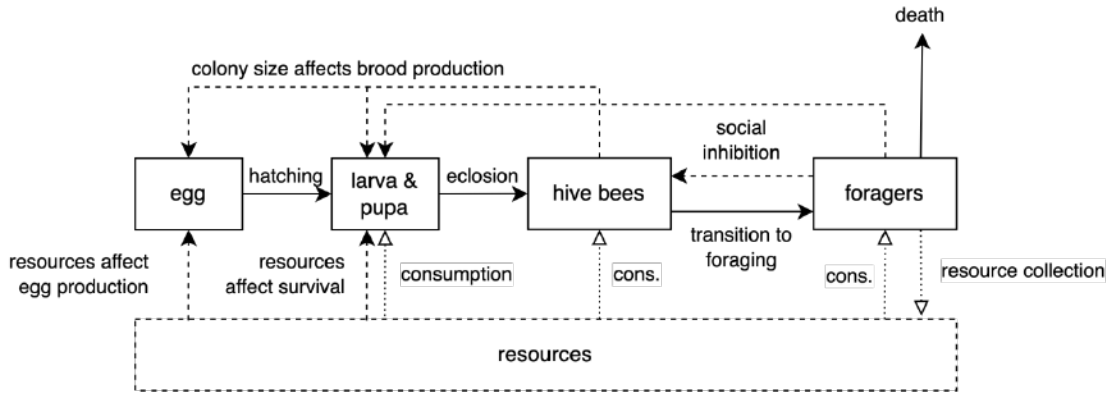


Figure 1: Graphical Representation of a Typical Honeybee Lifecycle

Factors such as resource availability and beehive population can alter the development cycle, as shown in Figure 1.

### 3.3 Mathematical Modeling of Population

We model the population trend by first considering the change in the population of each stage. Assuming that all eggs are hatched after 3 days they are laid, we can derive:

$$\frac{dn_E}{dt} = L(t) - L(t-3)$$

The laying rate per day  $L(t)$  has a positive correlation to the number of available resources, but the effect reaches a maximum when the laying rate reaches a maximum for the queen. Similarly, the laying rate also positively correlates to the number of hive bees in the colony, but the effect reaches a maximum when there are excess workers. Therefore, we can construct the following equation:

$$L(t) = l \cdot \frac{R_t}{R_t + \varepsilon} \cdot \frac{n_{Ht}}{n_{Ht} + w}$$

where  $w$  and  $\varepsilon$  are arbitrary parameters that control the dependency of egg laying on the number of resources and the number of hive bees, respectively.

We then consider the rate of change of population in the larval and the pupal stages. The rate is equivalent to the hatching rate deducted by the eclosion rate, as presented in Figure 1. We estimate that the time taken between hatching and eclosion is 16 days, which means that eclosion happens on the 19th day after an egg is laid. Thus, the following equation can be constructed:

$$\frac{dn_B}{dt} = L(t-3) - L(t-19)$$

where  $L(t-19)$  is derived from  $L((t-3)-16)$ .

Next, we try to model the rate of change of hive bees. The increase in the number of hive bees each day is equivalent to the eclosion rate from newly grown pupa subtracted by the net forager transition rate. The eclosion rate is dependent on the quantity of essential resources and the number of workers that take care of them during larval and pupal stages. We define the hive-bee population rate and eclosion rate as follows:

$$\frac{dn_H}{dt} = E(t) - F_R(t)$$

$$E(t) = \frac{R_t}{R_t + \alpha} \cdot L(t-19) \cdot \frac{n_{Ft} + n_{Ht}}{n_{Ft} + n_{Ht} + \beta}$$

where the remaining population of  $L(t-19)$  are assumed to be dead.  $\alpha$  and  $\beta$  are arbitrary parameters that determine the relevance of the factors of resources and population, respectively, on broods' survival.

The net forager transition rate is the rate of transition to foragers subtracted by the rate of inhibition, as shown in Figure 1. We represent this relationship in the following equation:

$$F_R(t) = n_{Ht} \cdot \left( r_{\min} + r \left( \frac{\gamma}{\gamma + R_t} \right) - \sigma \left( \frac{n_{Ft}}{n_{Ft} + n_{Ht}} \right) \right)$$

where  $\gamma$  and  $\sigma$  are arbitrary parameters that control the sensitivity of our model towards resources and forager-hive bees ratio, respectively.  $r_{\min}$  is the rate of forager recruitment when there are maximum resources but no foragers yet. This can be concluded when  $R_t \rightarrow \infty$  and  $n_{Ft} = 0$ :

$$F_R(t) = n_{Ht} \cdot (r_{\min} + r(0) - \sigma(0)) = n_{Ht} \cdot r_{\min}$$

The rate of increase in the population of foragers is modelled as the net forager transition rate deducted by the death rate, which is shown as follows:

$$\frac{dn_F}{dt} = \begin{cases} n_{Ht} \cdot F_R(t) - k_r k_{\max} k_s n_{Ft} & \text{if } R_{t-1} + \Delta R_t \geq 0 \\ n_{Ht} \cdot F_R(t) - \frac{1}{2} n_{Ft} & \text{otherwise} \end{cases}$$

where  $k_r$  is a random real number between  $[0.75, 1.25]$ .

The first condition is satisfied when there are still resources to meet the basic survival of foragers, so there is a natural death rate of  $k_r k_{\max} k_s$ .  $k_r$  simulates the random factors the foragers may encounter which affect their lifespan, fluctuating the daily death rate by  $\pm 25\%$ .  $k_s k_{\max}$  is the seasonality factor, which determines the death rate based on typical patterns observed throughout the year.

When no resource is available, we set the forager death rate to 0.5 to simulate an exponential decline in population.

Lastly, we establish an equation to describe the change in resources over time. The resource is collected by foragers but consumed by larvae, pupa, hive bees, and foragers. Therefore:

$$\frac{dR}{dt} = c_r c_{\max} c_s n_{Ft} - \theta_B n_{Bt} - \theta_H n_{Ht} - \theta_F n_{Ft}$$

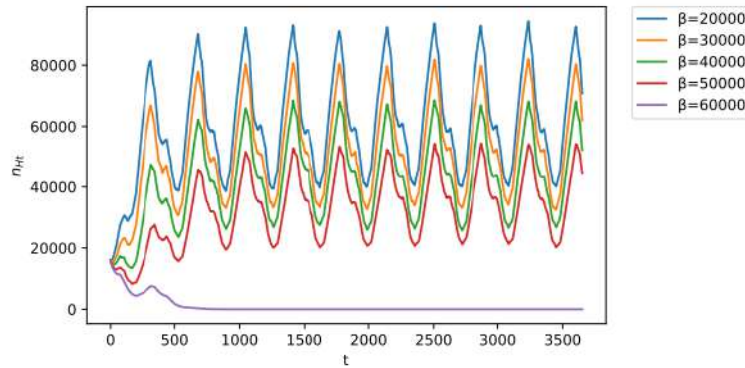
where  $c_r$  is a random real number between  $[0.75, 1.25]$ .

**Table 2: Table of  $k_s$  and  $c_s$  values**

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
$k_s$	0.29	0.29	0.54	0.79	0.68	0.57	0.78	1.00	0.83	0.67	0.29	0.29
$c_s$	0.03	0.03	0.48	0.93	0.97	1.00	0.60	0.19	0.32	0.45	0.25	0.10

**Table 3: Table of BCPM Results for Various  $\beta$  Values over 10 Years**

$\beta$ Value	Maximum $n_{Ht}(t > 1000)$	Minimum $n_{Ht}(t > 1000)$
20000	93635	38473
30000	82591	32149
40000	69531	25930
50000	54743	19674
60000	1	0

**Figure 2: BCPM Results for Various  $\beta$  Values Plotted over 10 Years**

Similar to the seasonality factor for death rate,  $c_s c_{\max}$  is the rate of resource collection subject to seasonal change.  $c_r$  is used to simulate the fluctuating resource collection by  $\pm 25\%$  due to unseen random factors.

### 3.4 Parameter Determination

We set the maximum laying rate of the queen  $l = 2000$ , in accordance with real-world observations and research [11]. According to Harbo [12], a brood typically requires 163 mg of resources to develop into an adult, which is equivalent to 10.2 mg per day over 16 days; an adult bee requires 6.7 mg resources per day. Thus, we take  $\theta_B = 0.0102$  and  $\theta_F = \theta_H = 0.0067$ .

We set  $r_{\min} = 0.25$  and  $\sigma = 0.75$  to assume that hive bees will only transform into foragers 4 days after their eclosion, and set  $r = 0.25$  to double forager recruitment when there is a shortage of resources, which is consistent with Khoury et al.'s suggestions. [13]

We also set  $w = 5000$  to represent the effect that half of the eggs fail to hatch when only 5000 hive bees take care of them. Then, we set  $\alpha = \gamma = \varepsilon = 1000$  to show the decreased effect when total resources stored exceeds 1 kg, as  $\frac{R_t}{R_t + \alpha} > \frac{1}{2}$ ,  $\frac{R_t}{R_t + \varepsilon} > \frac{1}{2}$ , and  $\frac{\gamma}{\gamma + R_t} < \frac{1}{2}$  when  $R_t > 1000$ .

According to Russell et al., the death rate  $k$  of foragers and the resource collection rate  $c$  is normalised to give maximum values  $k_{\max} = 0.102$ ,  $c_{\max} = 0.099$  and rates  $k_s$ ,  $c_s$  per month, as shown in Table 2.

Lastly, we conducted several tests using initial values introduced on our model to determine a reasonable  $\beta$  value. According to the results shown in Table 3 and Figure 2, since a typical honeybee hive contains between 20,000 and 80,000 honeybees, we observe that  $\beta = 40000$  fits our need for this model.

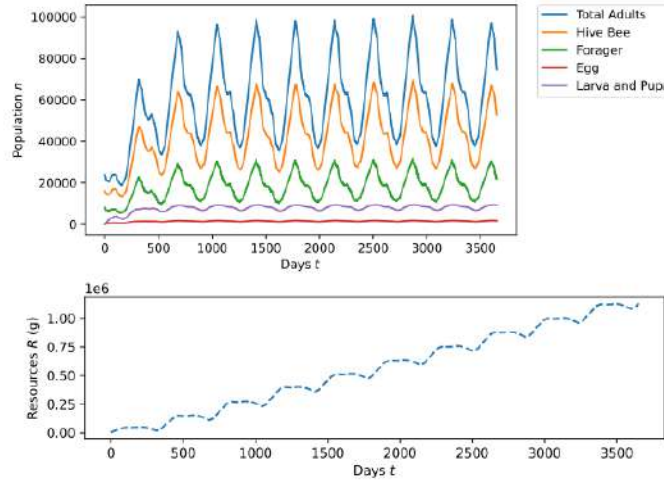
### 3.5 Typical Model Results

We define several initial values that represent a typical start-up phase of a honeybee colony. Initial eggs, larva and pupa number  $n_{E0} = n_{B0} = 0$  since the colony has not yet created any brood. Initial hive bee number  $n_{H0} = 16000$  and initial forager number  $n_{F0} = 8000$ , which add up to give a population slightly over the minimum of 20,000 and in an approximate ratio of 2 : 1. Initial resource store  $R_0 = 5000$  to imitate the situation of basic storage at start-up.

We also set the starting as the 90th day of the year (in spring), when there are enough resources to support the colony's survival

**Table 4: Table of Typical BCPM Results**

Name	Symbol	Maximum	Minimum	Average
Total (adult) population	$n$	100862	35676	64877
Hive bee count	$n_{Ht}$	69459	25277	45313
Forager count	$n_{Ft}$	31737	10247	19622
Egg count	$n_{Et}$	1780	1190	1544
Larva and pupa count	$n_{Bt}$	9468	6392	8234

**Figure 3: Typical BCPM Results**

and growth, and this is the typical period when broods start to form and develop [15]. The model is evaluated for 10 years ( $0 \leq t \leq 3650$ ) to give the results in Table 4 and Figure 3. All maximum, minimum and average values are calculated when  $t > 1000$ .

We observe a cyclical fluctuation in population and resource growth that occurs every year, which is reasonable due to the periodic seasonal death rate  $k_{\max}k_s$  and resource collection rate  $c_{\max}c_s$ .

### 3.6 Sensitivity Analysis

**3.6.1 Sensitivity Analysis on Parameters.** Firstly, we change the arbitrary parameters  $r_{\min}$ ,  $r$ ,  $\sigma$ ,  $\alpha$ ,  $\gamma$  and  $\varepsilon$  (the sensitivity analysis for  $\beta$  has been conducted by  $\pm 10\%$  to test the consistency of the repeating pattern shown by our model. The results are shown in Figure 4 and Figure 5.

We observe a generally consistent trend in all tests, so we consider our choices of parameters to be reasonable and stable overall.

**3.6.2 Sensitivity Analysis on Initial Values.** In order to identify the factors that have the most significant effect on the honeybee colony population, we conduct multiple simulations using our model while tweaking the initial values.

We notice that the initial resource stored  $R_0$  and initial hive bee population  $n_{H0}$  significantly affect the population compared to other factors: the honeybee population shrinks and eventually

collapses when the values of these two factors decrease. Thus, we perform another sensitivity analysis based on both factors changing simultaneously, where  $0 \leq R_0 \leq 10000$  and  $0 \leq n_{H0} \leq 32000$  for  $200^2 = 40000$  possible combinations for the two factors. We eliminated the effect of random terms  $k_r$  and  $c_r$  to reach a smoother graph in Figure 6.

We observe an unusual pattern– the honeybee population either maintains a relatively stable average number of around 60000 or the colony collapses completely. Therefore, we now construct an extra model which determines whether a colony collapses.

### 3.7 Simplified Colony Collapse Disorder Model (SCCDM)

We first set colonies with an average population below the threshold of  $n = 30000$  to collapse; otherwise, the average population is set to 60000. Dimension reduction is also performed to fit the relationship into a planar equation. The process is shown in Figure 7.

We now trace the edge shown in Figure 7 (right) by selecting the minimum red  $n_{H0}$  value for each  $R_0$ . We can then fit the following equation to the curve using non-linear least squares:

$$R_0 = ae^{-bn_{H0}} + c$$

where  $a$ ,  $b$  and  $c$  are constants to be determined. The results are plotted in Figure 8.

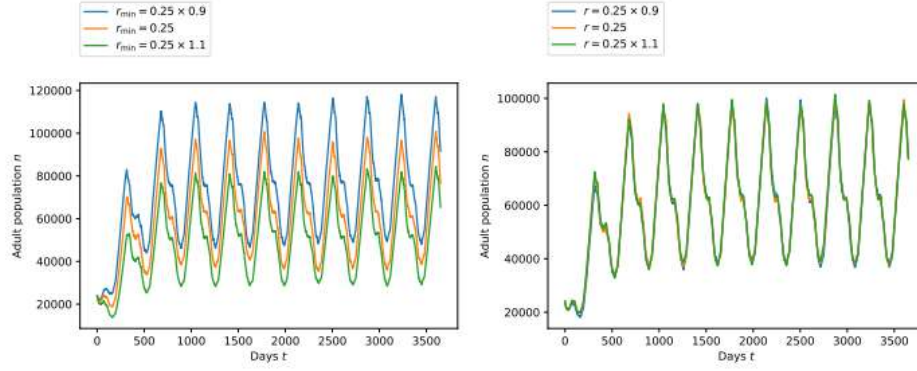


Figure 4: Sensitivity Analysis on Arbitrary Parameters

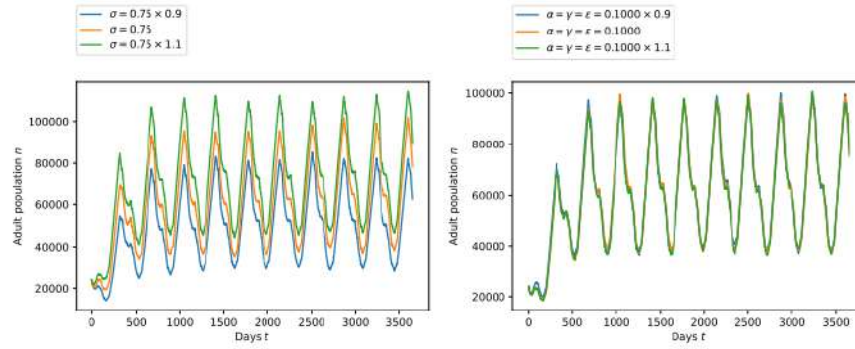
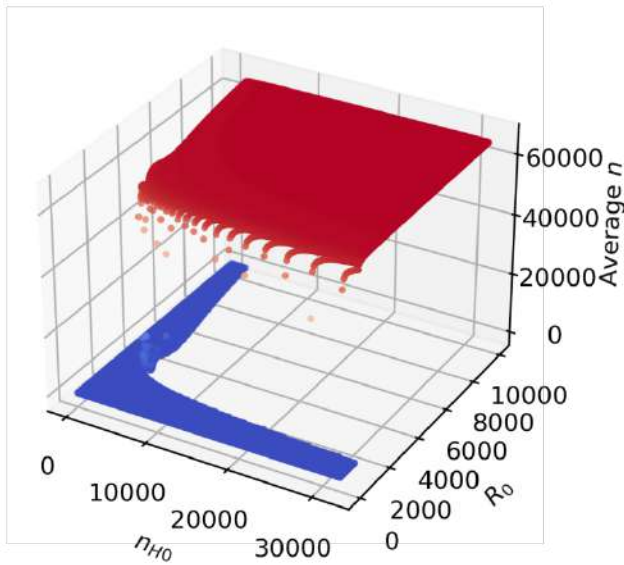


Figure 5: (cont.) Sensitivity Analysis on Arbitrary Parameters

Figure 6: Sensitivity Analysis on  $n_{H0}$  and  $R_0$ 

Therefore, the evaluated constants, together with the rearranged fitted equation, state that a honeybee colony will only thrive when

$$R_0 - a \exp(-bn_{H0}) > c$$

where  $a = 2.86660147 \times 10^4$ ,  $b = 5.3785166 \times 10^{-4}$  and  $c = 9.224787 \times 10^2$ .

## 4 HIVE DEPLOYMENT MODEL (HDM)

### 4.1 Model Establishment

We first attempt to establish the model through a simulation approach. Firstly, we model the 20-acre farmland into 20 nodes on an undirected, weighted graph, as shown in Figure 9. The number marked on each node is the remaining flowers waiting to be pollinated, and the number on each edge marks the cost. We denote the bottom-leftmost node as  $m_{00}$ , and the upper-rightmost node as  $m_{34}$ .

The model is evaluated multiple times for hive number  $H \geq 1$  until the number of hives can satisfy the pollination requirements. For each choice of  $H$ , we consider the "best" locations to place the hives so that the sum of distances from each node to their respective nearest hive  $\zeta$  is minimal. This relationship is expressed

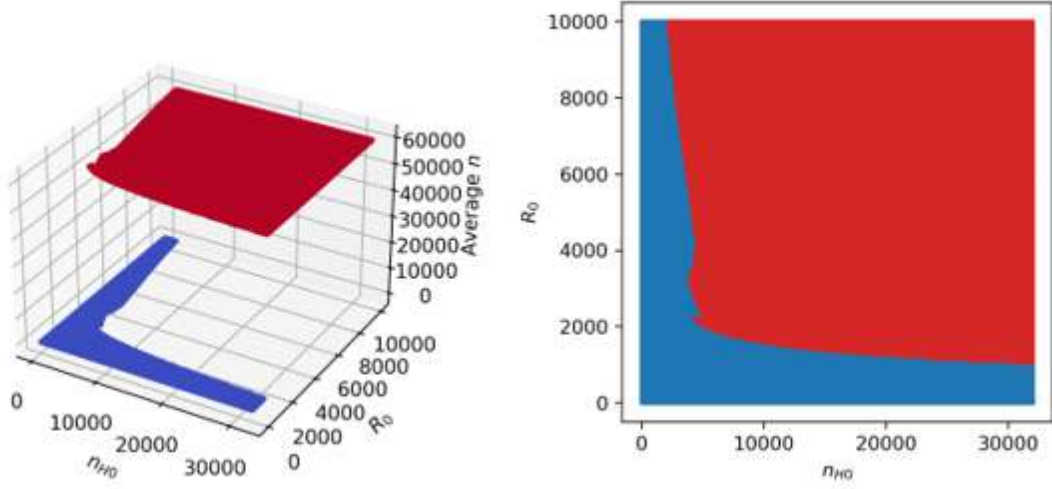


Figure 7: Thresholding and Dimension Reduction on Figure 6

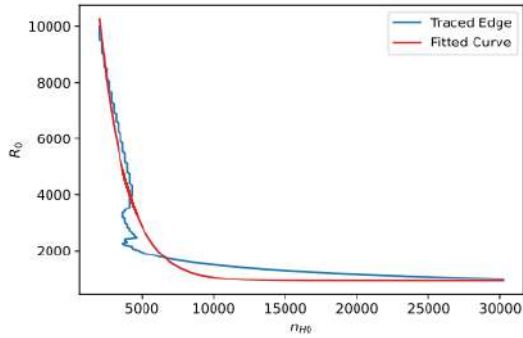


Figure 8: Fitted Function for SCCDM

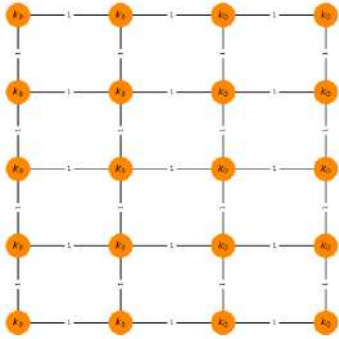


Figure 9: Initial construction of Graph

as the following equation:

$$\zeta = \sum_{x=0}^3 \sum_{y=0}^4 \min \{ |x - x_h| + |y - y_h| \mid (x_h, y_h) \in \mathbb{H} \}$$

where  $\mathbb{H}$  is the set of all hive locations. For  $1 \leq H \leq 6$ , we calculate  $\zeta$  for all possible  $\binom{20}{1} + \binom{20}{2} + \dots + \binom{20}{6} = 60459$  permutations to find out the best arrangement of the hives, where all nodes are represented in the structure same as Figure 9, and a blue circle indicates the presence of a hive. For  $7 \leq H \leq 20$ , an extra hive is put in a random location without a hive according to the arrangement of  $H - 1$ , since  $\zeta$  reaches  $20 - H$  in this range, and every extra hive only contributes a decrease of 1 in  $\zeta$  due to the decrease in distance of a current node from 1 to 0.

In order to reduce the time complexity of our model, we assume that  $p = 25$  visits are made to each node each time. We also group the foragers into groups of  $g = 200$  in which group members operate together. We find the total number of movements between nodes per day  $f = \lfloor v/p \rfloor$ , where  $v$  is the maximum flower visits per forager per day. Therefore, for each day  $t \in [t_0, t_0 + \Delta t]$  where  $t_0 = 365 \times 3 + t_s$ , we consider the followings for each hive at  $m_{x_h} y_h$ :  $d = \lfloor n_{Ft}/g \rfloor$  groups of foragers are formed and are all placed at  $m_{x_h} y_h$ . the number of flowers that require pollination increases by  $\frac{1}{50}(k_0 - k)$ , assuming that over 80% of flowers will require pollination again in 90 days if all flowers are pollinated.

For each group of foragers, all costs (weights on edges) are reset to 1 and steps 3-5 are repeated for  $fc_s$  times.  $c_s$  is the seasonality factor used in BCPM for simulating changing resource collection rates and is therefore used to represent the change in pollination due to seasonal changes such as temperature and weather.

The group pollinates  $p_r = g \cdot \frac{k}{k_0} \cdot p \cdot a \cdot s \cdot r$  flowers in current node, where  $\frac{k}{k_0}$  is the possibility that a randomly chosen flower is not pollinated. This factor shows the effect that a honeybee is more likely to visit a pollinated flower if the proportion of pollinated flowers is already large, decreasing successful pollination rates. The tiredness index is used to address the effect of falling pollination rate due to working strain during a single day. The distance attenuation  $s = \frac{1}{1+s_0^2}$  shows halved pollination rate when

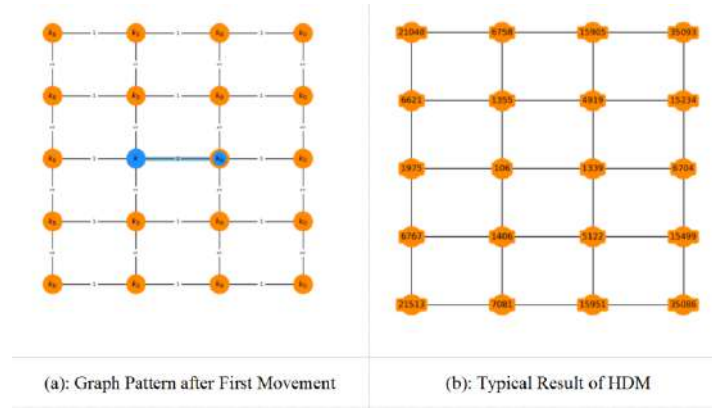


Figure 10: Results of HDM

Table 5: HDM Results for Various Species

Species	$k_0$	S [18]	$\Delta t$ [18]	Hives needed	Hives / acre
Alfalfas	266667	165	40	11	0.55
Almonds	860000	30	45	21	1.05
Apples	14400	135	9	2	0.10
Avocados	32000	90	60	3	0.15
Blueberries	3333333	97	49	83	4.15
Lilies	180000	157	35	9	0.45
Onion	80000	120	30	5	0.25
Roses	150000	150	90	10	0.50
Sunflowers	25000	200	84	2	0.10

the foragers are operating at neighbouring nodes from the hive, and gives an increasing effect as the distance increases. Since bees have a preference to work within 300 feet from their hive.

Figure 10 shows a typical pattern after simulating steps 1-6 from the initial graph in Figure 9, without any repetition involved. The blue circle indicates the hive location, the blue lines indicate the path the group has just passed, and the blue-in-orange circles indicate the resulting location of the group.

Following this algorithm, the number of yet-to-be-pollinated flowers on each node can be found for one specific number of hives. We repeat this algorithm with an increasing number of hives until we reach  $< 1\%$  of total not pollinated crops in the 20-acre area, and  $< 5\%$  of not pollinated crops in each individual node. If these conditions are not satisfied even when  $H = 20$ , we multiply the population in each hive by a factor  $q$  until the conditions are met, where we would conclude that  $H = 20q$  hives are needed.

As an example, for arbitrary parameters  $H = 1, v = 600, t_s = 90, \Delta t = 60$  and  $k_0 = 150000$ , we obtain a reasonable distribution of remaining not pollinated flowers, where the number increases as the distance from hive (1, 2) increases. In this stage, the hive is not considered to support pollination of the farmland, since the maximum  $k$  at (3, 4) is 35093, which is a number far larger than  $150000 \times 0.05 = 7500$ . Moreover, the total number of not pollinated flowers is 225482, and it is still far larger than the requirement  $150000 \times 20 \times 0.01 = 30000$ .

## 4.2 Evaluation and Results

We first set the maximum daily flower visits per forager  $v = 600$ , which is a crop species-inspecific parameter and is held constant across different evaluations of our model. Since a bee colony can pollinate 20 million flowers per day at maximum [17], we divide the number by the typical daily maximum forager count  $n_{F_{\max}} \approx 30000$ . Thus, it gives individual maximum flower visits  $v = 2 \times 10^7 \div 30000 = 600$ . Then, we choose 9 widely seen bee-pollinated plant species to consider their respective  $s$ ,  $\Delta t$  and  $k_0$ . The results are shown in Table 5.

## 4.3 Sensitivity Analysis

Lastly, we perform sensitivity analysis to ensure the stability and reliability of our model's outcomes. We alter the arbitrary variables  $r$  and daily increase in not pollinated flowers by:

Changing the lower bound for  $r$  by  $\pm 10\%$ , so that  $r \in [0.75 \pm 0.075, 1]$ .

Changing the factor of daily increase in not pollinated flowers by  $\pm 10\%$ , so that the increase becomes  $(\frac{1}{50} \pm \frac{1}{500})(k_0 - k)$ .

We randomly change the two variables in the  $\pm 10\%$  range for 50 random combinations, then test and compare the outcomes on lillies, whose  $k_0 = 180000, s = 157, \Delta t = 35$  and require 10 hives according to previous results. We notice that all tests evaluate an outcome of  $H = 9$ . Therefore, we consider the results to be stable and less prone to the arbitrary parameters mentioned above.

## REFERENCES

- [1] Jarriault, D., & Mercer, A. R. (2012). Queen mandibular pheromone: questions that remain to be resolved. *Apidologie*, 43(3), 292–307.
- [2] Carrington, D. (2013, February 28). Loss of wild pollinators serious threat to crop yields, study finds. *The Guardian*; *The Guardian*.
- [3] Medicine, C. (2019). Helping Agriculture's helpful honey bees. Retrieved November 15, 2022, from <http://www.fda.gov/animal-veterinary/animal-health-literacy/helping-agricultures-helpful-honey-bees>
- [4] Randall, B. (2022, June 06). The value of birds and bees. Retrieved November 15, 2022, from <http://www.farmers.gov/blog/value-birds-and-bees>
- [5] The need for bees. (2020). Clemson University. Retrieved November 15, 2022, from <http://www.clemson.edu/extension/pollinators/apiculture/importance.html>
- [6] Nearly one in 10 wild bee species face extinction in Europe while the status of more than half remains unknown. (2020, February 20). IUCN report. Retrieved November 15, 2022, from <http://www.iucn.org/content/nearly-one-10-wild-bee-species-face-extinction-europe-while-status-more-half-remains-unknown-iucn-report>
- [7] Sánchez-Bayo, F., & Wyckhuys, K. A. (2019). Worldwide decline of the entomofauna: A review of its drivers. *Biological Conservation*, 232, 8–27.
- [8] Burkle, L. A., Marlin, J. C., & Knight, T. M. (2013). Plant-pollinator interactions over 120 years: Loss of species, co-occurrence, and function. *Science*, 339(6127), 1611–1615.
- [9] The colony and its organization. Mid-Atlantic Apiculture Research and Extension Consortium. (2021). Retrieved November 6, 2022, from <https://canr.udel.edu/maarec/honey-bee-biology/the-colony-and-its-organization/>
- [10] Huang Z-Y, Robinson GE (1996) Regulation of honey bee division of labor by colony age demography. *Behavioral Ecology and Sociobiology* 39: 147–158.
- [11] Cramp D (2008) *A Practical Manual of Beekeeping*. London: How To Books. 304 p.
- [12] Harbo, J. R. (1993). Effect of brood rearing on honey consumption and the survival of worker Honey Bees. *Journal of Apicultural Research*, 32(1), 11–17.
- [13] Khoury, D. S., Barron, A. B., & Myerscough, M. R. (2013). Modelling food and population dynamics in honey bee colonies. *PLoS ONE*, 8(5).
- [14] Russell, S., Barron, A. B., & Harris, D. (2013). Dynamic modelling of Honey Bee (*apis mellifera*) colony growth and failure. *Ecological Modelling*, 265, 158–169.
- [15] Chen, Y. P., Pettis, J. S., Corona, M., Chen, W. P., Li, C. J., Spivak, M., ... Evans, J. D. (2014). Israeli acute paralysis virus: Epidemiology, pathogenesis and implications for honey bee health. *PLoS Pathogens*, 10(7).
- [16] University of Georgia. (2018). Managing bees for pollination. Managing Bees for Pollination - Protecting Pollinators. Retrieved November 12, 2022, from <https://bees.caes.uga.edu/bees-beekeeping-pollination/pollination/pollination-managing-bees-for-pollination.html>
- [17] Plantura. (2022). Bee pollination: How does it work? Plantura Magazine. Retrieved November 12, 2022, from <https://plantura.garden/uk/insects/bees/bee-pollination>
- [18] Missouri Botanical Garden. (2013). Bloom Times by Month. Retrieved November 12, 2022, from [https://www.missouribotanicalgarden.org/Portals/0/Gardening/Gardening Help/PDFs/Bloom summary by month.pdf](https://www.missouribotanicalgarden.org/Portals/0/Gardening/Gardening%20Help/PDFs/Bloom%20summary%20by%20month.pdf)

# Global-Local Framework for Medical Image Segmentation with Intra-class Imbalance Problem

Yifan Zhou\*

Department of Computer Science and  
Engineering, Southern University of  
Science and Technology  
Shenzhen, China  
11910311@mail.sustech.edu.cn

Bing Yang\*

Department of Computer Science and  
Engineering, Southern University of  
Science and Technology  
Shenzhen, China  
12031236@mail.sustech.edu.cn

Xiaolu Lin

Department of Computer Science and  
Engineering, Southern University of  
Science and Technology  
Shenzhen, China  
11911737@mail.sustech.edu.cn

Risa Higashita<sup>†</sup>

Department of Computer Science and  
Engineering, Southern University of  
Science and Technology  
Shenzhen, China  
Tomey Corporation  
Nagoya, Japan  
risa@mail.sustech.edu.cn

Jiang Liu<sup>†</sup>

Department of Computer Science and  
Engineering, Southern University of  
Science and Technology  
Shenzhen, China  
Research Institute of Trustworthy  
Autonomous Systems, Southern  
University of Science and Technology  
Shenzhen, China  
Guangdong Provincial Key  
Laboratory of Brain-inspired  
Intelligent Computation, Department  
of Computer Science and Engineering,  
Southern University of Science and  
Technology  
Shenzhen, China  
liuj@sustech.edu.cn

## ABSTRACT

Deep learning methods have been demonstrated effective in medical image segmentation tasks. The results are affected by data imbalance problems. The inter-class imbalance is often considered, while the intra-class imbalance is not. The intra-class imbalance usually occurs in medical images due to external influences such as noise interference and changes in camera angle, resulting in insufficient discriminative representations within classes. Deep learning methods are easy to segment regions without complex textures and varied appearances. They are susceptible to the intra-class imbalance problem in medical images. In this paper, we propose a two-stage global-local framework to solve the intra-class imbalance problem and increase segmentation accuracy. The framework consists of (1) an auxiliary task network(ATN) , (2) a local patch

network(LPN), and (3) a fusion module. The ATN has a shared encoder and two separate decoders that perform global segmentation and key points localization. The key points guide to generating the fuzzy patches for the LPN. The LPN focuses on challenging patches to get a more accurate result. The fusion module generates the final output according to the global and local segmentation results. Furthermore, we have performed experiments on a private iris dataset with 290 images and a public CAMUS dataset with 1800 images. Our method achieves an IoU of 0.9280 on the iris dataset and an IoU of 0.8511 on the CAMUS dataset. The results on both datasets show that our method achieves superior performance over U-Net, CE-Net, and U-Net++.

## CCS CONCEPTS

• **Computing methodologies** → *Image segmentation.*

## KEYWORDS

Deep Learning, Intra-class imbalance, Medical image segmentation

## ACM Reference Format:

Yifan Zhou, Bing Yang, Xiaolu Lin, Risa Higashita, and Jiang Liu. 2023. Global-Local Framework for Medical Image Segmentation with Intra-class Imbalance Problem. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590071>

\*Both authors contributed equally to this research.

<sup>†</sup>Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590071>

## 1 INTRODUCTION

Medical image segmentation plays a significant role in helping doctors with clinical diagnoses. U-Net[13] and its various derivations[5, 6, 12, 16] enhanced the efficiency and accuracy of the medical image segmentation task. One organ or tissue may have different features due to the noise inference, shielding of other organs, and variability of camera angle. The deep learning models will pay more attention to the dominant features and ignore the weak-represented ones. This phenomenon is known as the intra-class imbalance. Unlike the inter-class imbalance problem, which indicates the imbalance between classes, the intra-class problem focuses on the imbalance within a single class.

Previous works mainly focus on inter-class imbalance in the classification problems, such as the resampling strategies[1, 8], cost-sensitive methods[9], and ensemble methods[11]. The resampling strategies rebuild a more balanced dataset by over-sampling the minority classes[1, 8] or under-sampling the majority classes[10]. The cost-sensitive methods often perform as loss functions or cost matrices. They share the similar idea that increase the weights of the minority classes, such as Focal loss[9], MetaCost[3], and instance-weighting[14]. The ensemble methods like EasyEnsemble[11], BalanceCascade[11], and Adaboost[4] combines the results of multiple weak classifiers to get better results. These methods are more flexible for a wide range of tasks and can even be combined with some data-level resampling methods to achieve better results, depending on the situation[2]. For the intra-class imbalance, a classifier attempts to create multiple disjunct rules that describe the main concept[15]. In segmentation, a deep learning model with multi-channel kernels can be viewed as a non-linear classifier that forms the main concept. However, the deep learning models are easily dominated by the well-represented regions and result in poor generalization to the weak-represented cases.

However, to our best knowledge, there has been no research on the intra-class imbalance in segmentation tasks. The idea of boosting methods is borrowed from the inter-class imbalance to solve this problem. We propose a two-stage global-local framework that consists of an auxiliary task network(ATN), a local patch network(LPN), and a fusion module to solve this problem. In this paper, our contributions are in three-folds:

- (1) We propose a two-stage global-local framework to deal with the intra-class imbalance problem in medical images.
- (2) We adopt key points localization to guide the weak-represented parts and feed them into a local patch network to get more accurate results.
- (3) We demonstrate the effectiveness of our framework by conducting experiments on a private dataset and a public dataset compared with other methods.

## 2 METHODS

### 2.1 Pre-processing

**Key points localization.** We train a U-Net[13] to discriminate the weak-represented regions. We use the iris image as an example, shown in Figure 1. The weak-represented parts are distributed at the trailing end of the iris with a weak boundary. We place the key points on the tail of the iris in every image manually. The Gaussian

heatmaps are constructed according to the coordinates of key points. The heatmaps are used as the ground truth for localization in our framework. After the localization of key points, we crop the local patches from a couple of images and labels centered with key points with a fixed size.

### 2.2 Architecture

The proposed framework is shown in Figure 2. It consists of three modules: (1) an auxiliary task network (ATN), (2) a local patch network (LPN), and (3) a fusion module. Details of the networks are shown in Table 1.

**Table 1: The details of network layers. The K equals the number of the key points. conv3-N means a  $3 \times 3$  convolutional kernel with N channels.**

Layer Number	SE	GSD	KPLD	LPN-Encoder	LPN-Decoder	Fusion Module
1	conv3-16	up-conv-128	up-conv-128	conv3-16	up-conv-128	conv3-16
2	maxpool	---	---	maxpool	---	---
3	conv3-32	up-conv-64	up-conv-64	conv3-32	up-conv-64	conv3-16
4	maxpool	---	---	maxpool	---	---
5	conv3-64	up-conv-32	up-conv-32	conv3-64	up-conv-32	conv1-1
6	maxpool	---	---	maxpool	---	---
7	conv3-128	up-conv-16	up-conv-16	conv3-128	up-conv-16	---
8	maxpool	---	---	maxpool	---	---
9	conv3-256	conv1-1	conv1-K	conv3-256	conv1-1	---
10	---	sigmoid	sigmoid	---	sigmoid	sigmoid

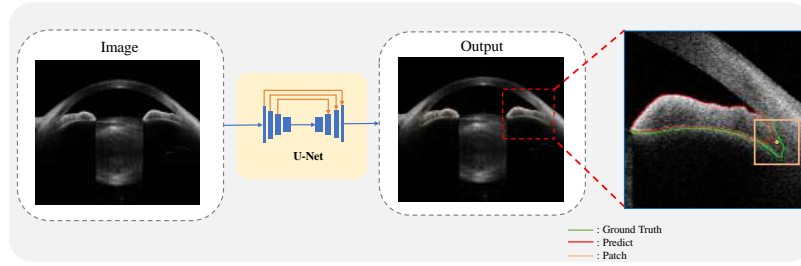
**Auxiliary task network.** The ATN has a shared encoder(SE), and two separate decoders denoted as global segmentation decoder (GSD) and key points localization decoder (KPLD). The GSD will cover the dominant features and perform global segmentation. The KPLD aims to localize the challenging patches with weak-represented features. The GSD and KPLD are auxiliary tasks to each other and improve the results together. The architectures of GSD and KPLD are similar, and the output channels correspond to the number of key points. The details of the architecture are shown in Table 1.

**Local patch network.** We adopt U-Net[13] as the LPN. The coordinates of critical points are calculated by argmax operation on the KPLD output. After that, we crop the local patches from the image with the fixed size and center with the key points. The patches are fed into the LPN to get more accurate results under the supervision of label patches.

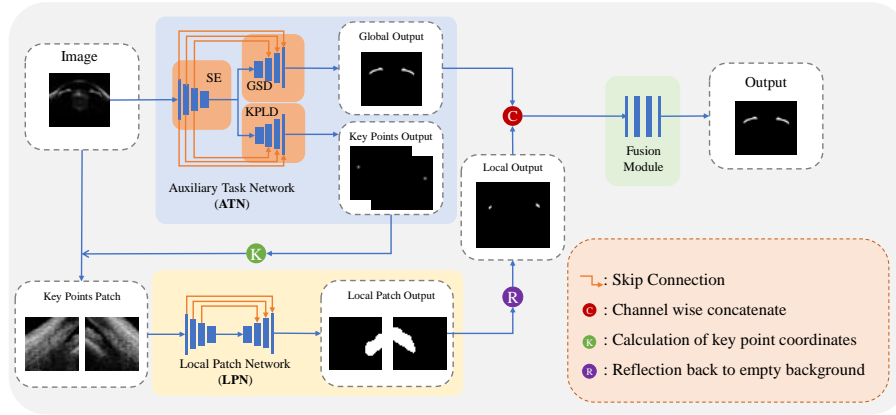
**Fusion module.** We reflect the local patch output to its original location with the coordinates, as shown in Figure. 2. The global output from ATN and local output from LPN are concatenated as the input of the fusion module. The fusion module consists of three CNN layers and ends with a sigmoid layer, which is shown in Table 1.

### 2.3 Training Process

We adopt binary cross-entry (BCE) as the loss function for the training process. There are three training steps for the framework. Firstly, we train ATN with ground truth for GSD and Gaussian heatmaps for KPLD. The Gaussian heatmaps come from the pre-processing step and indicate the coordinates of key landmarks. Secondly, we train LPN to segment local patches as compensation for ATN. Finally, we train the fusion module for some epochs and then fine-tune the whole framework.



**Figure 1: Key point localization.** The yellow point is the key point which is placed manually, and the orange rectangle is the local patch with a fixed size.



**Figure 2: Overview of our framework**

### 3 RESULT

#### 3.1 Datasets

In our experiments, we have selected both the private Iris-290 dataset and the public CAMUS [7] dataset to demonstrate the effectiveness of our method. The Iris-290 dataset is from four different data centers from 2017 to 2019, the Chinese University of Hong Kong(CUHK), Tokyo University(TU), University of California, San Francisco(UCSF), and Zhongshan Ophthalmic Center(ZOC). It contains eighteen B-scans evenly spread over 360 degrees in each eye under light and dark illuminating conditions, taken from 185 individuals. We selected 290 of these AS-OCT images and corresponding iris labels, of which every two were from an individual single-eye sample. In the CAMUS dataset, there are 450 patients' 1800 images containing two and four chambers in the end-diastolic and end-systolic periods. Moreover, we chose the left ventricle endocardium ( $LV_{Endo}$ ) segmentation task to compare. For both datasets, the ratio of the training and test sets is 8:2

#### 3.2 Result

We compared our framework to three state-of-the-art models: U-Net[13], CE-Net[5], and U-Net++[16]. Our experimental code was

implemented based on the Pytorch framework. Each set of experiments was trained with a learning rate of 0.00005 for 500 epochs. We adopt Dice and intersection-over-union(IOU) as the evaluation metrics to quantify the models' performance. The quantitative analysis is shown in Table 2. As can be seen, our method outperforms the three models on both datasets. Visual examples are shown in Figure 3. Our method performs better on the tail of the iris in Iris-290, and also better on the boundaries of the  $LV_{Endo}$ .

To demonstrate the effectiveness of the modules in our framework, we construct ablation experiments on both datasets. The quantitative results are shown in Table 3. We take U-Net[13] as the baseline and compare it with the global output from ATN and the final output from the whole framework. As it can be seen, the ATN slightly outperform the baseline. With the help of KPLD as the auxiliary task, the localization encourages the model to pay more attention to key point regions. The ATN+LPN(Fusion) is our framework that achieves the best results. The results show that the final output is further improved by fusing the outputs of ATN and LPN. Furthermore, the final output, with detailed information from LPN integration, is maintained in more areas of the weak-represented regions.

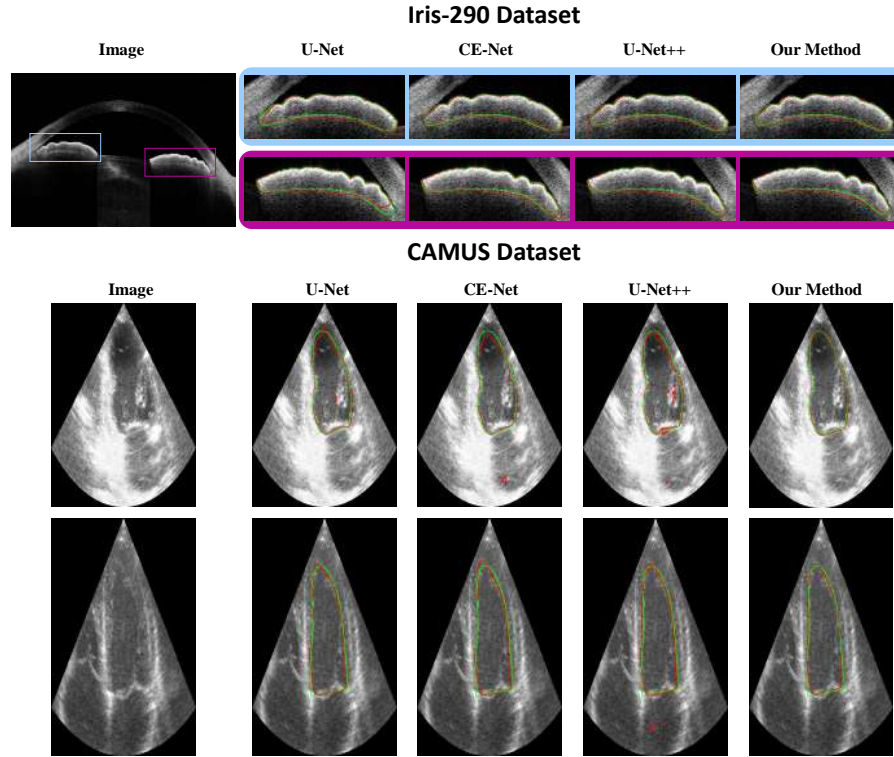


Figure 3: Examples of different models, the green line is the ground truth label, the red line is the inference output.

Table 2: The model result on Iris-290 and CAMUS dataset.

Methods	Iris-290		CAMUS	
	Dice	IoU	Dice	IoU
U-Net[13]	0.9463	0.8981	0.8947	0.8118
CE-Net[5]	0.9513	0.9073	0.9025	0.8246
U-Net++[12]	0.9572	0.9180	0.9000	0.8203
<b>Our Method</b>	<b>0.9626</b>	<b>0.9280</b>	<b>0.9185</b>	<b>0.8511</b>

to generate the final output. The framework separates intra-class imbalance problems as a two-stage workflow. The experiments on two datasets have shown the effectiveness of our framework. However, our approach has some limitations, as it is difficult to locate key points when there is a lack of commonality in the imbalanced regions within classes of the dataset. In future work, it may be possible to design a more general approach to locating regions to provide greater generalisability to the methodological framework.

Table 3: Ablation experiments on Iris-290 and CAMUS dataset.

Methods	Iris-290		CAMUS	
	Dice	IoU	Dice	IoU
Baseline	0.9463	0.8981	0.8947	0.8118
Baseline+ATN	0.9518	0.9054	0.9059	0.8280
<b>Baseline+ATN+LPN(Fusion)</b>	<b>0.9626</b>	<b>0.9280</b>	<b>0.9185</b>	<b>0.8511</b>

## 4 CONCLUSION

In this paper, we discuss the intra-class imbalance problem in medical images and propose a global-local framework. The auxiliary task network can generate global segmentation with dominated features and localize the patches with weak-represented parts. The local patch network performs accurate segmentation for the challenging patches. The fusion module combines the global and local results

## REFERENCES

- [1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W.P. Kegelmeyer. 2011. SMOTE: Synthetic Minority Over-sampling Technique. *arXiv: Artificial Intelligence* (2011).
- [2] Nitesh V. Chawla, Aleksandar Lazarevic, Lawrence O. Hall, and Kevin W. Bowyer. 2003. SMOTEBoost: Improving Prediction of the Minority Class in Boosting. *European conference on machine learning* (2003).
- [3] Pedro Domingos. 1999. MetaCost: a general method for making classifiers cost-sensitive. *knowledge discovery and data mining* (1999).
- [4] Yoav Freund and Robert E. Schapire. 1997. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *conference on learning theory* (1997).
- [5] Zaiwang Gu, Jun Cheng, Huazhu Fu, Kang Zhou, Huaying Hao, Yitian Zhao, Tianyang Zhang, Shenghua Gao, and Jiang Liu. 2019. Ce-net: Context encoder network for 2d medical image segmentation. *IEEE transactions on medical imaging* 38, 10 (2019), 2281–2292.
- [6] Steven Guan, Amir A Khan, Siddhartha Sikdar, and Parag V Chitnis. 2019. Fully dense UNet for 2-D sparse photoacoustic tomography artifact removal. *IEEE journal of biomedical and health informatics* 24, 2 (2019), 568–576.
- [7] Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Ostvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D’hooge, Lasse Lovstakken, and Olivier Bernard. 2019. Deep Learning for Segmentation Using an Open Large-Scale Dataset in 2D Echocardiography. *IEEE Transactions on Medical Imaging* (2019).
- [8] Junnan Li, Qingsheng Zhu, Quanwang Wu, and Zhu Fan. 2021. A novel over-sampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences* (2021).
- [9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).
- [10] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2006. Exploratory Under-Sampling for Class-Imbalance Learning. *international conference on data mining* (2006).
- [11] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 2 (2008), 539–550.
- [12] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018).
- [13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [14] Kai Ming Ting. 2002. An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering* (2002).
- [15] Gary M Weiss. 2004. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter* 6, 1 (2004), 7–19.
- [16] Zongwei Zhou, Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2018. Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support : 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, held in conjunction with MICCAI 2018, Granada, Spain, S...* (2018).

# End-to-end Parking Behavior Recognition Based on Self-attention Mechanism

Penghua Li\*

Dechen Zhu\*

liph@cqupt.edu.cn

15696236105@163.com

Key Laboratory of Intelligent Computing for Big Data, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065  
Chongqing, China

Yushan Tu

Key Laboratory of Intelligent Computing for Big Data, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065  
Chongqing, China  
1587021287@qq.com

Qiyun Mou

Key Laboratory of Intelligent Computing for Big Data, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065  
Chongqing, China  
s210301036@stu.cqupt.cn

Jinfeng Wu

Key Laboratory of Intelligent Computing for Big Data, College of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065  
Chongqing, China  
952427300@qq.com

## ABSTRACT

In response to the current problem of a large amount of abnormal data in parking behavior detection, this research proposes a network specialized in parking behavior identification, which identifies the background parking behavior data, classifies the data with high accuracy, reduces the cost of manually verifying the data in the background, speeds up the parking charging cycle of enterprises, and optimizes the user experience. The dynamic position embedding is introduced in the parking-transformer species, so that the self-attention within the transformer can dynamically model the structure of the input token and dynamically encode the input parking behavior sequence data to improve the accuracy of the model for parking behavior recognition. In addition, we created a self-collected parking behavior (SPB) dataset, which was acquired in a natural state and contained various behaviors, and manually classified the various behaviors within the data, and then randomly divided into a test set and a validation set for training and testing, respectively. Compared with the existing methods, indicate that parking-transformer hits acceptable trade-offs, namely, 97.14% accuracy for SPB dataset.

## CCS CONCEPTS

• **Software and its engineering** → *Software as a service orchestration system.*

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590072>

## KEYWORDS

Parking behaviour recognition, self-attention mechanism, gaze detection, transformer

### ACM Reference Format:

Penghua Li, Dechen Zhu, Qiyun Mou, Yushan Tu, and Jinfeng Wu. 2023. End-to-end Parking Behavior Recognition Based on Self-attention Mechanism. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590072>

## 1 INTRODUCTION

Smart on-street parking, similar to private [29] and off-street parking [23], is an effective way to ease traffic congestion and parking difficulties that intelligent transportation systems aim to solve [6]. Accurate monitoring of on-street parking provides the driver with reliable real-time knowledge about roadside parking occupancy, saving cruising time, economizing fuel costs, and reducing traffic flows. [17]. It also improves the operating efficiency and economic benefits for the companies that engage in the on-street parking business. However, compared to off-street parking, on-street parking is more challenging to monitor due to the complex urban traffic environment, harsh weather, poor light conditions, etc. [24] [3]. Valid data is obtained so that on-street parking can be monitored and managed through that data. This data includes real-time online monitoring and later offline evaluation. Still, later offline evaluation is labor-intensive, so there is a need to improve the high-quality automatic assessment of online data. [4] With the development of IoT, many technologies such as underground sensors (magnetic and infrared), cameras, or radars [24] [20] [1] have been proposed. However, these technologies are often expensive to install and maintain, require a large amount of parking space and are easily damaged [31].

With the development of deep learning, the detection of parking behavior of vehicles is becoming more and more automated, and various deep learning methods are applied to the recognition of parking behavior. [30] An essential component of deep learning

algorithms in computer vision, which can be applied to parking behavior recognition to improve the accuracy of parking behavior recognition, cope with multiple scene changes, and improve generalization[7].The camera has high accuracy, low cost, can be used in a wide range, will not cause damage to the parking facilities, and later can form a complete data chain, which is given to enterprises for use in other places[16].

The current high-level video parking behavior recognition still has a high error rate, and the data still needs to be verified manually in the later stage. There is a large amount of anomalous data in the current data that needs to be checked twice; for example, there are exits from the sidewalk, exits from intersections, in the form of parking areas, obscured by other vehicles, and failure to pull into parking spaces. The data examples of normal entry, normal exit, and abnormal parking are shown in Fig.1.

Since all data contains abnormal data, a large amount of manual verification of the data is required in the later stage, which will increase the corresponding expenditure of the enterprise. Therefore, it is necessary to develop an end-to-end network to perform secondary verification of data to reduce labor costs for enterprises.



**Figure 1: The original picture of the parking situation to be judged. (a) Drive out normally. (b) Drive in normally.(c)abnormal parking**

## 2 RELATED WORK

Although closed car parks can now be entirely unmanned, open car parks still require manual guarding and billing of incoming vehicles, and there is still a considerable challenge to improve the unmanned management of open parking. Much progress is being made in the study of motion recognition [18]. Among the traditional methods, there are physical model-based methods, which use the kinematics/citeschubert2008comparison and the vehicle's dynamic properties to predict the vehicle's state. Reachability studies [19] provide a formal method to measure uncertainty in the behavior of cars in control designs. All of the above methods require accurate vehicle modeling, which can greatly affect the inference of the model when road environment variables are increased.

In contrast, deep learning-based approaches can automatically extract features, and the models can be generalized to a high degree and subsequently migrated to various other new scenarios. Recurrent neural networks(RNN) and long and short term memory networks (LSTM) [22] are well known for their expertise in processing

sequences. Various research papers have also focused on modifying the connections within the network to improve the ability to interact with information between sequences and enhance the network's ability to extract feature information[2].Also available with CNN for behavioural recognition[8, 9] Design an end-to-end parking behavior recognition network, adaptable to multi-scene changes, multi-model changes, and enhanced robustness to camera shake. Since AlexNet[15] won the 2012 ImageNet competition, ConvNets accuracy is getting higher and higher. GoogleNet[27] won the 2014 ImageNet competition which achieves 74.8% top-1 accuracy. ResNet[11] can achieve deeper layers, with easy training, and can achieve higher accuracy. SEnet[13] achieves 82.7% top-1 accuracy with 145M parameters. Densnet[14] use densely connected ,which can enhance feature propagation,relieve the problem of wanishing-gradient. inception-v4[26] uses residual connections to reduce the error rate. VGGNet[25] use small receptive field convolution to improve performance by increasing network depth. EfficientNet[28] The benchmark network of EfficientNet is obtained through network structure search, which balances the width and depth of the network. Compared with other recognition networks, there is a qualitative breakthrough. Vision-Transformer[10] segments a single image into a sequence of linear embeddings to provide as input to the Transformer. Swin-transformer[21] uses sliding window operation and adopts hierarchical design, and achieves 86.4% accuracy on imagenet. Using the self-collected data set for training and testing, the designed network has an accuracy rate of up to 97.14%. Our method is robust to various real-world changes, such as ambient light and weather changes, including the most challenging: being obscured by other vehicles can be accurately identified.

## 3 METHODOLOGIES

The designed Parking-Transformer(Fig. 2) consists of the ResNet, Bi-LSTM, and Transformer Encoder. In the Parking-Transformer, We use Bi-LSTM to obtain time series features in parking data to obtain better patch-embedded. Compared with existing model efforts, the proposed Parking-Transformer demonstrates its novelty by the following aspects.

- Using resnetv2 to extract features, where the prior activation function can make model optimization easier, and using BN(Batch Normalization)'s prior activation of the network can reduce network overfitting.
- Use Bi-LSTM to obtain better sequence features from behavior recognition and dynamically position embedding.
- The introduction of transformer, which focuses on global information, can model the longer distance dependencies of the parking behavior sequence and avoid focusing only on local information.

The related symbols of Parking-Transformer and what they represent are given in the table 1.

### 3.1 Feature Extraction Stage

Feature extraction stage using resnetv2 as feature extraction network.

$$\mathbf{F}^f = \mathbf{U} + \sum_{i=1}^n \mathbf{H}(x_i, \omega_i) \quad (1)$$

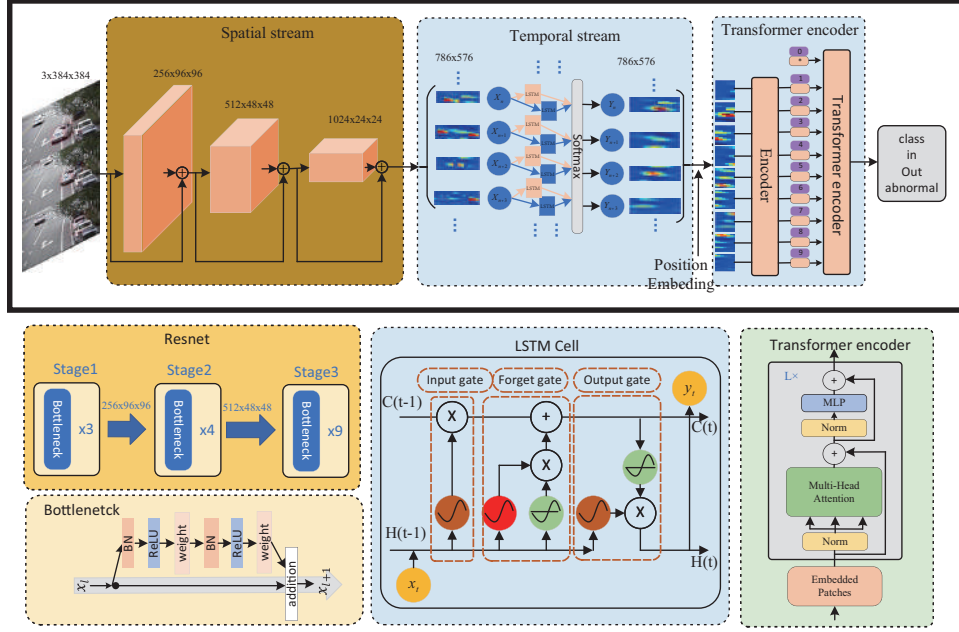


Figure 2: The design of Parking-transformer

Table 1: Related Mathematical Symbols and Their Representations

Symbol	Representation
$U \in \mathbb{R}^{h \times w \times c}$	The input image
$F^f \in \mathbb{R}^{H^f \times W^f \times C^f}$	Feature Extraction stage's feature map
$Q \in \mathbb{R}^{n \times h \times M \times w \times M \times c \times M}$	Attention Queries
$K \in \mathbb{R}^{n \times h \times M \times w \times M \times c \times M}$	Attention Keys
$V \in \mathbb{R}^{n \times h \times M \times w \times M \times c \times M}$	Attention Values

where  $x_i$ ,  $H$  represent the input of the  $n$ th bottleneck, the function of the residual branch contains batch normalization, vanilla convolution, ReLU activation functions.

### 3.2 Spatial-Temporal Feature Extraction State

The hidden layer of bi-LSTM contains two values, one is  $A$  and the other is  $A'$ . The final output  $O_t$  connects the forward vector with the reverse vector to the output, which can be obtained as a feature between the front and back. The specific calculation formula is as follows.

$$O_t = g(VA_t + V'A'_t) \quad (2)$$

$$A_t = f(Ux_t + WA_{t-1}) \quad (3)$$

$$A'_t = f(U'x_t + W'A'_{t+1}) \quad (4)$$

### 3.3 Last Stage

In the last stage, the transformer, by introducing an attention mechanism, obtains even better results than cnn. It including MSA(Multi-head-Self-Attention) contributes to the network capturing richer

features or information, MLP(Multi-Layer perceptron) improves model performance on non-linear data, LN(Layer Normalization) regularizes the model and reduces overfitting.

The input  $X \in \mathbb{R}^{n \times d}$  is linearly transformed into three parts,  $Q$ ,  $K$ , and  $V$ , by the self-attentive module for obtaining sequence features in the image to improve the overall sequence characterization, which is expressed as follows

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (5)$$

$$Mh(Q', K', V') = Concat(head_1, ..., head_h)W^O \quad (6)$$

here  $d_k$  is the dimensions of queries,  $Mh$  represent Multihead,  $head_n$  equal (5).

The linear layer completes the weighting of  $Q$ ,  $K$ , and  $V$  and transforms it into a linear combinatorial output.

In the MLP module, apply MLP to the self-attentive layer for feature transformation and nonlinearity.

$$MLP(X) = FC(\sigma(FC(X))) \quad (7)$$

$$FC(X) = XW + b \quad (8)$$

where  $W$  and  $b$  are the weight and bias term of fully-connected layer.  $\sigma$  represent GELU activation function[12]. Note that the GELU function is given as  $x\Phi(x)$ .

Accelerate network convergence by adding LN(layer normalization)[5]. LN is applied over each MLP output  $x \in \mathbb{R}^d$  as follows:

$$LN(x) = \frac{x - \mu}{\delta} \circ \gamma + \beta \quad (9)$$

where  $\mu \in \mathbb{R}$ ,  $\delta \in \mathbb{R}$  are the mean and standard deviation of the feature maps output from the MLP respectively,  $\circ$  is the dot product, and  $\gamma \in \mathbb{R}^d$ ,  $\beta \in \mathbb{R}^d$  are learnable parameters.

**Table 2: Related Mathematical Symbols and Their Representations**

Input	Operator	OutputSize	expandsize	layer
$384 \times 384 \times 3$	stem	64	64	1
$64 \times 96 \times 96$	Bottleneck	256	64	3
$256 \times 96 \times 96$	Bottleneck	512	128	4
$512 \times 48 \times 48$	Bottleneck	1024	256	9
$1024 \times 24 \times 24$	Conv2d	768	-	1
$576 \times 768$	BiLSTM	768	1536	2
$577 \times 768$	Transformer Block	768	-	12
768	head	$N_c$	-	1

**Table 3: The Training and Testing Details of Different Networks**

Network	Ours	3D resnet50	densenet101	efficientb0	inceptionv4	Resnet50	SwinT-B	VGG16	ViT16-B
Training	97.49	98.93	91.41	98.57	98.55	94.31	99.33	99.03	99.15
Testing	<b>97.14</b>	93.64	95.37	96.80	96.73	96.27	96.98	96.88	96.86

## 4 EXPERIMENT AND RESULT ANALYSIS

We conducted ablation experiments on the self-collected dataset SPB dataset and compared the results with ResNet-50, DenseNet, InceptionNet, vgg16, Swin-transformer, and Vision-transformer, to demonstrate the strengths and weaknesses of our model

### 4.1 Data Description

The self-collected parking behavior dataset (SPB) captures photos of various vehicle behaviors, including different light, temperature, and scenarios over a 24-hour period, from May 2021 to October 2021, through Hikvision HD cameras erected on actual street parking. All data are manually classified into 3 categories: normal entry(in), normal exit(out), and abnormal (ab), with a total of 40485 images. For SPB Dataset, 77%, 23% of such datasets are regarded as training and testing data.

### 4.2 Model Training

We divide the SPB dataset randomly and use the same software environment on the same too server for training as well as testing. We use cross-entropy as the loss function as follows We use cross-entropy as the loss function as follows:

$$E(L, S) = \sum_{i=1}^{N_c} l_i \log(s_i) \quad (10)$$

where  $l_i$  and  $s_i$  denote the likelihood and prediction scores of the  $i$ th class, respectively. According to the structure shown in Figure 2, we set the input size of the Parking-Transformer (including other comparison networks) to  $384 \times 384 \times 3$  and the output class to 3. We processed some SPB dataset experiments and gave the predefined structure of the Parking-Transformer on the dataset, as shown in Table 2. The structure of each bottleneck is shown as the bottleneck in Figure 2, and the expandsize is the hidden layer of the internal convolution of the bottleneck. Introduce a layer of vanilla convolution before BiLSTM for patch embedding. Class token after BiLSTM to make the output  $577 \times 768$ .

We train Set groupnorm's group to 32 and epsilon to  $1e-5$  in bottleneck. The group norm compensates for the shortcoming of batch normalization in increasing the accuracy of the model when the batch size is small. Select adaptive moment estimation (Adam) as the optimizer and set the learning rate to  $1e-5$  and weight decay to  $1e-5$ . Random data augmentation of the sample including small-angle rotation, random resize cropping, horizontal flipping, vertically flipping, color jitter, affine, random erasure to further increase diversity. Use label smoothing training to generate better calibrated networks for better generalization.

$$l_i = \begin{cases} 1 - \frac{N_c - 1}{N_c} \epsilon, & \text{if } (i = t) \\ \frac{\epsilon}{N_c}, & \text{if } (i \neq t) \end{cases} \quad (11)$$

where  $\epsilon$  is the confidence of the model on the training dataset and  $t$  is the ground truth corresponding to the target category. In this study,  $\epsilon$  is set as 0.1.

### 4.3 Model Testing

To validate the accuracy of the different methods and models on the dataset, we used 9287 images as a performance test. Table III shows the accuracy of the network trained on the SPB dataset. Figure 5 depicts the accuracy of each category using the blending down matrix image. We have found that our network has a higher accuracy rate than any other network. We found that the accuracy of our proposed network is higher than any other network. Among all the methods, only densenet121d has an accuracy rate below 90% for drive-out and below 80% for the anomaly. This is because densenet uses feature reuse, which reacts to the actual situation where feature interference occurs when the traffic volume within the road increases, causing the network to fail to correctly identify the behavior of the target vehicle. The accuracy of the models for identifying anomalous categories is below 90% due to the complexity of the situations contained in the abnormal behavior and the relatively small amount of data, which makes the models easy to over-fit.

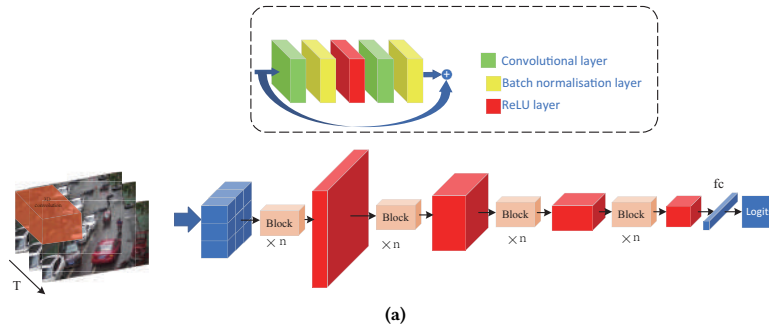


Figure 3: Using 3DResnet for vehicle behavior recognition

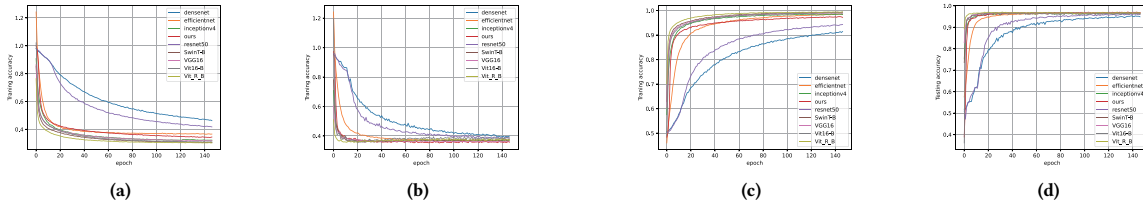


Figure 4: The Training and testing profiles on SPB dataset, where (a) and (b) present the training and testing loss, (c) and (d) show the training and testing accuracy

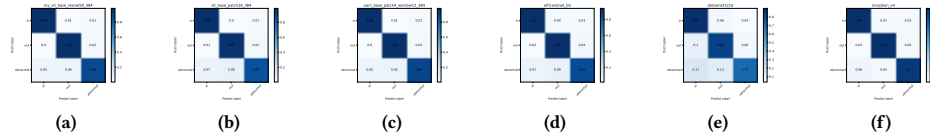


Figure 5: Examples of the CAMs generated from the predicted classes, where (a) are ours, (b) are ViT-B, (c) are ViT-resnet50, (d) are SwinT-B, (e) are Efficientnet-B, (f) are densenet121d, (g) are inception-v4

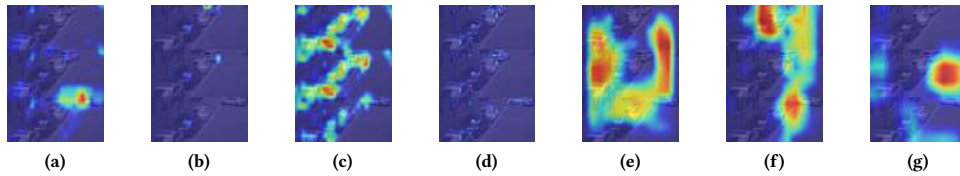


Figure 6: Examples of the CAMs generated from the predicted classes, where (a) are ours, (b) are ViT-B, (c) are SwinT-B, (d) are Efficientnet0, (e) are Resnetv2, (f) are densenet121d, (g) are inception-v4

Relying on the activation-like mapping (CAM) technique [32], Figure 5 gives a more detailed description of the results of Figure 4. Using a randomly selected image from SP's test set, we found that our proposed network accurately identifies the target vehicle to be discriminated against and determines which behavior the car belongs to based on the target vehicle's activity state. Compared to other networks, which also give correct results, the network does not focus on the key areas of the image so that other networks will be less precise than our network.

#### 4.4 Conclusion

This study aims to implement automatic parking behavior recognition for open car parks and achieve unmanned management of open parks for charging. To achieve such a goal, an end-to-end parking transformer is designed. The joint spatial and temporal modules acquire the target vehicle's spatial location and the target vehicle's time series, respectively, ultimately achieving an accuracy of 97.14%.

Manual verification of the anomalous data from the secondary inference of the model is expected to cost only 2.86% of the initial manual effort.

## ACKNOWLEDGMENTS

This work is supported by Chongqing Outstanding Youth Fund Project (cstc2021jcyj-jqX0001); Science and Technology Research Project of Chongqing Education Commission (KJZD-K202100603); National Natural Science Foundation of China (52272388);

## REFERENCES

- [1] Fadi Al-Turjman and Arman Malekloo. 2019. Smart parking in IoT-enabled cities: A survey. *Sustain. Cities Soc.* 49 (2019), 101608.
- [2] Alexandre Alahi, Kratharth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. 2016. Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 961–971.
- [3] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Claudio Vairo. 2016. Car parking occupancy detection using smart camera networks and deep learning. In *in Proc. IEEE Symp. Comput. Commun. (ISCC)*. IEEE, 1212–1217.
- [4] Behrang Assemi, Alexander Paz, and Douglas Baker. 2021. On-Street Parking Occupancy Inference Based on Payment Transactions. *IEEE T. Intell. Transp.* (2021).
- [5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [6] Fabian Bock, Sergio Di Martino, and Antonio Origlia. 2019. Smart parking: Using a crowd of taxis to sense on-street parking space availability. *IEEE T. Intell. Transp.* 21, 2 (2019), 496–508.
- [7] Harshitha Bura, Nathan Lin, Naveen Kumar, Sangram Malekar, Sushma Nagaraj, and Kaikai Liu. 2018. An edge based smart parking solution using camera networks and deep learning. In *2018 IEEE Int. Conf. on Cognit. Comput. (ICCC)*. IEEE, 17–24.
- [8] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2090–2096.
- [9] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, and Jeff Schneider. 2018. Short-term motion prediction of traffic actors for autonomous driving using deep convolutional networks. (2018).
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [12] Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).
- [13] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.
- [14] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [15] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and < 0.5 MB model size. *arXiv preprint arXiv:1602.07360* (2016).
- [16] Hidetomo Ichihashi, Tatsuya Katada, Makoto Fujiyoshi, Akira Notsu, and Katsuhiko Honda. 2010. Improvement in the performance of camera based vehicle detector for parking lot. In *Int. Conf. Fuzzy Syst.* IEEE, 1–7.
- [17] Ludovic Leclercq, Alméria Sénécat, and Guilhem Mariotte. 2017. Dynamic macroscopic simulation of on-street parking search: A trip-based approach. *Transport. Res. B-Meth.* 101 (2017), 268–282.
- [18] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. 2014. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH journal* 1, 1 (2014), 1–14.
- [19] Karen Leung, Edward Schmerling, Mengxuan Zhang, Mo Chen, John Talbot, J Christian Gerdes, and Marco Pavone. 2020. On infusing reachability-based safety assurance within planning frameworks for human–robot vehicle interactions. *The International Journal of Robotics Research* 39, 10–11 (2020), 1326–1345.
- [20] Trista Lin, Hervé Rivano, and Frédéric Le Mouél. 2017. A survey of smart parking solutions. *IEEE T. Intell. Transp.* 18, 12 (2017), 3229–3253.
- [21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10012–10022.
- [22] Yuexin Ma, Xinge Zhu, Sibao Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. 2019. Trafficpredict: Trajectory prediction for heterogeneous traffic agents. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 6120–6127.
- [23] Tooraj Rajabioun and Petros A Ioannou. 2015. On-street and off-street parking availability prediction using multivariate spatiotemporal models. *IEEE T. Intell. Transp.* 16, 5 (2015), 2913–2924.
- [24] Cristian Roman, Ruizhi Liao, Peter Ball, Shumao Ou, and Martin de Heaver. 2018. Detecting on-street parking spaces in smart cities: Performance evaluation of fixed and mobile sensing systems. *IEEE T. Intell. Transp.* 19, 7 (2018), 2234–2245.
- [25] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [26] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.
- [27] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.
- [28] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.
- [29] Lingling Wang, Xiaodong Lin, Eugene Zima, and Chunguang Ma. 2020. Towards airbnb-like privacy-enhanced private parking spot sharing based on blockchain. *IEEE T. Veh. Technol.* 69, 3 (2020), 2411–2423.
- [30] Shuguan Yang, Wei Ma, Xidong Pi, and Sean Qian. 2019. A deep learning approach to real-time parking occupancy prediction in transportation networks incorporating multiple spatio-temporal data sources. *Transp. Res. C, Emerg Technol* 107 (2019), 248–265.
- [31] Shuguan Yang and Zhen Sean Qian. 2017. Turning meter transactions data into occupancy and payment behavioral information for on-street parking. *Transp. Res. C, Emerg. Technol.* 78 (2017), 165–182.
- [32] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.

# Foreign object recognition method of transmission line based on improved outlier rate method

Dongmei Liu\*

School of Electrical and Automation  
Engineering, Hefei University of  
Technology, E-mail:  
dmliu100@hfut.edu.cn

Zhongwang Zhu

School of Electrical and Automation  
Engineering, Hefei University of  
Technology, E-mail:  
1186817256@qq.com

Bo Chen

School of Electrical and Automation  
Engineering, Hefei University of  
Technology, E-mail:  
82180126@qq.com

## ABSTRACT

Foreign matters hanging on the transmission line can be regarded as a potential risk of the transmission system, which will not only affect the normal power supply of the transmission line, but also pose a greater threat to pedestrians and vehicles under the line. Aiming at the low efficiency and high false detection rate of traditional foreign object recognition methods for hanging foreign objects, this paper proposes a foreign object recognition method for transmission lines based on improved outlier rate method. It proposes to use Hough line transformation to extract the transmission line, and then conduct convolution operation on the area where the transmission line is located and the non-transmission line area, and set the corresponding outlier rate in combination with the actual error to identify the foreign matters in the transmission line.

## CCS CONCEPTS

• **Computing methodologies** → Computer graphics; Image manipulation; Image processing.

## KEYWORDS

Transmission line, Foreign matter detection, Outlier method

### ACM Reference Format:

Dongmei Liu, Zhongwang Zhu, and Bo Chen. 2023. Foreign object recognition method of transmission line based on improved outlier rate method. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590073>

## 1 INTRODUCTION

Transmission line is the main carrier of power transmission, and also an important part of the power system, which plays a vital role in the safe and stable operation of the system. The transmission line will be exposed to nature and be affected by various factors. The foreign matters on the transmission line will greatly affect the normal operation of the transmission line, and even cause power

failure or safety accidents. For example, in February 4, 2021, there were 15 accidents of foreign objects flying on the high-speed rail catenary caused by citizens in Zhengzhou, Henan Province, and 8 high-speed rail trains were delayed. Therefore, it is extremely important to identify foreign objects on the transmission line.

In recent years, with the development of UAV technology, the combination of UAV and image processing technology to identify foreign matters in transmission lines has become a hot spot. Literature [1] uses the parallel characteristics of the detection line and the transmission line to extract the transmission line and detect foreign matters. The existence of foreign objects in the transmission line is detected by moment invariants and Adaboost algorithm. When the fusion degree between the suspended foreign objects and the background is high, false detection is easy; Literature [2] extracts transmission lines by detecting straight lines and transmission lines, locates transmission lines by Hough straight line detection and parallel characteristics, and detects the existence of foreign objects within the transmission line. However, this method ignores the existence of suspended foreign matters except in the area near the transmission line; Literature [3] judges whether there is foreign matter by making a pixel matrix around the transmission line, and then calculating the pixel value outlier rate of each transmission line; The outlier rate method determines the existence of foreign matters by convolving the extracted transmission line direction and setting the outlier rate. This method has a good effect on foreign matters such as attached power lines, but it has a very low detection rate for the existence of suspended foreign matters (kites, balloons, etc.).

To solve the problem of poor detection of foreign objects hanging on the transmission line in the foreign object recognition in the above images, this paper aims at the difference between foreign objects hanging on the transmission line and foreign objects attached to it, sets up a convolution kernel scanning full image and determines the location of foreign objects by setting outlier rate in different areas to solve the problem of poor detection of foreign objects hanging on the transmission line;

Section 2 of this paper will introduce the principle of the outlier rate method and the improved outlier rate method; In section 3, we give the image pre-processing before foreign object recognition of foreign object digital image of transmission line; In section 4, we give the research on the detection method of foreign matters on the transmission line, extract the full picture scanning along the angle of the transmission line, set the pixel outlier rate to judge the detection method of foreign matters, set the threshold screening box to detect foreign matters on the transmission line, and give the comparison between the outlier rate algorithm and the improved

\*corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590073>

outlier rate algorithm; Finally, section 5 is the conclusion of the article.

## 2 BASIC PRINCIPLE OF IMPROVED OUTLIER RATE METHOD

Outlier rate method is used to detect the existence of foreign matters in transmission lines. It extracts the normal transmission line by acquiring all the line sequences in the image, and convolves the extracted normal transmission line. That is, a rectangle is made around the transmission line, which has a pixel with a height of 10 and a width of the absolute value of the difference between the x coordinates of the two endpoints of the line; Select the "lowest point" of the two endpoints of the line (that is, the point with the lowest x and y coordinates), and use the upper 4 pixels of the point and the lower 5 pixels of the point to form the height of the rectangle. Next, the total number of points n with a pixel value of 255 is calculated for each normal transmission line, and it is agreed that if n divided by the product of the height and width of the rectangle exceeds a certain value *RI* (called "outlier rate"), it means that foreign objects appear. Then the formula for judging foreign matters is:

$$n \div s = \begin{cases} > RI & \text{Foreignmatters} \\ < RI & \text{Noforeignmatters} \end{cases} \quad (1)$$

$$ni \div s = \begin{cases} > Rli & \text{Thereisforeignbody} \\ < Rli & \text{Therearenoforeignbodies} \end{cases} \quad (2)$$

$$i = \begin{cases} 0 & \text{Theconvolution kernel is not on the powerline} \\ 1 & \text{Theconvolution kernel is on the powerline} \end{cases} \quad (3)$$

where *n* is the total number of pixels with a value of 255, and *s* is the area of the convolution kernel; *RI* is the rate of outliers.

The improved outlier rate method is to determine the existence of foreign matters through convolution operation along the transmission line after obtaining the required transmission line. This method is not effective for the detection of foreign matters hanging on the transmission line. The improved outlier rate method adds *m* convolution kernels, where *m* is the integer of *M/10*, *M* is the number of lines in the image, and the convolution kernel scans the global image along the slope direction of the transmission line; And because the foreign objects hanging on the transmission line are usually kites, balloons, plastic bags and other large volumes, we can design the image part of the extracted transmission line and the image part outside the transmission line with different outlier rates. The parameter *i* is introduced as the basis for judging whether the convolution kernel is on the transmission line after the extraction of the transmission line.

The formula of improved outlier rate is:

$$ni \div s = \begin{cases} > Rli & \text{Foreignmatters} \\ < Rli & \text{Noforeignmatters} \end{cases} \quad (4)$$

$$i = \begin{cases} 0 & \text{Convolutioncore is not on the transmissionline} \\ 1 & \text{Convolutioncore is on the transmissionline} \end{cases} \quad (5)$$

The Figure 1 flow chart of improved outlier rate method is as follows:

## 3 IMAGE RECOGNITION PREPROCESSING OPERATIONS

### 3.1 Image Preprocessing

All transmission line images will be subject to different degrees of external interference in the process of generation and transmission. In order to eliminate these external interferences, it is necessary to perform corresponding preprocessing operations on the transmission line images. The main preprocessing operations in this paper include weighted grayscale, medium Value filtering and mathematical morphology processing.

A common color image is composed of three RGB channels superimposed, and the RGB color is used as the three primary colors. The magnitude of the value in the digital image represents the intensity of the three colors. Color images can express the distribution and characteristics of chromaticity and brightness levels of the entire image, and grayscale images can also express, so grayscale processing will not reduce the information of the image. Generally, the weighted average method based on human eye sensitivity is used, and its formula is:

$$I_{xy} = 0.299R_{xy} + 0.587G_{xy} + 0.114B_{xy} \quad (6)$$

Median filtering is a kind of nonlinear filtering. This filtering method is very effective in eliminating salt and pepper noise, and it also has excellent protection of edges and features. It is a classic smoothing noise method; median filtering cannot be represented by a core matrix. The pixel and its neighbors form a set, and then the median value of this set is calculated as the value of the current pixel, and the noise point is eliminated by this value. The output of the two-dimensional median filter is:

$$g(x, y) = \text{med}\{f(x - k, y - i), (k, i \in W)\} \quad (7)$$

After the image is preprocessed and threshold, although the edge of the transmission line is reflected, there are still many large or small white area noises, so it is necessary to find a way to remove these noise areas. The basic operations of mathematical morphology processing include: erosion, dilation, opening operation, and closing operation. A represents the set of images to be processed, B represents processing structural elements, and morphological processing is to use B to operate on A. Mathematical morphology can solve this problem.

### 3.2 Iterative threshold segmentation

The iterative threshold algorithm is a deepening of the histogram bimodal method, and can also automatically estimate the threshold. Given an initial estimate of *T*; dividing an image by *T* will yield two parts: *G1* and *G2*. *G1* contains all pixels whose gray value is greater than or equal to *T*, and *G2* consists of all pixels whose gray value is less than *T*; obtain all pixel gray values *m1* of *G1* part and analyze *G1* part for the whole image. Say the weight occupied  $\lambda_1$ ; find all the pixel gray values *m*

*G1* of the *G2* part and analyze the weight occupied by the *G*  $\lambda_2$  part for the entire image; find the new threshold according to the following formula:  $T = m_1\lambda_1 + m_2\lambda_2$ ; Repeat the above two steps, until the difference between the *T* values in successive iterations is less than a predetermined parameter  $\Delta T$ .

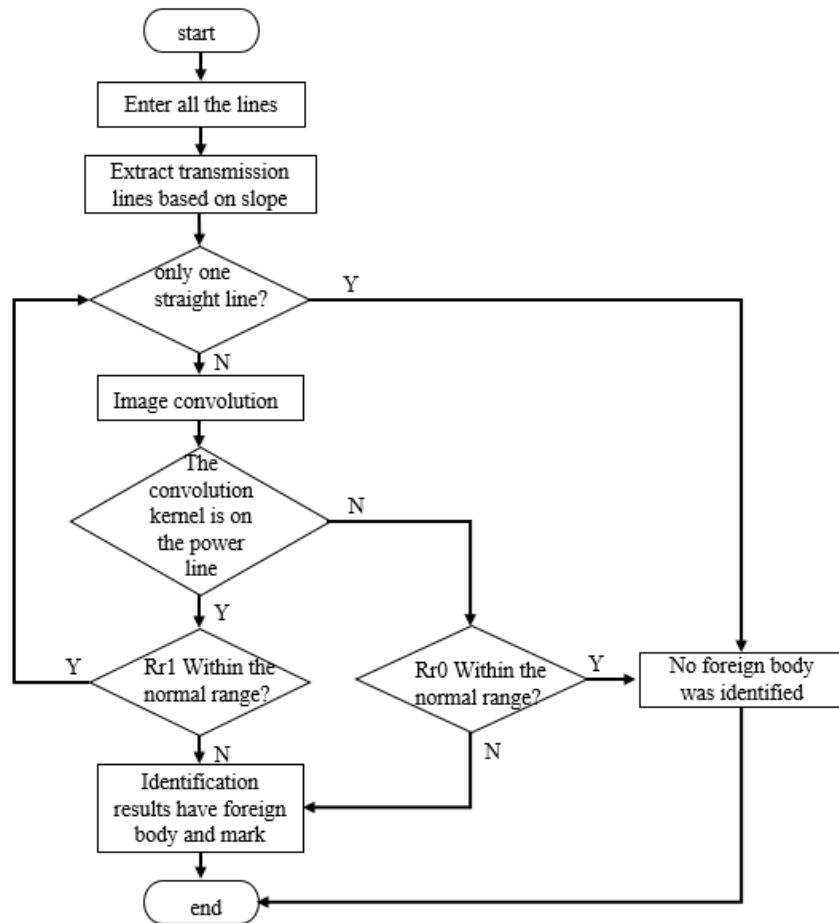


Figure 1: The flow chart of improved outlier rate method

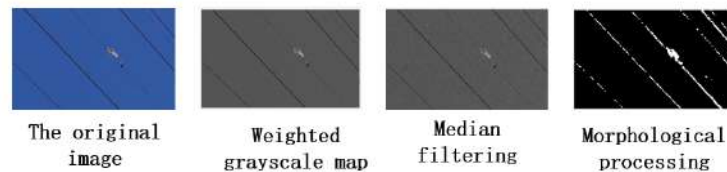


Figure 2: Early preprocessing image

### 3.3 Probabilistic Hough Transform Line Extraction

Probabilistic Hough Transform (PPHT) is an improvement over the Standard Hough Transform (SHT). It does not transform in the whole range, but defines the scale coefficient through the consistent probability density function of image pixels to make it transform within a certain range, and reduces the amount of calculation and shortens the calculation time by calculating the direction and range

of individual line segments. Defining the scale coefficient  $\beta \in [0,1]$  Assuming  $\beta = 0.3$ , then the calculation amount of Hough transform is only 30% of the original, thus eliminating useless data very well.

**Table 1: Probabilistic Hough transform to extract straight line table**

	Pixel stacking number	$\theta$ parameter
1	1658	49
2	1085	45
3	1421	48
4	1926	50
5	2053	51
6	2098	51
7	1606	46
8	1693	51
9	1391	46
10	2024	49

#### 4 IDENTIFICATION OF FOREIGN OBJECTS IN TRANSMISSION LINES BASED ON IMPROVED OUTLIER RATE METHOD

##### 4.1 Angle extraction of power lines

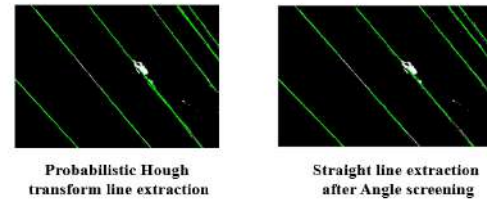
When extracting a normal transmission line, it is necessary to determine the slope angle of the transmission line, so as to design the convolution kernel along the convolution direction of the transmission line. The basic principle of Hough line transformation is to traverse each pixel in the image, and then to Perform frequency statistics on all straight lines passing through this point, and a straight line with a higher frequency is used as the detected straight line in the image. In practice, the slope of the straight line tends to infinity, so we will use polar coordinates to identify the straight line and control the angle within the range of  $\theta$  ( $0^\circ, 90^\circ$ ) to solve the problem this disadvantage.

It can be seen from Table 1 that the linear polar coordinate angle parameters obtained by Hough transform are roughly in the range of (49~51). The suspension of foreign objects will lead to factors such as omission of transmission lines and misdetection when designing too few transmission lines to extract when dealing with straight line extraction. Therefore, the desired transmission lines are obtained through the extraction of ten transmission lines and subsequent screening. According to the angle parameters in the above table, most of the polar coordinate angles of transmission lines tend to be within a certain range, and straight lines beyond this range can be eliminated. We will set the power line angle selection formula to filter the value that retains the most angles  $\theta$ .

The angle selection formula is:

$$-\theta_0 \leq \arctan\left(\frac{k \times 180^\circ}{\pi}\right) - \theta \leq \theta_0 \quad (8)$$

The figure below is the straight line screened by the angle limit. It can be seen from the figure that some overlapping and other interfering straight lines have been removed very well, and the remaining straight line segment is in good agreement with the transmission line. The convolution kernel production provides the basis.

**Figure 3: Hough Line Transformation and Screening Graphs**

##### 4.2 Foreign body contour extraction

After the pre-processing of the transmission line image, the retained transmission line image contains the binary image of the transmission line, the binary image of foreign objects, and the binary image of a very small amount of noise. Then the foreign body contour extraction detection process is as follows:

**Global foreign body contour scanning:** After the convolution kernel is produced, the image used in this paper has a pixel size of  $712 \times 1148$  after preprocessing. Therefore, the number of convolution kernels is set to  $M/10=71$  to perform global scanning on the foreign object image of the transmission line. Next, the total number of points  $n_1$  containing the pixel value 255 is calculated for the normal power line, and it is agreed that if  $n_1$  divided by the product of the height and width of the rectangle exceeds a certain value  $rR1$  (called "the convolution kernel on the power line"). If it does not exceed  $rR1$ , it will return to judge whether it is the location of the straight line segment of the separate transmission line, if it is, it will output as no foreign object; for the total number of points  $n$  that do not include the pixel value of 255 for the calculation of the transmission line 0, and it is agreed that if the product of  $n_0$  divided by the height and width of the rectangle exceeds a certain value  $rR0$  (called "the rate of outliers of the convolution kernel not on the transmission line"), it means that there is a foreign body. If it does not exceed  $rR0$ , it will be directly output as no foreign matter. Since the convolution kernel in the power transmission line is usually aimed at detecting foreign matter attached to the power transmission line,  $rR1 < rR0 = 0.315$ . After scanning the global foreign body contour, the convolution kernel is selected and marked in the area where the foreign body is determined.

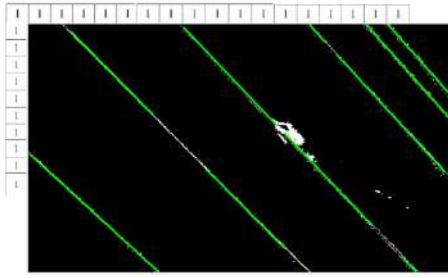


Figure 4: Power Line Image Coordinate Design

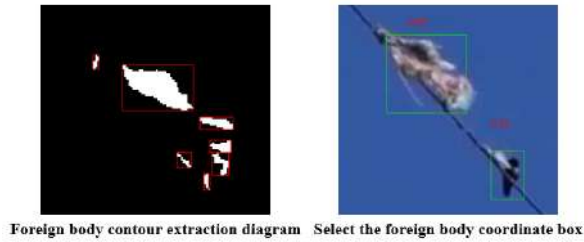


Figure 5: Foreign objects on the transmission line

#### 4.3 Determination of foreign body coordinates

An image containing only power lines and foreign objects is successfully extracted, the next step is to pinpoint the location of the foreign objects. First, a template with a value of 1 of the same length and width as the image is generated, and its size is  $1 \times N$  and  $1 \times M$ . Where  $N$  is the image width pixel value,  $M$  is the image height pixel value.

After the global coordinate design, the coordinate statistics of the area where the foreign body is judged by the improved anomaly rate in the above, in the filtered frame selection convolution rectangle, correspond to all the convolution rectangle collections ( $x, y$ ). The minimum and maximum coordinates are set as the final coordinates of  $(x_{max}, y_{max})$  the box-selected foreign object rectangle,  $(x_{min}, y_{min})$  This point is the coordinate of the position of the foreign object, thus drawing a rectangular frame where the foreign object is located.

#### 4.4 Comparative analysis of easy point rate method and improved outlier rate method

It can be seen from the above improved anomaly rate method that the improved anomaly rate method has the advantage of considering the anomaly rate method globally when detecting foreign objects in suspended transmission lines, and has a better detection effect on suspended foreign objects. And because the foreign objects hanging on the power line are usually kites, balloons, plastic bags and other large-scale existences, the feasibility and robustness of the easy point rate method are increased according to this information. After the detected abnormal point exists, the determination of the coordinates of the foreign object is analyzed, the maximum and minimum coordinates of the convolution kernel box

selection rectangle are selected, and the final box selection output of the foreign object is carried out through the determination of the coordinates, so as to improve the overall transmission line. Foreign body detection.

In the actual training test of 100 foreign object images of transmission lines, the training analysis results of the outlier rate method and the improved outlier rate are obtained, including the result table of the accuracy rate, missed detection rate, false detection rate, and the comparison of the detection results of a certain group. picture.

#### REFERENCES

- [1] T. F. Garbelim Pascoalato, P. Torrez Caballero and S. Kurokawa, "Application of the lumped parameter line model to simulate electromagnetic transients in three-phase transmission lines with vertical symmetry," in *IEEE Latin America Transactions*, vol. 20, no. 3, pp. 379-385, March 2022, doi: 10.1109/TLA.2022.9667135.
- [2] G. Yin, X. -D. Cai, D. Secker, M. Ortiz, J. Cline and A. Vaidyanath, "Impedance Perturbation Theory for Coupled Uniform Transmission Lines," in *IEEE Transactions on Electromagnetic Compatibility*, vol. 57, no. 2, pp. 299-308, April 2015, doi: 10.1109/TEM.2014.2377050.
- [3] Ricardo Justo de Araujo, R. Cleber da Silva and S. Kurokawa, "Using Universal Line Model (ULM) for Simulating Electromagnetic Transients in Three-Phase Transmission Lines," in *IEEE Latin America Transactions*, vol. 12, no. 2, pp. 190-196, March 2014, doi: 10.1109/TLA.2014.6749537.
- [4] Y. Liu, B. Wang, X. Zheng, D. Lu, M. Fu and N. Tai, "Fault Location Algorithm for Non-Homogeneous Transmission Lines Considering Line Asymmetry," in *IEEE Transactions on Power Delivery*, vol. 35, no. 5, pp. 2425-2437, Oct. 2020, doi: 10.1109/TPWRD.2020.2968191.
- [5] T. Tan *et al.*, "Research on Monitoring the Transmission Line Tension and Galloping Based on FBG Fitting Sensor," in *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-8, 2022, Art no. 7008108, doi: 10.1109/TIM.2022.3216598.
- [6] L. Xie, L. Zhao, J. Lu, X. Cui and Y. Ju, "Altitude Correction of Radio Interference of HVDC Transmission Lines Part I: Converting Method of Measured Data," in *IEEE Transactions on Electromagnetic Compatibility*, vol. 59, no. 1, pp. 275-283, Feb. 2017, doi: 10.1109/TEM.2016.2597300.

**Table 2: Foreign body detection results of two detection methods**

Detection method	Accuracy	Missing detection rate	false detection rate
anisotropy method	61.0% ( 61/100)	26.0% 26/100)	13.0% ( 13/100)
Improved outlier rate method	72.0% ( 72/100)	12.0% ( 12/100)	16% ( 16/100)

# Improved YOLOv5 UAV Target Detection Algorithm by Fused Attention Mechanism

Yan, YH, He  
Shanghai Normal University Tianhua  
College AI School, No. 1661 North  
Sheng Xin Road  
heyang\_886@163.com

Yanni, YNz, Zhao\*  
Shanghai Normal University Tianhua  
College AI School, No. 1661 North  
Sheng Xin Road  
zhao.yan.ni@126.com

Hongfei, Hfn, Nie  
Application Engineering Department  
of Special Products, Shanghai  
Microelectronics Equipment Group  
Co., Ltd  
niehf2003@163.com

## ABSTRACT

This paper proposes a modified YOLOv5 UAV target detection algorithm for the low detection accuracy caused by the dense target distribution and too small size in the UAV image. Firstly, the coordinate attention mechanism (Coordinate Attention, CA) is introduced in the backbone network CSPDarknet53 to enhance the feature extraction capability of the network; secondly, the multi-size feature pyramid network is designed to introduce a larger resolution feature map for feature fusion and prediction, and to improve the accuracy of small target detection. Experiments on the VisDrone2021 dataset, the results show that the average detection accuracy (Mean Average Precision, *mAP*) of the improved YOLOv5 algorithm reached 43.0%, 5.8 percentage points higher than the original algorithm, which fully proves the high efficiency of the proposed improved algorithm on the ground target detection of the UAV.

## CCS CONCEPTS

• **Theory of computation** → Logic; Modal and temporal logics.

## KEYWORDS

UAV target detection, multi-scale feature pyramid, YOLOv5, Coordinate Attention

## ACM Reference Format:

Yan, YH, He, Yanni, YNz, Zhao\*, and Hongfei, Hfn, Nie. 2023. Improved YOLOv5 UAV Target Detection Algorithm by Fused Attention Mechanism. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590074>

## 1 INTRODUCTION

In recent years, drone technology with its portability, mobility, high efficiency is widely used in civil fields, in the intelligent traffic management, using drones for low cruise, HD to observe image data, based on target detection technology, can be completed in the drone platform of ground vehicles, pedestrians and other target detection

identification and tracking, real-time grasp traffic information, conducive to traffic management, save a lot of manpower. Due to the flight height and the shooting perspective, the target size in the UAV image is small, and the distribution is relatively dense, which poses a great challenge to the target detection algorithm [1].

With the rapid development of stochastic deep learning in computer vision, object detection algorithms based on convolutional neural networks have become the mainstream methods. In the universal object detection algorithm, it can be divided into two categories [2], One is single-stage object detection. This algorithm uses deep convolutional neural networks (such as VGG, ResNet) to extract high-level semantic features of the image, and directly predicts the position and coordinates of the target on the feature map to realize end-to-end training, including SSD [3], RetinaNet [4], YOLO [5–8]. The other algorithm is the two-stage object detection algorithm. Compared with the single-stage algorithm, this algorithm first predicts the possible target position in the image, and then it is based on Region Proposal Network [9](RPN) Extract the features of the target position for secondary prediction. The whole process is divided into two steps: the candidate box extraction and the candidate box optimization, including the RCNN [10], Fast R-CNN [11], Cascade R-CNN [12], Mask R-CNN [13] scheduling algorithm. Compared with the two types of detection algorithms, the two-stage algorithm has better detection accuracy, but the single-stage reasoning speed is faster.

In recent years, computer vision, image and video processing and pattern recognition technologies have developed rapidly, the target detection technology in natural images is becoming more and more mature, the detection accuracy is constantly improving, and the related detection algorithms are gradually used in the target detection tasks of UAV images. Jian-xiu Yang [13] The lightweight feature extraction network and semantic information fusion module to enhance the multi-scale features of the vehicle; Zhang Ruiqian [14] On the basis of Faster R CNN, add a multiscale cavity convolutional module to increase the field perception domain, and improve the network's ability to learn the target distribution and size difference in the UAV image; Ma Jun [15] The compression-excitation module is proposed and integrated into the PP-YOLO detection algorithm, while introducing the Mish activation function, to alleviate the gradient disappearance problem during backpropagation, and further improve the detection accuracy; Liu Fang [16] The improved convolutional attention module is introduced into the residual network, and the multi-scale feature fusion module is designed to enhance the model's feature expression ability to the multi-scale targets.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590074>

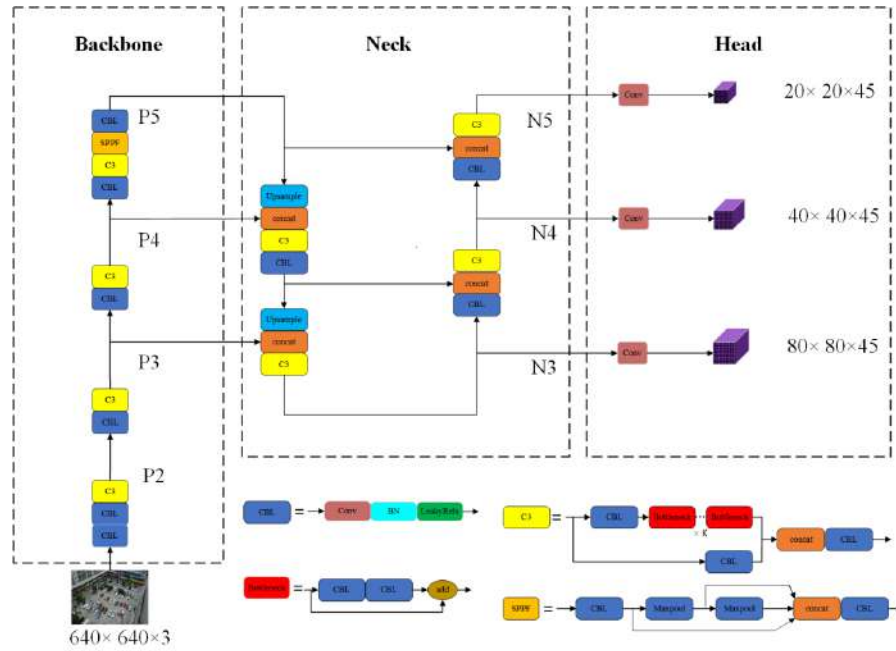


Figure 1: The YOLOv5 overall network structure diagram

Combining the above literature, the existing UAV image target detection algorithm is mainly under the natural scene of general target detector applied to the UAV image, and introduce better convolution module, but the research data set is relatively simple, and for the shielding target, small size target, the target of complex background conditions, these methods are difficult to achieve good detection effect. This paper presents a novel UAV image object detection algorithm for improving YOLOv5 for the challenging VisDrone 2021 dataset. According to the problems of dense target distribution and serious occlusion, improve the backbone network, fuse the coordinate attention mechanism, improve the feature extraction ability of the algorithm; build the four-layer feature graph fusion network, introduce the lower-level target semantic information to enrich the small target features, and improve the positioning ability of the network to the small object detection. Extensive experiments on VisDrone 2021 show that the average detection accuracy of the improved YOLOv5 algorithm reaches 43.0%, significantly improving the detection accuracy of both small and dense targets.

## 2 YOLOV5 ALGORITHM

YOLOv5 is a single-stage target detection algorithm, whose main structures include Backbone, Neck, and Head, and its overall network structure is shown in Figure 1.

YOLOv5 uses CSPDarknet53 as the backbone network, and fused the Focus structure in other detection algorithms, while maintaining the feature extraction ability to significantly reduce the computation, the overall structure mainly consists of CBL, C3 and SPPF basic units in the form of series, the network output three layers of features, including P3, P4 and P5, the downsampling multiple of 8, 16 and 32, respectively.

In the Neck network, PANet is used as the feature pyramid fusion network, and the P3, P5 and P4 feature maps output from the backbone network are used for feature fusion, and the fused feature maps N3, N4 and N5 are output. The PANet construction of top-down path enables high-level semantic information to fully interact with low-level characteristics. Meanwhile, the bottom-up path further improves the diversity of characteristics, shortens the transmission path of low-level information to the predictor head, and improves the optimization efficiency of the model.

The Head of YOLOv5 is the same as YOLOv4, based on a single convolution of the anchor frame mechanism, and the output matrix contains information about the coordinate position, category, and confidence, and so on of the target. The GIOU Loss was used during the training session [17] As a loss function, as well as post-prediction box filtering using DIOU-NMS.

## 3 IMPROVED THE YOLOV5 ALGORITHM

This paper proposes a multi-scale feature pyramid detection network based on YOLOv5 algorithm, first introducing the coordinate attention network into the backbone network [18] to improve the feature extraction capability of the network; add the P2 feature map in the triple-input Neck to construct the multi-layer feature graph fusion network. The overall framework of the algorithm is shown in Figure 2.

### 3.1 Feature extraction network combined with GA

In the VisDrone 2021 data set, there are a large number of dense vehicles and people, and the spacing between different targets is too close. The backbone network tends to introduce redundant

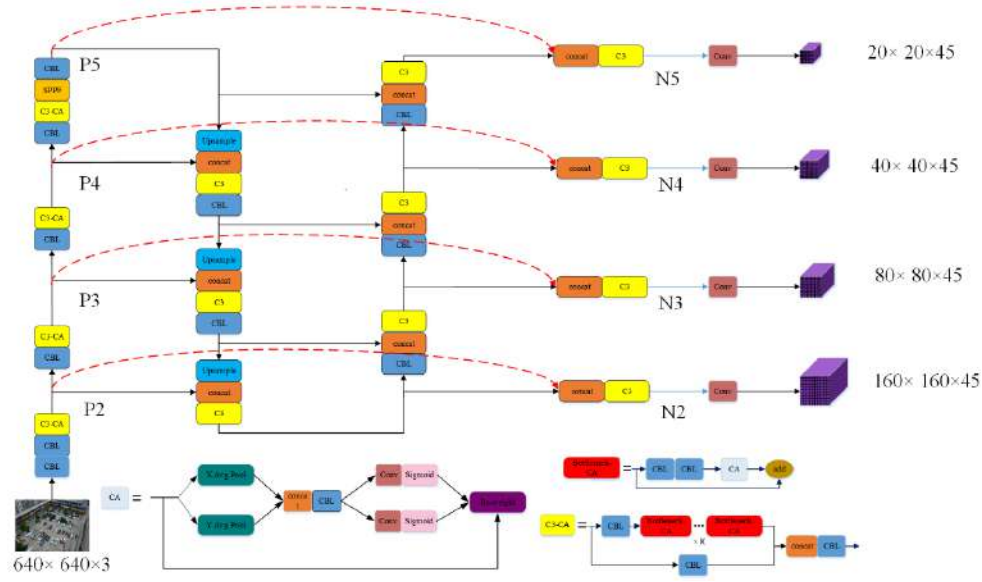


Figure 2: Improved network structure of the YOLO v 5 algorithm

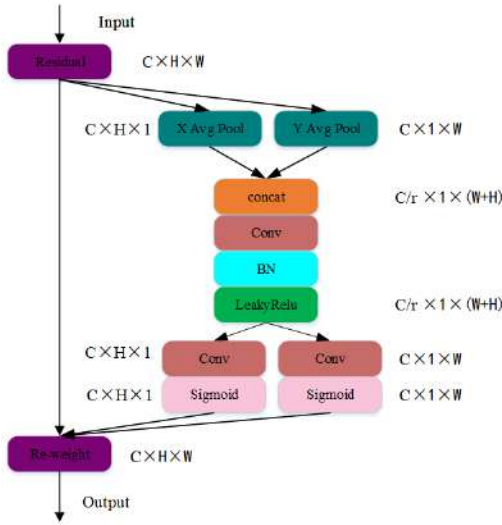


Figure 3: Coordinated attention module structure diagram

semantic information from nearby targets when extracting target features, resulting in mixed features and seriously affecting the detection accuracy. The coordinate attention network encodes the spatial attention information, and integrates it into the channel attention features, so that the location information can not only be retained in the generated feature map, but also integrates the long-range dependence along the spatial direction into the feature map, which helps the network to more accurately locate the target of interest, and enhance the feature extraction ability of the network.

The coordinate attention network is shown in Figure 3, consisting of two structures, coordinate information embedding and

coordinate attention generation. In the coordinate information embedding, in order to strengthen the remote information interaction capability of attention module in space, CA module decomposes global pooling into a pair of 1D feature coding operations. Specifically, given the input  $x$ , we use pooling kernel  $(H, 1)$  to pool all channels horizontally; similarly, all channels are pooling vertically  $(1, W)$ , and the output of  $C$  channel at height and width is expressed as:

$$z_c^h = \frac{1}{W} \sum_{0 \leq i < w} x_c(h, i) \quad (1)$$

$$z_c^w = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w) \quad (2)$$

Coordinate attention generates feature aggregate maps generated based on both Equation 1 and equation 2, first stitching them together and then passing into the shared  $1 \times 1$  convolution transformation function  $F_1$  to obtain:

$$f = \delta \left( F_1 \left( \left[ z^h, z^w \right] \right) \right) \quad (3)$$

Where  $[z^h, z^w]$  represents the channel merging of the feature plots  $z^h$  and  $z^w$ ,  $f$  represents the feature plots containing the encoding of the target spatial information, and  $\delta$  is a nonlinear activation function. The  $f$  is then split into two separate components  $f^h$  and  $f^w$  along the spatial dimension, with the two  $1 \times 1$  convolution transformations  $F_h$  and  $F_w$  being used for  $f^h$  and  $f^w$ , respectively, yielding:

$$g^h = \sigma \left( F_h \left( f^h \right) \right) \quad (4)$$

$$g^w = \sigma \left( F_w \left( f^w \right) \right) \quad (5)$$

Where,  $\sigma$  is the Sigmoid function, expand  $g^h$  and  $g^w$  separately as the attention weight of the input to get the output  $x$ :

$$y_c(i, j) = x_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (6)$$

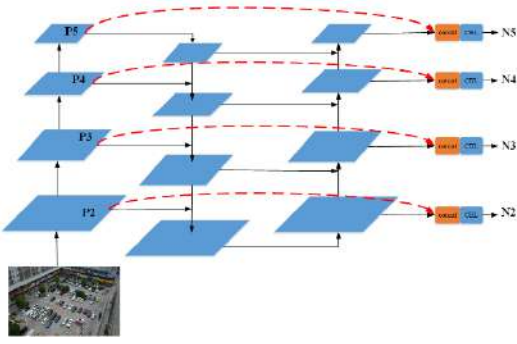


Figure 4: An improved feature fusion network

Where  $y$  represents the output feature map of the coordinate attention module,  $g_c^h$  and  $g_c^w$  applies attention in different directions to the input feature map  $x$ .

When the CA module is integrated into the backbone network, as shown in Figure 2, the CA module is added after two tandem CBL modules based on the Bottleneck, and its output is summed with the residual connection to obtain the output of Bottleneck-CA. In addition, the Bottleneck in the C3 module was replaced with Bottleneck-CA, and a backbone network of fused coordinate attention was constructed based on C3-CA.

### 3.2 Improved feature fusion networks

In UAV object detection, there are a large number of small objects, including pedestrians, and achieving a high accuracy of small object detection remains a challenging problem due to the limited image resolution. The PAN was used in the Neck of the YOLOv5 [19] as a feature fusion network, it has excellent performance, but it still has the following disadvantages: (1) only uses three-layer feature maps (P3, P4, P5) for feature fusion, its sampling rate is 8, 16 and 32, respectively, losing more detail features; (2) PAN uses a large amount of high-level semantic information, resulting in the original input feature information is submerged, which is not conducive to the detection of small targets.

In view of the problems existing in the above small target detection, improve PAN network: (1) adds high-resolution feature map with richer detail information to the feature pyramid network, introduces feature map P2 with sampling rate of 4 to feature fusion, improves the detection accuracy of small target; (2) builds cross-level connection between pyramid input feature map and output feature map, makes full use of the detail information of input feature map, enhances the interaction between high-level semantic information and low-level information, shortens the gradient back propagation path, and improves the network optimization efficiency.

The improved feature fusion networks are shown in Figures 1 and 4, Adding the P2 feature map to the feature fusion network input, For the given input feature plots P2, P3, P4, and P5, First with top-down feature fusion, For each layer, the feature map was upsampled by 2 x by nearest neighbor upsampling, Splice it with the feature map at the next level, After the C 3 module for fusion, Repeat and iterating the process down, The characteristic map after

the initial fusion is obtained; Through the bottom-up passage, The feature maps were twice upsampled, Then with the corresponding next-level feature map through the channel splicing and C3 module to obtain the secondary fusion feature map, Repeat the iteration up; Finally, by building the cross-level connections, P2, P3, P4 and P5 were stitched with the feature maps after secondary fusion, And using the C3 module to fuse the features, Obtain the final output feature figure N2, N3, Both N4 and N 5 respectively.

## 4 EXPERIMENTAL RESULTS AND ANALYSIS

### 4.1 Data set and Evaluation Indicators

This paper is based on the latest UAV dataset VisDrone 2021 [20], three objectives of pedestrian, ordinary car and bus were selected as the research objects. The training set included 6,471 images with more than 270k annotated instances, and the test set included 548 images.

To accurately assess the detection performance of the model, precision 1, recall 2,3 and 4 were used as evaluation indicators.5 To predict the ratio of the correct number to all predicted instances, 6 is the ratio of the predicted correct instances to all annotated instances, 7 is the detection accuracy of a certain category of target, which is obtained by calculating the P-R curve; 8 is the average detection accuracy, representing the average of all categories 9.

To accurately evaluate the detection performance of the model, precision  $P$ , recall  $R$ ,  $AP$  and  $mAP$  were used as evaluation indicators.  $P$  is the ratio of the number of predicted instances to the number of all predicted instances,  $R$  is the ratio of the predicted correct instances to all annotated instances,  $AP$  is the detection accuracy of a certain class of targets, obtained by calculating the P-R curves;  $mAP$  is the average detection accuracy, representing the mean of all categories  $AP$ .

The exact rate is calculated as follows:

$$P = \frac{TP}{TP + FP} \quad (7)$$

The recall rate is calculated as follows:

$$R = \frac{TP}{TP + FN} \quad (8)$$

The calculation formula of  $mAP$  is as follows:

$$mAP = \int_0^1 p(R) dR \quad (9)$$

Where  $TP$  indicates the target correctly detected,  $FN$  indicates the target not detected, and  $FP$  indicates the false alarm target falsely detected.

### 4.2 Setting of the Experimental Parameters

Experiments were performed with the Ubuntu18.04 system, processor Intel Xeon E5-2680, RTX 3090×2 accelerated neural network computation, and python3.8 and pytorch1.8 as developing language and deep learning frameworks. In the experiment, scaling, color space adjustment and mosaic enhancement were used as the data enhancement method, and each image was scaled to 640×640 size by maintaining the aspect ratio. The training process used stochastic gradient descent (SGD) for 120 epoch, with 0.002 as the initial learning rate, and the number of batch images was set to 16 plots. In the initial training stage, the learning rate preheating

**Table 1: Results of ablation experiments for modified YOLOv5 (✓ and × represent the improved module, respectively)**

Network	CA	Improved PAN	$P$	$R$	$mAP$
YOLOv5-m	×	×	78.1	54.7	37.2
	✓	×	75.6	56	39.2
	×	✓	82.6	61	42.5
	✓	✓	80.5	62.3	43.0

strategy was used, and the cosine annealing learning strategy was adopted. Use the YOLOv5 in COCO [21] Pre-trained weights on the dataset initialize the network, and perform weight-normal random initialization using the default method for the new network layers.

### 4.3 Ablation experiments

This paper performed ablation experiments on the VisDrone2021 dataset to verify the effectiveness of the proposed improved method, and the experimental results are shown in Table 1.

The analysis table shows that:

(1) Comparing the data in the first and second rows of Table 1, it can be seen that the fusion CA module increased by 2.0 and 1.3 percentage points in  $mAP$  and recall  $R$  indicators respectively than the original algorithm. The experimental results proved that the coordinate attention module can improve the global perception ability of the model and reduce the number of missed targets;

(2) Comparing the data in the first and third rows of Table 1, it can be seen that the YOLOv5 algorithm of the improved feature fusion network is improved by 6.3, 4.5 and 5.3 percentage points over the original algorithm in recall  $R$ , accuracy  $P$  and  $mAP$  indicators, respectively, indicating that the improved feature pyramid network can effectively improve the accuracy of small target detection;

(3) As shown in Table 1, line 4, combined with the two improved modules, the detection performance  $mAP$  reached 43.0%, indicating that the optimal detection performance can be achieved by using both the CA module and the improved feature fusion module.

The detection performance improvement effect of different improvement modules on different categories of targets is specifically analyzed specifically, as shown in Table 2. For the pedestrian category, Improved PAN in precision  $P$ , recall  $R$  and average detection

accuracy  $mAP$  three achieved 3.5%, 3.3% and 3.2% performance improvement, illustrates the fusion of large resolution features to detect small target is a very effective strategy, using P2 rich details and texture features enhance pedestrian target positioning, while connecting across the level of information circulation, to ensure that the original feature layer information is not submerged by high-level semantic information, improve the detection ability; On the contrary, the CA module decreased the  $mAP$  by 1.4 percentage points, and both the precision rate and the recall rate decreased to varying degrees, indicating that the CA module cannot improve the detection performance of small targets, but damages its detection accuracy to a certain extent.

For both the car and bus targets, the appearance size is larger than the pedestrian. Both improved modules can significantly improve the detection accuracy of these two targets, and the Improved PAN module is better than the CA module. In the high-resolution feature map, medium and large size targets occupy a wider range of features and have rich detailed features, which information is more important for the positioning of the target. By embedding the position information into the channel attention, CA can not only capture the local features of the target area, but also obtain the long-range dependence information, so that the model can more accurately locate and identify the overall target region.

### 4.4 Visualization of the Detection Results

The detection results of the improved YOLO v 5 algorithm and the original algorithm on the test set are compared, as shown in Figure 5 (the yellow box indicates the target area of the missing detection). In the first and second groups, the pedestrian target is shaded, and the YOLOv5 algorithm fails to identify the target. In the fourth and fifth groups, the UAV takes pictures at different overlooking angles, and the algorithm can detect more pedestrian targets and have stronger detection ability for images with different overlooking angles. From the above comparison, it can be seen that the improved algorithm has higher accuracy, significantly reduces the number of missed detection targets, and is better robust to occlusion, light, and angular changes.

### 4.5 Contrast Experiment

In this paper, experiments are conducted on the VisDrone 2021 dataset to analyze and compare the proposed improved algorithms

**Table 2: Comparison of the detection performance of various types of targets**

type	CA	Improved PAN	$P$	$R$	$mAP$
people	×	×	72.2	48.1	22.1
	✓	×	69.3	45.7	20.7
	×	✓	75.7	51.4	25.3
car	×	×	84.3	73.1	54.3
	✓	×	83.6	75	56.0
	×	✓	86.6	79.8	59.5
bus	×	×	77.7	42.9	35.1
	✓	×	73.9	47.7	40.8
	×	✓	85.5	51.8	42.7

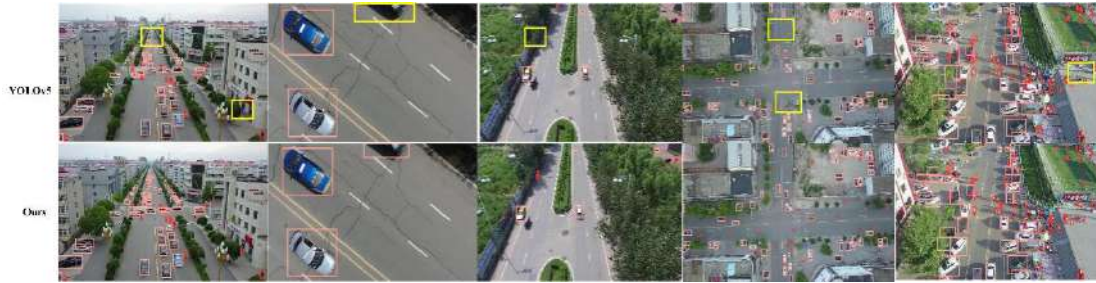


Figure 5: Visual comparison of the test results

Table 3: Comparison of the detection accuracy of the mainstream algorithms

Model	One-stage	Backbone	$mAP$
Faster-RCNN	×	Res net50	35.4
Cascade-RCNN	×	Res net50	36.0
RetinaNet	✓	Res net50	29.2
FCOS	✓	Res net50	25.4
YOLOv4	✓	CSP D arknet53	32.4
Improved YOLOv5(ours)	✓	CSP D arknet53	43.0

with other mainstream single-stage and two-stage target detection algorithms, as shown in Table 3.

In Table 3, the detection accuracy of this algorithm (43.0%) is increased by 7.6% and 7.0% compared with Faster-RCNN and Cascade-RCNN, respectively; The  $mAP$  of the single-stage detection algorithm RetinaNet with Resnet50 as the backbone network and FCOS is only 29.2% and 25.4%, much lower than the accuracy of the dual-stage target detector; Under the same backbone network conditions, Improved YOLOv5 increased by 10.6 percentage points over YOLOv4.

We show that the proposed algorithm has higher detection accuracy compared to existing both dual-and single-stage methods. Compared with the existing single-stage detection algorithm, it uses the anchor frame-based prediction method, with higher recall rate, stronger backbone network, higher robustness, and can achieve higher detection accuracy. Compared with the existing two-stage detection algorithm, the two-stage method is slower because of the strategy of secondary optimization, but the two-stage algorithm has higher accuracy compared with the dual-stage algorithm and the single-stage algorithm of the same backbone network. The proposed algorithm can achieve good results on the authoritative data set, and has a high accuracy for the UAV target detection, which is better than the comparison algorithm.

## 5 CONCLUSION

This paper proposes a modified YOLOv5 detection algorithm, which has great performance advantages over the mainstream anchor frame-based and anchor frame-free detection algorithms. The coordinate attention mechanism is introduced into the backbone network to embed the position information into the channel attention, so that the network can obtain more complete information, improve the detection accuracy of the model; improve the feature pyramid

fusion network, introduce a larger resolution feature map, further integrate the feature information of high and low layers, make full use of the underlying detail features to optimize the detection effect of pedestrians, vehicles and other small targets, enhance the original feature information, and improve the information transmission efficiency. The detection performance of the proposed algorithm has been significantly improved, which provides reference ideas for the subsequent UAV target detection algorithm, and is highly practical.

## REFERENCES

- [1] Liu Pengfei, Zhou Hai, Feng Shuichun, Bian Chunjiang. Multi-scale low-altitude UAV detection based on a modified SSD [J]. Computer Engineering and Design, 2021,42 (11): 3277-3285.
- [2] Zou Z, Shi Z, Guo Y, *et al*. Object detection in 20 years: A survey[J]. arXiv preprint arXiv: 1905.05055, 2019.
- [3] Liu W, Anguelov D, Erhan D, *et al*. Ssd: Single shot multibox detector [C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
- [4] Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [5] Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [6] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
- [7] Redmon J, Farhadi A. YOLOv3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.
- [8] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.
- [9] Ren S, He K, Girshick R, *et al*. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28: 91-99.
- [10] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [11] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [12] He K, Gkioxari G, Dollár P, *et al*. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

- [13] Yang Jianxiu, Xie Xuemei, Shi Guangming, Li Fu. The feature information-enhanced UAV real-time vehicle detection algorithm [J / OL]. Signal processing: 1-15 [2021-12-20]
- [14] Zhang Ruiqian, Shao Zhenfeng, Aleksei Portnov, Wang Jiaming. UAV image object detection method for multiscale void convolution [J]. Wuhan University Journal of Science Information Science Edition, 2020,45 (6): 895-903.
- [15] Ma Jun, Yao Zhen, Xu Cuifeng, *et al.* Multi-UAV real-time tracking algorithm based on the improved PP-YOLO and Deep-SORT [J]. Computer Application, 2021:1-10.
- [16] Liu Fang, Pu Zhaohui, Zhang Shuichao. UAV Multi-object Tracking Algorithm based on attention feature fusion [J]. Control and Decision-making, 2021:1-9.
- [17] Zheng Z, Wang P, Liu W, *et al.* Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence.2020, 34(07): 12993-13000.
- [18] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.2021: 13713-13722.
- [19] Liu S, Qi L, Qin H, *et al.* Path aggregation network for instance segmentation [C]// Proceedings of the IEEE conference on computer vision and pattern recognition.2018: 8759-8768.
- [20] Cao Y, He Z, Wang L, *et al.* VisDrone-DET2021: The Vision Meets Drone Object detection Challenge Results[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision.2021: 2847-2854.
- [21] Lin T Y, Maire M, Belongie S, *et al.* Microsoft coco: Common objects in context[C]//European conference on computer vision.Springer, Cham, 2014: 740-755.

# Haze video image Clarity Processing Based on Optical Flow Threshold

Chen Ru

College of Information Engineering, Shaanxi Institute of  
International Trade & Commerce, Xi'an 712046, Shaanxi,  
China  
24131266@qq.com.

Wang Xijuan

College of Information Engineering, Shaanxi Institute of  
International Trade & Commerce, Xi'an 712046, Shaanxi,  
China  
308332017@qq.com.

## ABSTRACT

In view of the problem of haze weather on the visual effect of video image, which causes the picture distortion, image quality degradation and definition blur of video image, a defogging processing method of haze video image based on optical flow threshold is proposed so as to restore the real and natural color image. Firstly, extract the image of the  $t$  frame at time  $t$ , track the characteristics of the image at time  $t + 1$  to time  $t + n$ , extract the image of the  $t+n$  frame, then calculate the optical flow values of the  $t$  frame and the  $t + n$  frame, make a difference between the obtained optical flow values to obtain the optical flow threshold, compare the obtained optical flow threshold with the given threshold, if the value is greater than or equal to the given threshold, take the optical flow threshold intermediate frame image, and the middle frame and  $t+n$  frame images are processed by Retinex algorithm, and this operation is performed iteratively. Finally, the processed single frame video sequence is merged into a whole and output. The experiment shows that the processing speed of the algorithm is 0.07, much lower than other processing methods, which verifies the effectiveness and innovativeness of the proposed algorithm.

## CCS CONCEPTS

• **Computing methodologies** → Computer graphics; Image manipulation; Image processing.

## KEYWORDS

Haze video, Video frame, Optical flow threshold

### ACM Reference Format:

Chen Ru and Wang Xijuan. 2023. Haze video image Clarity Processing Based on Optical Flow Threshold. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590075>

## 1 INTRODUCTION

In recent years, hazy weather occurs from time to time, and hazy weather has caused serious harm to people's daily lives. Various

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590075>

videos in traffic, daily life and the ones required in other industrial fields have been affected by haze to different degrees, resulting in the degradation of imaging mechanism and leading in the blurring visual effect of images and reduced contrast [1]. Therefore, it is of great significance to clarify the videos taken in hazy weather [2]. A video is a sequence consisting of a series of still images, and when processed as a signal, the video is essentially decomposed into a series of images for defogging; having been defogged, these images are merged into a video image. Because the video image is more complex and more tricky to deal with, if video image is defogged directly, there will be the problem of video discontinuity and flicker after processing. Therefore, defogging the video image is much more difficult than defogging a single image, and requiring high real-time performance of the algorithm.

## 2 RETINEX THEORY

Retinex theory was proposed by Edwin Land and John McCann [3] based on human color vision. They maintain that, when adjusting the color and brightness of an object, the color of the object is obtained by the reflection of lights in different colors including red, green and blue. According to the consistency principle of color perception, the image can be decomposed into an illumination component image and a reflected component image [4], as shown in Figure 1.

$$S(x, y) = L(x, y)R(x, y) \quad (1)$$

In this equation,  $S(x, y)$  is the original image,  $L(x, y)$  is the illumination component,  $R(x, y)$  the reflection component [5].

SSR is a common image enhancement algorithm, which can be expressed as:

$$r(x, y) = \log S(x, y) - \log [F(x, y) * S(x, y)] \quad (2)$$

Here  $r(x, y)$  is the output image,  $*$  is the convolution symbol, and  $F(x, y)$  is the central surround function, which can be expressed as [6]

$$F(x, y) = e^{-\frac{(x^2 + y^2)}{c^2}} \quad (3)$$

In this equation,  $C$  represents the Gaussian surround scale, which is a scale, and the value must meet the following requirements:

$$\iint F(x, y) dx dy = 1 \quad (4)$$

## 3 OPTICAL FLOW THRESHOLD

According to optical flow method, each pixel in the video image is treated as an individual, and the position of the pixel changes with the video sequence, and its essence is to estimate the motion of the

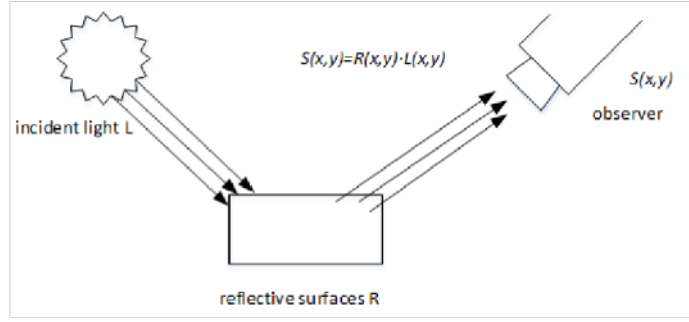


Figure 1: Image composition in Retinex theory

object through the image data [7]. Assume that the gray value at the pixel point  $(x, y)$  at time  $t$  is  $I(x, y, t)$ , the time taken to move the distance  $(dx, dy)$  to the next frame is  $dt$ , and the gray value is  $I(x + \Delta x, y + \Delta y, t + \Delta t)$ . According to the image consistency assumption, there is the following equation:

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \quad (5)$$

By Taylor expansion on the right side of equation (5), we can get:

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \frac{\partial I}{\partial x} dx + \frac{\partial I}{\partial y} dy + \frac{\partial I}{\partial t} dt + \epsilon \quad (6)$$

Of which:  $\epsilon$  is a infinite small term. Bring equation (6) into equation (5), we can be obtain:

$$\frac{\partial I}{\partial x} \frac{dx}{dt} + \frac{\partial I}{\partial y} \frac{dy}{dt} + \frac{\partial I}{\partial t} \frac{dt}{dt} = 0 \quad (7)$$

Suppose  $u = \frac{dx}{dt}$ ,  $v = \frac{dy}{dt}$ ,  $u$  represents the displacement vector of the optical flow along the X axis, and  $v$  represents the displacement vector of the optical flow along the Y axis. And  $I_x = \frac{\partial I}{\partial x}$ ,  $I_y = \frac{\partial I}{\partial y}$ ,  $I_t = \frac{\partial I}{\partial t}$  respectively represents the partial derivatives of the gray level of pixels at each position in the image along the X, Y and T directions. (7) The formula can be written as:

$$I_x u + I_y v + I_t = 0 \quad (8)$$

$I_x, I_y, I_t$  can be obtained from the image data,  $(u, v)$  is the optical flow vector.

$$\begin{aligned} u^{n+1} &= \bar{u}^n - \frac{I_x (I_x \bar{u}^n + I_y \bar{v}^n + I_t)}{\lambda + (I_x^2 + I_y^2)} \\ v^{n+1} &= \bar{v}^n - \frac{I_y (I_x \bar{u}^n + I_y \bar{v}^n + I_t)}{\lambda + (I_x^2 + I_y^2)} \end{aligned} \quad (9)$$

Through the iteration of equation (9),  $u$  and  $v$  can be calculated. Find the difference between the two optical flow values:

$$\Delta \vec{m} = |u - v| \quad (10)$$

Assume that the error threshold is  $\lambda$ , If  $\Delta \vec{m}$  is greater than or equal to  $\lambda$ , The Retinex algorithm should be used to defog the image in the middle frame of two optical flow values.

## 4 HAZE VIDEO DEFOGGING PROCESSING

In the haze video processing, multiple single images are not simply integrated into a whole, but the correlation between video frame-also should also be considered. so the processing results of video frames will directly affect the finished video.

Processing process are as follows:

- (1) The first frame image at the extraction time;
- (2) The features of the image at the first extraction time are tracked to the ones at the next extraction time, and the image of the final frame is extracted;
- (3) Calculate the optical flow values of frame 1 and frame n, and calculate the optical flow threshold;
- (4) If  $\Delta \vec{m} \geq \lambda$ , the intermediate frames are processed by using the Retinex algorithm, and if  $\Delta \vec{m} < \lambda$ , they are left unprocessed. As shown in Figure 2.

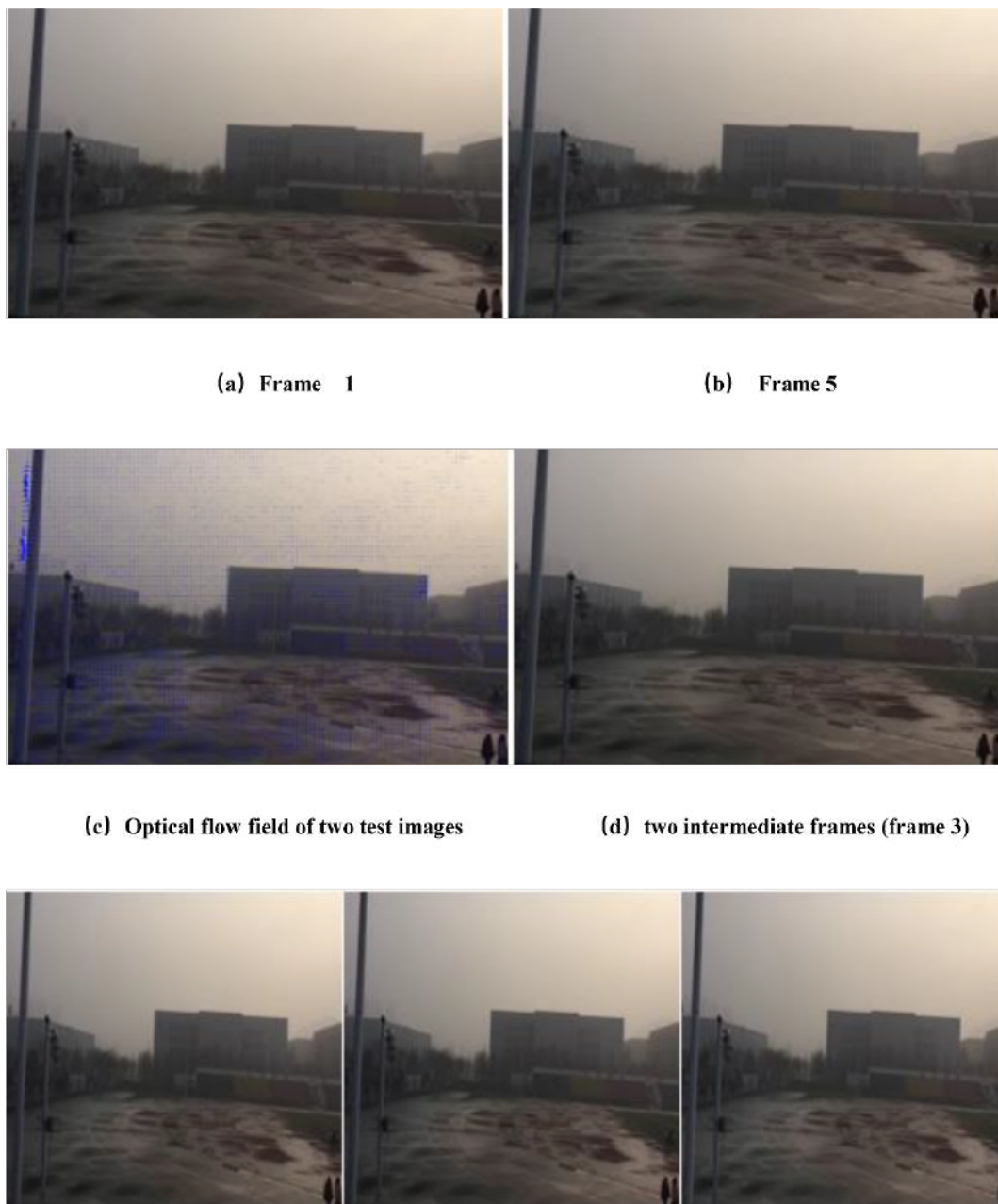
The single image was processed by Retinex algorithm, and finally the processed single frame video sequence was synthesized into the whole video output, as shown in Figure 2.

## 5 EXPERIMENTAL ANALYSIS

In order to better illustrate the real-time processing and effectiveness of the algorithm proposed in this paper, qualitative analysis is carried out through specific data to make a more accurate evaluation. When using DCP algorithm and AHE algorithm for video processing, it is necessary to defog each frame, and the whole processing takes quite a long period of time. If algorithm proposed in this paper is adopted, the processing of frames is reduced and therefore the processing speed is faster. As shown in Table 1.

## 6 CONCLUSION

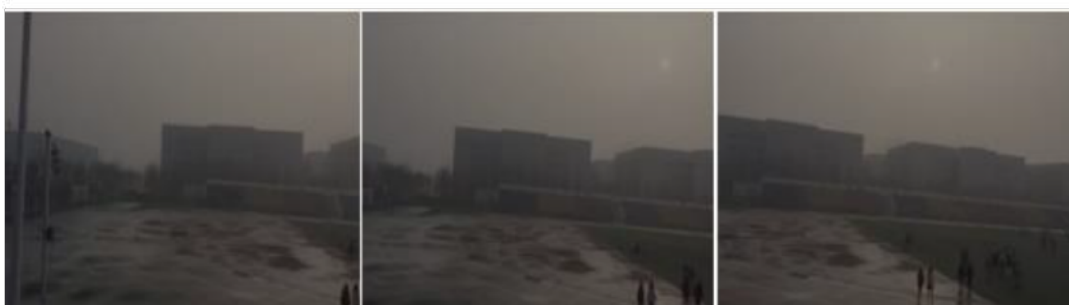
Research work has been carried out to tackle the problems of the low clarity of video images taken in haze weather and the inter-frame impact on video image processing as single images. A defogging processing method of haze video image has been proposed based on optical flow threshold. According to the image consistency assumption, light intensity of the image along the motion trajectory is constant. Having calculated the optical flow between single images, whether the image defogging processing is needed depends on the optical flow threshold, which solves the correlation between video frames. The research done in this paper is limited only to the defogging processing of images taken in the hazy weather, and

**Figure 2: Video defogging processing****Table 1: video processing speed**

Video sequence	Video frame size	DCP algorithms	AHE algorithms	algorithm adopted in this paper
Video	568*320	7.38	7.19	0.07



**Original video sequence**



**DCP algorithms**



**AHE algorithms**



**Figure 3: finished product of haze video processing**

the video defogging processing in rainy and snowy weather has not been covered yet. Future work will be done to optimize the algorithm on the basis of this research, and make it possible to process the video image in rainy and snowy weather.

## ACKNOWLEDGMENTS

Science and Technology Innovation Team of Shaanxi Institute of International Trade & Commerce "Computer Vision and Image Processing Technology"; Shaanxi Province Natural Science Basic Research Program(2021JM-539).

## REFERENCES

- [1] Wang Keping, Yang Yi, Fei Shumin. A review of haze image clarification algorithms [J/OL]. *Journal of Intelligent Systems*:1-16.
- [2] Shi Y, Xiang XG. A haze video compression method based on motion estimation sharing[J]. *Computer and Digital Engineering*, 2021, 49(03):550-555.
- [3] L and E H. Recent advances in the Retinex theory and some implications for cortical computations: Color vision and the natural image[J]. *Proceedings of the National Academy Sciences of the United States of America*, 1983, 80(16): 5163-5169.
- [4] Li Wang, YANG Jinbao, SUN Ting, FU Lingling. Retinex-based multi-scale single-image defogging network[J]. *Journal of Qingdao University (Natural Science Edition)*, 2022, 35(04):26-32.
- [5] Liu Weihua, Xue Yansong, Yichen, Wang Fuping. A low-light image enhancement algorithm combining multi-scale deep learning network and Retinex theory[J/OL]. *Signal Processing*:1-12.
- [6] Li, Can-Lin, Zhu, Jin-Juan, Liu, Jin-Hua, Bi, Li-Hua. An adaptive SSR method for low illumination image enhancement in foggy days[J]. *Computer Application and Software*, 2022, 39(09):233-239+268.
- [7] Shao X. Q., Yang Y., Liu Y. L.. A review of research on optical flow algorithms for fluid motion estimation[J]. *Chinese Journal of Graphics*, 2021, 26(02):355-367.

# Gaussian-guided character erasure for data augment of industrial characters

Hongchao Gao  
gaohongchao@optmv.com  
OPT Machine Vision Tech Co  
Dongguan, China  
South China University of Technology  
Guangzhou, China

Chao Yao  
chaomi\_yc@163.com  
South China Normal University  
Dongguan, China

Zhennan Wang  
wangzhennan@optmv.com  
OPT Machine Vision Tech Co  
Dongguan, China

## ABSTRACT

The application of scene text erasure technology in privacy protection, camera-based virtual reality translation and image editing has attracted more and more research interests. Recent efforts on scene text erasing have shown promising results. We utilize text removal methods as a component of industrial characters generation procedure to generate large-scale synthetic character images so as to mitigate the issue of insufficient samples in the recognition task of industrial characters. Existing character erasure models have achieved good performance in natural scenes. However, in industrial scenes, these erasure networks are easily affected by salient non-character regions leading to the attention shift. To overcome this limitation, we proposed a character erasure network based on attention mechanism which embed an additional region awareness layer to guide attention to the correct character regions. Meanwhile, we devise a gaussian heat map supervision method for learning additional region awareness layer. The experiments show that the proposed method performs favourably on four industrial character datasets.

## CCS CONCEPTS

• **Theory of computation** → *Unsupervised learning and clustering*.

## KEYWORDS

character erasure, neural networks, attention shift, region awareness

### ACM Reference Format:

Hongchao Gao, Chao Yao, and Zhennan Wang. 2023. Gaussian-guided character erasure for data augment of industrial characters. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590077>

## 1 INTRODUCTION

Inspired by the notable success of deep learning in image transform, GAN [6]-based text erasing approaches have attracted increasing

interest because of its wide range of applications such as personal information protection [9], image restoration [18], image data augment [17, 19].

Scene text erasing research can be roughly divided into two groups: one-step and two-step methods. One-step methods [14, 21] were end-to-end way to character locating and erasure function. EnsNet [25] is the first end-to-end framework to remove text at a whole image level. The drawback is that the text localization mechanism of these networks is weak, and the text-erasing process is not controllable. To overcome the shortcoming in locating character, MTRNet++ [21] utilizes an auxiliary mask to improve the text detection branch. After that, we found that the U-GAT-IT [11] was effective in focusing on character regions and achieving good erasure results in the character erasure task under the help of attention mechanism. However, unlike natural scenes, industrial scenes often have highlights and high contrast areas, which can affect the model's judgement of character region. In this case, we believe that character region features are easily ignored when utilizing the above algorithms for industrial scene text generation. In general, character erasure models employ semi-supervised learning, which leaves the model without targeted instruction and frequently leaves it open to the effect of prominent non-character regions during the learning process, leading to attention shift, as shown in the Figure 1. The two-step approaches decouple the character erasure task into two main parts: text detection and background inpainting. MTRNet [22] presents a two-stage coarse-to-fine network with an additional segmentation head. Although they have obtained remarkable improvements, their erasing quality strongly relies on a great amount of annotated data, which requires significant economic and labor costs.

In this paper, we proposed a new end-to-end model for text erasure. The algorithm uses a Gaussian heat map [3, 12, 15, 24] to guide the model to focus more on character regions. The Gaussian heat map is obtained by encoding the probability of the center of the character using a Gaussian distribution. In contrast to the binary partition map that discrete identifies each pixel, the Gaussian heat map is extremely versatile in dealing with character regions without rigid constraints. Additionally, we incorporate a region awareness mechanism into the model upsampling procedure by adding an additional channel to the generated image to calculate a region score, indicating the probability that a given pixel is a character's center. The model attention region is corrected by loss between Gaussian heat map and this region-score layer. The region awareness mechanism not only focuses on the information about the region to which the string belongs, but also contains more

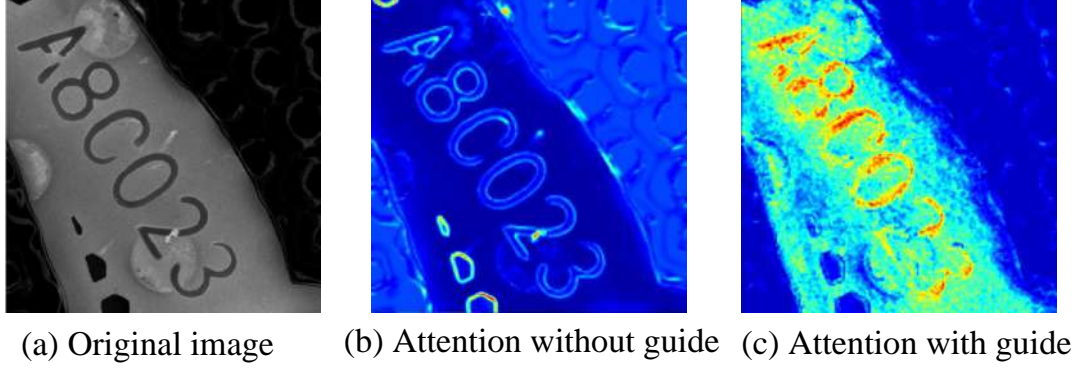
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590077>



**Figure 1: Visualization of attention.**

detailed local information about the character. The combination of Gaussian heat map and attention domain makes the model pay more attention to the domain character region and thus has better erasure effect.

In summary, the contribution points of this paper as follows:

1) we design a new character erasure model, which introduces some targeted guidance to the model by introducing a Gaussian heat map to avoid the model being influenced by saliency features, and by adding a region awareness mechanism to the model, which makes the model focus more on the regions where characters are present and eventually generates erasure images with better results.

2) On this basis, we combine character style migration and character region fusion to design a character generation strategy for industrial scenes, which can generate large batches of high-quality known and unknown industrial scene text data.

In this section, the specifics of the character erasure approach put forward in this study are presented. The network architecture is schematically illustrated in Figure 2. The character erasure model consists of two generators  $G_{s \rightarrow t}$  and  $G_{t \rightarrow s}$ , two discriminators  $D_s$  and  $D_t$ , as well as the Gaussian Guide. We integrate the attention module into both generator and discriminator. The generator in turn consists of Encoder  $\epsilon_s$ , Decoder  $\sigma_s$  and Auxiliary classifier  $\eta_s$ . In the character erasure model, the model structure of generator and discriminator is similar, so we only show the generator and the Gaussian Guide. In Figure 2,  $X_s$ ,  $X_t$  are the sets of original and erased character images, respectively, such that  $x \in \{X_s, X_t\}$  is passed into the model as a set of source and target domain samples for training. The incoming images are first passed through  $\epsilon_s$  to obtain the encoded feature maps  $f_c$ . Then, an auxiliary classifier based on the attention mechanism will extract the high semantic vectors from encoded feature maps. After that, the  $\sigma_s$  decodes the results generated by  $\eta_s$  into the generated output  $\mathcal{Y}$ . Where  $\mathcal{Y}$  consists of three parts, Generating Images  $GI$ , Attention Mask  $AM$ , and Region Score  $RS$ .  $AM$  combined with  $GI$  can effectively obtain the region of character erasure, and the remaining region is filled with the original input image. Then, we go into detail about three essential parts of our approach, i.e. the Auxiliary Classifier (Sec.2.1), the Decoder (Sec.2.2), and the Gaussian Guide (Sec.2.3).

## 2 METHODOLOGY

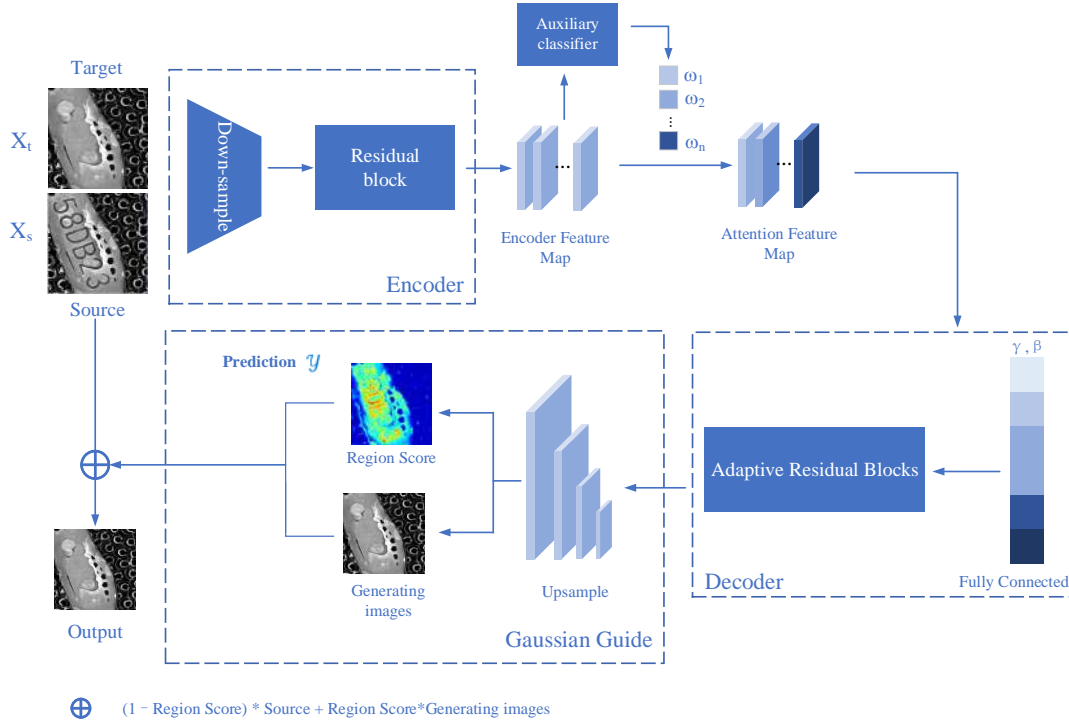
In this section, the specifics of the character erasure approach put forward in this study are presented. The network architecture is schematically illustrated in Figure 2. The character erasure model consists of two generators  $G_{s \rightarrow t}$  and  $G_{t \rightarrow s}$ , two discriminators  $D_s$  and  $D_t$ , as well as the Gaussian Guide. We integrate the attention module into both generator and discriminator. The generator in turn consists of Encoder  $\epsilon_s$ , Decoder  $\sigma_s$  and Auxiliary classifier  $\eta_s$ . In the character erasure model, the model structure of generator and discriminator is similar, so we only show the generator and the Gaussian Guide. In Figure 2,  $X_s$ ,  $X_t$  are the sets of original and erased character images, respectively, such that  $x \in \{X_s, X_t\}$  is passed into the model as a set of source and target domain samples for training. The incoming images are first passed through  $\epsilon_s$  to obtain the encoded feature maps  $f_c$ . Then, an auxiliary classifier based on the attention mechanism will extract the high semantic vectors from encoded feature maps. After that, the  $\sigma_s$  decodes the results generated by  $\eta_s$  into the generated output  $\mathcal{Y}$ . Where  $\mathcal{Y}$  consists of three parts, Generating Images  $GI$ , Attention Mask  $AM$ , and Region Score  $RS$ .  $AM$  combined with  $GI$  can effectively obtain the region of character erasure, and the remaining region is filled with the original input image. Then, we go into detail about three essential parts of our approach, i.e. the Auxiliary Classifier (Sec.2.1), the Decoder (Sec.2.2), and the Gaussian Guide (Sec.2.3).

### 2.1 Auxiliary Classifier Module

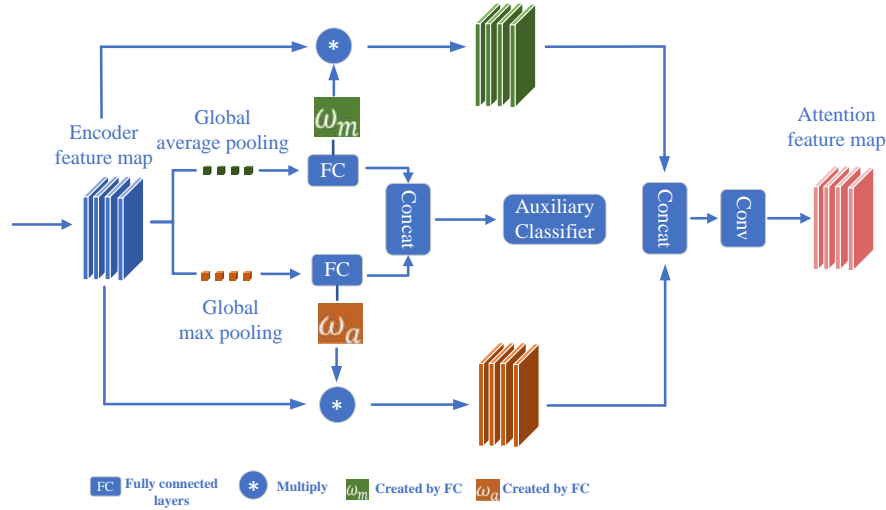
An overview of the Auxiliary Classifier Module inspired by CAM [1, 26] is depicted in Figure 3. The module is to generate attention feature map focusing on more important regions and ignore minor regions. We first aggregate information of the Encoder feature map by using both average-pooling and max-pooling operations, generating two different context descriptors, and then, use fully connected layer to produce two attention mask  $\omega_m$  and  $\omega_a$  by reducing the dimensionality of two different context descriptors to one dimension. After that, the Encoder feature map is modulated by two attention mask respectively to produce two sets of feature maps which is connected to generate attention feature map.

### 2.2 Decoder

Decoding process as shown in Figure 4, where  $\mu_I$ ,  $\mu_L$  and  $\sigma_I$ ,  $\sigma_L$  are channel-wise, layer-wise mean and standard deviation respectively,



**Figure 2: The overall structure of our character erasing module.**



**Figure 3: The pipeline of auxiliary classifier module. Where FC represents full connect,  $\omega_m$  and  $\omega_a$  are parameters generated by the fully connected layer.**

$\gamma$  and  $\beta$  are dynamically computed by a fully connected layer from the attention map. The value of  $\rho$  is initialized to 1 in the residual blocks of the decoder and 0 in the up-sampling blocks of the decoder. To decode attentional information in the attention feature map, we weight AdaLin normalized results with parameters  $\gamma$  and  $\beta$ . Explicitly, AdaLIN [13] (Yellow Box) combines AdaIN [8] and LN [2] selectively retain or change the content information and maintain

the content structure of the original domain while changing the image features.

### 2.3 Gaussian guide

We propose the Gaussian map approach to address the attention drift generated by the model during the attention mechanism. Specifically, we set an additional feature layer, the region-score layer *RSL*, to give a region score, which represents the probability

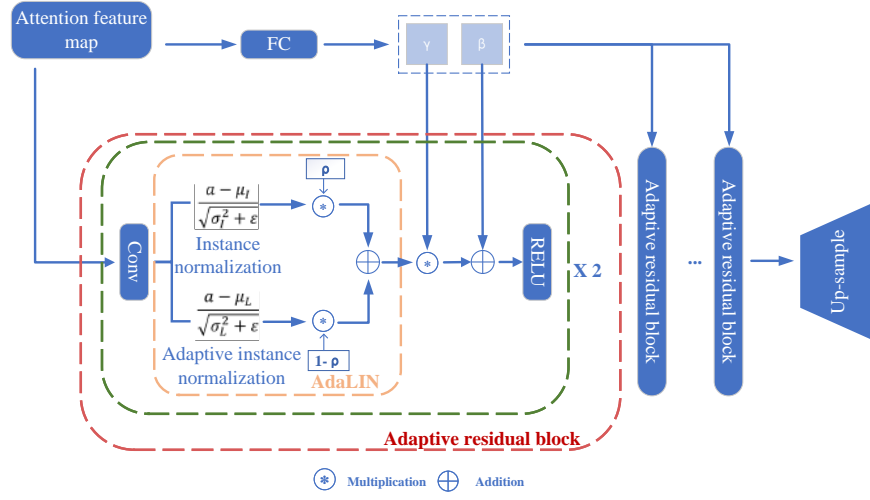


Figure 4: The pipeline of decoder.

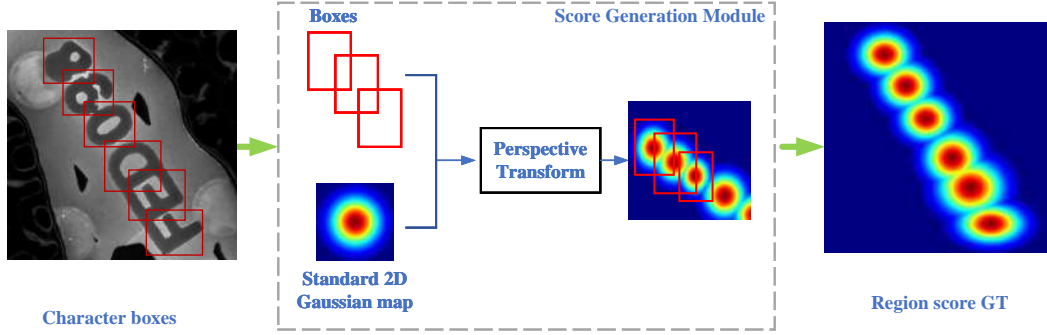


Figure 5: Illustration of gauss image generation procedure in our framework.

that a given pixel is the center of a character. The Gaussian heat map has been used in other applications, such as the pose estimation work [4, 5]. It provides a high level of flexibility when working with target regions that lack clearly defined borders. Therefore, we use a Gaussian heat map to characterize the character region.

For each training image, we generate the ground truth label for the region score, as shown in Figure 5. Since the character boxes on the image usually overlap together, causing deformation of the graph, we warp a 2-dimensional isotropic standard Gaussian head map to the character box selection region.

Visualize the update process of region score as training progresses is provided in Figure 6. In the early stages of training, the text regions in the industrial scene images are not familiar, so the region scores are relatively low. As the training progresses, the background generation module can handle the character regions more accurately and the confidence scores gradually increase.

## 2.4 Loss function

The overall objective of the model consists of five loss functions. We use least squares [16] to stabilize the training by taking generated or real images as input and trying to distinguish between them.

**Adversarial loss**  $G_{s \rightarrow t}$  is trained to minimize  $L_{lsgan}^{s \rightarrow t}$  and  $D_t$  trained to maximize it. The goal is to match the distribution of the generated images to the target domain.

$$L_{lsgan}^{s \rightarrow t} = - \left( E_{x \sim X_t} [(D_t(x))^2] + E_{x \sim X_s} [(1 - D_t(G_{s \rightarrow t}(x)))^2] \right) \quad (1)$$

**Cycle loss** We apply cyclic consistency to ensure that source domain images transferred to the target domain can be restored to the source domain. This approach mitigates the model crash problem.

$$L_{cycle}^{s \rightarrow t} = E_{x \sim X_s} [|x - G_{t \rightarrow s}(G_{s \rightarrow t}(x))|_1] \quad (2)$$

**Identity loss** Ensures continuity between the output image and the input image.

$$L_{identity}^{s \rightarrow t} = E_{x \sim X_t} [|x - G_{s \rightarrow t}(x)|_1] \quad (3)$$

**CAM loss** Ensures continuity between the output image and the input image.

$$L_{cam}^{s \rightarrow t} = - (E_{x \sim X_s} [\log(\eta_s(x))] + E_{x \sim X_t} [\log(1 - \eta_s(x))]) \quad (4)$$

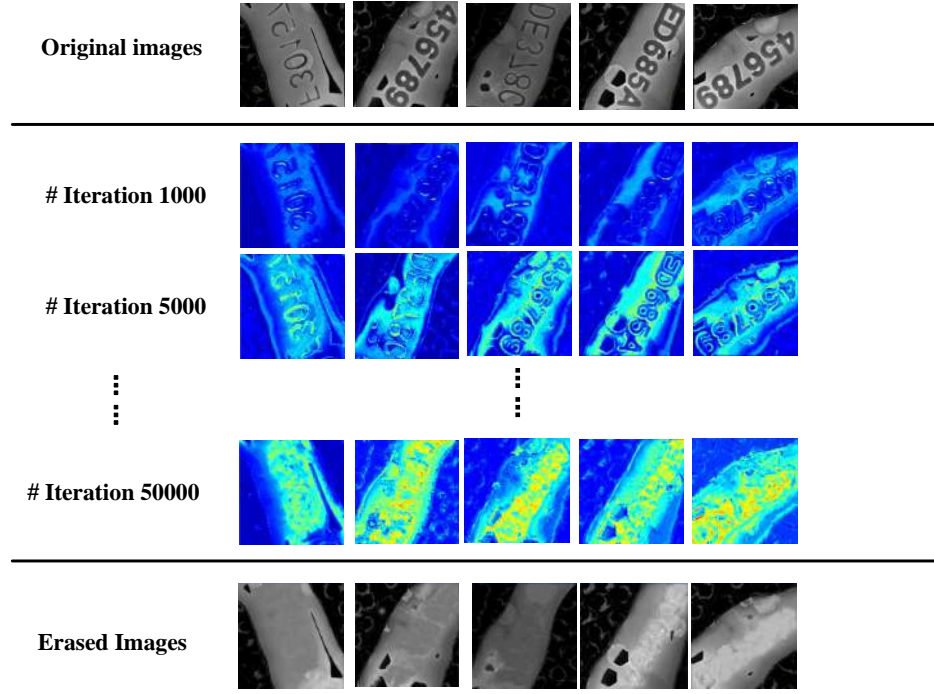


Figure 6: Visualize the update process of region score as training progresses

$$L_{cam}^{D_t} = E_{x \sim X_t} [(\eta D_t(x))^2] + E_{x \sim X_s} [(1 - \eta D_t(G_{s \rightarrow t}(x)))^2] \quad (5)$$

**Region-score loss** Used to correct the model’s region of interest to avoid attention shift.

$$L_{\text{region-score}} = E_{x \sim X_t} \left[ \sum_p \|x - G_{s \rightarrow t}(p)\|_2^2 \right] \quad (6)$$

**Full loss** Finally, we jointly train all modules to optimize the final goal.

$$L = \lambda_1 L_{lsgan}^{s \rightarrow t} + \lambda_2 L_{\text{cycle}}^{s \rightarrow t} + \lambda_3 L_{\text{identity}}^{s \rightarrow t} + \lambda_4 L_{cam} + \lambda_5 L_{\text{region-ware}} \quad (7)$$

### 3 EXPERIMENT

In this part, we will assess the character erasure method proposed. Before comparing our method with other methods, we first show the dataset we use. After that, the ablation study assesses the efficiency of every module of our approach and how it impacts the outcomes.

#### 3.1 Datasets and evaluation metrics

In this paper, we conducted experiments on four different classes of industrial character data: Plastic Surface (PS), Charger Shell (CS), SIM Card (SC), and SIM Card Slot (SCS). The four datasets contain 157, 40, 3400, and 4160 original images, respectively. In the character erasure experiments, we select 15 character region images of the dataset and the corresponding non-character backgrounds as the input of the character erasure module for training, where the non-character backgrounds are erased manually. Then, we measure the quality of the image erasure according to the common metrics FID [7], SSIM [23], and L2 loss.

#### 3.2 Character Erasure Experiment Results

We compare our approach with the previous methods, including CycleGAN [27], pix2pix [10], AttentionGAN [20] and U-GAT-IT [11]. We retrain all models on these four datasets to fully evaluate the efficacy of our approach.

Table 1 shows the erase performance of our approach and other competitors. We conducted the experiment on four datasets, our approach is superior to previous methods in all similarity metrics from pixel-level to perceptual-level.

We can know that the attention-based AttentionGAN and U-GAT-IT do not perform well in scenarios with salient features, such as PS and CS datasets. The reason is that the semi-supervised learning lacks certain guidance, and those salient regions tend to occupy larger weights when extracting features, therefore, it is easy to cause attention drift during the learning process, which eventually affects the image generation. Despite the fact that our approach relies on the attention mechanism, when we specify specific guidelines, the model will pay more attention to the targeted regions and produce better experimental findings.

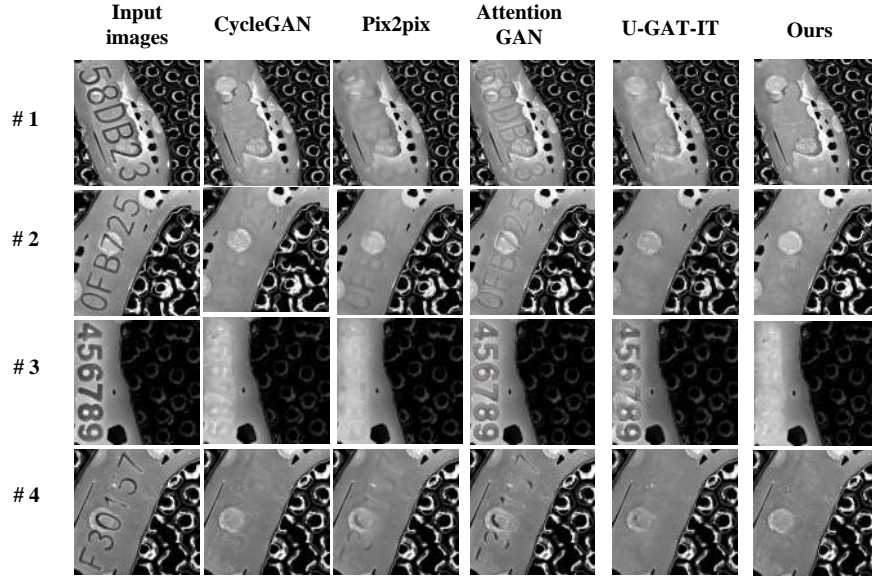
In order to intuitively reflect the performance difference between the proposed model and the comparison model, we show the test results of different methods on the four images of the Plastic Surface dataset in Figure 7. The ability of the model we proposed to remove the text from an image and restore the background is excellent compared with other models.

#### 3.3 Ablation studies

In this part, we have examined the impacts of the Gauss Map (GM), Region Awareness Layer (RWL) and Attention Mask (AM) components on character erasure. The overall evaluation results on

**Table 1: Results on 4 different datasets, such as Plastic surface and Charger shells ect . Where FID, SSIM and L2 loss used as the metrics on generating images.**

Methods	Plastic surface			Charger shell			SIM card			Card slot		
	FID	SSIM	L2 loss	FID	SSIM	L2 loss	FID	SSIM	L2 loss	FID	SSIM	L2 loss
CycleGAN	42.99	0.8462	36.42	160.56	0.867	56.12	44.84	0.9659	29.98	73.36	0.969	16.27
Pix2pix	110.25	0.7328	115.46	259.08	0.7171	224.94	27.74	0.992	8.48	153.16	0.920	256.15
AttentionGAN	134.29	0.530	78.41	157.23	0.556	36.57	21.52	0.996	2.80	110.13	0.914	161.80
U-GAT-IT	103.49	0.753	57.18	165.69	0.798	63.03	61.22	0.966	89.99	70.31	0.961	63.61
<b>Ours</b>	35.22	0.951	22.20	44.95	0.9491	18.27	22.34	0.995	3.23	54.19	0.975	23.01

**Figure 7: Test results of different methods on Plastic Surface dataset.****Table 2: The quantitative results of ablations of our method.**

Plastic surface					
Mask	RWL	GM	FID	SSIM	L2 loss
✓	✓	✓	35.22	0.951	22.20
✓	✓	✗	44.91	0.879	32.84
✓	✗	✗	69.78	0.878	36.51
✗	✗	✗	103.49	0.753	57.18

Plastic surface datasets are shown in Table 2. By gradually eliminating the GM, RWL, and AM components, we can see that each of these components has a favorable impact on the original model. The model with three modules has the best performance in all evaluation metrics.

For the purpose of assessing how GM affected the model, we swapped out the Gaussian heat map for a binary segmentation map, i.e., the region where the character’s target box is selected is set to white and the rest is assigned to black. The results demonstrate that the binary partition map is useful as a guide for character erasure, but this rigid regional constraint renders the model less effective for erasing the features of the region, which may lead to the characters staying in the local erasure. The difference between binary segmentation map and Gaussian map is shown in Figure 8.

When we do not place any constraints on attention, the model is influenced by the saliency features of the surrounding background. In Figure 9, by comparing the model’s attention during character erasure, we find that unconstrained attention causes the model to focus more on salient non-character regions, thus limiting the effect of character erasure.

## 4 CONCLUSION

It is experimentally shows that salient non-character regions can affect the performance of the model in erasing characters. In order to prevent the effects of salient non-character regions, we propose a new character erasure model in this study that is successfully steered model to focus on character regions under an attention mechanism. The limitation of manually erasing the character image indirectly affects the effect of the final erasing character of the model. When there is a gradient light in the image, the erased part cannot be blended naturally with the background manually, and therefore the results generated by the model in this region cannot be blended naturally into the background image.

## ACKNOWLEDGMENTS

This work was supported by the Scientific and Technological Innovation Program (No. 2022A0505020028), jointly funded by Guangdong and Macao.

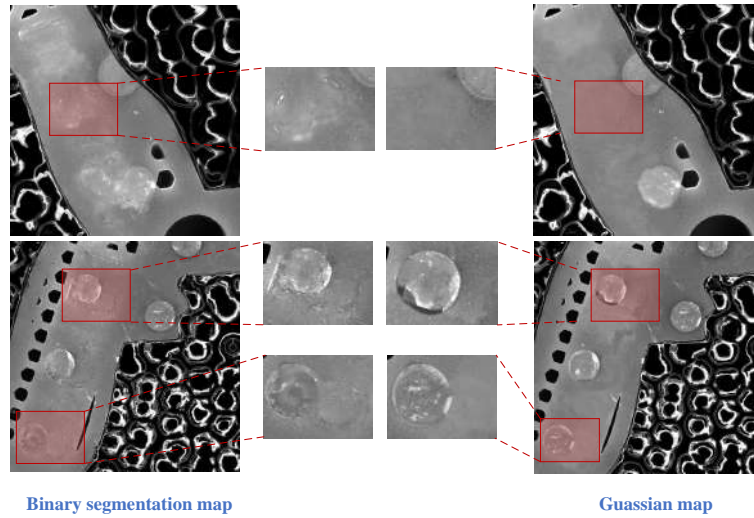


Figure 8: Difference between binary segmentation map and Gaussian map

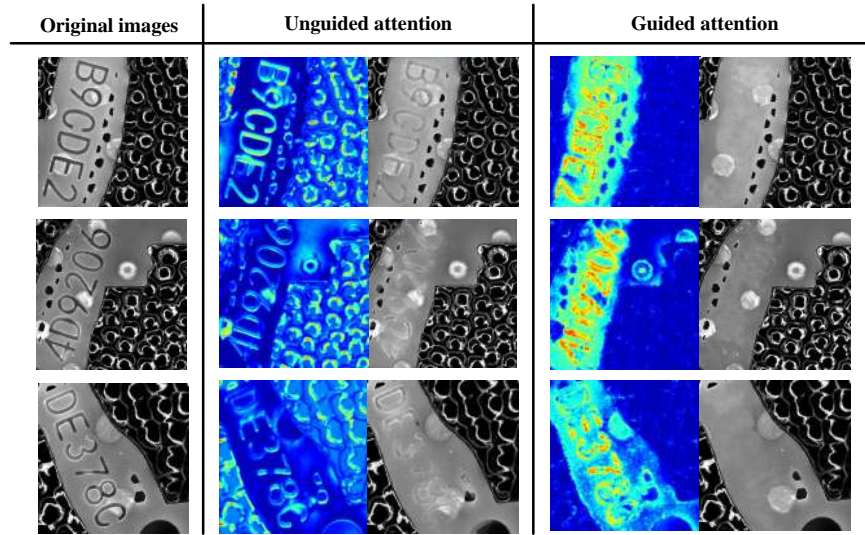


Figure 9: Visualization of attention

## REFERENCES

- [1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Benetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [3] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9365–9374.
- [4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.
- [5] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2969–2978.
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).
- [7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [8] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.
- [9] Kohei Inai, Mårten Pålsson, Volkmar Frinken, Yaokai Feng, and Seiichi Uchida. 2014. Selective concealment of characters for privacy protection. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 333–338.
- [10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the*

- IEEE conference on computer vision and pattern recognition*. 1125–1134.
- [11] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. 2019. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830* (2019).
  - [12] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. 2020. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11474–11481.
  - [13] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9361–9370.
  - [14] Chongyu Liu, Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Yongpan Wang. 2020. EraseNet: End-to-end text removal in the wild. *IEEE Transactions on Image Processing* 29 (2020), 8760–8775.
  - [15] Shangbang Long, Xin He, and Cong Yao. 2021. Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision* 129, 1 (2021), 161–184.
  - [16] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.
  - [17] Giovanni Mariani, Florian Scheidegger, Roxana Istrate, Costas Bekas, and Cristiano Malossi. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655* (2018).
  - [18] Toshiaki Nakamura, Anna Zhu, Keiji Yanai, and Seiichi Uchida. 2017. Scene text eraser. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Vol. 1. IEEE, 832–837.
  - [19] Sungho Suh, Haebom Lee, Paul Lukowicz, and Yong Oh Lee. 2021. CEGAN: Classification Enhancement Generative Adversarial Networks for unraveling data imbalance problems. *Neural Networks* 133 (2021), 69–86.
  - [20] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. 2021. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems* (2021).
  - [21] Osman Tursun, Simon Denman, Rui Zeng, Sabesan Sivapalan, Sridha Sridharan, and Clinton Fookes. 2020. MTRNet++: One-stage mask-based scene text eraser. *Computer Vision and Image Understanding* 201 (2020), 103066.
  - [22] Osman Tursun, Rui Zeng, Simon Denman, Sabesan Sivapalan, Sridha Sridharan, and Clinton Fookes. 2019. MTRNet: A generic scene text eraser. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 39–44.
  - [23] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
  - [24] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. 2020. Gliding vertex on the horizontal bounding box for multi-oriented object detection. *IEEE transactions on pattern analysis and machine intelligence* 43, 4 (2020), 1452–1459.
  - [25] Shuaitao Zhang, Yuliang Liu, Lianwen Jin, Yaoxiong Huang, and Songxuan Lai. 2019. Ensnet: Ensconce text in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 801–808.
  - [26] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929.
  - [27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.

# A 3D discrete memristive chaotic map and its application in image encryption

Junwei Shen

shen\_jun\_wei@163.com

School of Computer Science & School of Cyberspace Science, XiangTan University  
Xiangtan, Hunan, China

## ABSTRACT

In recent years, researchers proposed many discrete memristive models. And the performance of chaotic map can be improved by using discrete memristor. In this paper, a kind of discrete chaotic map is studied. First, the map is cascaded with memristor to generate a new discrete memristor chaotic map. The dynamic behavior of discrete memristor chaotic map is analyzed. Numerical simulations demonstrate that the proposed map has complex dynamics, like hyperchaos and coexisting attractors. Then, based on the proposed memristive map and DNA coding, an image encryption algorithm is designed and its security and robustness are analyzed. Experimental results show that the algorithm can effectively resist plaintext attacks and has good robustness.

## CCS CONCEPTS

• Security and privacy → Block and stream ciphers.

## KEYWORDS

discrete memristor, chaotic map, cascading, image encryption, DNA code

### ACM Reference Format:

Junwei Shen. 2023. A 3D discrete memristive chaotic map and its application in image encryption. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3590003.3590078>

## 1 INTRODUCTION

With the development of science and technology, images become the main form of information transmission. More secure and efficiently image encryption algorithm is a popular reach topic nowadays in the field of personal privacy protection, cloud storage, medical privacy, etc[12, 15, 17]. Data encryption is the last barrier of information security. Due to the characteristics of image information relative to text information, large amount of information, high correlation between pixels, and high redundancy, traditional encryption algorithms are not suitable for image encryption[4, 19, 25]. In recent decades, plenty of image encryption algorithms are proposed [16, 24, 31].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590078>

Chaotic sequences produced by chaotic map have characteristics such as pseudo-randomness, initial value sensitivity, and unpredictability[27, 30]. The more complicated the chaotic sequence is, the more secure the encrypted image is. The discrete chaotic maps mostly exhibit the disadvantages of low Lyapunov index and small key space. These defects are the hidden danger of image encryption algorithm based on chaotic map [21]. In order to improve the security of the algorithm, a more complicated hyperchaotic map is needed to constructed to increase the complexity of sequences [1].

In 1971, professor Chua proposed the concept of memristor [6], and decades later, Hewlett-Packard made the first nanoscale memristor, which received academic attention [26]. In practical applications, memristor has the characteristics of nonlinearity, memory and switching mechanism, so it has a very good application prospect in the field of artificial intelligence(AI) [14], neural networks [29], logic in memory [9]. Researchers found that memristors have broad application prospects in the field of chaotic secure communication. And discrete memristors are easier to implement in hardware than continuous memristors. Peng et al. [20] applied the discrete memristor to Hénon map and found that the performance of Hénon map was improved. Yuan et al. [32] introduced a method for constructing chaotic and hyperchaotic cascaded memristor discrete map. The new map has larger chaotic area, more control parameters and more complex chaotic behavior.

Bao et al. [2] proposed a general-purpose discrete memristor model and gives four discrete memristors based on this model. In article [3], the cosine memristor is coupled with five maps, and the chaotic properties of the coupled maps are enhanced. Qian et al. [23] presented an image encryption algorithm with one round of bit-level permutation and two rounds of DNA (Deoxyribonucleic acid) level confusion and diffusion based on the chaotic sequence. Li et al. [11] proposed a color image encryption method, which a random DNA matrix generated by a newly hyperchaotic system is used to diffusion. However, some scholars proved that some image encryption algorithm still has problems in security and reliability [10, 13, 22, 28]. Therefore, chaotic map and image encryption need to be enhanced to improve the security and reliability of image algorithms.

Section 2 introduces the construction of hyperchaotic map, and analyzes the performance of chaotic map from the Lyapunov exponent, bifurcation, basin of attraction. Section 3 introduces an image encryption algorithm that combines plain information with chaotic sequences, which uses DNA encode and compute technology. Section 4 analyzes the security of the encryption algorithm from the aspects of histogram, pixel correlation, key sensitivity, plaintext sensitivity and information entropy. In Section 5, the robustness

of the encryption algorithm is analyzed from two aspects: noise resistance and shear resistance. Section 6 concludes this paper.

## 2 CHAOTIC MAP

In this section, several chaotic maps are mainly introduced through Lyapunov exponent etc. The Lyapunov exponent is the exponential separation rate of the initial approach trajectory over time. It is an essential tool in studying chaotic signal.

### 2.1 Seed map

The seed map is a two-dimensional chaotic map, and the mathematical formula is as follows,

$$\begin{cases} x_{i+1} = x_i + a \sin(y_i), \\ y_{i+1} = y_i + b \sin(y_i) \sin(x_i), \end{cases} \quad (1)$$

where  $a$  and  $b$  are control parameters. Its Lyapunov exponent and bifurcation plots are shown in Fig. 1 and Fig. 2. According to these figures, when  $b = 4.4$ ,  $a \in (-1.1, 1.1)$ ,  $a = -1$ ,  $b \in (4.01, 4.27) \cup (4.36, 4.63)$  and the initial value  $x_0 = -4.6$ ,  $y_0 = -5$ , the map is chaotic. This map has a good complexity but no hyperchaotic properties, so this map needs to be improved.

### 2.2 S-DM map

The memristor model and the discrete memristor map (2) where  $M(q_n) = \sin(\pi q_n)$  is obtained from article [2]. Its Lyapunov exponent diagram and bifurcation diagram are shown in Fig. 3. Within a certain range of parameters, the discrete memristor map exhibits hyperchaotic behavior.

$$\begin{cases} v_n = M(q_n) i_n, \\ q_{n+1} = q_n + i_n, \end{cases} \quad M(q_n) = \sin(\pi q_n) \Rightarrow \begin{cases} x_{n+1} = h \sin(\pi q_n) x_n, \\ q_{n+1} = q_n + x_n. \end{cases} \quad (2)$$

### 2.3 New map after cascading

In general, cascading and coupling are common methods to obtain higher-dimensional maps. Cascading is which one input signal of one circuit is replaced by the output signal of another circuit. Coupling is which the output signals of the two circuits do sum.

This paper chose the cascading to build a new map. The  $v_i$  signal of the memristor replaces the  $y_i$  in the seed map as input. The memristor is cascaded to the seed map to obtain a new discrete map as follows,

$$\begin{cases} x_{i+1} = x_i + a \sin(h(\sin(\pi q_i)) y_i), \\ y_{i+1} = h(\sin(\pi q_i)) y_i + b \sin(x_i) \sin(h(\sin(\pi q_i)) y_i), \\ q_{i+1} = q_i + y_i, \end{cases} \quad (3)$$

where  $a$ ,  $b$  and  $h$  are control parameters. The Lyapunov exponential and bifurcation plots of this map where  $x_0 = 1.42$ ,  $y_0 = 0.6$ ,  $q_0 = 3.01$  are shown in Fig. 4, 5, and 6. In Fig. 4, when  $b = 0.11$ ,  $h = 1.84$  and  $a \in (5.2, 7.5)$ , the map is hyperchaotic. In Fig. 5, when  $a = 5.7$ ,  $h = 1.84$  and  $b \in (-0.14, 0.14)$ , the map is hyperchaotic. In Fig. 6, when  $a = 5.7$ ,  $b = 0.11$  and  $h \in (1.75, 1.87)$ , the map is hyperchaotic. The attractor of this map is shown in Fig. 7. Chaotic performance that varies with initial values is illustrated in Fig. 8.

The conclusion that the new map exhibits hyperchaotic dynamics in a specific range of parameters and initial values is obtained

**Table 1: the DNA code.**

	1	2	3	4	5	6	7	8
00	A	A	T	T	C	C	G	G
01	C	G	C	G	A	T	T	A
10	G	C	G	C	T	A	A	T
11	T	T	A	A	G	G	C	C

**Table 2: DNA operation rule(XOR).**

	A	G	C	T
A(00)	A	G	C	T
G(01)	G	A	T	C
C(10)	C	T	A	G
T(11)	T	C	G	A

form these numerical simulation results. And the state of this system is affected by the initial value. These figures can be used as a reference for selected a initial values and parameters.

## 3 ENCRYPTION ALGORITHM

In this section, an image encryption algorithm based on DNA coding and this 3D discrete memristor chaotic map is designed. The flowchart of encryption algorithm is shown in Fig. 9. Here, DNA coding is the compilation of one byte into four DNA base. There are eight specific compilation methods as shown in Table 1. For example, a binary number 11011000 can be coded as DNA code TCGA in the first way. According to the binary evaluation ruler, the DNA encoded XOR is shown in Table 2. Algorithm 1 is a detailed description of the encryption algorithm.

**Step 1:** The sequences  $x$  and  $y$  of length  $N$  and  $z$  of length  $M * N$  are obtained.

**Step 2:** The plain image is encoded by DNA to C. Sequences  $z$  is encoded by DNA to obtained D.

**Step 3:** If  $y_i$  is greater than 0, then row  $i$  of C is shifted to the right by  $r$  units and column  $i + m * k$  is shifted up by  $c$  units.

$$r = \text{mod}(\text{floor}(\text{abs}(x_i * 100000000)), N);$$

$$c = \text{mod}(\text{floor}(\text{abs}(x_i * 100000000)), M);$$

$$i < 512, m = M/4, k = \{1, 2, 3, 4\}.$$

**Step 4:** Initializes the initial value of  $C_1$ .

**Step 5:** Convert C into a one-dimensional DNA sequence of length  $M * N$ . Spread from front to back.

**Step 6:** Initializes the initial value of  $C_{MN}$ .

**Step 7:** Spread from back to front.

**Step 8:**  $\hat{C}$  is decoded by DNA, return Cipher image C.

The decryption algorithm is the inverse process of the encryption algorithm. The key consists of  $[a, b, h, x_0, y_0, q_0]$ . In the simulation experiment, the  $Key = [5.5, 0.11, 1.84, 1.42, 0.6, 3.01]$  is selected, which the map is in hyperchaotic state. The plain, cipher and recover image of Lena, orangutan, balloon and peppers are shown in Fig. 10 from top to bottom.

## 4 SECURITY ANALYSIS

### 4.1 Histogram analysis

Plain image has obvious statistical characteristics, and histogram can be intuitive to see the distribution of pixel value. By encrypting the image, this feature should be eliminated. After encryption, the number of pixel value should be evenly distributed in each gray

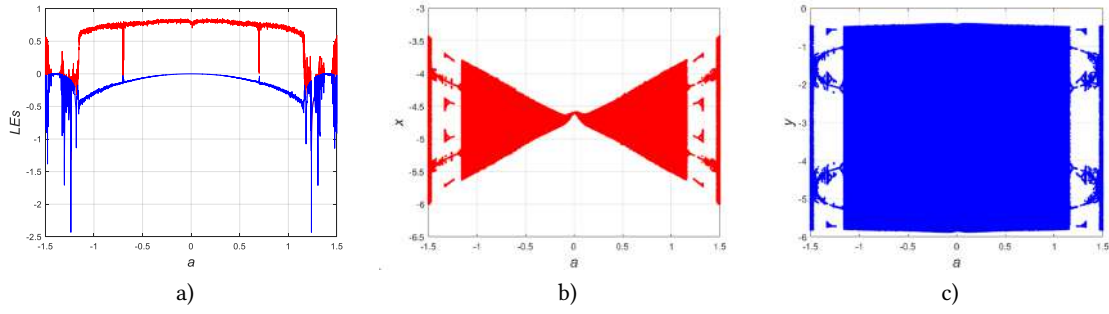


Figure 1: LEs vs. Bifurcation diagram where  $b = 4.4$  and  $a \in (-1.5, 1.5)$  for a) LEs, b) bifurcation of  $x$ , c) bifurcation of  $y$ .

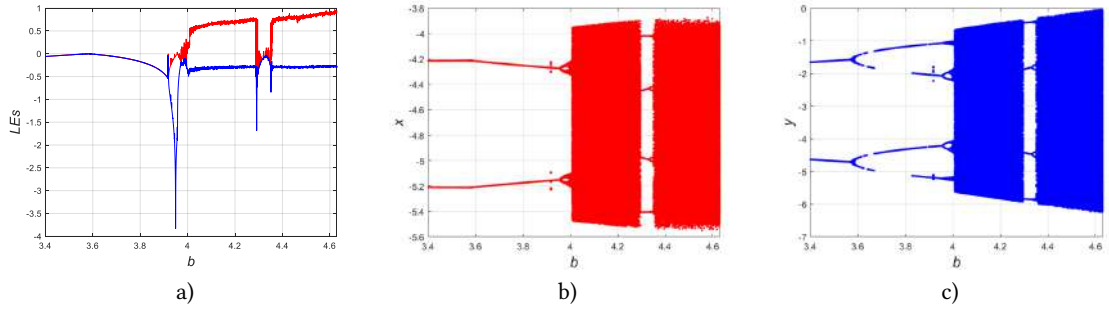


Figure 2: LEs vs. Bifurcation diagram where  $a = -1$  and  $b \in (3.4, 4.63)$  for a) LEs, b) bifurcation of  $x$ , c) bifurcation of  $y$ .

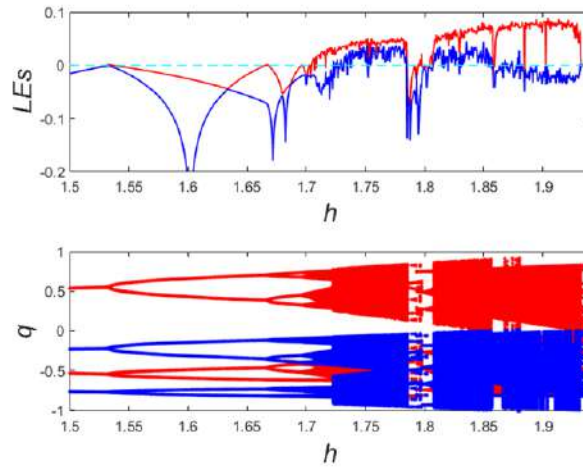


Figure 3: LEs vs. Bifurcation diagram where  $x_0 = 0.5$ ,  $q_0 = -0.6$  and  $h \in (1.44, 1.94)$ .

value. Original, encrypted and recovered histogram features of images Lena, orangutan, balloon and peppers are shown in Fig. 11.

## 4.2 Correlation analysis

Due to the strong correlation between plaintext pixels, attackers often use this feature as a breakthrough in attacks. Generally, the

gray values between image pixels are relatively close, and the correlation coefficient is close to 1. A secure encryption algorithm should eliminate the correlation between adjacent pixels. The security of cipher image can be guaranteed by having a weak correlation between cipher image pixels. So the correlation coefficient between cipher image pixel values is an important indicator to measure the

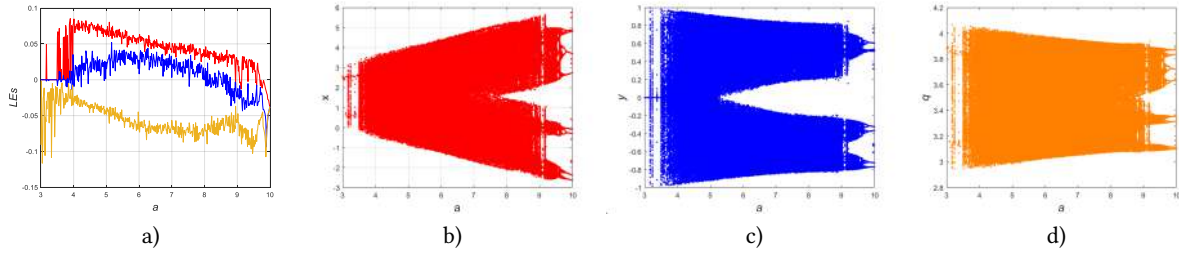


Figure 4: LEs vs. Bifurcation diagram where  $b = 0.11$ ,  $h = 1.84$  for a) LEs, b) bifurcation of  $x$ , c) bifurcation of  $y$ , d) bifurcation of  $q$ .

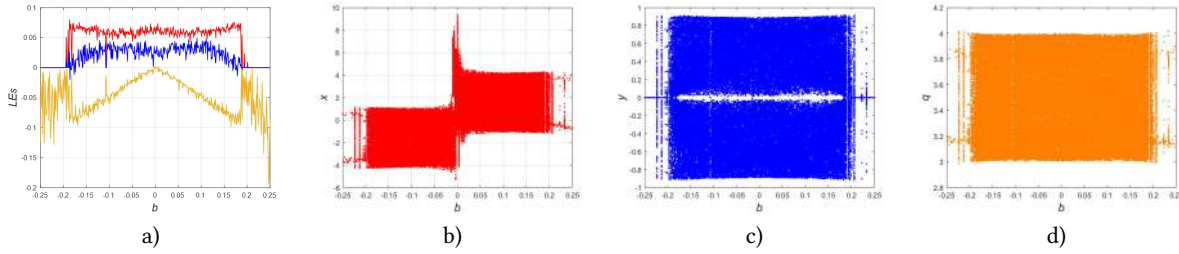


Figure 5: LEs vs. Bifurcation diagram where  $a = 5.7$ ,  $h = 1.84$  for a) LEs, b) bifurcation of  $x$ , c) bifurcation of  $y$ , d) bifurcation of  $q$ .

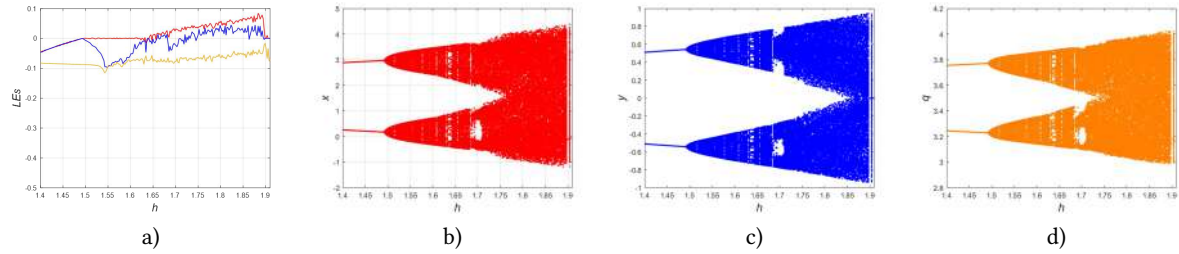


Figure 6: LEs vs. Bifurcation diagram where  $a = 5.7$ ,  $b = 0.11$  for a) LEs, b) bifurcation of  $x$ , c) bifurcation of  $y$ , d) bifurcation of  $q$ .

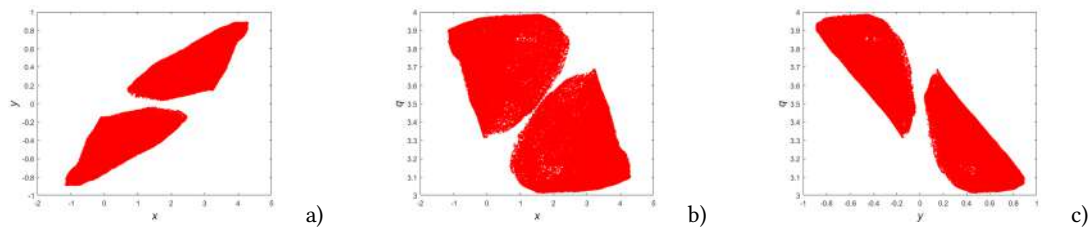


Figure 7: Attractor where  $x_0 = 1.42$ ,  $y_0 = 0.6$ ,  $q_0 = 3.01$ ,  $a = 5.5$ ,  $b = 0.11$ ,  $h = 1.84$  for a)  $x - y$ , b)  $x - q$ , c)  $y - q$ .

safety of the algorithm. The correlation coefficient is calculated as:

$$\begin{cases} C_{\vec{u}, \vec{v}} = \frac{\text{cov}(\vec{u}, \vec{v})}{\sqrt{D(\vec{u})} \sqrt{D(\vec{v})}} \\ \text{cov}(\vec{u}, \vec{v}) = \sum_{i=1}^N \frac{(u_i - E(\vec{u}))(v_i - E(\vec{v}))}{N} \\ D(\vec{u}) = \sum_{i=1}^N \frac{(u_i - E(\vec{u}))^2}{N} \\ E(\vec{u}) = \frac{1}{N} \sum_{i=1}^N u_i. \end{cases} \quad (4)$$

Here, 1000 pixels are randomly selected to calculate. Without losing universality, a set of keys is randomly taken in the key space

to encrypt the image Lena. As shown in Fig. 12, (a) (b) (c) are the correlation between adjacent pixels in the vertical, diagonal and horizontal directions of image Lena, respectively, and (d) (e) (f) are the correlation between adjacent pixels in the vertical, diagonal and horizontal of encrypted image Lena, respectively. Before encryption, adjacent pixels in each direction are distributed near diagonal line. And the correlation coefficient values of plain and cipher images is listed in Table 3.  $C\_Lena$  means cipher image Lena. Combining

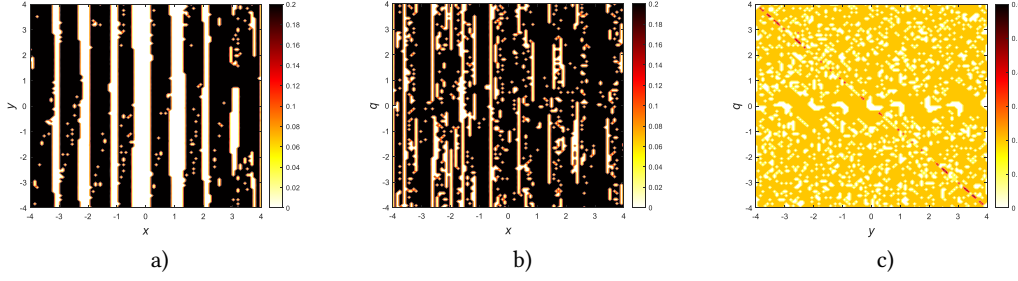


Figure 8: LEs plots with initial values as variable where  $a = 5.5$ ,  $b = 0.11$ ,  $h = 1.84$  for a)  $x - y$ , b)  $x - q$ , c)  $y - q$ .

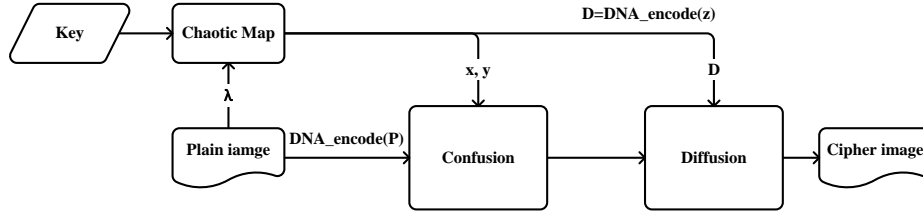


Figure 9: flowchart of encryption algorithm.

the figure and table, it can be concluded that the correlation of ciphertext image adjacent pixels in all directions is eliminated.

#### 4.3 Resistance to plain attack analysis

Currently, attackers primarily use selective plain attacks and known plain attacks to attack cryptographic algorithms. Cipher image sensitivity to plain is an inherent requirement of algorithms, which strengthen the sensitivity of plain pixels and the introduction of scrambled keys. The following is mainly by randomly selecting a pixel value plus one or minus one to obtain the test cipher image, the two cipher images are treated as random images, and then analyze the number of pixels change rate(NPCR) [5], unified average changing intensity(UACI)[18] in

$$NPCR(C_1, C_2) = \sum_{i=1}^M \sum_{j=1}^N \frac{|sign(C_1(i, j) - C_2(i, j))|}{M \times N}, \quad (5)$$

$$UACI(C_1, C_2) = \sum_{i=1}^M \sum_{j=1}^N \frac{|C_1(i, j) - C_2(i, j)|}{M \times N \times 255}, \quad (6)$$

between the original and revised cipher image. In addition, the block average change intensity(BACI) [33] is express as

$$BACI(C_1, C_2) = \frac{1}{(M-1)(N-1)} \sum_{k=1}^{(M-1)(N-1)} \frac{m_k}{255}, \quad (7)$$

where  $m_k = \frac{1}{6}(|d_{i,j} - d_{i,j+1}| + |d_{i,j} - d_{i+1,j}| + |d_{i,j} - d_{i+1,j+1}| + |d_{i,j+1} - d_{i+1,j}| + |d_{i,j+1} - d_{i+1,j+1}| + |d_{i+1,j} - d_{i+1,j+1}|)$  and  $d_{i,j} = C_1(i, j) - C_2(i, j)$ . The BACI of these two cipher images are also analyzed. As shown in Table 5,  $C_1$  is ciphertext image encrypted by original image, and  $C_2$  is obtained by encrypting the image after randomly selecting a pixel value in clear image to add or subtract one. The indicators of two cipher images are close to the theoretical values and higher than several recently proposed encryption schemes.

So the algorithm theoretically has good plain sensitivity and can effectively resist plain attacks.

#### 4.4 Key sensitivity analysis

A relatively secure encryption system should be sensitive to the key, and the sensitivity should be measured by the precision of the key value. Here a key is randomly selected  $key = \{a, b, h, x_0, y_0, q_0\}$ ,  $key_1 = \{a+1 \times 10^{-13}, b, h, x_0, y_0, q_0\}$ ,  $key_2 = \{a, b+1 \times 10^{-13}, h, x_0, y_0, q_0\}$ ,  $key_3 = \{a, b, h+1 \times 10^{-13}, x_0, y_0, q_0\}$ ,  $key_4 = \{a, b, h, x_0+1 \times 10^{-13}, y_0, q_0\}$ ,  $key_5 = \{a, b, h, x_0, y_0+1 \times 10^{-13}, q_0\}$ , and  $key_6 = \{a, b, h, x_0, y_0, q_0+1 \times 10^{-13}\}$ . These six slightly altered keys are used to encrypt the Lena image. The image encrypted with the original key is denoted as C, and the image encrypted with the changed key is denoted as C1. In Table 4, a statistical analysis between the original cipher image and the new cipher image while  $a, b, h, x_0, y_0$  and  $q_0$  minor changes occur. As given in Table 4, the values of evaluation criteria NPCR, UACI and BACI are all greater than or close to the ideal value. It can be concluded that the two cipher image are completely different after a slight change in the key. That is to say, after the key changes slightly, the ciphertext changes greatly. On the other hand, key sensitivity can be analyzed intuitively. Here, only image Lena and  $key_1$  are selected as the reference. As shown in Fig. 13, (a), (b), (c), (d) are respectively image Lena, cipher image Lena with  $key$ , cipher image Lena with  $key_1$  and difference image between (b) and (c). The difference graph of two ciphertext images presents the noise style. It can be seen that the key is changed slight, and the resulting cipher image is significantly different from the original cipher image.

**Algorithm 1:** Image Encryption

---

**Input:** Key, P  
**Output:** C

```

1  $C = DNA\_encode(P)$ ;
2  $[x, y, z] = DMSM(Key)$ ; Three chaotic sequences are obtained;
3  $[x, y, z] += \frac{1}{255MN} \sum_1^N \sum_1^M P(i, j)$ ; The key is associated with plain image information;
4  $D = DNA\_encode(z)$ , the chaotic sequence  $z$  is transformed into a byte array, and then  $z$  is encoded according to DNA coding rules to get DNA matrix  $D$ ;
5  $[M, N] = size(D)$ ;
6  $\boxplus$  stands for XOR between DNA codes;
7 for  $k = 1 : 4$  do
8   for  $i = 1 : 512$  do
9     if  $y_i \geq 0$  then
10       $r = \lfloor \lfloor x_i 10^8 \rfloor \rfloor \bmod N$ ;  $\lfloor \cdot \rfloor$  and  $\lfloor \cdot \rfloor$  stand for  $abs()$  and  $floor()$ ;
11       $c = \lfloor \lfloor x_i 10^8 \rfloor \rfloor \bmod M$ ;
12       $C[i, :] \rightarrow r$ ;
13       $C_k[:, i] \rightarrow c$ ;
14    end
15    else
16       $r = \lfloor \lfloor x_i 10^8 \rfloor \rfloor \bmod N$ ;
17       $c = \lfloor \lfloor x_i 10^8 \rfloor \rfloor \bmod M$ ;
18       $C[i, :] \leftarrow r$ ;
19       $C_k[:, i] \leftarrow c$ ;
20    end
21  end
22 end
23  $C'_1 = C'_1 \boxplus D_1 \boxplus D_{MN}$ ; Get the initial nucleobase of the cipher, and then spread from front to back;
24 for  $i = 2 : MN$  do
25    $C'_i = C'_i \boxplus D_i \boxplus C'_{i-1}$ ;
26 end
27  $C'_{MN} = C'_{MN} \boxplus D_1 \boxplus D_{MN}$ ; Update the last nucleobase of the cipher, then spread back and front;
28 for  $i = MN - 1 : 1$  do
29    $C'_i = C'_i \boxplus D_i \boxplus C'_{i+1}$ ;
30 end
31  $C = DNA\_decode(C')$ ;

```

---

**Table 3: Analysis of correlation coefficient.**

Image	Vertical	Diagonal	Horizontal
Lena	0.9471	0.9210	0.9704
Orangutan	0.8565	0.7825	0.8337
Balloon	0.8592	0.8384	0.9402
Peppers	0.9613	0.9410	0.9665
C_Lena	-0.0268	0.0340	-0.0191
C_Orangutan	0.0031	0.0866	-0.0108
C_Balloon	-0.0260	0.0476	0.0050
C_Peppers	0.0081	-0.0167	0.0084

## 4.5 Information entropy

To evaluate the randomness of ciphertext, researchers usually calculate the information entropy of images generated by encryption algorithm. Information entropy is actually a mathematical expectation of the amount of information contained in an image, the

higher value indicates the higher randomness of the image. The ideal value of information entropy is eight. Information entropy  $E$  can express by

$$E = - \sum_{i=1}^{255} p(i) \log_2 p(i). \quad (8)$$

The information entropy of encrypted images in this document is listed in Table 5 compared to other algorithms, and the encrypted object is Lena. Experimental results show that information entropy of cipher image obtained by this algorithm is closer to 8. And information entropy of Cipher image Lena is closer to ideal value than other algorithms, which has greater randomness.

## 5 ROBUST ANALYSIS

### 5.1 Noise analysis

Since the cipher images are polluted by various factors such as the environment and the channel during transmission, the pixel value may be changed to affect the decryption. Anti-noise analysis is the

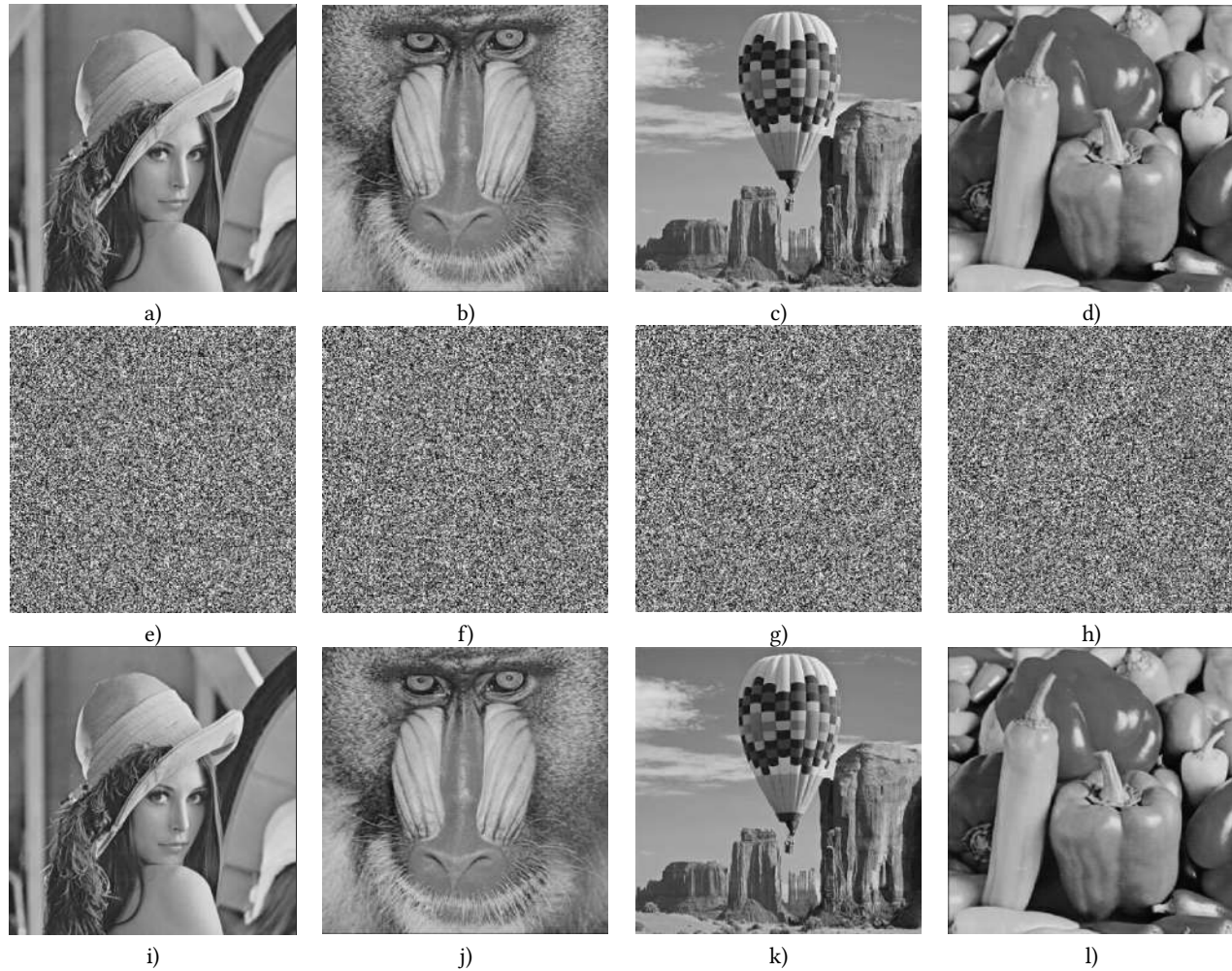


Figure 10: Original, ciphered and decrypted images of a) e) i) Lena, b) f) j) orangutan, c) g) k) balloon and d) h) l) peppers.

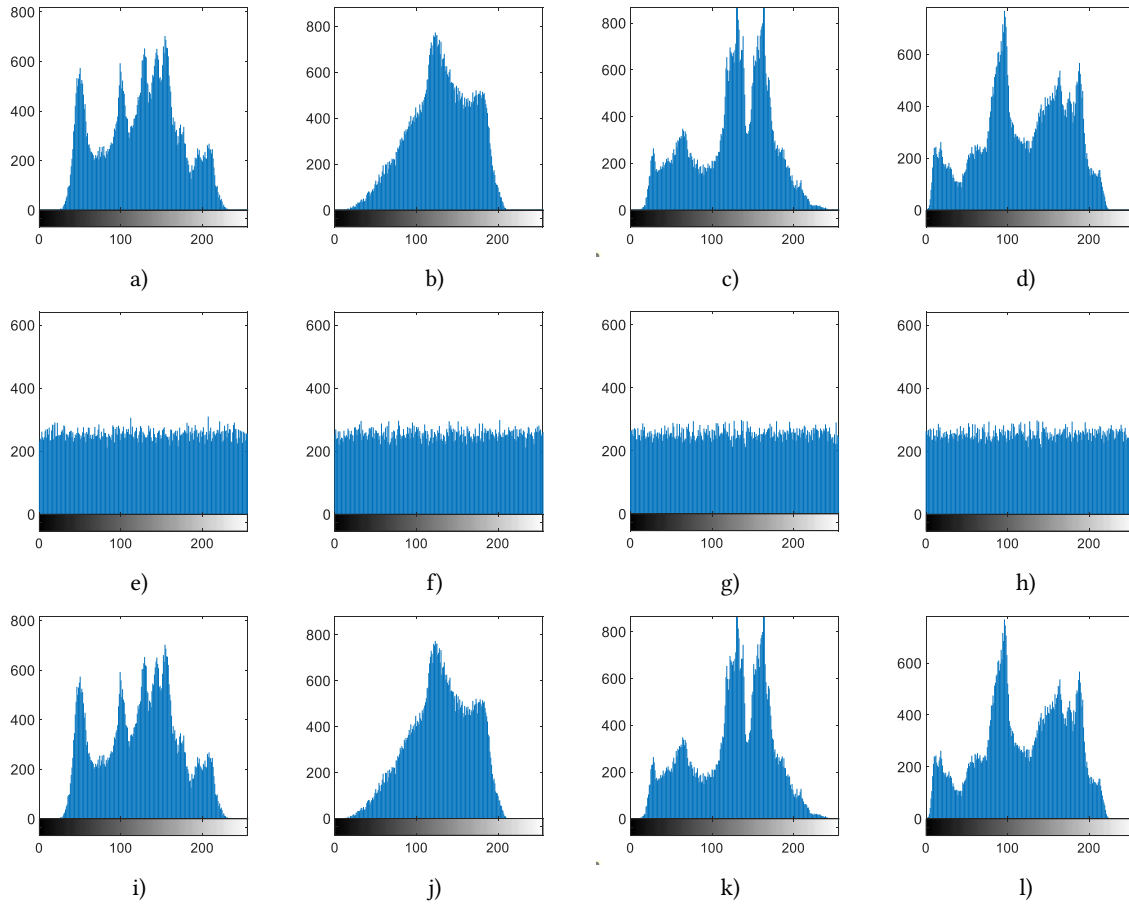
Table 4: Numerical analysis of key sensitivity with Lena.

keys	$E(C, C1)$	$(key, key_1)$	$(key, key_2)$	$(key, key_3)$	$(key, key_4)$	$(key, key_5)$	$(key, key_6)$
$NPCR_{C,C1}$	99.6094	99.5544	99.5865	99.6475	99.6048	99.5789	99.6124
$UACI_{C,C1}$	33.4635	33.5695	33.4965	33.4796	33.5128	33.5623	33.4908
$BACI_{C,C1}$	26.7712	26.9268	26.8104	26.7359	26.7760	26.8378	26.8338

evaluation of the ability to decrypt ciphertext images under the influence of noise. Here, salt and pepper noise is added to simulate the noise during transmission, and different intensities of the noise are added to compare the effect on decryption. As shown in Fig. 14, the noise intensity from front to back is 0.005, 0.05, 0.1, respectively. It can be seen that when the noise intensity is 0.05, the plain can be restored, and the impact on decryption is greater as the noise intensity increases. Finally, when the noise intensity is 0.1, the recovered image, although blurry, can still identify the outline of the original image. According to this, it can be judged that the algorithm can have good anti-noise performance.

## 5.2 Resistance to shear performance analysis

In practical, ciphertext images also carry the risk of data loss in transit. And ciphertext images can also be modified by attackers. Resistance to shear performance analysis refers to recovery degree of a cipher image while the cipher image is modified or partially cut. If the correlation between redaction pixels is high, a redaction that is missing some of the correct information will cause decryption to fail. In this algorithm, DNA level scrambling and diffusion are realized through DNA coding, which reduces the interaction between ciphertext pixels. Since the plain pixels are randomly disturbed, the distribution of the pixels in the cipher images shows a random and uniform distribution. And the clear text corresponding to the



**Figure 11: Histograms analysis of image Lena, orangutan, balloon and peppers for a) b) c) d) original images, e) f) g) h) encrypted images and i) j) k) l) decrypted images.**

**Table 5: Numerical analysis of plain image sensitivity and information entropy.**

Images	$NPCR(C_1, C_2)$	$UACI(C_1, C_2)$	$BACI(C_1, C_2)$	Information Entropy( $C_1$ )
balloon	99.5697	33.5367	26.8840	7.9968
orangutan	99.6033	33.4618	26.8465	7.9968
peppers	99.6124	33.2265	26.6584	7.9972
white	99.8123	34.9288	26.4747	7.9973
black	99.6857	33.8833	25.0458	7.9968
Lena	99.6460	33.4668	26.9149	7.9976
[34](Lena)	99.5894	33.3259	-	7.9976
[35](Lena)	99.6246	33.4115	-	7.9971
[7](Lena)	99.5925	33.3849	-	7.9968
[8](Lena)	99.6063	33.2985	-	7.9975

cut cipher image is also evenly distributed in the plain image, so the destroyed cipher image still carries some valid information. In order to mimic the shear attack, the  $1/16$ ,  $1/4$ , and  $1/2$  parts of the cipher image are replaced by 0 respectively to obtain three tampered cipher images, such as Fig. 15(a), (b), and (c). The three decrypted images are obtained as shown in Fig. 15(d), (e), and (f). It can be seen that when cutting  $1/16$ , the decrypted image is roughly clear. When

cutting  $1/4$ , the decrypted image is slightly blurry. When cutting  $1/2$ , the image is more blurry, but it can still clearly identify the plain image. According to the results of simulation experiments, it can be concluded that when the cipher image is attacked by shear, it still has the ability to recover part of the plain, and the proposed algorithm has a certain ability to resist the shear attack.

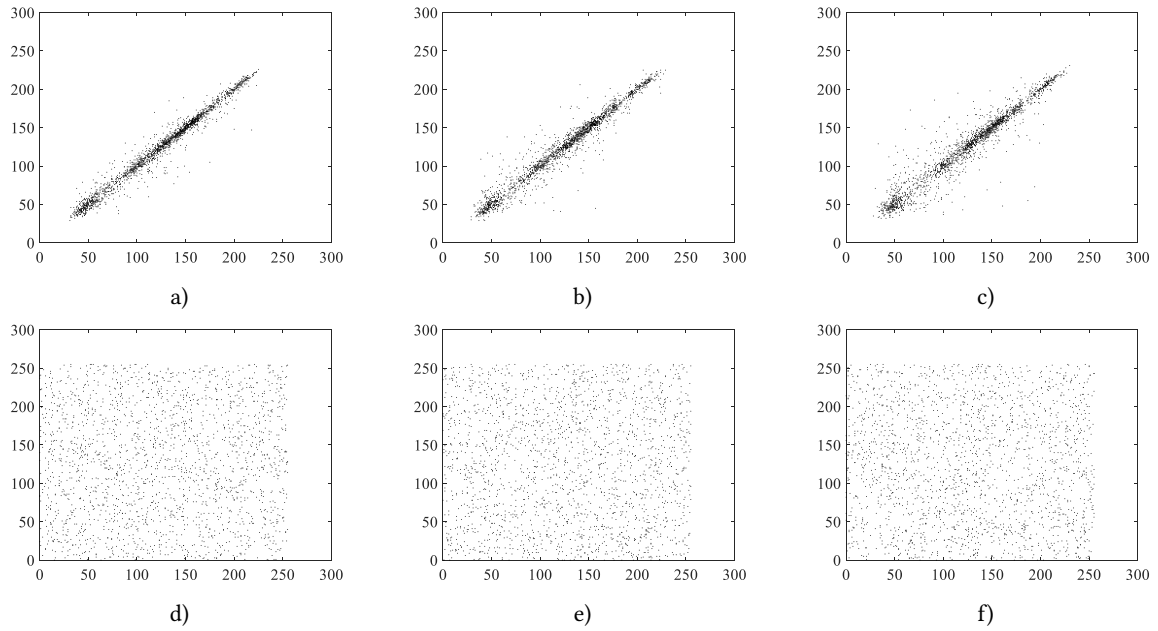


Figure 12: Correlation between adjacent pixels of Lena for a), b), c) vertical, diagonal, horizontal correlation of plain image and d), e), f) vertical, diagonal, horizontal correlation of encrypted images.

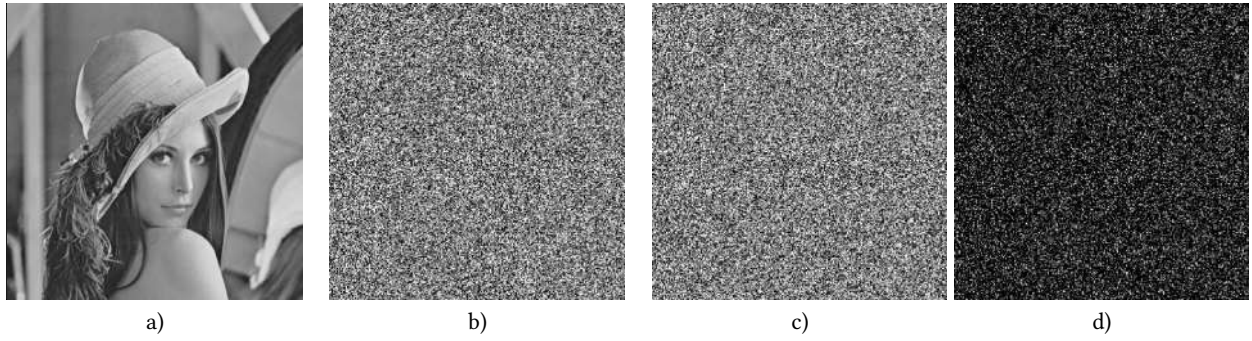


Figure 13: Key sensitive in the encryption where a) image Lena, b) C of encrypted image Lena, c) C1 of encrypted image Lena with key1, d)  $|C - C1|$ .



Figure 14: Noise attack with noise density of a) 0.005, b) 0.05 and c) 0.1.

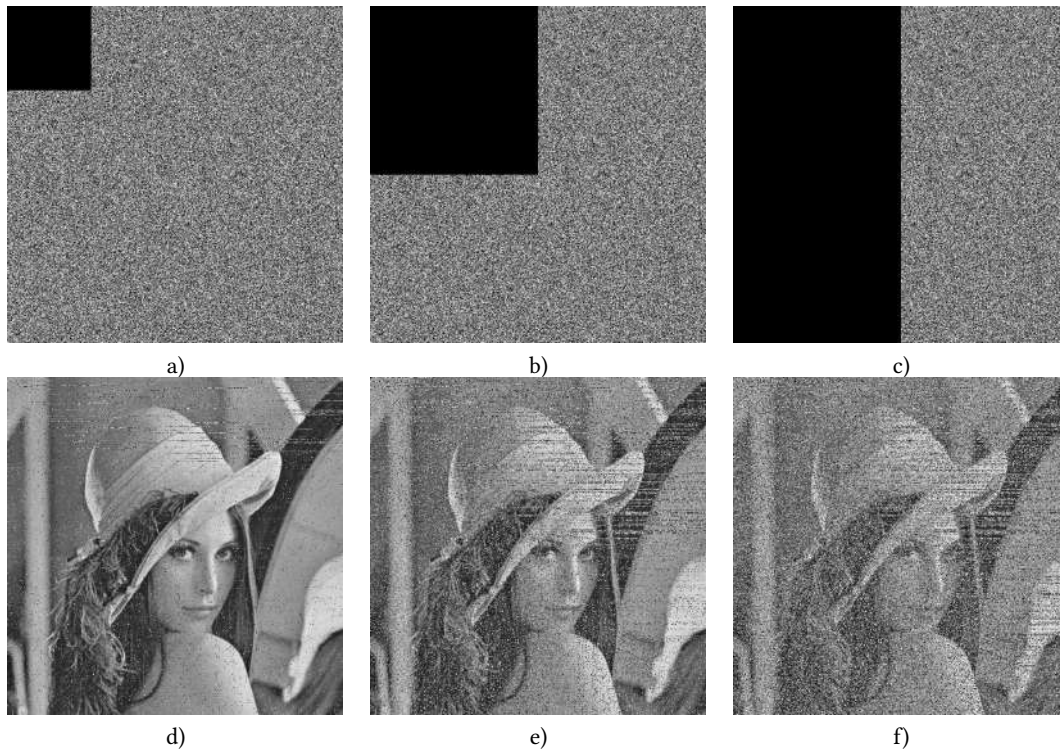


Figure 15: Cropping attack with losing data of a) 0.0625, b) 0.25 and c) 0.5, and d), e), f) are decrypted image respectively.

## 6 CONCLUSION

In this paper, a new chaotic map is designed by cascades of discrete memristor and discrete chaotic map. The new map has multiple positive Lyapunov exponents and coexisting attractors. Then, we propose an image encryption algorithm based on discrete memristor chaotic map and DNA coding. Compared with other encryption schemes, UACI and information entropy are closer to the ideal values. It proves that the sequence generated by the discrete memristor chaotic system has good security and reliability, and discrete memristors have good application prospects in the field of information security. In the future, we will explore a high-dimensional discrete memristor hyperchaotic map with a higher Lyapunov index, and promote the discrete memristor chaotic map to video encryption, audio encryption and other fields.

## REFERENCES

- [1] G. Alvarez, F. Montoya, M. Romera, and G. Pastor. 2003. Cryptanalysis of a chaotic secure communication system. *Physics Letters A* 306, 4 (2003), 200–205. [https://doi.org/10.1016/S0375-9601\(02\)01502-5](https://doi.org/10.1016/S0375-9601(02)01502-5)
- [2] Han Bao, Zhongyun Hua, Houzhen Li, Mo Chen, and Bocheng Bao. 2021. Discrete Memristor Hyperchaotic Maps. *IEEE Transactions on Circuits and Systems I-Regular Papers* 68 (2021), 4534–4544. <https://doi.org/10.1109/TCSI.2021.3082895>
- [3] Han Bao, Zhongyun Hua, Houzhen Li, Mo Chen, and Bocheng Bao. 2022. Memristor-Based Hyperchaotic Maps and Application in Auxiliary Classifier Generative Adversarial Nets. *IEEE Transaction on Industrial Informatics* 18 (2022), 5297–5306. <https://doi.org/10.1109/TII.2021.3119387>
- [4] M. A. Ben Farah, A. Farah, and T. Farah. 2020. An image encryption scheme based on a new hybrid chaotic map and optimized substitution box. *Nonlinear Dynamics* 99, 4 (MAR 2020), 3041–3064. <https://doi.org/10.1007/s11071-019-05413-8>
- [5] Guanrong Chen, Yaobin Mao, and Charles K. Chui. 2004. A symmetric image encryption scheme based on 3D chaotic cat maps. *Chaos, Solitons & Fractals* 21, 3 (2004), 749–761. <https://doi.org/10.1016/j.chaos.2003.12.022>
- [6] L. Chua. 1971. Memristor-The missing circuit element. *IEEE Transactions on Circuit Theory* 18 (1971), 507–519. <https://doi.org/10.1109/TCT.1971.1083337>
- [7] Mehmet Demirtas. 2022. A novel multiple grayscale image encryption method based on 3D bit-scrambling and diffusion. *Optik* 266 (2022). <https://doi.org/10.1016/j.ijleo.2022.169624>
- [8] Nadeem Iqbal, Muhammad Hanif, Zia Ul Rehman, and Muhammad Zohaib. 2022. On the novel image encryption based on chaotic system and DNA computing. *Multimedia Tools and Applications* 81 (2022), 8107–8137. <https://doi.org/10.1007/s11042-022-11912-5>
- [9] Dahye Kim, Sunghun Kim, and Sungjun Kim. 2021. Logic-in-memory application of CMOS compatible silicon nitride memristor. *Chaos Solitons & Fractals* 153, 2 (2021). <https://doi.org/10.1016/j.chaos.2021.111540>
- [10] Chengqing Li, Dongdong Lin, Bingbing Feng, Jinhu Lu, and Feng Hao. 2018. Cryptanalysis of a Chaotic Image Encryption Algorithm Based on information Entropy. *IEEE Access* 6 (2018), 75834–75842. <https://doi.org/10.1109/ACCESS.2018.2883690>
- [11] Xinyu Li, Jian Zeng, Qun Ding, and Chunlei Fan. 2022. A Novel Color Image Encryption Algorithm Based on 5-D Hyperchaotic System and DNA Sequence. *Entropy* 24, 9 (2022). <https://doi.org/10.3390/e24091270>
- [12] Yinghua Li, He Yu, Bin Song, and Jinjun Chen. 2021. Image encryption based on a single-round dictionary and chaotic sequences in cloud computing. *CONCURRENCY AND COMPUTATION-PRACTICE & EXPERIENCE* 33, 7, SI (APR 10 2021). <https://doi.org/10.1002/cpe.5182>
- [13] Sheng Liu, Chengqing Li, and Qiao Hu. 2022. Cryptanalyzing Two Image Encryption Algorithms Based on a First-Order Time-Delay System. *IEEE Multimedia* 29 (2022), 74–84. <https://doi.org/10.1109/MMUL.2021.3114589>
- [14] Sijia Liu, Yanzhi Wang, Makan Fardad, and Pramod K. Varshney. 2018. A Memristor-Based Optimization Framework for Artificial Intelligence Applications. *IEEE Circuits and Systems Magazine* 18, 1 (2018), 29–44. <https://doi.org/10.1109/MCAS.2017.2785421>
- [15] Wenhao Liu, Kehui Sun, and Congxu Zhu. 2016. A fast image encryption algorithm based on chaotic map. *Optics and Lasers in Engineering* 84 (SEP 2016), 26–36. <https://doi.org/10.1016/j.optlaseng.2016.03.019>
- [16] Olfa Mannai, Rabei Bechikh, Houcemmedine Hermassi, Rhouma Rhouma, and Safya Belghith. 2015. A new image encryption scheme based on a simple first-order time-delay system with appropriate nonlinearity. *Nonlinear Dynamics* 82 (2015), 107–117. <https://doi.org/10.1007/s11071-015-2142-x>

- [17] Ali Mansouri and Xingyuan Wang. 2020. A novel one-dimensional sine powered chaotic map and its application in a new image encryption scheme. *Information Sciences* 520 (MAY 2020), 46–62. <https://doi.org/10.1016/j.ins.2020.02.008>
- [18] Yaobin Mao, Guanrong Chen, and Shiguo Lian. 2004. A Novel Fast Image Encryption Scheme Based on 3D Chaotic Baker Maps. *Int. J. Bifurc. Chaos* 14 (2004), 3613–3624. <https://doi.org/10.1142/S021812740401151X>
- [19] Chanil Pak and Lilian Huang. 2017. A new color image encryption using combination of the 1D chaotic map. *Signal Processing* 138 (SEP 2017), 129–137. <https://doi.org/10.1016/j.sigpro.2017.03.011>
- [20] Yuexi Peng, Kehui Sun, and Shaobo He. 2020. A discrete memristor model and its application in Henon map. *Chaos Solitons & Fractals* 137 (2020). <https://doi.org/10.1016/j.chaos.2020.109873>
- [21] Yuexi Peng, Kehui Sun, and Shaobo He. 2020. An Improved Return Maps Method for Parameter Estimation of Chaotic Systems. *International Journal of Bifurcation and Chaos* 30, 4 (2020). <https://doi.org/10.1142/S0218127420500583>
- [22] Mario Preishuber, Thomas Huefter, Stefan Katzenbeisser, and Andreas Uhl. 2018. Depreciating Motivation and Empirical Security Analysis of Chaos-Based Image and Video Encryption. *IEEE Transactions on Information Forensics and Security* 13 (2018), 2137–2150. <https://doi.org/10.1109/TIFS.2018.2812080>
- [23] Kun Qian, Wei Feng, Zhentao Qin, Jing Zhang, Xuegang Luo, and Zhengguo Zhu. 2022. A novel image encryption scheme based on memristive chaotic system and combining bidirectional bit-level cyclic shift and dynamic DNA-level diffusion. *Frontiers in Physics* 10 (2022). <https://doi.org/10.3389/fphy.2022.963795>
- [24] Arslan Shafique and Junaid Shahid. 2018. Novel image encryption cryptosystem based on binary bit planes extraction and multiple chaotic maps. *European Physical Journal plus* 133, 8 (2018). <https://doi.org/10.1140/epjp/i2018-12138-3>
- [25] Chunyan Song and Yulong Qiao. 2015. A Novel Image Encryption Algorithm Based on DNA Encoding and Spatiotemporal Chaos. *Entropy* 17 (2015), 6954–6968. <https://doi.org/10.3390/e17106954>
- [26] Dmitri Strukov, Gregory Snider, Duncan Stewart, and Stanley Williams. 2008. Memristor-the missing circuit element. *Nature* 453 (2008), 80–83. <https://doi.org/10.1038/nature06932>
- [27] Simiao Wang, Qiqi Peng, and Baoxiang Du. 2022. Chaotic color image encryption based on 4D chaotic maps and DNA sequence. *Optics and Laser Technology* 148 (APR 2022). <https://doi.org/10.1016/j.optlastec.2021.107753>
- [28] Heping Wen, Chongfu Zhang, Lan Huang, Juxin Ke, and Dongqing Xiong. 2021. Security Analysis of a Color Image Encryption Algorithm Using a Fractional-Order Chaos. *Entropy* 23, 2 (2021). <https://doi.org/10.3390/e23020258>
- [29] Ailong Wu, Shiping Wen, and Zhigang Zeng. 2012. Synchronization control of a class of memristor-based recurrent neural networks. *Information Sciences* 183, 1 (2012), 106–116. <https://doi.org/10.1016/j.ins.2011.07.044>
- [30] Lu Xu, Zhi Li, Jian Li, and Wei Hua. 2016. A novel bit-level image encryption algorithm based on chaotic maps. *Optics and Lasers in Engineering* 78 (MAR 2016), 17–25. <https://doi.org/10.1016/j.optlaseng.2015.09.007>
- [31] Guodong Ye, Chen Pan, Xiaoling Huang, Zhenyu Zhao, and Jianqing He. 2018. A Chaotic Image Encryption Algorithm Based on Information Entropy. *International Journal of Bifurcation and Chaos* 28, 1 (2018). <https://doi.org/10.1142/S0218127418500104>
- [32] Fang Yuan, Cheng-Jun Bai, and Yu-Xia Li. 2021. Cascade discrete memristive maps for enhancing chaos\*. *Chinese Physics B* 30, 12 (2021). <https://doi.org/10.1088/1674-1056/ac20c7>
- [33] Yong Zhang. 2018. The unified image encryption algorithm based on chaos and cubic S-Box. *Information Sciences* 450 (JUN 2018), 361–377. <https://doi.org/10.1016/j.ins.2018.03.055>
- [34] Jiming Zheng and Tianyi Lv. 2022. Image encryption algorithm based on cascaded chaotic map and improved Zigzag transform. *IET Image Processing* (2022). <https://doi.org/10.1049/ipr2.12600>
- [35] Chengye Zou, Xingyuan Wang, Changjun Zhou, Shujuan Xu, and Chun Huang. 2022. A novel image encryption algorithm based on DNA strand exchange and diffusion. *Appl. Math. Comput.* 430 (2022). <https://doi.org/10.1016/j.amc.2022.127291>

# CBAM-based Method in YOLOv7 for Detecting Defective Vacuum Glass Tubes

Zeyu Sheng  
High School Affiliated to Fudan  
University  
13162851183@qq.com

Haiguang Chen  
Shanghai Normal University  
19276368@qq.com

Zifeng Qi  
Shanghai Normal University  
zifengqi7@163.com

## ABSTRACT

The vacuum glass tube is one of the most important materials in the physical industry, and the inspection rate of its production is crucial to the production of subsequent products. We propose a CBAM-based target detection method for YOLOv7 to detect defects in transparent glass tubes, which are not easily detectable due to their transparent walls. We replace all pooling layers in YOLOv7 with CBAM to enable it to better grasp target features. The experimental results show that the recall rate for defective product detection reaches 98.34% and the accuracy rate reaches 96.33% in the simulated industrial inspection environment. It can meet the accuracy requirements of detecting defects of transparent glass tubes in industrial sites.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Object detection;

## KEYWORDS

YOLOv7, CBAM, defect detection, deep learn

### ACM Reference Format:

Zeyu Sheng, Haiguang Chen, and Zifeng Qi. 2023. CBAM-based Method in YOLOv7 for Detecting Defective Vacuum Glass Tubes. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590079>

## 1 INTRODUCTION

As one of the core materials of basic industrial materials, vacuum glass tube has the characteristics of high temperature resistance, corrosion resistance, good thermal stability, good electrical insulation properties, good light transmission properties, etc., so it is also used in various fields such as electric light sources, semiconductors. For the manufacturing and processing of some tiny semiconductor chips, the glass is used in large quantities because of its good physical properties. In order to improve the quality and production efficiency of integrated circuits and devices, the high purity and

high temperature resistance of glass tubes are particularly important. However, during the manufacturing process, bubbles, stains on the tube wall, and unevenness of the tube wall often occur due to contamination in the fabrication process and aging of the fabrication tools. Due to its importance, the inspection technique for defective products with these defects is particularly important. Usually, it is difficult for a person to identify defects in glass tubes because of the large size of vacuum glass tubes. In the era of Industry 4.0, the use of computer vision technology for defective products detection is a good choice. Among them, YOLO technology is particularly prominent. For the current development status of YOLO and the difficulty level of vacuum glass tube detection, we propose the target detection algorithm of YOLOv7 with added attention mechanism.

## 2 RELATED WORK

Glass tube quality inspection is the most important part of glass quality inspection, most defects such as bubbles and stains can significantly degrade the quality of glass tubes affecting their practical use. [1] Today, in manufacturing industry, manual inspection is not up to the task of detecting defects in clear glass tubes with high demand and high generation. However, with the advancement of computer vision inspection technology, it is widely used in artificial intelligence, medical monitoring and industrial automation. Contact inspection methods do not achieve the need for high-speed inspection, so the application of computer vision inspection technology together with image processing to achieve monitoring means on production line products is the most effective way to ensure the quality of transparent glass tubes.

As early as 1984 researchers began to experiment with glass tube defect detection, Mitra.S.K and Parker.J.M designed a glass tube defect detection system that could identify and determine at least 40 glass tube defects. [2] In 1985, Magers.C, Latham.V and Nixon.M designed an automated optical detection equipment to perform the detection task by characterizing each defect of glass tubes based on their optical characteristics. [3] In 1991, Kysztrof.J.Cios et al. started to improve the detection algorithm and they used neural network technology and applied it to the glass tube defect detection industry to achieve better defect recognition and classification results. [4] In 2005, Dae Cheol Lim et al. proposed a review image classification method for image quality issues that can affect image defects, i.e., focusing on quantitative defects for feature analysis and then removing other redundancies to improve defect recognition and classification efficiency. [5] Due to the inspiration of the previous work, researchers started to study the analysis of defect features in depth. In 2013, Sachdeva et al. designed a defect detection system that used a grayscale co-occurrence matrix for defect feature

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590079>

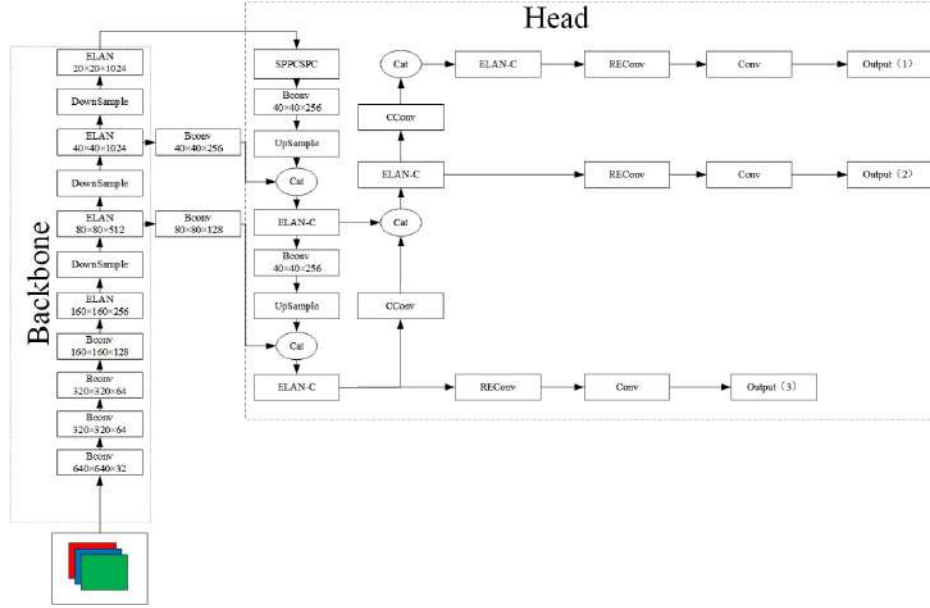


Figure 1: YOLOv7 main frame diagram

extraction to achieve defect classification recognition. [6] In 2014, Soukup, Masci et al. used CNN networks to identify surface defects on steel strips, and In 2015, Tolba et al. proposed a multi-scale structural similarity index approach for detecting glass tubes, which can achieve fast detection and localization of surface defects. [7–9] In 2017, Afshar et al. introduced a local variance rotation invariant measurement operator for detecting the edges of glass tube defects, and the classifier was designed using a support vector machine to achieve defect classification recognition. [10]

### 3 YOLOV7 MODEL BASED ON CBAM MECHANISM

YOLOv7 is mainly derived by further optimization of the model architecture of YOLOv5. [11, 12] The optimization direction is mainly directed to the optimization problem of re-referencing the model structure and the optimization problem of dynamic label assignment. Based on the high performance of YOLOv7 in target detection, we further add a CBAM attention mechanism to it and use it to replace the internal pooling layer to obtain higher industrial performance for its detection in vacuum glass tubes. [13] The main framework of the model is shown in Figure 1.

#### 3.1 YOLOv7 model

First, YOLOv7 will resize the input image to a 640×640 size image, and then input it into the backbone network, which will enter into the head network through the SPPCSPC module, and its output will be three different size feature maps, and then output the prediction result through rep and conv.

The input section will first go through four BConv modules while the feature map size will change to 160×160×128. The BConv modules are shown in Figure 2. Then it will pass through four ELAN modules. Each ELAN module consists of seven BConv modules,

and its structure is shown in Figure 3. In the BConv module, where the convolution layer action formula can be expressed as (1):

$$x_j^i = f \left( \sum_{i \in M_j} x_i^{i-1} * w_{ij}^k + b_j^i \right) \quad (1)$$

where  $x_j^i$  is the feature mapping of the current layer,  $x_i^{i-1}$  is the feature mapping of the previous layer,  $w_{ij}^k$  is the weight of the location coordinate  $(i, j)$  of the  $k$ th layer,  $b_j^i$  is the bias, and  $f$  is the activation function SiLU, whose formula is shown in (2).

$$f(x) = x \cdot \text{Sigmoid}(x) \quad (2)$$

The results of the Backbone network are fed into the Head network, which eventually forms the prediction results.

#### 3.2 CBAM attention mechanism

CBAM is a simple and efficient attention mechanism, which is a combination of spatial attention mechanism and channel attention mechanism. We take the two parts in parallel to combine the attention mechanism, so that for any feature map in the convolutional neural network, the two attention mechanisms can be computed in independent dimensions, and then combined for feature mapping, and finally for feature refinement. the architecture diagram of CBAM is shown in Figure 4.

For the feature map input into CBAM, it can be expressed as (3):

$$F \in R^{C \times H \times W} \quad (3)$$

After the feature map is fed into the channel attention mechanism, it passes through parallel maximum pooling and average pooling layers, which change the dimension of the feature map to  $C \times 1 \times 1$ . Then it is fed into the fully connected layer for computation and the results are summed up by the sigmoid function to

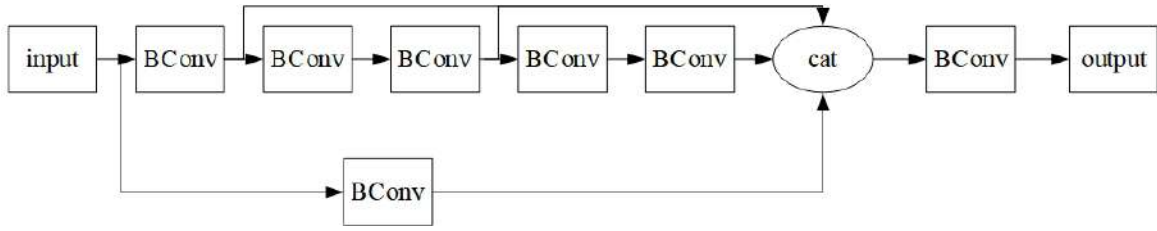


Figure 2: ELAN model diagram

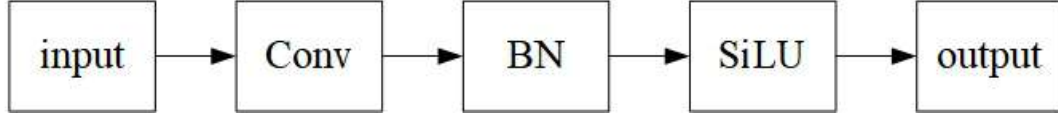


Figure 3: BConv model diagram

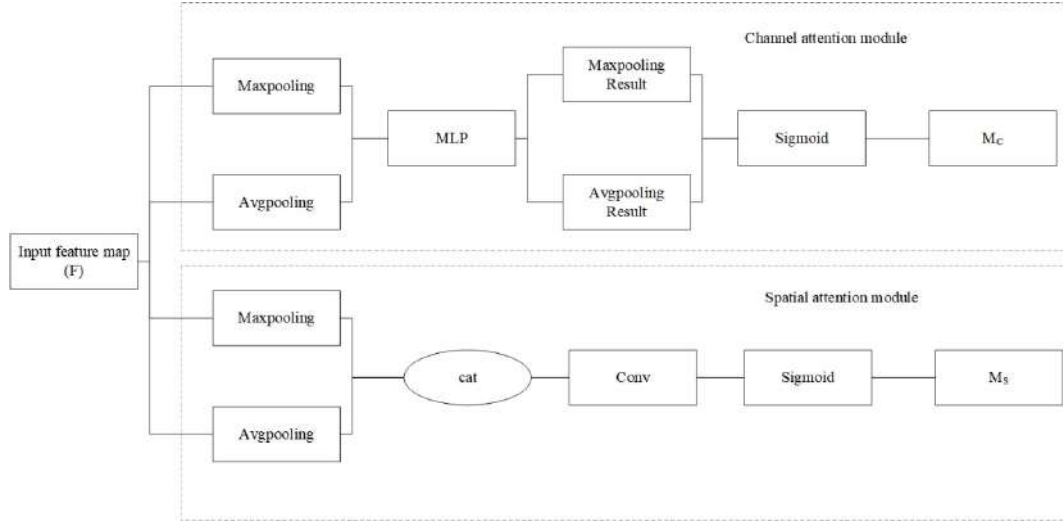


Figure 4: Architecture diagram of CBAM

obtain the one-dimensional channel attention  $M_C$ . Then  $M_C$  is multiplied with  $F$  to obtain the adjusted feature map  $F'$ .  $F'$  is fed into the spatial attention mechanism. In which the maximum pooling and mean pooling are first performed separately, the feature maps of  $1 \times H \times W$  size generated by the two results are stitched together and then convolved, and  $M_S$  is obtained by the sigmoid function, and finally  $F''$  is obtained by multiplying  $M_S$  with  $F$ .

In the channel attention mechanism, the spatial dimension will be compressed while keeping the channel dimension unchanged, which is calculated as (4):

$$M_C(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ = \sigma(w_1(w_0(F_{avg}^C)) + w_1(w_0(F_{max}^C))) \quad (4)$$

where MLP is the fully connected layer,  $w_0$  is the weight value of the first layer of MLP,  $w_1$  is the weight value of the second layer of MLP, and  $\sigma$  is the sigmoid function, which is given by (5):

$$\sigma(x) = \frac{1}{1 + e^x} \quad (5)$$

In the spatial attention mechanism, the channel dimension will be compressed while keeping the spatial dimension unchanged, which is calculated as (6):

$$M_S(F) = \sigma(f([AvgPool(F); MaxPool(F)])) \\ = \sigma(f([F_{avg}^A; F_{max}^A])) \quad (6)$$

where  $f$  is the operation of stitching and convolving. Finally, the results of its two attention modules are summed so that they are concatenated, as in equation (7), to obtain all the required feature information.

$$M = M_C(F) + M_S(F) \quad (7)$$

We replaced all pooling layers in YOLOv7 network with CBAM module to prevent some loss of valid information when passing through the convolutional layer. For example, the max-pooling layers in SSPCSPC were replaced. The revised module's structure is shown in Figure 5.

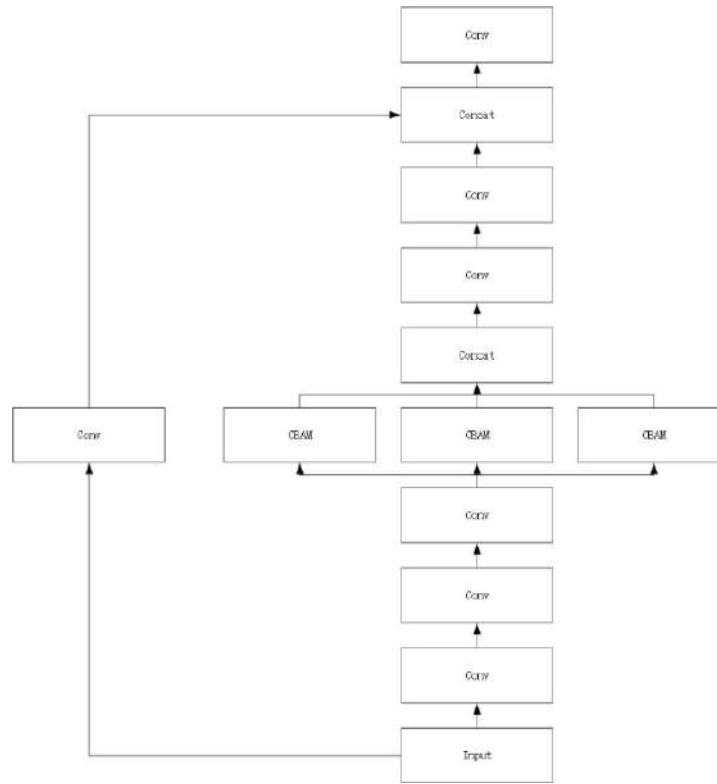


Figure 5: SSPCSPC module's max-layers were replaced with CBAM

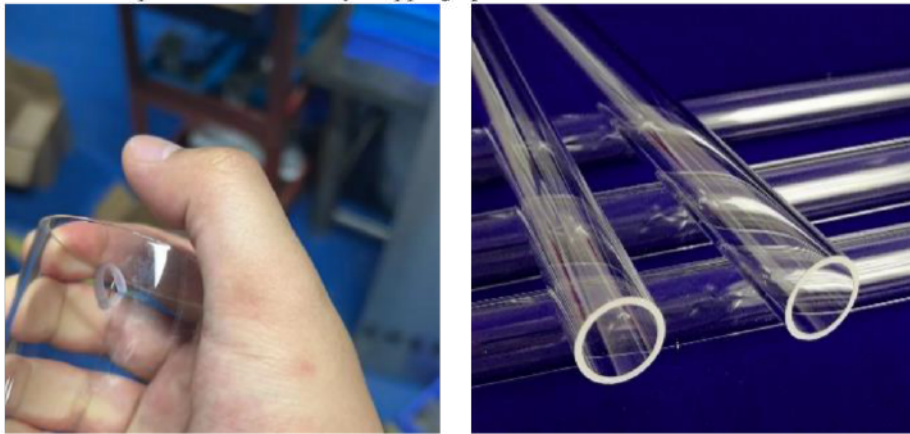


Figure 6: Architecture diagram of CBAM

## 4 EXPERIMENTS AND ANALYSIS OF RESULTS

### 4.1 Experimental data set and parameter settings

The dataset for this experiment is a dataset consisting of both qualified clear glass tubes and defective clear glass tubes produced in actual industrial environments together. This dataset includes four categories: qualified products, defective products with bubbles

on the tube wall, defective products with stains on the tube wall, and defective products with uneven tube wall. In this dataset, there are 1150 original samples, and after the image enhancement process, it becomes 11,500 total samples, and then each category is randomly divided into a total of 10,350 training samples and 1150 validation samples respectively. An example of the source data of the dataset is shown in Figure 6. Thereafter, the input data will be kept at the same size by cropping operation.

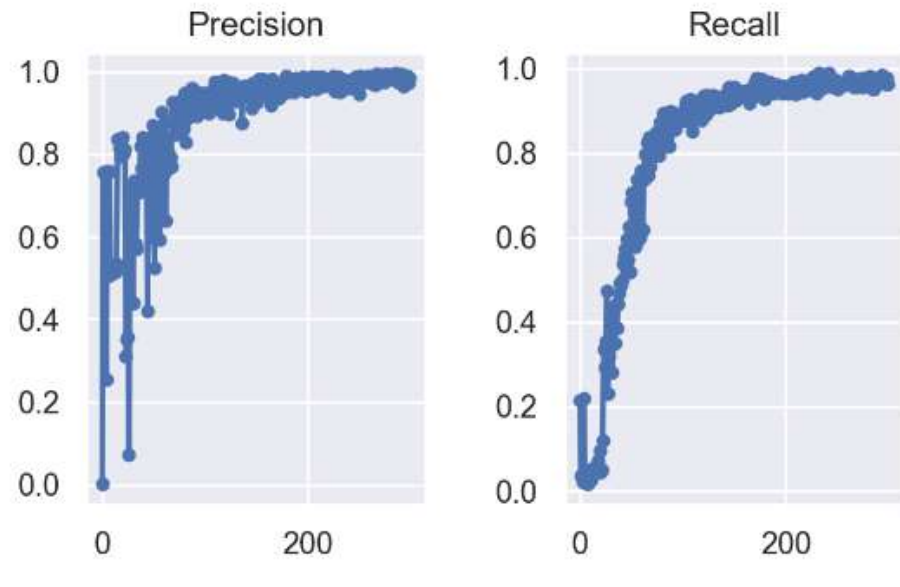


Figure 7: The change curves of Precision and recall during the training process



Figure 8: The results in the industrial inspection

The experiments were conducted in a windows environment with python version 3.6.0 and the deep learning framework was the pytorch framework with version 1.9.1+cu111. The hardware environment is AMD R7 5800H CPU and NVIDIA RTX3070 GPU.

#### 4.2 Training results and analysis

Epoch is set to 300, batch size is set to 16, and the initial learning rate is 0.01. After training, the Precision and recall curves are shown in Figure 7 with a recall rate of 98.34% and an accuracy rate of 96.33% after training is completed.

The actual results in industrial inspection are shown in Figure 8. The model obtained after training can already achieve industrial inspection level results and can detect defective products and effectively label their locations in field inspection.

## 5 CONCLUSION

We propose to combine CBAM and YOLOv7, using CBAM instead of the original pooling layer in YOLOv7, to further enhance its sensitivity to target detection, enabling it to show better results in industry, which can greatly reduce the occurrence of defective products missing detection. Channel attention and spatial attention can better enable the network to accurately capture the defective features of vacuum glass tubes. From the experimental results, the defective products recall rate of detecting vacuum glass tubes reached 98.34% and the accuracy rate reached 96.33%. Although the detection effect has reached the industrial level, improvements can still be made in the detection speed to achieve the instant warning function.

## REFERENCES

- [1] 吴房胜,徐金秀,李如平.基于数字图像处理的玻璃瓶瑕疵秀测系统[J].宜宾学院学报,2014,14(06):103-107
- [2] Mitra. S.K, Perker. J. M. Expert System for Identifying Defects:Stones and Cord in Glass[J]. Proceeding of the Sixth European Conference on Artificial Intelligence,1984,8(5):362-363
- [3] Latham,V,Nixon.M,Mayars.C.Automatic Optical Inspection of Table Glass[J].Glass Technology,1986,27(6):188-194
- [4] Kysztol. J.C, Robert.E.T, Ning L. Study of Continuous ID3 and Radial Basis Function Algorithms for the Recognition of Glass Defects[C].Neural Networks IJCNN-91-Seattle International Joint Conference,Seattle:IEEE,1991:49-54
- [5] Dae Cheol. L, Dae Gyu. S, Dae Hwa J. Defect Classification for the Inspection of TFT:CD Glass[J].Proceeding of the SPIE,2005,60(51):605-610
- [6] Sachdeva K, Girdhar. A.A Technique for Glass Defect Detection[J].International Journal of Innovative Research and Development,2013,2(13):25-31
- [7] Soukup. D.R, Huber. M. Convolutional Neural Networks for Steel Surface Defect Detection from Photometric Stereo Images[J].AIT Austrian Institute of Technology GmbH,2014,88(87):668-677
- [8] Masci. J, Meier. U, Ciresan. D *et al.* Steel Defect Classification with Max-Pooling onvolutional Neural Networks[C].The 2012 International Joint Conference on Neural Networks(IJCNN),Australia:IEEE,2012:10-15
- [9] Tolba A.S, Raafat. H.M. Multiscale Image Quality Measures for Defect Detection in Thin Films[J].The International Journal of Advanced Manufacturing Technology,2015,79(1-4):113-122
- [10] Afshar. A, Hanzaei. S.H, Barazandeh. F. Automatic Detection and Classification of the Ceramic Tiles Surface Defects[J].Pattern Recognition,2017,66(12):174-189
- [11] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv: 2207.02696, 2022
- [12] JOCHER G, STOKEN A, BOROVEC J, *et al.* Ultralytics/YOLOv5: V3.1 - bug fixes and performance improvements [EB/OL].<https://doi.org/10.5281/zenodo.4154370>, 2020. doi: 10.5281/zenodo.4154370,2020
- [13] Woo S, Park J, Lee J Y, *et al.* Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19

# Image Generation Model Applying PCA on Latent Space

Myung Keun Song

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
mksong@vim.cau.ac.kr

Asim Niaz

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
asim@vim.cau.ac.kr

Kwang Nam Choi\*

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
knchoi@cau.ac.kr

## ABSTRACT

Image generation is an important area of artificial intelligence that involves creating new images from existing datasets. It involves learning the distribution of target images from randomly generated vectors. Like other deep learning models, the image generation model requires a vast refined data set to produce high-quality results. When there is little data, there is a problem that the diversity and quality of generated images are compromised. In this paper, we propose a new generative model that applies PCA to the generator of the least square error adversarial generative network that, in turn, generates high-quality images even with a small data set. Unlike the existing models that generate target data from randomly generated noise, in the proposed method the direction of the image to be generated is guided by extracting the features of the target data through PCA. The results section shows the superior performance of the proposed model against a different number of images in datasets.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Reconstruction;

## KEYWORDS

Generative Adversarial Network, Principal Component Analysis, Least Square Error

### ACM Reference Format:

Myung Keun Song, Asim Niaz, and Kwang Nam Choi. 2023. Image Generation Model Applying PCA on Latent Space. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590080>

## 1 INTRODUCTION

Artificial intelligence and deep learning have recently been hot research topics, especially computer vision and natural language processing. Deep learning models require large amounts of refined data to improve the accuracy of target tasks and result data quality.

\*Corresponding author: Kwang Nam Choi (knchoi@cau.ac.kr)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590080>

Accordingly, many data sets have been created in various fields, such as CIFAR-10, COCO, and STS-B. However, in the case of face image data, the dataset is smaller than in other domains because it is difficult to collect and pre-process, such as portrait rights and personal information protection. Various studies, such as data augmentation techniques [1], are being conducted to perform smooth learning even with small data. However, it was not effective in generative models that learn the distribution of data. Therefore, we proposed a method for generating high-quality target images in this paper; the proposed method can generate compliant image data with a smaller amount of data than existing image generation models. Since the generated image is a face of a non-existent person, it can contribute to the creation of a face image dataset.

To extract the features of the generated target image through principal component analysis [2], 2,000 images corresponding to about 1 percent of the CelebA data set [3] were randomly selected and used. The feature map image obtained from 2,000 face images is applied to the latent space of the generator of the adversarial generative network that randomly generates target data and suggests the target's direction. The Large Age-Gap [4] is used for network training to compare the existing and the proposed method. It has face images of 1,010 celebrities of various ages from childhood to the present and consists of 3,828 images, which are significantly smaller than other data sets. [5] To prove the validity of the experiment, the LSGAN model, which uses the least square error [6] as a loss function to enable stable learning, was adopted as the backbone model for the Deep Convolution GAN model [7], which applies a convolutional neural network to GAN [8]. The structure of the adversarial generative neural network is shown in Figure 1 below. It consists of a generator model that learns the target data distribution and generates an image from a random vector and a discriminator model that distinguishes the generated image from the actual image.

The proposed model applies the feature map projected through PCA to the latent space of the generator of LSGAN, learns the target image more quickly compared to previous models, and is able to generate a target image that conforms with only a small amount of image data. In addition, we compared our model to the previous models, trained on 202,599 images of the CelebA dataset and the LAG datasets with 3800 images. The proposed model shows excellent results outclassing previous models when comparing the initial training image, the final generated image, and the initial generator error.

The rest of the paper describes the model proposed in Section 2, followed by experiments and results in Section 3 and Section 4, respectively. Finally, in Section 5, the conclusion of the paper is presented.

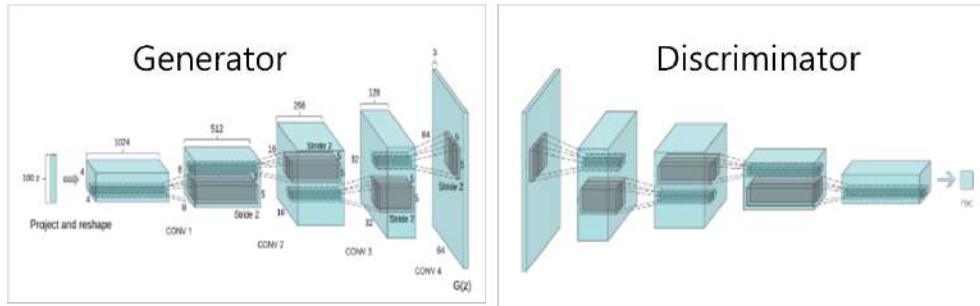


Figure 1: Deep Convolutional Generative Adversarial Network Model Architecture

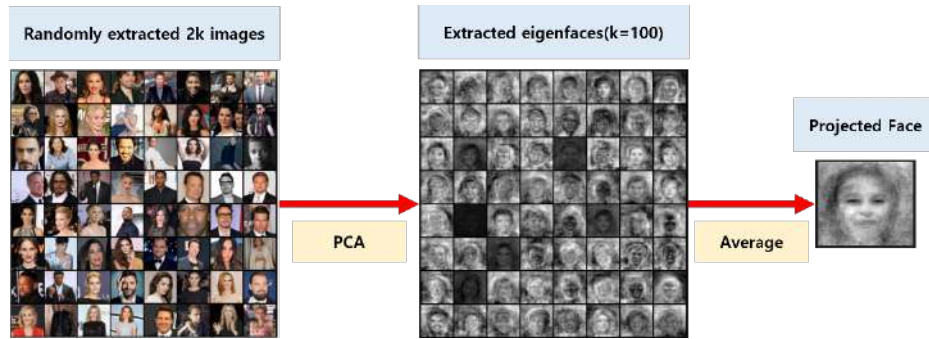


Figure 2: Extracting Projected face through Principal Component Analysis

## 2 PROPOSED METHODS

The last decade has embraced continuous research on image generation networks, including 2D and 3D image generation models. In most studies, a method of changing the structure of a generator or classifier was mainly proposed targeting the plausible data generation in the target domain. Since these studies require paired datasets or enormous data, there are several limitations in the generation process. To overcome these limitations, networks that more precisely adjust the feature condition of the generation target have been proposed. In this paper, we present the learning direction of the target data to the Generator by adding general features of the human face to the latent space, which is the input part of the generator model.

### 2.1 Feature Extractor

In this paper, we control the generation of target images by applying the condition of the projected face to the Generator’s latent space. The process of extracting the features of the face image and obtaining the projected face is shown in Figure 2. The projected face is extracted through principal component analysis from 2,000 images, corresponding to about 1 percent of the CelebA data set. Furthermore, we reduced the size of extracted projection face to  $10 \times 10$ , which is a low resolution, for two reasons. The first is to perform multiplication without adding layers to the model because the resolution of the latent space of the Generator of the LSGAN input is  $1 \times 1 \times 100$ . Second, because a high resolution includes a large number of features, the variety of generated face images may

be insufficient. Since the extracted image contains the morphological features of a human face, it is possible to generate a human face image that is faster and more natural than the existing model generated from random vectors through a convolutional layer.

### 2.2 Model

**2.2.1 Generator.** Image generation and deep learning-based model training require high-quality data for good results, and a lot of time and money is consumed in data collection and pre-processing. When the image generation model is trained with a small amount of data, a model collapse phenomenon or image quality degradation problem may occur. In this study, a generative model is trained with a small amount of data. Still, to prevent model collapse, the features of the target image extracted using PCA are applied to the Generator of LSGAN. LSGAN is a model characterized by relatively stable learning by applying least square error as a loss function to a DCGAN model in which a convolutional neural network is applied to a GAN model. The structure of the LSGAN model is the same as that of the DCGAN model, and the proposed model also uses the same model structure, as shown in Figure 1.

The LSGAN model Generator extends the randomly generated latent space size by using five transposed convolutional layers, the same as DCGAN. A transposed convolutional layer is one of the up-sampling methods that increase the size of an image, as opposed to a convolutional layer. At this time, the proposed model multiplies the projected face obtained through PCA to the latent space so that the randomly generated vectors have directionality. Through this, unlike the existing method, which is randomly initialized and

generated, the learning speed increases during initial learning. By suggesting the direction of the target image, it is possible to create a more natural image.

Fine-tuning has been made in various studies to find the appropriate resolution of the latent space, which is commonly declared to have a resolution of  $1 \times 1 \times 100$ . Then, the target image is generated while increasing the resolution of the image through the transposed convolutional layer. At this time, the features calculated through the convolutional product layer are called receptive fields. The more receptive fields there are, the more diverse features can be obtained from the resulting image, but the disadvantage is that the quality of the image deteriorates if it is increased too much at once. Therefore, finding a structure with an appropriate number of layers is important. The structure of the proposed method follows the generator structure of LSGAN and DCGAN and finally generates a  $64 \times 64 \times 3$  image through 5 transposed convolutional layers.

**2.2.2 Discriminator.** The structure of the discriminator is the same as that of the existing LSGAN. The proposed model creates the image by applying the projection face condition to the latent space of the Generator of LSGAN. Therefore, the discriminator is trained in the same way as the existing LSGAN, and whether the proposed method is effective is checked through each model's training results.

**2.2.3 Loss Function.**

$$\min \max V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

General GAN and DCGAN apply GAN Loss with Sigmoid Cross Entropy as the objective function. Equation 1.

$$\min V_{LSGAN}(D) = \frac{1}{2} E_{x \sim P_{data}(x)} [(D(x) - b)^2] + \frac{1}{2} E_{z \sim p_z(z)} [(D(G(zf)) - a)^2] \quad (2)$$

$$\min V_{LSGAN}(G) = \frac{1}{2} E_{z \sim p_z(z)} [(D(G(zf)) - c)^2] \quad (3)$$

The proposed method uses an equation obtained by multiplying the input part of the Generator by an additional feature to the LSGAN loss applied with the least square error as the objective function. (2), (3) The general GAN loss of Equation 1 consists of a min-max game of discriminator D, and generator G of value function V. D learns to maximize the probability of correctly labeling samples of the training data and samples of G as real or fake. Having G learned to minimize  $\log(1 - D(G(z)))$  (maximize  $D(G(z))$ ) (minimize  $V(D, G)$ ) is the key to GANs and min-max loss functions. At this time, discriminator D judges only real and fake, so Binary Cross Entropy loss is used. Here, the binary cross-entropy function has a problem of low learning stability due to the disappearing gradient problem.

This paper uses the objective function of LSGAN using the least square error function to overcome this problem. In Equations 2 and 3, f denotes a projection face that controls the input of the Generator. The proposed objective function aims to minimize by squaring the difference between the actual value and the predicted value rather than maximizing the formula of the discriminator and minimizing the formula of the Generator, which is the goal of the existing GAN Loss. Here, b in the discriminator formula means a real label, a in the generator formula means a fake label, and c is a value that G wants D to determine that it is real after seeing the fake

data. To minimize the min problems of Equation 2 and Equation 3,  $D(x)$  must have a value close to b, and  $D(G(zf))$  must have a value close to a. The experiments of this study were conducted through the methods mentioned above and are examined in detail in the following section, Experiments.

### 3 EXPERIMENTS

The CelebA data set consisting of 202,599 celebrity face images and the Large Age Gap data set consisting of 3,828 celebrity faces of various ages were used to train the proposed model. All experimental environments were conducted using the same Linux Ubuntu OS and Nvidia 3090 GPU.

Since the GAN model aims to generate similar data by learning the target data distribution, there is no precise evaluation index. There are methods to measure how similar the original image is to the image created through the Inception Score [9] based on the quality and variety of the image. However, there are limitations in that there are cases where a high score is obtained even if the image data is not good, such as when an image that is not in the training data set is generated, or a feature related to the quality of the generated image cannot be detected. Therefore, qualitative evaluation where individual differences exist, such as an image survey, is conducted in addition to quantitative evaluation. In this paper, along with qualitative evaluation, the evaluation was conducted through the error comparison of GAN Loss described in the cost function.

#### 3.1 Feature Extractor

First, a feature image that will be multiplied by the latent space of the generator to create a non-random direction of the generated image is obtained through principal component analysis. The data used for feature extraction is a random sample of 2,000 images, approximately 1 percent of the CelebA dataset. When dimension reduction is performed through PCA, it is possible to obtain various feature images by reducing the dimension by setting the number of principal components of the features of the original data. Figure 3 is images created by maintaining k numbers of principal components from 2,000 images. As the k value decreases, the unique features of the sample image disappear, and only minimum contour features such as eyes, nose, and mouth are displayed. Feature images with many principal components can be applied to the latent space of the generator, but the risk of damaging the diversity of generated images, which is important for evaluating the performance of the generator model, increases. Therefore, the face image generated as the average value of the eigenfaces [10], maintaining 100 principal components, was set as the feature image in this experiment.

#### 3.2 Generator Training

As mentioned earlier, the generator of LSGAN is composed of 5 transposed convolutional layers. It aims to create an image similar to the actual image by updating the weights of each layer. As explained in the objective function section above, the discriminator learns so that  $D(G(zf))$  or  $D(x)$  has a value close to the real label and a value close to the fake label, respectively. At this time, the error is obtained using the least square error function, and each layer's weights are updated through the Adam Optimizer [11].

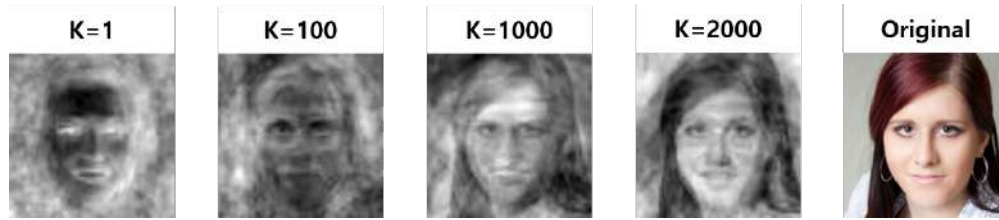


Figure 3: Image results according to K component values

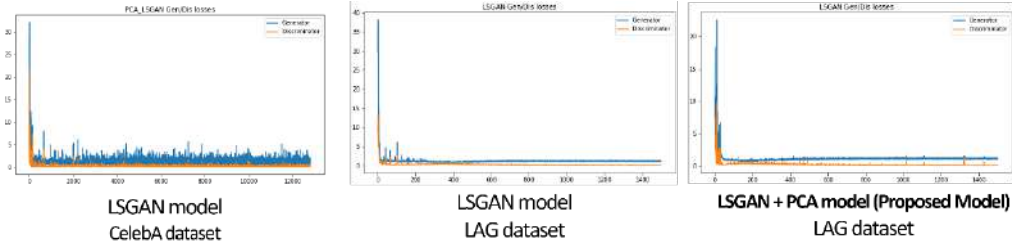


Figure 4: Loss values of the three models

First, to compare the performance with the existing LSGAN model, the CelebA data set was trained on the existing model. Since it consists of 202,599 images, it took a total of 8 hours and 30 minutes to learn 200 epochs. After that, we trained the LAG data set consisting of 3,828 images, which is about 50 times smaller than the same existing model for 500 epochs. The total learning time was 1 hour and 40 minutes, and learning was completed quickly as the number of data was small. Each model saved the images generated by the generator in the current state every 10 epochs and saved the Gan Loss. To evaluate the results, we compared GAN Loss and average GAN Loss at the initial stage for quantitative comparison. For qualitative assessment, we compared the diversity and quality of the images generated at the initial stage and the resulting images after all learning was completed. Detailed results are discussed in the result section below.

## 4 RESULTS

Figure 4 graphs the three models' generator and discriminator loss rates. Reflecting the characteristics of the least squares error function in which the weights are modified to minimize the error, most loss values converge to 1 in the later stages of learning. The average loss rate of the three models was high in the order of LSGAN using the CelebA data set, LSGAN using the LAG data set, and the proposed model (LSGAN applying PCA), and the loss rate of the generator at the beginning of training was also the lowest in the proposed model. Detailed numerical values are explained in Table 1.

Figure 5 compares the number of images learned in the initial stage of learning until the target data, the shape of the human face, is created. The original model using the CelebA dataset was trained on 64,862 image data until it could determine the shape of the face, eyes, nose, and mouth. The existing model using the LAG dataset created a face shape after training 76,944 image data. Finally, the proposed model generated face shapes after learning 46,204 image

data. We found that the proposed model generates target data the fastest in the initial training stage among the three models.

Figure 6 is the resulting image created by the generators of each model after training is done. First, comparing the results of the proposed model with the existing model trained with a small data set, we noticed that the quality of the generated image of the existing model is inferior in terms of noise or artifacts. In addition, we found no significant difference in quality and variety when comparing the generated images of the proposed model trained with a small data set and the existing model trained with a large image data set.

As shown in Table 1, the experiment results are compared with the quality of the generated image, the error rate during learning, and the image generation speed at the initial stage. The proposed model outperforms the existing model in all aspects.

## 5 CONCLUSION

In this paper, we propose a model that applies principal component analysis to the generator's latent space, which reduces the number of data sets required in a deep learning-based image generation network. In the proposed method, the latent space is multiplied by the projected face obtained through principal component analysis to LSGAN. Further, an adversarial generative model is applied to the convolutional neural network. Through this, it was possible to better generate images with facial features rather than randomly generated during image generation. Compared to the existing model, the proposed model showed similar or better image generation results than the existing model, even with relatively little training data. In addition, as a result of generating images with a small amount of training data, both the proposed model and the existing model produced images with more variety and higher quality than the existing model, and the initial stage and average error rate were small. The proposed model suggests a direction for research on image generation models using small data sets. It

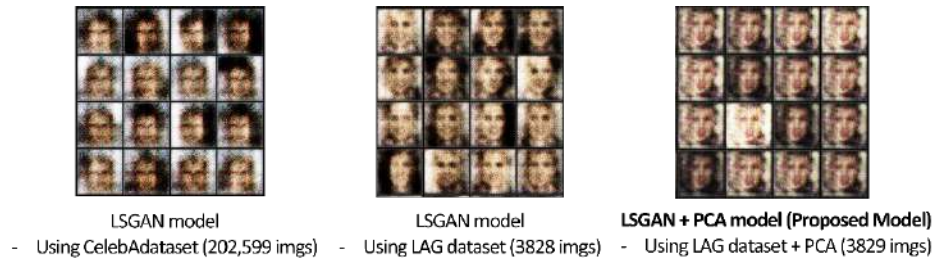


Figure 5: Result images of initial stage of training

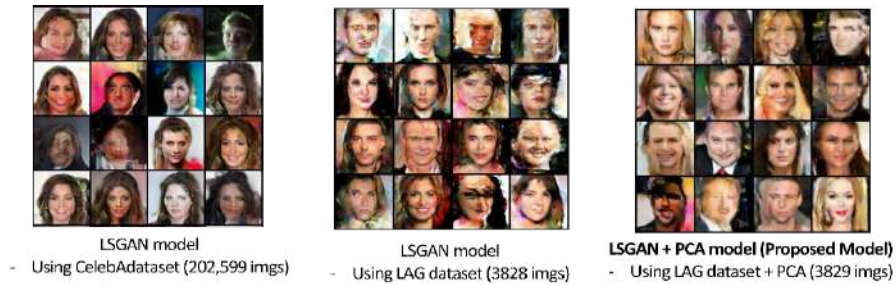


Figure 6: Result images generated by the generators

Table 1: Comparison results by model and dataset

Model	Dataset	G Loss(average)	D Loss(average)	G loss in initial step	Training images for human feature
LSGAN	CelebA	3.261123	0.869752	12.6	64,862
LSGAN	LAG	1.956428	0.671219	11.2	76,944
<b>LSGAN + PCA</b>	<b>LAG</b>	<b>1.937428</b>	<b>0.552874</b>	<b>10.2</b>	<b>46,204</b>

could serve as a research base for models that generate new data for domain data sets that are difficult to collect in large quantities, such as human face data. In the future, if domain adaptation technology is applied in the pre-PCA processing stage, we expect that data can be effectively generated in various domains, not limited to faces.

## ACKNOWLEDGMENTS

This research was supported by the 'High Performance Computing Support' project of the Ministry of Science and ICT and the National IT Industry Promotion Agency. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00107, Development of the technology to automate the recommendations for big data analytic models that define data characteristics and problems

## REFERENCES

- [1] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6(1), 1-48.
- [2] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- [3] Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15(2018), 11.
- [4] Bianco, S. (2017). Large age-gap face verification by feature injection in deep networks. *Pattern Recognition Letters*, 90, 36-42.
- [5] Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2794-2802).
- [6] Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3), 469-475.
- [7] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- [8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144.
- [9] Barratt, S., & Sharma, R. (2018). A note on the inception score. *arXiv preprint arXiv:1801.01973*.
- [10] Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1), 71-86.
- [11] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- [12] Cha, G. S., Asim, U., Song, M. K., Niaz, A., & Choi, K. N. (2022, August). Image Generation Network Model based on Principal Component Analysis. In *2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE)* (pp. 76-80). IEEE.

# An Interpretable Brain Network Atlas-Based Hybrid Model for Mild Cognitive Impairment Progression Prediction

Xianglong Guan  
gxl@stu.gxnu.edu.cn  
Guangxi Normal University  
Guilin, Guangxi, China

Li Ma  
122012017@glmc.edu.cn  
Guilin Medical University  
Guilin, Guangxi, China

Suqin Tang  
Guangxi Normal University  
Guilin, China

Tinghui Li  
tinghuili@gxnu.edu.cn  
Guangxi Normal University  
Guilin, Guangxi, China

Yunyou Huang\*  
Guangxi Normal University  
Guilin, China  
huangyunyou@gxnu.edu.cn

## ABSTRACT

The process of Alzheimer's disease (AD) is irreversible, but reasonable medical intervention for preclinical AD can delay AD's onset. Progressive mild cognitive impairment (pMCI) is the most critical stage for AD preclinical intervention. Therefore, accurate identification of pMCI will significantly improve patient benefits. Functional MRI is a neuroimaging modality that has been widely utilized to study brain activity related to AD. However, it is challenging to obtain functional MRI data, and a small amount of data will easily lead to the overfitting of the identification model. In addition, the current pMCI identification model lack interpretability leads to difficulty in acceptance by clinicians. In this work, we propose an interpretable hybrid model based on a brain network atlas to identify pMCI subjects. First, the hybrid model utilizes multi-layer perceptron to obtain categorical global features to help graph neural networks reduce overfitting. Second, the attention mechanism is introduced into the model to explain the recognition behavior of the model. The results show that our model outperforms the comparison models on multiple metrics.

## CCS CONCEPTS

• Computing methodologies → Artificial intelligence; Neural networks.

## KEYWORDS

Alzheimer's Disease, Graph attention network, Mild cognitive impairment, Personalized regions selection, rs-fMRI analysis

## ACM Reference Format:

Xianglong Guan, Li Ma, Suqin Tang, Tinghui Li, and Yunyou Huang. 2023. An Interpretable Brain Network Atlas-Based Hybrid Model for Mild Cognitive Impairment Progression Prediction. In *2023 2nd Asia Conference*

\*Correspondence authors: Yunyou Huang (huangyunyou@gxnu.edu.cn)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590081>

on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590081>

## 1 INTRODUCTION

According to new research, the number of people with Alzheimer's worldwide is set to triple by 2050[12]. Mild cognitive impairment (MCI) is a transitional stage between normal cognition (CN) and AD. At the same time, MCI can be further divided into stable MCI (sMCI) and progressive MCI (pMCI) according to whether MCI transforms into AD [20]. Although there is no cure for pMCI, patients with pMCI can take medications to slow cognitive decline[11]. Therefore, the pMCI is the most critical stage for AD preclinical intervention.

Currently, most pMCI identification research is based on structural MRI (sMRI) since sMRI is readily available[3]. However, changes in a patient's brain structure usually mean that the patient has developed into the clinical stage of AD[6]. Therefore, the researchers hope there can be some way to predict pMCI before the brain structure changes significantly. The fMRI image in the resting state reflects the fluctuation of the brain, and these fluctuations usually change in the pre-or early stages of the disease[7, 10]. Thus, researchers began to turn their attention to functional MRI (fMRI), expecting to find the preclinical features of AD for identifying pMCI patients. With the development of artificial intelligence, more deep learning algorithms have been proposed to use fMRI data features for diagnosing preclinical Alzheimer's disease[4]. Convolutional Neural Networks (CNNs) and Graph Neural Network (GNN) have achieved promising results in medical image analysis, and previous studies have shown that both CNN and GNN has promising potential. Kam et al. decomposed rs-fMRI into multiple static brain functional networks as features and used 3D CNN for feature learning[8]. Sima et al. proposed a 3D CNN, which realized the effective prediction of MCI through the analysis of multi-modal data[13]. Zhao et al. used the functional connection network of rs-fMRI and other information of the subjects to construct a graph. Then, they used the graph convolutional network for MRI diagnosis[19]. Wen et al. proposed a multi-view graph convolutional neural network (MVS-GCN), which combines graph structure learning and multi-task graph embedding learning. This method is conducive to the embedding learning of brain networks and can improve the diagnostic effect[16]. However, CNN requires much data to train the model, and fMRI data is challenging to obtain. Therefore, training CNN

models with fMRI data usually leads to the problem of overfitting. At the same time, CNN’s lack of interpretability makes it difficult for people to judge the model’s behavior intuitively. GNN also has problems such as overfitting and low interpretability. Although GNN has improved interpretability compared with CNN, it is still tricky to intuitively judge the behavior and changes of the model.

In order to reduce the model’s overfitting and increase the model’s interpretability, this paper proposes a new interpretable brain network atlas-based hybrid model for mild cognitive impairment progression prediction. First, we convert fMRI data into networks based on brain network atlases. This method enhances the model’s interpretability. Secondly, we constructed a hybrid model for MCI diagnosis to obtain more feature information and mitigate the risk of model overfitting. Finally, we introduce an attention mechanism to explain the model’s behavior and reveal the disease’s pathological features.

In summary, our work has the following contributions:

- A new interpretable hybrid model is proposed to classify brain disorders. It uses Graph Attention Networks(GAT) Layers to learn the relationship between local Region of Interest(ROIs) and uses Multi-Layer Perceptron(MLP) to learn global features, which can be used to identify MCI-related brain regions.
- By analyzing the network attention coefficients, we identify and visualize ROIs that significantly impact MCI.
- Experiments demonstrate that our model achieves state-of-the-art performance.

## 2 METHODS

The framework of the MCI predictions model consists of three components: image preprocessing, functional connectivity matrix acquisition, and model, as shown in Figure 1.

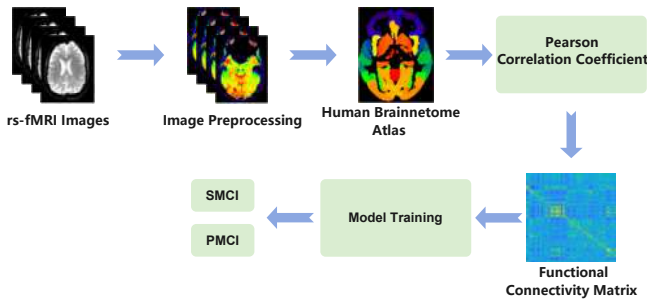


Figure 1: Overall flowchart of MCI diagnosis.

### 2.1 Image Preprocessing and Functional Connectivity Matrix FC

In this work, we preprocessed the dataset using SPM12 (Statistical Parametric Mapping 12) and BRANT[18] (A Versatile and Extendable Resting-State fMRI Toolkit) toolbox. The preprocessing process of fMRI data includes data format conversion, discarding the first ten time points for each subject’s magnetization balance, time layer correction, head movement correction, spatial normalization to the

MNI standard space, resample to a  $3 \times 3 \times 3 \text{mm}^3$  voxel size and noise removal.

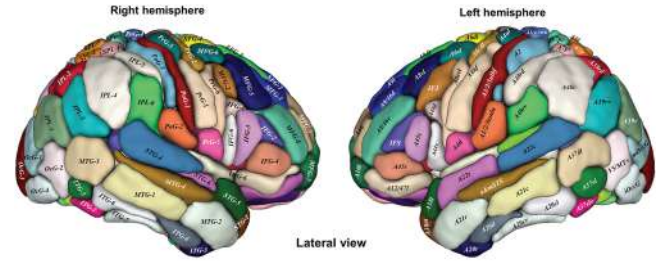


Figure 2: Parcellation scheme of the human brain in the Brainnetome Atlas. Blocks of different colors represent different brain regions generated by the Human Brainnetome Atlas[5].

We performed time series extraction of each brain ROI on the preprocessed rs-fMRI images (at this time, each data contained 130 time points). Here we use the human brainnetome atlas[5] provided by Fan et al. to select the ROI. The atlas contains 246 brain regions, as shown in Figure 2. Next, we use the Brant toolbox to calculate the average time series matrix ( $246 \times 130$ ) of 246 ROIs. Finally, using the average time series obtained from rs-fMRI data, we calculated the Pearson correlation coefficient[2] between each ROI. The Pearson correlation coefficients of the two time series are shown in Equation 1:

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sigma_X \sigma_Y} \quad (1)$$

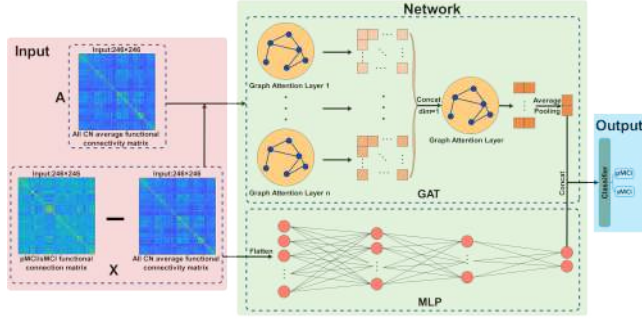
After calculating the Pearson correlation coefficient, it will generate a  $246 \times 246$  functional connectivity matrix. The results of the Pearson correlation coefficient  $\rho_{X,Y}$  are in the range of  $-1 \leq \rho_{X,Y} \leq 1$ . Through the functional connectivity matrix, we can obtain whether the functions of the two brain regions are synergistic or antagonistic within a certain period of time[14].

### 2.2 The proposed model

The framework of the proposed model consists of four components: feature input, Graph Attention Networks Layers, MLP Information Compensation Layers, and a simple classifier, as shown in Figure 3.

**2.2.1 Input Feature Set.** Different from the current general use of the preprocessed functional connection matrix as input (such as the input used by Sun[14]), we used X(the functional connection matrix of pMCI and sMCI minus the average functional connection matrix of all CN we selected) as the input of node feature of the graph attention neural network layer and the input of the fully connected neural network layer. Specifically, both X and A are symmetric matrices of  $246 \times 246$ ; we regard  $X[i]$  as the feature of the  $i$ th node of GAT Layers and express  $A[i][j]$  as the edge (connection strength) between the  $i$ th and the  $j$ th nodes.

**2.2.2 Graph Attention Networks Module.** This module consists of a multi-head graph GAT and a separate (graph attention layer) GAL. The graph attention layer has an edge-sharing mechanism, does not depend on the global graph structure, and can calculate the



**Figure 3: Overview of our model framework.** We constructed a new input feature, using feature X as the node feature of GAT and the input feature of MLP, and using A as the edge feature of GAT. At the same time, we construct a network of two channels, a Multi-head graph attention network is used to obtain the interaction and classification features of local ROIs, and MLP is used to get global classification features.

importance of graph nodes in the neighborhood[15]. Here we use a feed-forward neural network to learn the attention coefficient:

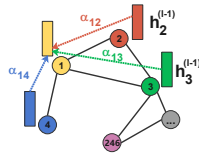
$$\alpha_{ij} = \frac{\exp(\text{Leaky ReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]))}{\sum_{v_k \in \tilde{N}(v_i)} \exp(\text{Leaky ReLU}(\mathbf{a}^T [\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_k]))} \quad (2)$$

Where  $\|$  is a series connection, and  $\mathbf{W} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$  is the weight parameter of the node feature transformation. Then we update the feature information of this node by aggregating the information of neighbor nodes:

$$\mathbf{h}'_i = \sigma \left( \sum_{v_j \in \tilde{N}(v_i)} \alpha_{ij} \mathbf{W}\mathbf{h}_j \right) \quad (3)$$

This process is shown in Figure 4. We show the schematic diagram of the node feature update of node 1. The expansion formula of this process is:

$$\mathbf{h}_1^{(l)} = \sigma(\alpha_{(12)} \mathbf{W}^{(l)} \mathbf{h}_2^{(l-1)} + \alpha_{(13)} \mathbf{W}^{(l)} \mathbf{h}_3^{(l-1)} + \alpha_{(14)} \mathbf{W}^{(l)} \mathbf{h}_4^{(l-1)}) \quad (4)$$



**Figure 4: Node features update process.**

**2.2.3 MLP Information Compensation Module.** We construct an MLP module for learning global classification features. The MLP module can significantly improve the classification accuracy of pMCI and sMCI. Here we flatten feature X as input to this module.

**Table 1: The demographic statistics of the datasets used in this work**

Study	Number of Subject	Number of Scans	Age	Gender(M/F)
pMCI	23	92	72.79(6.39)	52/40
sMCI	46	167	72.12(7.02)	89/78
CN	45	166	75.28(6.38)	74/92

**Table 2: The performance of models**

Study	ACC(%)	AUC(%)	F1(%)
KNN	69.29(6.48)	67.61(3.13)	81.09(4.13)
RF	72.83(5.00)	76.64(3.42)	82.92(2.86)
CNN	70.37(8.38)	72.31(1.75)	80.65(4.06)
RL-GCN[9]	73.08(2.72)	69.55(3.27)	79.79(2.66)
BC-GCN[1]	76.93(6.08)	65.95(5.26)	84.59(5.25)
<b>Ours</b>	<b>81.89(4.13)</b>	<b>86.97(3.55)</b>	<b>86.13(3.96)</b>

### 3 RESULTS AND DISCUSSION

#### 3.1 Data set

Data used in this paper were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. Patients were labeled pMCI if they converted to AD within 36 months. Otherwise, Patients were labeled sMCI. We downloaded rs-fMRI images of pMCI\sMCI\some CN categories from the ADNI cohort (in ADNI2 and ADNI GO), which contains 425 fMRI scans (92 pMCI, 167 sMCI, and 166 CN) from 114 subjects (23 pMCI, 46 sMCI, and 45 CN). The details of the rs-fMRI dataset we used are shown in Table 1. The rs-fMRI images were scanned using Philips Medical systems scanner with the following parameters:140 time points, field strength = 3.0, TE(Echo Time) = 30.001, TR(Repetition Time) = 3000.0, matrix (X, Y) = (64.0, 64.0), flip angle = 80.0, number of slices = 6720.0 and slice thickness = 3.313. The data was obtained in DICOM format.

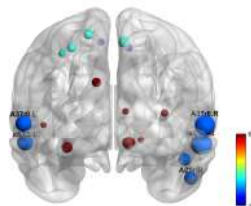
#### 3.2 Models for comparison

We validate the effectiveness of our model by comparing it with related works: the K-Nearest Neighbor(KNN) and the Random Forest(RF), the Convolution Neural Networks(CNN), the RL-GCN[9] and the BC-GCN[1].

#### 3.3 The performance of models

To demonstrate the effectiveness of our hybrid model, we compare it with several mainstream models. As shown in Table 2: our model achieves the best score compared to other models and has achieved more than 80% results in ACC, AUC, and F1 evaluation indicators. Although the BC-GCN model also achieved acceptable ACC scores, it achieved the lowest AUC value. At the same time, we can observe that the variance values (4.13, 3.55, 3.96) of various indicators of our model have achieved good results, indicating that our model has good stability. In conclusion, our model is effective and performs well.

### 3.4 The interpretability of models



**Figure 5: Visualization of ROIs with high impact on MCI. We extracted the top 20 ROIs with the influence on MCI and displayed them in the brain map. Each node represents an ROI, the node's color represents different brain regions, and the node's size means the degree of influence of the ROI on MCI. We visualization of ROIs with high impact on MCI by the MATLAB-based BrainNet toolbox.[17].**

To explain the behavior of our model, we introduce graph attention layers to address this issue. First, we train and obtain the best-performing model. Then we calculated the attention coefficient to the subject's node. The size of the attention coefficient represents the importance of the node. Because the nodes represent medical brain regions, the size of the attention coefficient also means the degree of influence of the ROI on the MCI. Finally, we identified 20 ROIs with top attention coefficient values and visualized these 20 ROIs. As shown in Figure 5, the areas that rank relatively high are the middle temporal gyrus anterior superior temporal sulcus (aSTS\_r, aSTS\_l), the brain middle temporal gyrus dorsolateral area 37 (A37dl\_r, A37dl\_l) and the right brain middle temporal gyrus rostral area 21 (A21r\_r) and other brain regions. Most of them concentrate in the Temporal lobe, the Occipital lobe, and the Parietal lobe.

## 4 CONCLUSIONS

This paper proposes an interpretable hybrid model based on a brain network atlas to identify pMCI. The experimental results show that our model has achieved more advanced results than other classical algorithms and can effectively alleviate the overfitting problem. In addition, according to the attention coefficient value obtained in the experiment, we obtained the brain regions that have a significant impact on pMCI, mainly in the temporal lobe, occipital lobe, and parietal lobe. Our paper has the following contributions: First, we propose an effective interpretable mixed model to identify pMCI. The model can effectively avoid the overfitting problem caused by the small sample size. Second, we introduce an attention mechanism, which not only captures high-contribution ROIs but also explains the recognition behavior of the model. At the end of the paper, we call researchers to pay more attention to the impact of relevant brain regions on the disease, and the different patterns of patients also need to be further explored.

## ACKNOWLEDGMENTS

This work is supported by the Project of the National Natural Science Foundation of China (Grant No.61967002), the Project of

Guangxi Science and Technology (Grant No. GuiKeAD20297004), and the Key Program of the National Natural Science Foundation of China (Grant No.U21A20474).

## REFERENCES

- [1] Abdulaziz Alorfi and Muhammad Usman Ghani Khan. 2022. Multi-label classification of Alzheimer's disease stages from resting-state fMRI-based correlation connectivity data and deep learning. *Computers in Biology and Medicine* 151 (2022), 106240. <https://doi.org/10.1016/j.cmpbiomed.2022.106240>
- [2] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Springer* (2009), 1–4. [https://doi.org/10.1007/978-3-642-00296-0\\_5](https://doi.org/10.1007/978-3-642-00296-0_5)
- [3] Elie Dolgin. 2018. Alzheimer's disease is getting easier to spot. *Nature* 559, 7715 (2018), S10–S10.
- [4] Mr Amir Ebrahimighahnavieh, Suhui Luo, and Raymond Chiong. 2020. Deep learning to detect Alzheimer's disease from neuroimaging: A systematic literature review. *Computer methods and programs in biomedicine* 187 (2020), 105242. <https://doi.org/10.1016/j.cmpb.2019.105242>
- [5] Lingzhong Fan, Hai Li, Junjie Zhuo, Yu Zhang, Jiaojian Wang, Liangfu Chen, Zhengyi Yang, Congying Chu, Sangma Xie, Angela R Laird, et al. 2016. The human brainnetome atlas: a new brain atlas based on connectonal architecture. *Cerebral cortex* 26, 8 (2016), 3508–3526. <https://doi.org/10.1093/cercor/bhw157>
- [6] Kilian Hett, Vinh-Thong Ta, José V Manjón, Pierrick Coupé, Alzheimer's Disease Neuroimaging Initiative, et al. 2018. Adaptive fusion of texture-based grading for Alzheimer's disease classification. *Computerized Medical Imaging and Graphics* 70 (2018), 8–16. <https://doi.org/10.1016/j.compmedimag.2018.08.002>
- [7] Ronghui Ju, Chenhui Hu, Quanzheng Li, et al. 2017. Early diagnosis of Alzheimer's disease based on resting-state brain networks and deep learning. *IEEE/ACM transactions on computational biology and bioinformatics* 16, 1 (2017), 244–257. <https://doi.org/10.1109/TCBB.2017.2776910>
- [8] Tae-Eui Kam, Han Zhang, Zhicheng Jiao, and Dinggang Shen. 2019. Deep learning of static and dynamic brain functional networks for early MCI detection. *IEEE transactions on medical imaging* 39, 2 (2019), 478–487. <https://doi.org/10.1109/TMI.2019.2928790>
- [9] Jiyeon Lee, Wonjun Ko, Eunsong Kang, Heung-Il Suk, Alzheimer's Disease Neuroimaging Initiative, et al. 2021. A unified framework for personalized regions selection and functional relation modeling for early MCI identification. *NeuroImage* 236 (2021), 118048. <https://doi.org/10.1016/j.neuroimage.2021.118048>
- [10] R Mohtasib, J Alghamdi, A Jobeir, A Masawi, N Pedrosa de Barros, T Billiet, H Struyfs, TV Phan, W Van Hecke, and A Ribbens. 2022. MRI biomarkers for Alzheimer's disease: the impact of functional connectivity in the default mode network and structural connectivity between lobes on diagnostic accuracy. *Heliyon* 8, 2 (2022), e08901.
- [11] Edward D Plowey, Thierry Bussiere, Raj Rajagovindan, Jennifer Sebalusky, Stefan Hamann, Christian von Hehn, Carmen Castrillo-Viguera, Alfred Sandrock, Samantha Budd Haerberlein, Christopher H van Dyck, et al. 2022. Alzheimer disease neuropathology in a patient previously treated with aducanumab. *Acta Neuropathologica* (2022), 1–11. <https://doi.org/10.1007/s00401-022-02433-4>
- [12] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E Teunissen, Jeffrey Cummings, and Wiesje M van der Flier. 2021. Alzheimer's disease. *The Lancet* 397, 10284 (2021), 1577–1590. [https://doi.org/10.1016/S0140-6736\(20\)32205-4](https://doi.org/10.1016/S0140-6736(20)32205-4)
- [13] Ahmad Shalbaf Sima Ghafoori. 2022. Predicting conversion from MCI to AD by integration of rs-fMRI and clinical information using 3D-convolutional neural network. *International Journal of Computer Assisted Radiology and Surgery* 17 (2022), 1245–1255. <https://doi.org/10.1007/s11548-022-02620-4>
- [14] Haijing Sun, Anna Wang, and Shanshan He. 2022. Temporal and Spatial Analysis of Alzheimer's Disease Based on an Improved Convolutional Neural Network and a Resting-State fMRI Brain Functional Network. *International Journal of Environmental Research and Public Health* 19, 8 (2022), 4508. <https://doi.org/10.3390/ijerph19084508>
- [15] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017). <https://doi.org/10.48550/arXiv.1710.10903>
- [16] Guangqi Wen, Peng Cao, Huiwen Bao, Wenju Yang, Tong Zheng, and Osmar Zaiane. 2022. MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis. *Computers in Biology and Medicine* 142 (2022), 105239. <https://doi.org/10.1016/j.cmpbiomed.2022.105239>
- [17] Mingrui Xia, Jinhui Wang, and Yong He. 2013. BrainNet Viewer: a network visualization tool for human brain connectomics. *PloS one* 8, 7 (2013), e68910. <https://doi.org/10.1371/journal.pone.0068910>
- [18] Kaibin Xu, Yong Liu, Yafeng Zhan, Jiaji Ren, and Tianzi Jiang. 2018. BRANT: a versatile and extendable resting-state fMRI toolkit. *Frontiers in neuroinformatics* 12 (2018), 52. <https://doi.org/10.3389/fninf.2018.00052>

- [19] Xin Zhao, Feng Zhou, Le Ou-Yang, Tianfu Wang, and Baiying Lei. 2019. Graph convolutional network analysis for mild cognitive impairment prediction. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 1598–1601. <https://doi.org/10.1109/ISBI.2019.8759256>
- [20] Bowen Zheng, Ang Gao, Xiaona Huang, Yuhua Li, Dong Liang, and Xiaojing Long. 2022. A modified 3D EfficientNet for the classification of Alzheimer's disease using structural magnetic resonance images. *IET Image Processing* (2022). <https://doi.org/10.1049/ipr2.12618>

# Improved Convolutional Neural Networks by Integrating High-frequency Information for Image Classification

Chengyuan Zhuang

Xiaohui Yuan\*

Xuan Guo

Department of Computer Science and Engineering,  
University of North Texas  
3940 N. Elm, Denton, Texas, USA 76207

Zhenchun Wei

Juan Xu

Yuqi Fan

School of Computer and Information, Hefei University of  
Technology  
193 Tunxi Road, Hefei, Anhui, China, 230009

## ABSTRACT

Deep convolutional neural networks are powerful and popular tools as deep learning emerges in recent years for image classification in computer vision. However, it is difficult to learn convolutional filters from the examples. The innate frequency property of the data has not been well considered. To address this problem, we find high-frequency information import within deep networks and therefore propose our high-pass attention method (HPA) to help the learning process. HPA explicitly generates high-frequency information via a stage-wise high-pass filter to alleviate the burden of learning such information. Strengthened by channel attention on the concatenated features, our method demonstrates consistent improvements upon ResNet-18/ResNet-50 by 1.36%/1.60% and 1.47%/1.39% on the ImageNet-1K dataset and the Food-101 dataset, respectively, as well as the effectiveness over a variety of modules.

## CCS CONCEPTS

• Computing methodologies → Object recognition.

## KEYWORDS

classification, deep convolutional neural networks, high frequency, attention

### ACM Reference Format:

Chengyuan Zhuang, Xiaohui Yuan, Xuan Guo, Zhenchun Wei, Juan Xu, and Yuqi Fan. 2023. Improved Convolutional Neural Networks by Integrating High-frequency Information for Image Classification. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590082>

## 1 INTRODUCTION

Deep convolutional neural networks (DCNNs) have become powerful and popular tools for image classification in computer vision. The power lies in a large number of parameters (typically dozens of

millions) for the representation ability of the model in classification tasks, especially convolutional filters as feature extractors [11, 15], within deep layers. However, deep networks are difficult to train via backpropagation to learn all the convolutional filters. Studies have been conducted to address this learning difficulty. Although techniques such as initialization [3], normalization [7], data augmentation [5], and complex network structures [4, 14] alleviate the problem to some extent, deep network learning remains a challenging task.

Recent studies start to rethink the value and explore designs from the frequency perspective, which is popular in traditional methods before deep learning. Anti-aliasing [17] inserts low-pass filters into the deep network before the downsampling step to mitigate aliasing artifacts of high-frequency signals. Octave Convolution (OctConv) [1] generates low-frequency feature maps in reduced resolution from the original feature maps while enabling communication within and between them. High-frequency residual learning [2] reuses low-frequency features from an auxiliary parallel network with low-resolution input, to focus on learning high-frequency residuals. As suggested by Wang et al. [12], high frequency components are critical for correct deep network predictions, as well as training heuristics such as BatchNorm [7] and Mix-up [16]. However, in the above methods, learning high-frequency information for convolutional filters purely relies on backpropagation, leaving all the burdens to the learning process.

We find high-frequency information at different scales valuable, and explicitly generating such information within deep networks leads to improved performance. As shown in Fig. 1, high-frequency information on a fine scale provides more details, while on a coarse scale provides more global structural information. In this paper, we propose our high-pass attention method (HPA), which explores high-pass filters in a stage-wise fashion to explicitly generate high-frequency information starting from the original scale for improved network learning. The integrated high-pass filter with a clear physical meaning is capable of capturing high-frequency information, which could alleviate the burden of learning such information within the network and help the following convolutional filters focus on further extracting discriminative features from different perspectives.

Our main contributions include 1) exploring the integration of high-pass filter within the deep network starting from the original scale, to explicitly generate high-frequency information for helping the learning process; 2) applying the attention mechanism to make

\*Corresponding author. Email: xiaohui.yuan@unt.edu

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590082>



Figure 1: Original image (left) and high-frequency information on three scales (right).

better use of high-frequency information; and 3) proposing a stage-wise design that is compatible and effective on a variety of modules, as well as the choice of frequencies.

## 2 RELATED WORK

**High-frequency studies.** High-frequency information is visualized [15] in feature maps of the initial layer [5]. Visualization in higher layers also demonstrates high-frequency information, such as boundaries and contours. High-frequency components are found to be critical [12] for deep networks, from cases of correct predictions on high-frequency components of images and incorrect predictions on low-frequency ones. It is also discovered that training heuristics BatchNorm [7] and Mix-up [16] tend to catch more high-frequency components.

**Wavelet transform.** Multilevel wavelet CNN (MWCNN) [9] for image restoration extends Discrete wavelet transform (DWT) to the image classification task. DWT replaces downsampling, by decomposing feature maps into four subbands in reduced resolution: one low-frequency and three high-frequency bands (in horizontal, vertical, and diagonal directions). However, DWT only targets certain scales and suffers information loss by compression and high computational cost.

**Anti-aliasing.** Anti-aliasing method [17] integrates the low-pass filter into deep networks to mitigate high-frequency signals aliasing into low-frequency ones during downsampling. By splitting max/average pooling or strided convolution into two steps of computation step (stride of 1) and downsampling step, it shows inserting a low-pass filter between the two would increase accuracy and robustness. WaveCNet [8] for robustness is also antialiasing, which uses DWT to replace downsampling, but keeps only the low-frequency band in reduced resolution and replaces max/average pooling.

**Low frequency features.** Octave Convolution (OctConv) [1] factorizes the feature maps into groups of high and low frequencies and obtains the latter by average pooling from the former (original) feature maps. It enables convolutions within and between groups and finally keeps the former, to help the learning of it. The best accuracy adopts the ratio of low-frequency channels as 0.125. High-frequency residual learning [2] applies two branches with

identical network structures for the high (original) and low resolution of the same input, respectively. The high-resolution network reuses the features of the low-resolution network by upsampling and summation to focus on learning high-frequency residuals.

**Attention.** Squeeze-and-Excitation (SE) [6] archives channel reweighting by the squeeze of context via the global average pooling for a descriptor per channel and then excitation via two fully connected layers and the sigmoid function. Bottleneck Attention Module (BAM) [10] computes channel attention similarly and computes spatial attention through  $1 \times 1$ , two dilated  $3 \times 3$  convolutions and  $1 \times 1$  convolution, with the arrangement of channel and spatial attention in parallel by summation. Convolutional Block Attention Module (CBAM) [13] adopts the sequential order of channel spatial attention for the best performance. It uses channel-wise global max and average pooling, with shared multilayer perceptron (MLP), concatenation, and sigmoid function for channel attention first. It then uses max and average pooling along the channel axis, with concatenation,  $7 \times 7$  convolution, and sigmoid function for spatial attention.

## 3 PROPOSED METHOD

Our method consists of two components: the high-pass filter and the attention module. First, the high-pass filter is integrated at different scales within the deep network to explicitly generate high-frequency information. Then, the attention module is applied to the concatenated features (the original and generated high-frequency information) to adjust the amplitude, since high-frequency information is usually small in value. Unlike most proposed methods, which target only convolutional blocks, our design is stage-wise. This maximizes compatibility, making our method capable of working on a variety of modules to help network learning.

### 3.1 High-Pass Filtering

Our high-pass filter is placed before strided convolution inside the deep network. Taking ResNet [4] as an example, this classic design has strided convolution for the initial convolution and the first convolutional block of several consequential stages. Our high-pass filter is then placed right before these places (the initial convolution and the stage of convolution 2 ~ 4), as illustrated in Fig. 2.

Because the initial convolution has limited learning capacity, we set up an auxiliary branch with the same convolution setting to

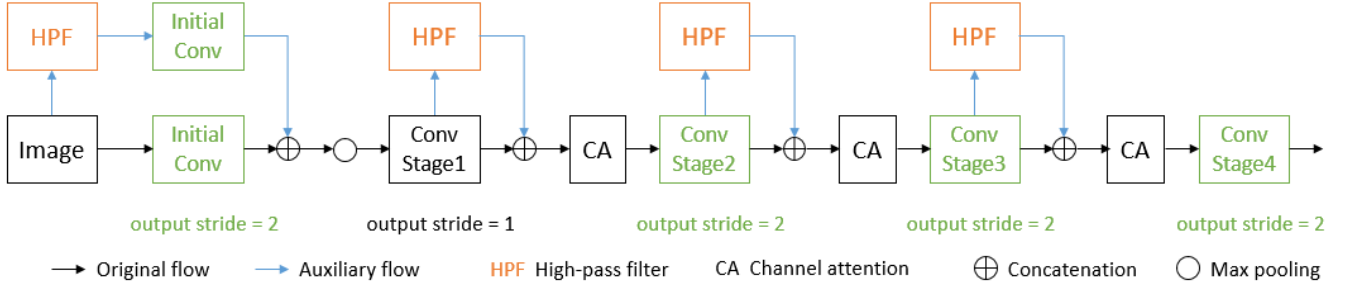


Figure 2: Network architecture of our proposed method on ResNet.

learn from the high-frequency components and the original image separately, to alleviate the learning burden. This separate learning turns out to be easier than mixing them, with better performance. For the stage of convolution 2 ~ 4 with strided convolution, no auxiliary branch is added due to the high computational cost. Instead, we use the high-pass filter to generate high-frequency information from the input features of that stage and concatenate them as the new input. The generated high-frequency information could be eventually explored throughout the stage.

For the details of our high-pass filter, we use the Gaussian high-pass filter. Here, we denote the high-pass function as  $H(f)$ , the low-pass function as  $L(f)$ , the Gaussian high-pass filter as  $H$ , the low-pass filter as  $G$ , the input features as  $f$ , the output features as  $O$ , convolution as  $Conv$ , and concatenation as  $Concat$ . The 2D low-pass filter ( $G$ ) is calculated using the following formula (normalized with the weighted sum as 1), and the 2D high-pass filter ( $H$ ) is the difference between a unit impulse matrix ( $E$ ) and the low-pass filter ( $G$ ). Here  $x$  and  $y$  denote the coordinates with respect to the center of the filter, and  $\sigma$  denotes the standard deviation to control the blurriness (we set filter size as 3 and  $\sigma$  as 1):

$$H(x, y) = E - G(x, y), \text{ where } G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

$$E = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad (2)$$

High-pass filtering results are computed by the convolution of the Gaussian high-pass filter with the input features.

$$H(f) = Conv(H, f) \quad (3)$$

This is equivalent to the difference between the input features and the low-pass filtering results.

$$L(f) = Conv(G, f), \text{ and } H(f) = f - L(f). \quad (4)$$

The output features are generated by the concatenation of the input and the high-pass filtering results.

$$O(f) = Concat(f, H(f)). \quad (5)$$

### 3.2 Attention Mechanism

We apply channel attention to adjust the amplitude of the concatenated features (original features and high-pass filtering results) since the amplitude of high-frequency information is usually small.

In our implementations, we use Squeeze-and-Excitation (SE) [6], while other attention modules could also be applied. For SE attention, the input features go through global average pooling, two fully connected layers, and the sigmoid unit to obtain the global channel descriptors and then the channel weights for recalibration. SE channel attention is illustrated in Fig. 3.

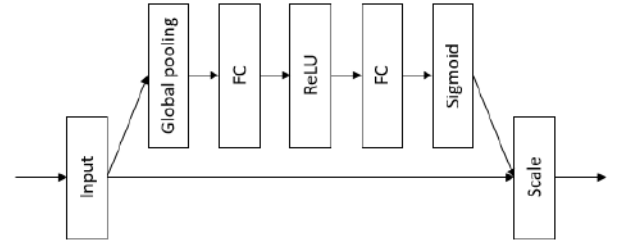


Figure 3: Structure of the attention module.

## 4 EXPERIMENTS

We evaluate our proposed design for image classification, using the ImageNet-1k dataset and the Food-101 dataset. We add our proposed design to a variety of deep network models proposed in corresponding papers, to see the improvements. We first train all models on the ImageNet-1K dataset and then perform fine-tuning on the Food-101 dataset. ImageNet-1K dataset contains 1,000 object categories, with 1.2 million training images and 50,000 validation images. Food-101 dataset contains 101 food categories, with 75,750 training images and 25,250 validation images.

In model training, images are  $224 \times 224$ . The learning rate starts at 0.1 and decays every 30 epochs by a factor of 10. Momentum is 0.9 and weight decay is 0.0001. We use random horizontal flipping, random scale [0.08, 1.0], and ratio change [0.75, 1.33] before image resizing for data augmentations. Training for each method is 100 epochs on 4 GPUs, with a batch size of 128 and batch accumulation of 2. The best model is selected among 100 epochs using validation data.

### 4.1 Performance Evaluation

Tables 1 and 2 show that our high-pass channel attention design (HPA) leads to consistent accuracy improvements upon ResNet18

and ResNet50 on both datasets. The improvements are 1.36% and 1.60% for the ImageNet dataset, respectively, while 1.47% and 1.39% for the Food-101 dataset, all in absolute value. Given the high performance of the latter dataset, error rates are further reduced by 8.78% and 10.73%. HPA also brings noticeable accuracy gains over a variety of modules such as anti-aliasing [17], OctConv [1], SE [6] and CBAM [13], demonstrating its compatibility and effectiveness. Our design (HPA) outperforms other block-wise attention (SE, CBAM) on ResNet, and contains only stage-wise channel attention, suggesting an alternative way of block-wise attention. Our design archives the best performance among the comparison methods (except for the small network ResNet18 compared to anti-aliasing [17]).

The results also demonstrate that explicitly generating high-frequency information is quite complementary to the anti-aliasing method, as combining our design with the anti-aliasing method archives the best or very competitive performance on the ImageNet dataset and the Food-101 dataset. This is intuitive because high-frequency information suffers from aliasing artifacts, which severely affects recognition performance.

**Table 1: Top-1 accuracy (%) of ResNet18 models on ImageNet dataset and Food-101 dataset.**

Method	ImageNet		Food-101	
	Acc.	Imp.	TAcc.	Imp.
ResNet18 (baseline)	70.30	-	83.26	-
+Wavelet [9]	70.99	0.69	84.50	1.24
+AA [17]	71.95	1.65	84.78	1.52
+SE [6]	71.24	0.94	84.12	0.86
+CBAM [13]	70.97	0.67	84.11	0.85
+HPA (Our)	71.66	1.36	84.73	1.47
+AA [17] +HPA (Our)	<b>73.28</b>	<b>2.98</b>	<b>85.86</b>	<b>2.60</b>
+SE [6] +HPA (Our)	72.19	1.89	85.14	1.88
+CBAM [13] +HPA (Our)	71.96	1.66	84.97	1.71

**Table 2: Top-1 accuracy (%) of ResNet50 models on ImageNet dataset and Food-101 dataset.**

Method	ImageNet		Food-101	
	Acc.	Imp.	Acc.	Imp.
ResNet50 (baseline)	76.00	-	87.05	-
+Wavelet [9]	76.56	0.56	87.75	0.70
+AA [17]	76.92	0.92	87.68	0.63
+OctConv [1]	76.94	0.94	87.77	0.72
+SE [6]	77.20	1.20	88.14	1.09
+CBAM [13]	77.48	1.48	88.35	1.30
+HPA (ours)	77.60	1.60	88.44	1.39
+AA [17] +HPA (Our)	77.71	1.71	<b>89.23</b>	<b>2.18</b>
+OctConv [1] +HPA (Our)	<b>77.93</b>	<b>1.93</b>	88.73	1.68
+SE [6] +HPA (Our)	77.76	1.76	88.63	1.58
+CBAM [13] +HPA (Our)	77.76	1.76	88.95	1.90

## 4.2 Ablation Study

We explore the contribution of the high-pass filter (only) with different numbers of scales in our design. Table 3 lists corresponding top-1 accuracy on the ImageNet dataset. Consistent improvement is achieved with an increasing number of scales upon ResNet18, from the initial scale (0.63%) to all scales (1.01%). Using a low-pass filter (LP) with all scales leads to inferior performance. This is intuitive, as blurriness missing details such as shape or texture information, which is important for recognition.

**Table 3: Top 1 accuracy using different HP scales on ImageNet dataset. HP denotes a high-pass filter, and LP denotes a low-pass filter.**

Method	ResNet18	+HP 1-scale	+HP 2-scales
Acc.	70.30	70.93	71.07

Method	+HP 3-scales	+HP 4-scales	+LP 4-scales
Acc.	71.20	<b>71.31</b>	71.08

Table 4 compares different settings of the kernel (filter) size and standard deviation in our high-pass attention design on the ImageNet dataset. We can see that filter size  $5 \times 5$  gives the best accuracy (71.89%) among the three filter sizes, and a small standard deviation (1 or 2) is better. Due to limited resources and time, we only apply a common filter size of  $3 \times 3$  (to be the same as the convolutional filter size in ResNet) for the experiments with our design on both datasets.

**Table 4: Top-1 accuracy (%) of ResNet18 models with different settings of kernel size (k) and standard deviation (s) for our high-pass attention design (HPA) on the ImageNet dataset. Here k3-s1 means kernel size is 3 and standard deviation is 1.**

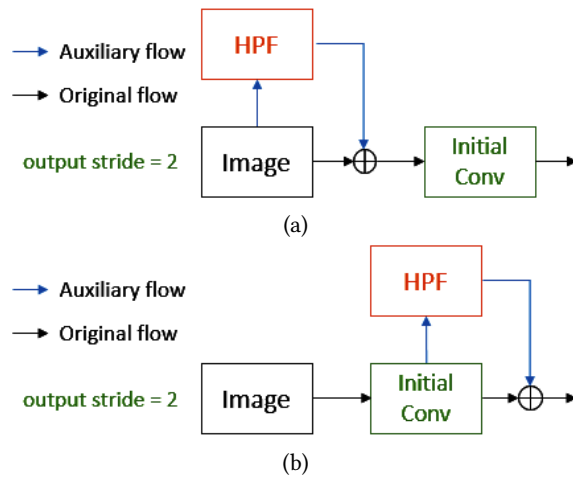
Method	ResNet18	k3-s1	k3-s2	k5-s1	k5-s2
Acc.	70.30	71.66	71.66	<b>71.89</b>	<b>71.89</b>

Method	k5-s3	k7-s1	k7-s2	k7-s3
Acc.	71.79	71.72	71.87	71.74

## 4.3 Design Choice

In Table 5, we explore variations of the initial high-pass filter for ResNet18 in our design on the ImageNet dataset. We can see that placing a high-pass filter on the image and then concatenating the image with a high-pass filtering result before the initial convolution (named init-before, as shown in Fig. 4 (a)) leads to the least performance (70.70%). In this way, raw input and high-frequency components are mixed, making it difficult to learn the simple initial convolution with limited capacity. Placing a high-pass filter after the initial convolution (named init-after, as shown in Fig. 4 (b)) leads to higher performance (71.18%). This is better because the high-pass filter works on learned convolutional features instead of



**Figure 4: Position variations of the high-pass filter (before and after) for the initial convolution in our design.**

raw pixels. The best design (named init-two-branch, used for our final design) archives 71.31%. This is done by adding an auxiliary branch with the same convolution setting to learn from the high-frequency components and the original image separately, which makes the initial convolution focus on individual input instead of mixed input.

**Table 5: Top-1 accuracy (%) of ResNet18 models with variations of the initial high-pass filter on ImageNet dataset.**

Method	ResNet18	+HP init-before
Acc.	70.30	70.70
Method	+HP init-after	+HP init-two-branch
Acc.	71.18	<b>71.31</b>

#### 4.4 Visualization

In Fig. 5, we visualize a few filtering results from the initial high-pass branch of our design. We can see that high-frequency information has been extracted from different perspectives (such as contour and texture patterns in different orientations), which are helpful cues for human visual understanding. The features of the high-pass filter at higher levels also demonstrate high-frequency information and clear details. We can observe aliasing artifacts due to downsampling, which also suggests the need for an antialiasing method.

## 5 CONCLUSION

In this paper, we propose our high-pass attention method (HPA) to help network learning. HPA explicitly generates high-frequency information through high-pass filters on all scales to alleviate the burden of learning such information. Strengthened by the attention mechanism on concatenated features, we demonstrate consistent improvements upon ResNet-18/ResNet-50 by 1.36%/1.60% and 1.47%/1.39% on the ImageNet-1K dataset and the Food-101 dataset,

respectively, as well as the effectiveness of our stage-wise design over a variety of modules. Future work of this research includes more efficient incorporation of multi-scale information and more robust and explainable architecture.

## ACKNOWLEDGMENTS

The authors declare that there is no conflict of interest regarding the publication of this paper.

## REFERENCES

- [1] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of CVPR*. 3435–3444.
- [2] Bowen Cheng, Rong Xiao, Jianfeng Wang, Thomas Huang, and Lei Zhang. 2020. High frequency residual learning for multi-scale image classification. In *30th British Machine Vision Conference, BMVC 2019*.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*. 770–778.
- [5] Geoffrey E Hinton, Alex Krizhevsky, and Ilya Sutskever. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1106–1114.
- [6] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of CVPR*. 7132–7141.
- [7] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of Machine Learning Research*, Vol. 37. 448–456.
- [8] Qiufu Li, Linlin Shen, Sheng Guo, and Zhihui Lai. 2020. Wavelet integrated cnns for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7245–7254.
- [9] Pengju Liu, Hongzhi Zhang, Wei Lian, and Wangmeng Zuo. 2019. Multi-level wavelet convolutional neural networks. *IEEE Access* 7 (2019), 74973–74985.
- [10] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In-So Kweon. 2018. BAM: Bottleneck Attention Module. In *British Machine Vision Conference (BMVC)*. British Machine Vision Association (BMVA).
- [11] Zhinan Qiao, Xiaohui Yuan, and Mohamed Elhoseny. 2020. Urban Scene Recognition via Deep Network Integration. In *Urban Intelligence and Applications*. Singapore, 135–149.
- [12] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. 2020. High-frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8684–8694.
- [13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of ECCV*. 3–19.
- [14] Xiaohui Yuan, Zhinan Qiao, and Abolfazl Meyarian. 2022. Scale Attentive Network for Scene Recognition. *Neurocomputing* 492 (2022), 612–623.
- [15] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of ECCV*. 818–833.
- [16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.
- [17] Richard Zhang. 2019. Making Convolutional Networks Shift-Invariant Again. In *International Conference on Machine Learning*. 7324–7334.

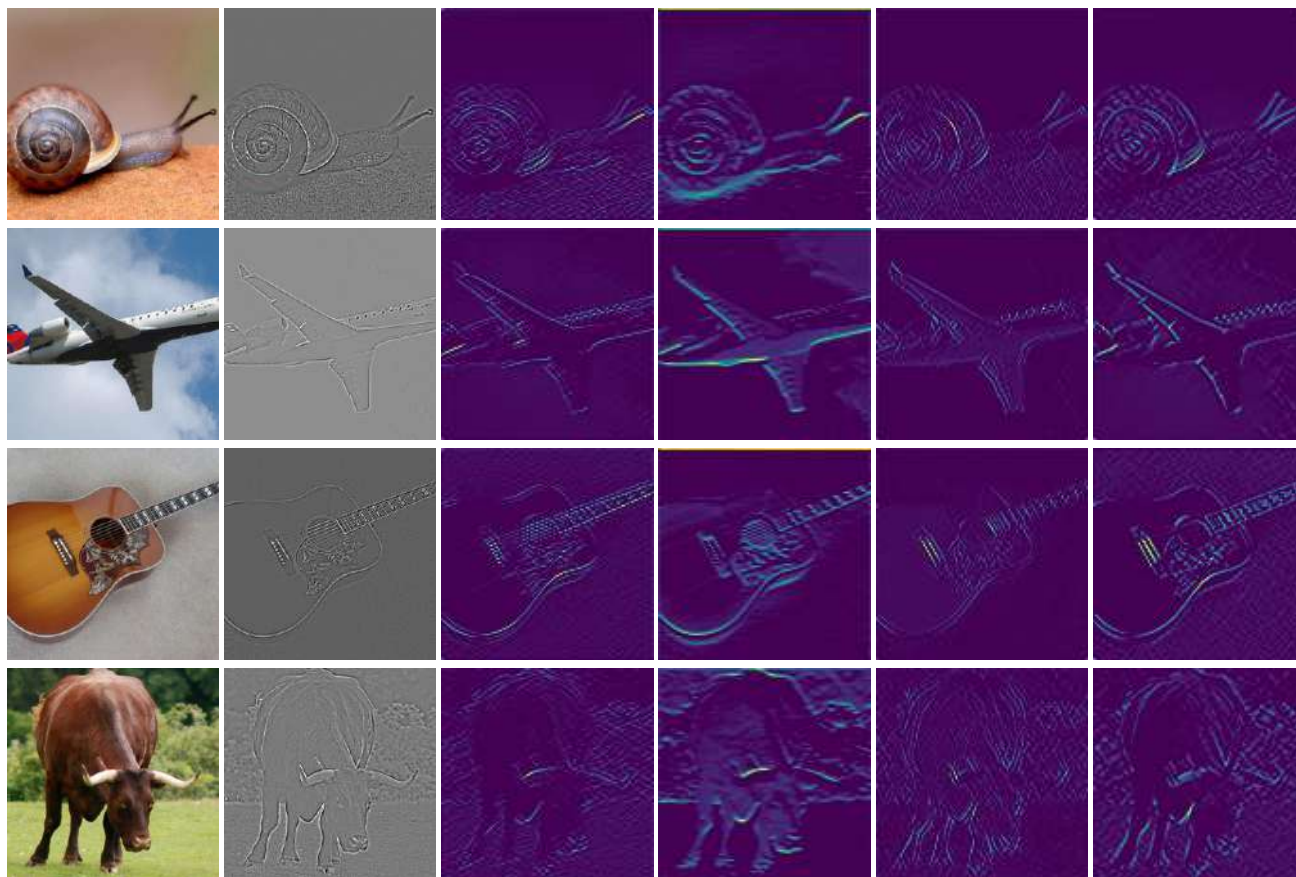


Figure 5: Original image, high-pass filter, and initial high-pass branch filtering results.

# Comparison of regional monitoring methods for grassland degradation based on remote sensing images

Haoran Wang

Inner Mongolia Key Laboratory of Wireless Networking  
and Mobile Computing, Inner Mongolia University  
Hohhot, China  
445250534@qq.com

Zhaoran Wang

Inner Mongolia Key Laboratory of Wireless Networking  
and Mobile Computing, Inner Mongolia University  
Hohhot, China, 1418050439@qq.com

Tianyu Xue

Inner Mongolia University of Science and Technology  
Baotou, China  
1120545385@qq.com

Xiangyu Bai

Inner Mongolia School of Computer Science, Inner  
Mongolia University, Hohhot, China  
bxy@imu.edu.cn

## ABSTRACT

As an integral part of the ecosystem, grassland plays an important role in protecting water and soil, preventing wind and fixing sand and protecting biodiversity. However, some grasslands are degraded at this stage, so a grassland monitoring method is urgently needed to prevent desertification from spreading. With the rapid rise of deep learning, it is more and more popular to apply artificial intelligence methods to grassland degradation monitoring. This paper systematically and comprehensively analyzes that almost all semantic segmentation methods have been applied to relevant research on grassland degradation areas since semantic segmentation methods were applied to grassland monitoring. Then, according to the different algorithm structures of grassland extraction methods, the principles of representative algorithms are introduced in turn. Then we made a statistical analysis of the publication status, research space distribution and the number of citations of papers in this field. Finally, the analysis results are discussed, and the possible research hotspots in the future are discussed.

## KEYWORDS

Grassland monitoring, Remote sensing images, Deep learning, Semantic segmentation

### ACM Reference Format:

Haoran Wang, Tianyu Xue, Zhaoran Wang, and Xiangyu Bai. 2023. Comparison of regional monitoring methods for grassland degradation based on remote sensing images. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590083>

## 1 INTRODUCTION

As a component of the ecosystem, grassland plays an important role in protecting water and soil, preventing wind and sand and

protecting biodiversity<sup>[1]</sup>. Due to the dual impact of predatory development of the ecological environment and global warming, the original grassland ecosystem has been seriously damaged. The areas affected by desertification account for 1/3 of the total land area in China, Grassland desertification has caused serious ecological and social problems<sup>[2]</sup>. Therefore, to prevent desertification from spreading, it is very necessary to accurately monitor the degraded grasslands and protect them accordingly<sup>[3]</sup>.

In recent years, with the development of remote sensing technology, the impact of UAV remote sensing images and satellite remote sensing is a grassland feature Extraction of a hot topic<sup>[4]</sup>. However, grassland monitoring is not easy to distinguish, with wide area, long cycle and high cost, which makes the task very challenging. Traditional monitoring methods are time-consuming and labor-intensive<sup>[5]</sup>. Most monitoring methods require manual intervention, and the extraction effect is not ideal. In recent years, deep learning has emerged in natural language processing, speech recognition, computer vision and other fields, developed rapidly and achieved remarkable results<sup>[6]</sup>. His rise has also attracted extensive attention in the field of remote sensing<sup>[7]</sup>. Many remote sensing related researches have also applied deep learning to scene classification, target detection, change detection and other remote sensing image analysis fields, and made significant breakthroughs. It has been widely used in remote sensing, grassland monitoring, and has achieved rapid development<sup>[8]</sup>.

At present, semantic segmentation has also made many achievements in grassland monitoring, and is also very popular in this field. Therefore, it is necessary to review its application in this field. There are some summary studies in this field. The extraction method of UNet network model in UAV grassland remote sensing image is summarized in<sup>[9]</sup>. Use special grassland species to predict grassland degradation. Based on the fact that there is no article to analyze the latest semantic segmentation method of the extraction task of the deep learning method in the field of grassland monitoring in detail. On this basis, this study comprehensively reviewed and analyzed the effectiveness of this specific task method, and finally provided further insights into the future development of this field.

The structure of the discussion is as follows: outlines the classification of grassland feature extraction methods, and focuses on the principle of representative grassland extraction algorithms based on the semantic segmentation method of remote sensing images.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590083>

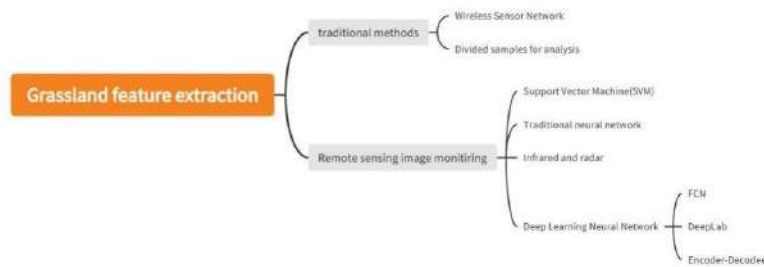


Figure 1: Classification of grassland monitoring

Describes how to The process of studying quantitative analysis results using meta-analysis methods and the whole field. Introduces the statistics and analysis results of the literature content. Discusses the conclusions and future developments [10]. We hope that this review can help readers become familiar with grassland feature extraction based on deep learning semantic segmentation task from a new perspective, and provide some possible hints for future learning.

## 2 METHODOLOGY AND DISCUSSION

This briefly summarizes and classifies the traditional methods for grassland monitoring and the monitoring methods using remote sensing images. At present, most of the studies on monitoring grassland ecological conditions only focus on single type of grassland ecological problems. Benefit Use hardware facilities or divided samples to analyze, evaluate and estimate the ecological status of the study area. However, the grassland has the characteristics of wide area, long research period and high research cost, so there are various difficulties in the practical application of traditional monitoring methods. In the aspect of remote sensing, in view of the difficulties in feature extraction and complex texture of traditional remote sensing image classification methods, a method of using depth convolution neural network to classify high- resolution remote sensing images is proposed. Compared with support vector machine (SVM) classification method and traditional neural network (NN) classification method, convolution neural network can effectively improve the classification accuracy of high- resolution remote sensing images.

### 2.1 traditional method

At present, most studies on grassland ecological monitoring only focus on single types of grassland ecological problems, and use hardware facilities or divided samples to analyze, evaluate and estimate the ecological status of research areas [11]. Literature proposed a grassland environment monitoring system scheme based on ZigBee wireless sensor network for serious grassland degradation and reduced stocking capacity. By monitoring environmental temperature and humidity signals, it can grasp the growth status of grassland grass in real time, which can better meet the application requirements of grassland environment monitoring in some areas. Based on the problems of grassland ecosystem destruction, literature proposed to use wireless sensor network to establish grassland ecological protection and monitoring system [12].

### 2.2 remote sensing image analysis

Lisitian proposed a features-matching registration method for remote sensing images [13]. The sampling set and verification set were determined by using two different thresholds, and more matching points were generated with fewer iteration times. Then the convolutional neural network was used for change detection. The change detection results are obtained by the final model, and this method has strong detection ability and good robustness. As for the classification of remote sensing images, many scholars proposed to classify high-resolution remote sensing images by using deep convolutional neural network, aiming at the difficulties in feature extraction and complex texture of traditional remote sensing image classification methods [14], and to integrate multi- source and multi-feature features of remote sensing images such as spectral features [15], texture features and spatial structure features in the form of vector or matrix. The non-subsampled contour wave transform algorithm is used to improve the extraction method of texture features. Compared with the support vector machine (SVM) classification method and the traditional neural network (NN) classification method, the deep learning network can effectively improve the classification accuracy of high- resolution remote sensing images [16].

In recent years, data- supported deep learning methods have been widely used in grassland feature extraction, such as semantic segmentation tasks. It mainly relies on the color, geometry and texture features of remote sensing images to extract features from images, and then classifies pixels and finally completes the extraction of grassland features. In this process, convolutional neural network plays a decisive role. The most basic convolutional neural network structure is usually composed of five parts: input layer, convolutional layer, nonlinear activation layer, pooling layer and complete connection layer. The input layer is the input of the whole neural network, which usually refers to the pixel matrix of the RGB image. Feature extraction is carried out through the convolutional layer, through the nonlinear activation layer, and then through the pooling layer to compress the data to extract the main features. Finally, fully connected layers are used to complete the classification task. After continuous development, deep convolutional neural networks have been perfected and many different structures have been generated. Next, we look at some of the most cited and influential studies in the field and present some of the landmark network models.

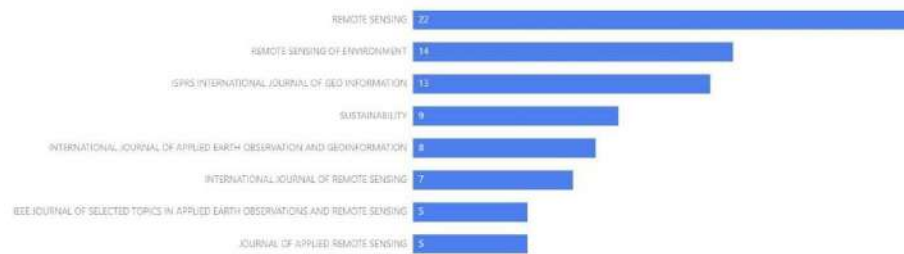


Figure 2: Source of 131 selected papers

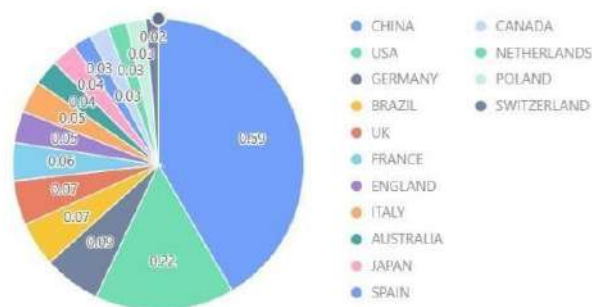


Figure 3: Country affiliation of the lead author

### 3 META ANALYSIS

#### 3.1 data source

In order to systematically screen out articles that use deep learning semantic segmentation to extract grassland from high-resolution remote sensing images, We use "remote sensing images" and "grassland recognition" in the Web of Science database to search the topic. An extensive search was conducted to prevent transition screening from affecting the quality of the analysis, followed by further subdivided screening. Retrieved from January 1, 2023, 2015-2023, from the Web of Science database, with a total of 1051 articles. Not including review papers, books, editorial materials and data reports, there are 1008 articles. Then, based on the standard title and abstract, articles that are clearly not relevant to the topic of the study are further excluded. Contributions in languages other than English are excluded. The remaining 131 articles are read, and finally 131 articles are included in a peerreviewed meta-analysis.

#### 3.2 Quantitative analysis

After a series of search and screening, firstly, the paper data is quantitatively analyzed to understand the relevant content, research hotspot and trend in this field.

**3.2.1 Source of paper.** Among the 131 papers, the top three are "REMOTE SENSING", "REMOTE SENSING OF ENVIRONMENT" and "ISPRS INTERNATIONAL JOURNAL OF GEO INFORMATION". Of these, 131 were published in REMOTE SENSING, accounting

for 58% of all relevant peer-reviewed papers. Figure 2 shows the number of publications by journal.

**3.2.2 Study spatial distribution.** As shown in Figure 3, the largest contribution of the first author is from China, accounting for 59%, followed by the United States, Germany, Brazil and the United Kingdom.

According to Figure 4, it can be seen that the top five institutions with great influence in this field are CHINESE ACADEMY OF SCIENCES, CHINESE ACAD SCI and UNIVERSITY OF CHINESE ACADEMY OF SCIENCES CAS, UNIV CHINESE ACAD SCI, INSTITUTE OF GEOGRAPHIC SCIENCES NATURAL RESOURCES RESEARCH CAS.

### 4 CONCLUSION

#### 4.1 Algorithm structure and convolutional backbone

We have reviewed the selected papers in their entirety and conducted a comprehensive statistical analysis of their contents to provide a more systematic overview of the domain.

The vast majority of relevant papers in this field are studies of deep learning semantic segmentation algorithm architectures that have been proposed and proven, so Figure 5 statistics the algorithm architectures used in the papers or further improved on this basis.



Figure 4: Overview of the lead author's affiliation

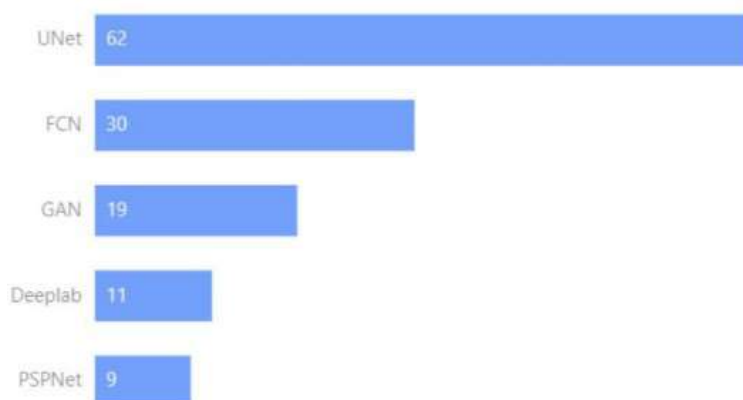


Figure 5: The situation of the deep learning network model used

Among the papers that mention the use of convolutional backbone network, ResNet is the most commonly used backbone network. There are 51 papers that use this kind of backbone network. The results obtained by ResNet working effect are often more ideal, followed by VGG. See Figure 6.

## 4.2 Grassland Dataset

For deep learning to be successful, data is very important. In every machine vision application, it is important to use "high quality" image data. Whether using classification, physical examination, segmentation or outlier detection in image processing, deep learning networks must be trained using data. Table 1 summarizes some data sets commonly used in grassland detection. We emphasize that since there is no data set specifically for grassland, most data sets need to be processed on the basis of them, leaving some required categories. In addition, about onethird of the papers studied used their own data sets.

## 5 CONCLUSION

We analyzed 131 papers and systematically reviewed the research on the use of semantic segmentation to extract grassland features from high-resolution remote sensing images, mainly the research on methods and algorithm models. When RGB images are only used without using near-infrared and lidar, the ecological environment in the grassland is complicated and the tall and short plants on the grassland are similar, so it is a challenge to correctly identify the grassland. Grassland monitoring is divided into traditional methods and remote sensing image recognition. Monitoring using remote sensing images is divided into SVM, traditional network model, near infrared and radar methods and deep neural network methods. In this paper, we systematically analyze the research results of using deep neural network to analyze the features of grassland, and mainly review and summarize the semantic segmentation algorithm and structure based on artificial intelligence. Models are divided into three categories: FCN-based methods, Deeplab and encoder-decoder. In the future, semantic segmentation can be combined with

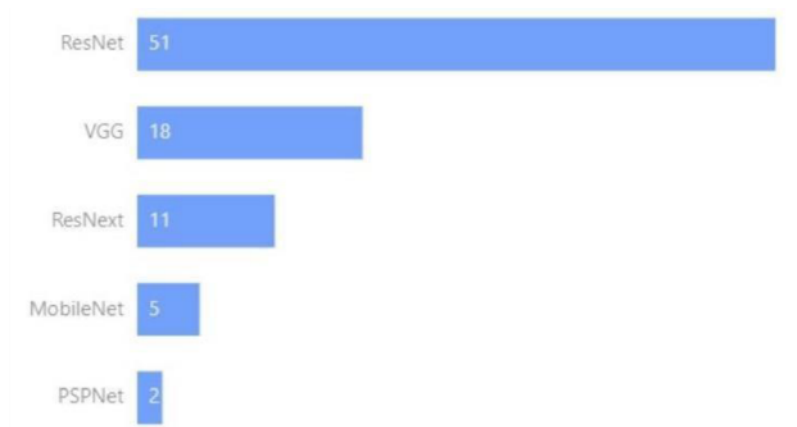


Figure 6: Basic Network Usage

Table 1: Data sets on grasslands

Dateset	Source
Slovenia Dataset[29]	<a href="http://www.sciencedirect.com/science">www.sciencedirect.com/science</a>
Cityscapes[30]	<a href="http://www.cityscapes-dataset.com/">www.cityscapes-dataset.com/</a>
SYNTHIA	<a href="http://Synthia-dataset.net/">Synthia-dataset.net/</a>
Gaofen Image Dataset[31]	<a href="http://Captain.whu.edu.cn/GID/">Captain.whu.edu.cn/GID/</a>
Vaihingen and Postdam	<a href="http://www2.isprs.org/commissions/comm3/wg4">www2.isprs.org/commissions/comm3/wg4</a>
Aerial Image Segmentataion Dataset	<a href="http://Zenodo.org/record/1154821#.XH6HtygzblU">Zenodo.org/record/1154821#.XH6HtygzblU</a>
DeepGlobe Land Cover Classification	<a href="http://Deepglobe.org/index.html">Deepglobe.org/index.html</a>
Semantic Drone Datatset	<a href="http://www.tugraz.at/index.php?id=22387">www.tugraz.at/index.php?id=22387</a>

near infrared and radar methods for detailed grassland monitoring. This paper mainly analyzed the papers on grassland monitoring by using artificial intelligence, provided theoretical basis and data basis, promoted the process of grassland digitization, and provided support for future researchers.

## ACKNOWLEDGMENTS

We would like to thank the editors and anonymous reviewers as well as the sponsors of our project.

## REFERENCES

- [1] SU J, ZHU X, LI S, *et al.* AI meets UAVs: a survey on AI empowered UAV perception systems for precision agriculture [J]. 2022.
- [2] VASWANI A, SHAZEER N, PARMAR N, *et al.* Attention is all you need [J]. 2017, 30.
- [3] WANG Q, MA Y, ZHAO K, *et al.* A comprehensive survey of loss functions in machine learning [J]. 2020: 1-26.
- [4] WOLF T, DEBUT L, SANH V, *et al.* Transformers: State-of-the-Art Natural Language Processing; proceedings of the Conference on Empirical Methods in Natural Language Processing, F, 2019 [C].
- [5] WOO S, PARK J, LEE J-Y, *et al.* Chm: Convolutional block attention module; proceedings of the Proceedings of the European conference on computer vision (ECCV), F, 2018 [C].
- [6] XU C, YANG J, LAI H, *et al.* UP-CNN: Un-pooling augmented convolutional neural network [J]. 2019, 119: 34-40.
- [7] XU K, LI C, TIAN Y, *et al.* Representation learning on graphs with jumping knowledge networks; proceedings of the International conference on machine learning, F, 2018 [C]. PMLR.
- [8] YAN X, JIANG Y, CHEN S, *et al.* Automatic Grassland Degradation Estimation Using Deep Learning; proceedings of the IJCAI, F, 2019 [C].
- [9] YANG B, WANG S, LI S, *et al.* Research and application of UAV-based hyperspectral remote sensing for smart city construction [J]. 2022, 2: 255-66.
- [10] YE M, JI L, TIANYE L, *et al.* A Lightweight Model of VGG-U-Net for Remote Sensing Image Classification [J]. 2022, 73(3): 6195-205.
- [11] YU X, GAODI X, LIN Z, J O R, *et al.* Framework Design of Eco-Technology Evaluation Platform and Integration System [J]. 2017, 8: 325 - 31.
- [12] YUAN L, CHEN Y, WANG T, *et al.* Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet [J]. 2021: 538-47.
- [13] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks; proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, F, 2014 [C]. Springer.
- [14] ZHANG C, YUE P, TAPETE D, *et al.* A multi-level context-guided classification method with object-based convolutional neural network for land cover classification using very high resolution remote sensing images [J]. 2020, 88: 102086.
- [15] ZOU R, SONG C, ZHANG Z. The devil is in the details: Window-based attention for image compression; proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, F, 2022 [C].
- [16] 石纯一, 黄昌宁. 人工智能原理[M]. 清华大学出版社有限公司, 1993.

# Multi-Modal Fusion Object Tracking Based on Fully Convolutional Siamese Network

Ke Qi

qikersa@163.com.com

School of Computer Science and Cyber Engineer,  
Guangzhou University  
Guangzhou, China

Yicong Zhou

School of Computer and Information Science, University  
of Macau  
Macau, China

Liji Chen\*

rhichardChan@gmail.com

School of Computer Science and Cyber Engineer,  
Guangzhou University  
Guangzhou, China

Yutao Qi

School of Computer and Information Engineering,  
Guangzhou Huali College  
Guangzhou, China

## ABSTRACT

RGBT tracking incorporates thermal infrared data to achieve more accurate visual tracking. However, the efficiency of RGBT tracking may be diminished by some bottlenecks, such as thermal crossover, illumination variation and occlusion. To address the aforementioned problems, we propose a fully-convolutional Siamese-based Multi-modal Feature Fusion Network (SiamMFF) that integrates RGB and thermal features. In our work, visible and infrared images are initially processed by the Multi-Modal Feature Fusion framework (MFF) at the search and template sides, respectively. Then, the attribute-aware fusion module is introduced to conduct feature extraction and fusion for the major challenge attributes. In particular, we design a skip connections guidance module to prevent the propagation of noise and to enrich the feature information so that we can improve the tracker's discriminative ability for modality-specific challenges. The proposed SiamMFF method has been evaluated in a great number of trials on two benchmark datasets GTOT and RGBT234, and the precision rate and success rate can reach 90.5%/73.6% and 81.2%/57.3%, respectively, demonstrating the superiority of our method over existing state-of-the-art methods.

## CCS CONCEPTS

• Computing methodologies → Artificial intelligence.

## KEYWORDS

Object Tracking, RGBT, Multi-modal fusion, Siamese network

### ACM Reference Format:

Ke Qi, Liji Chen, Yicong Zhou, and Yutao Qi. 2023. Multi-Modal Fusion Object Tracking Based on Fully Convolutional Siamese Network. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590084>

## 1 INTRODUCTION

Visual object tracking based on RGB[1, 12, 16] is an important task in computer vision. However, it is difficult for RGB trackers to obtain precise tracking results in the presence of challenging problems such as low illumination, weather changes, motion blur, fast motion,

and object occlusion. Although there are ways to improve on a challenging factor[3], they may still fall short when faced with multiple challenging factors, and even greatly limit the application scope of visual tracking. Therefore, RGBT tracking [2, 7, 8, 18] has become a popular research direction in recent years.

RGBT tracking can take advantage of the complementing benefits of RGB and thermal infrared data to get better object-tracking results. Therefore, it can play a significant role in the fields of automatic driving, video surveillance, intelligent transportation and other fields. At present, most of the existing work based on RGBT is to study the fusion model[7, 18] to solve all challenges. The fusion of feature information method[7] for various challenging attributes such as illumination change and thermal cross can learn a target representation under a specific attribute with a small number of parameters. Although the fusion is innovative and has already achieved remarkable success in RGBT tracking, the utilization of modality-shared and modality-specific attribute information is still limited and like most RGBT trackers, it requires a lengthy processing period to fuse multi-modal features.

In this work, we propose a fully-convolutional Siamese-based Multi-modal Feature Fusion Network (SiamMFF) for fusing RGBT object tracking to enrich the feature information while fusing the modality-shared and modality-specific information. In particular, we design a multi-modal feature fusion module to integrate RGB and thermal features. Note that there are challenges with both shared and unique issues with RGB and RGBT data, and thus we design the multi-modal feature fusion module based on CAT[7] to address the challenges that are specific to each modality and the problems that they share. And finally, we design a skip connections guidance module to enrich the information of feature images and avoid the propagation of noisy information.

This paper's contributions are summarized below.

(1) We design a multi-modal feature fusion module to integrate visible and infrared features for RGBT object tracking. The module can extract the data that are specific to each modality and the information that they share with varying difficulties to fully use the complementing properties of visible and infrared images.

(2) We design a skip connections guidance module to prevent the propagation of noise and to enrich the feature information to improve the tracker's discriminative ability for modality-specific challenges.

\*Corresponding author.

(3) Extensive experiments have been conducted on the common RGBT datasets GTOT and RGBT234, and the precision and success rate can reach 90.5%/73.6% and 81.2%/57.3%, respectively, which show the effectiveness of our method in comparison to other state-of-the-art methods.

## 2 RELATED WORK

### 2.1 RGBT Tracking

With the advancement of deep learning, RGBT tracking has gradually become more widespread. It is a challenging task that is frequently influenced by factors such as thermal crossover, scale variation, and fast motion. While most of the current RGBT tracking is to manually extract attribute features from the model, EBT[20] designs an objective measurement based on edge features to generate high-quality object proposals to improve object detection accuracy and quickly locate the objects. The current three large-scale datasets, GTOT [5], RGT210 [9], and RGBT234 [6], are also released for RGBT tracking. Meanwhile, Li et al [8] use deep learning for RGBT tracking and propose a two-stream CNN and a fusion subnetwork to extract features from two modalities respectively. CAT[7] uses parallel and hierarchical challenge-aware branches to depict how an object appearance changes under specific challenges.

Although these approaches have obtained good performance on the above datasets, there is still a lot of feature information in RGBT that has not been mined, and they are unable to fully utilize feature information in multi-modality, limiting their ability to increase performance.

### 2.2 Attribute-aware fusion

The essence of attribute-aware fusion is to design five attributes including illumination variation (IV), thermal crossover (TC), scale variation (SV), occlusion (OCC) and fast motion (FM) according to the five main challenges of RGBT. FM, SV, and OCC are modality-shared, and IV and TC are modality-specific. Li et al [18] propose the challenge-aware RGBT tracker for a modality-shared challenge, learning object appearance representations under different challenges. And they propose a guidance module to transmit discriminative features between modalities, which can improve some weak modalities' ability to discriminate while keeping the computational complexity low. Zhang et al [19] propose an attribute-driven representation network (ADNet) with an attribute-driven residual branch to mine the attribute-specific characteristic and develop effective residual representations, which achieve good performance in both precisions and recall on RGBT datasets.

### 2.3 Fully-convolutional Siamese networks

The use of fully-convolutional networks allows online operability to stay at a fast speed. Bertinetto et al [1] propose SiamFC, using the fully-convolutional Siamese architecture for RGB tracking, in which the correlation of the two inputs is computed through bilinear layers to achieve a dense and efficient sliding window evaluation. Wang et al [16] propose SiamMask which improves the offline training procedure of fully-convolutional Siamese methods for object tracking. It adds a mask branch to the siamese network to obtain the most accurate box by directly predicting the mask of

the object. Then it uses a vector to encode a response of a candidate window (RoW) of masks and perform depth-wise convolution followed by cascading 1x1 convolutions to increase the dimension and achieve efficient operation. It also proposes a top-down refine module to enhance the precision of segmentation.

## 3 MULTI-MODAL FUSION OBJECT TRACKING

### 3.1 SiamMFF Network

As shown in Fig.1, we present a fully-convolutional Siamese-based Multi-Modal Feature Fusion Network (SiamMFF) for RGBT object tracking, and the primary component of SiamMFF is the module of multi-modal feature fusion(MFF), which comprises of five attribute-specific fusion, an adaptive aggregation layer and a skip connections guidance module. In specific, we adopt SiamMask network[16] as the backbone network to allow fast speed and operability. On each side of the input, we embed MFF module to extract modality-shared and modality-specific information in visible and infrared images which is crucial for tracking. First, we input visible and thermal images and crop them to generate search images and template images centered on the tracked target object. Then the MFF module extracts modality-specific features, and the skip connections guidance module guides the fusion of modality-specific information at the same time. Next, the attributes fused by five challenge branches and the convolutional features of two template images are concatenated. The classification results and regression results are then generated by deep cross-correlation of the template and search features, and a more accurate object mask is generated following the strategy of [14], using multiple upsampling layers and a skip-connected refinement module to merge low and high-resolution features. After generating the binary mask, the Box module automatically generates bounding boxes from it using the minimum bounding rectangle strategy.

### 3.2 Multi-modal Feature Fusion

In GTOT[5] and RGBT234[6] datasets, the five major challenge attributes contained in each video frame are manually annotated. We design an attribute-aware fusion module based on CAT[7] to better extract and fuse the modality-shared and modality-specific branches for the tracking task. Specifically, the MFF component removes the RoIAlign layer and the fully connected layer of CAT, and only uses the convolutional layers modified from VGG-M[2] for feature extraction. Among them, the three convolutional layers have a kernel size of 7×7, 5×5 and 3×3, respectively. We remove maximum pooling layers to retain more feature information in the next layer. In the second block with the padding as 2 to retain boundary information and maintain the image output size into the next layer of convolution, we modify the stride as 2 in the third block to 2 to boost the model's computational efficiency.

As shown in Fig. 2, two 255×255 pixel RGB and RBGT images are input to the MFF module. After the convolution operation of each stage, 96×125×125, 256×63×63 and 512×31×31 are output in turn, concatenating the two modalities' feature maps and output a feature map of 1024×31×31.

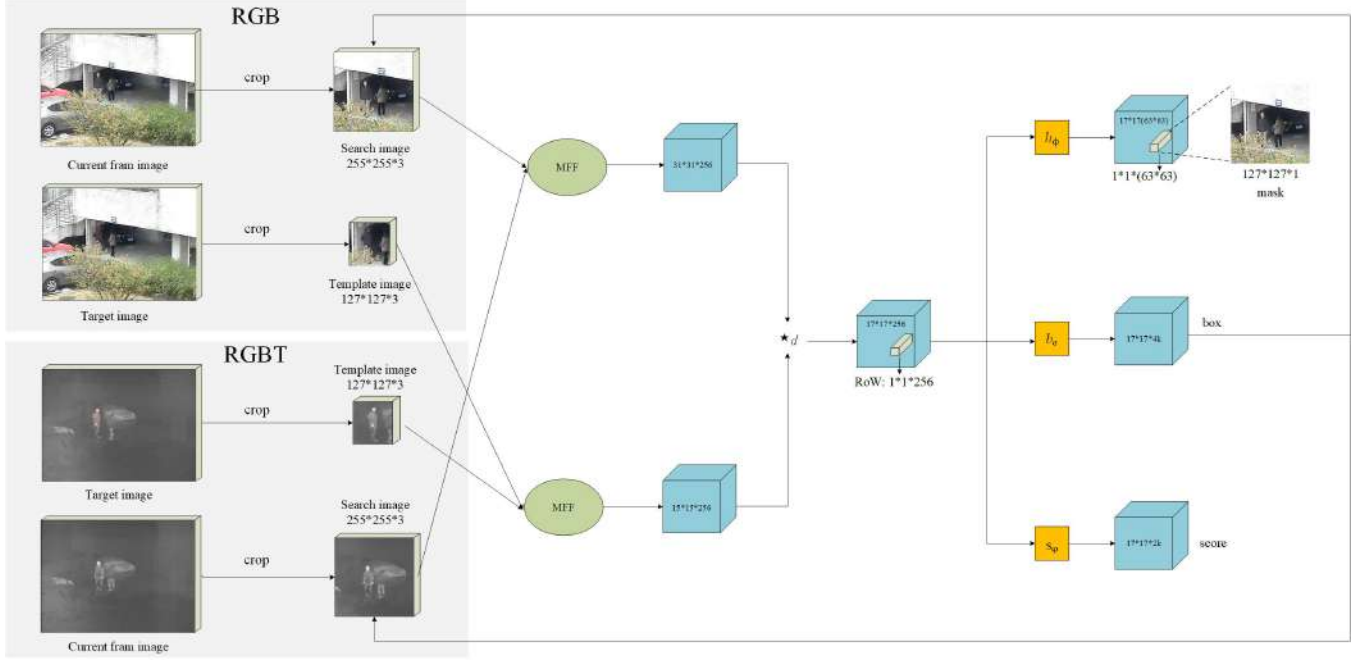


Figure 1: The structure of SiamMFF network framework. MFF represents a multi-modal feature fusion module. Mask module combines low and high resolution features through the use of several refinement modules composed of upsampling layers and skip connections. Box generation uses the minimum bounding rectangle strategy. Herein,  $\star d$  denotes depth-wise cross-correlation.

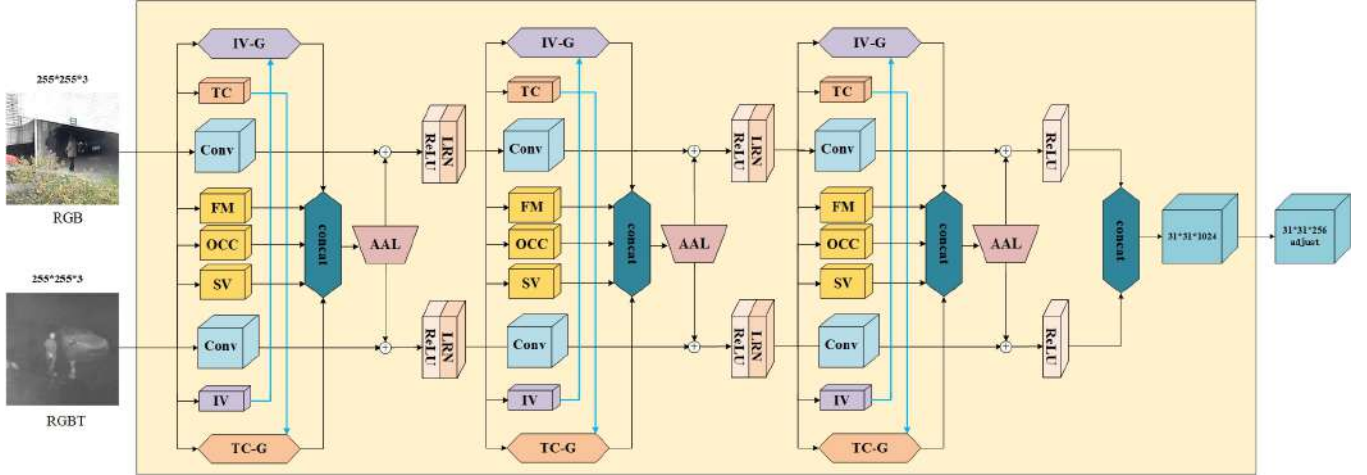


Figure 2: The structure of Multi-Modal Feature Fusion model (MFF). Herein,  $+$  represents an element-wise addition operation. AAL stands for adaptive aggregation layer. FM, OCC and SV are abbreviations for fast motion, occlusion, and scale variation. IV-G and TC-G are illumination variation and thermal crossover with the guidance module.

### 3.3 Skip connections guidance module

For TC and IV challenges, we use the skip connections guidance module to improve the discrimination of guided modality during tracing. As shown in Fig.3, we design the skip connections guidance module motivated by the guidance module on CAT[7]. Unlike only

using feature shift in CAT, we use skip connections, which are commonly used to enrich image details, such as U-Net[15], and the results of the studies support our design's efficacy. Additionally, in order to avoid noisy information, we introduce a gate mechanism. Specifically, we use a convolutional layer with  $1 \times 1$  kernel size and in order to learn a nonlinear mapping, we use a layer of nonlinear

activation. The gate operation is accomplished using element-wise sigmoid activation.

Here is the formulation of our skip connections module:

$$\begin{aligned}\alpha &= \omega_1 * \mathbf{x} + b_1 \\ \beta &= \omega_2 * \mathbf{z} + b_2 \\ \beta &= \alpha + \beta \\ \theta &= \omega_3 * \text{ReLU}(\beta) + b_3 \\ \tilde{\theta} &= \sigma(\theta) * \beta \\ \mathbf{z} &= \mathbf{z} * \tilde{\theta}\end{aligned}$$

The convolutional layer's weight and bias are denoted by  $\omega_i$  and  $b_i$ .  $\sigma$  denotes the sigmoid function. Point-by-point feature conversion without gate operation is represented by  $\alpha$  and  $\beta$ , and  $\tilde{\theta}$  represents point-by-point feature conversion with gate operation. The feature maps of the preceding and guided modalities are indicated, respectively, by  $\mathbf{x}$  and  $\mathbf{z}$ .

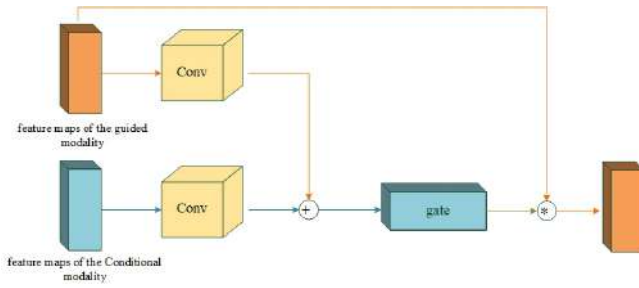


Figure 3: Skip connections Guidance modules.

## 4 EXPERIMENTS

### 4.1 Evaluation Data

We put our SiamMFF to the test on two RGBT datasets, i.e., GTOT[5] and RGBT234[6]. The GTOT dataset comes from the research group of Chenglong Li of Anhui University which consists of 50 statistically biased video sequences. The sequences of grayscale image-thermal infrared image pairs are divided into 7 subsets according to different properties. The RGBT234 dataset includes 234 RGBT video sequence pairs and their corresponding ground truth values. There are 234K frames in all. There are 12 attributes in the video sequence annotation, which are useful for evaluating the efficacy of various tracking algorithms for various challenging attributes.

### 4.2 Evaluation metrics

On the GTOT and RGBT234 datasets, our evaluation metrics use success rate (SR) and precision rate (PR) via one-pass evaluation. The fraction of successfully tracked frames with overlaps greater than thresholds is measured by SR. PR is the proportion of all frames in which the true distance between the tracking result's center point and the ground truth is smaller than the threshold. Since GTOT dataset contain mostly small objects, the threshold is set to 5 pixels on GTOT and 20 pixels on RGBT234.

### 4.3 Quantitative Comparison

We put our SiamMFF to the test on two benchmark datasets, GTOT and RGBT234. We train the challenge-aware branches in our multi-modal feature fusion module with corresponding challenge-based training data collected from the RGBT234 dataset by attribute labels in the GTOT dataset test. Then, we train MFF using the whole RGBT234 dataset. The training dataset for RGBT234 testing is GTOT, and the whole training procedure is similar to what we have described above.

To evaluate the usefulness of our SiamMFF, we run it through two RGBT datasets and compare its performance to that of numerous state-of-the-art trackers, such as ADRNet[19], MANet++[11], CAT[7], MANet[10], DAFNet[4], mFDiMP[18], DAPNet[21], SGT[9], FANet[22], MaCNet[17], RT-MDNet+RGBT [12] and MDNet[13]+RGBT.

Fig.4 depicts the comparison findings on GTOT and RGBT234 datasets. As shown in Fig.4(a), our SiamMFF surpasses CAT by 1.6%/1.9% in PR/SR which also use attribute-aware fusion. We also achieve comparable results when compared with the state-of-the-art approach ADRNet, where SiamMFF is 0.3% lower in SR but 0.1% higher in PR. And as shown in Fig.4(b), SiamMFF achieves the best tracking performance on RGBT234 dataset. Compared with ADRNet, which is the top advanced tracker in RGBT234, our PR/SR is 0.5%/0.3% higher than it and when compared with CAT, we advance them by 0.8%/1.2% in PR/SR. Furthermore, we outperform MDNet+RGBT in PR/SR by more than 9%/7.8%. These findings totally support the efficacy of our strategy.

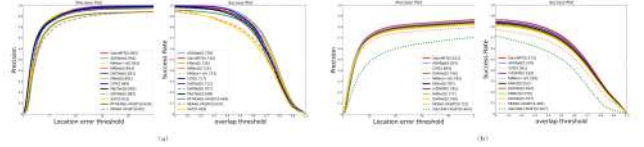


Figure 4: The assessment curve on GTOT and RGBT234 datasets. The legend displays the PR/SR representative scores.(a)GTOT, (b)RGBT234

### 4.4 Ablation Study

We conduct an ablation study on GTOT and RGBT234 to validate the efficiency of the major components of our approach. Aim to test the effectiveness of the proposed network, we implement the two special variants of our method described below: 1. SiamMFF-AAF, all attribute-aware branches are combined by elements addition, deleting the structure of the Attribute-aware fusion module, in order to test the efficacy of the proposed modal feature fusion approach. 2. APFNet-SCGM, the special challenge branches are fused by elements addition, removing the skip connections guidance modules to verify the effectiveness of the proposed skip connections guidance module. Table1 shows the experimental findings on the GTOT and RGBT234 datasets. On the GTOT and RGBT234 datasets, we take SiamMFF without the skip connections guidance module as the baseline and compare CAT's guidance module with our skip connections module on our SiamMFF model to validate the proposed method's efficacy. As shown in Table 2, we can see our

skip connections guidance module outperforms CAT on the two datasets.

**Table 1: The PR/SR scores of various versions induced by our SiamMFF on two RGBT benchmark datasets are used to verify the usefulness of our model.**

		SiamMFF-AAF	SiamMFF-SCGM	SiamMFF
GTOT	PR	0.853	0.861	<b>0.905</b>
	SR	0.685	0.720	<b>0.736</b>
RGBT234	PR	0.755	0.793	<b>0.812</b>
	SR	0.501	0.561	<b>0.573</b>

**Table 2: Compare the performance of the skip connections guidance module and CAT’s guidance module on SiamMFF on two RGBT benchmark datasets.**

		Baseline	SiamMFF-CAT	SiamMFF
GTOT	PR	0.861	0.884	<b>0.905</b>
	SR	0.720	0.725	<b>0.736</b>
RGBT234	PR	0.793	0.798	<b>0.812</b>
	SR	0.561	0.567	<b>0.573</b>

## 4.5 Qualitative Results

Fig. 5 displays the qualitative results of 7 different trackers on two sequences under different challenging circumstances, such as scale variation(e.g. Otcvbs) and thermal crossover(e.g. Motorbike). We can see that our SiamMFF is more robust under these challenging situations and gets better results under the qualitative comparison of bounding boxes, which demonstrates the effectiveness of our method.



**Figure 5: Qualitative results of 7 trackers on two sequences in the GTOT dataset. (a) Otcvbs, (b) Motorbike.**

## 5 CONCLUSION

In this paper, we propose a fully-convolutional Siamese-based Multi-Modal Feature Fusion Network (SiamMFF) to fully integrate multi-modal feature information and introduce the attribute-aware module into the modified fully-convolution Siamese network SiamMask, to enhance the fusion at the attribute feature level. We also design a skip connections guidance module to enrich feature image information. The proposed algorithm performs well when evaluated on commonly used benchmark datasets and compared with mainstream algorithms. The experimental results validate the proposed

method’s effectiveness and feasibility. In further exploration, we plan to try more fusion structures under challenging circumstances such as motion blur and size change so that the model can fully fuse the information between multi-modalities. We will also further consider adding visual reasoning of optical flow direction under the presumption of guaranteeing the model runs light.

## REFERENCES

- [1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*. Springer, 850–865.
- [2] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531* (2014).
- [3] Siyu Di and Wensheng Sun. 2022. Research on Low Illumination Image Processing Algorithm Based on Adaptive Parameter Homomorphic Filtering. In *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*. 681–685. <https://doi.org/10.1109/CACML55074.2022.00118>
- [4] Yuan Gao, Chenglong Li, Yabin Zhu, Jin Tang, Tao He, and Futian Wang. 2019. Deep adaptive fusion network for high performance RGBT tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.
- [5] Chenglong Li, Hui Cheng, Shiyi Hu, Xiaobai Liu, Jin Tang, and Liang Lin. 2016. Learning collaborative sparse representation for grayscale-thermal tracking. *IEEE Transactions on Image Processing* 25, 12 (2016), 5743–5756.
- [6] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2019. RGB-T object tracking: Benchmark and baseline. *Pattern Recognition* 96 (2019), 106977.
- [7] Chenglong Li, Lei Liu, Andong Lu, Qing Ji, and Jin Tang. 2020. Challenge-aware RGBT tracking. In *European Conference on Computer Vision*. Springer, 222–237.
- [8] Chenglong Li, Xiaohao Wu, Nan Zhao, Xiaochun Cao, and Jin Tang. 2018. Fusing two-stream convolutional neural networks for RGB-T object tracking. *Neurocomputing* 281 (2018), 78–85.
- [9] Chenglong Li, Nan Zhao, Yijuan Lu, Chengli Zhu, and Jin Tang. 2017. Weighted sparse representation regularized graph learning for RGB-T object tracking. In *Proceedings of the 25th ACM international conference on Multimedia*. 1856–1864.
- [10] Cheng Long Li, Andong Lu, Ai Hua Zheng, Zhengzheng Tu, and Jin Tang. 2019. Multi-adapter RGBT tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.
- [11] Andong Lu, Chenglong Li, Yuqing Yan, Jin Tang, and Bin Luo. 2021. RGBT tracking via multi-adapter network with hierarchical divergence loss. *IEEE Transactions on Image Processing* 30 (2021), 5613–5625.
- [12] Yuwei Lu, Yuan Yuan, and Qi Wang. 2020. A dense connection based network for real-time object tracking. *Neurocomputing* 410 (2020), 229–236.
- [13] Hyeonseob Nam and Bohyung Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4293–4302.
- [14] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. 2016. Learning to refine object segments. In *European conference on computer vision*. Springer, 75–91.
- [15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [16] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 1328–1338.
- [17] Hui Zhang, Lei Zhang, Li Zhuo, and Jing Zhang. 2020. Object tracking in RGB-T videos using modal-aware attention network and competitive learning. *Sensors* 20, 2 (2020), 393.
- [18] Lichao Zhang, Martin Danelljan, Abel Gonzalez-Garcia, Joost van de Weijer, and Fahad Shahbaz Khan. 2019. Multi-modal fusion for end-to-end rgb-t tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.
- [19] Pengyu Zhang, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 2021. Learning adaptive attribute-driven representation for real-time rgb-t tracking. *International Journal of Computer Vision* 129, 9 (2021), 2714–2729.
- [20] Gao Zhu, Fatih Porikli, and Hongdong Li. 2016. Beyond local search: Tracking objects everywhere with instance-specific proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 943–951.
- [21] Yabin Zhu, Chenglong Li, Bin Luo, Jin Tang, and Xiao Wang. 2019. Dense feature aggregation and pruning for rgbt tracking. In *Proceedings of the 27th ACM International Conference on Multimedia*. 465–472.
- [22] Yabin Zhu, Chenglong Li, Jin Tang, and Bin Luo. 2020. Quality-aware feature aggregation network for robust RGBT tracking. *IEEE Transactions on Intelligent Vehicles* 6, 1 (2020), 121–130.

# An Ensemble Model using Face and Pose Tracking for Engagement Detection in Game-based Rehabilitation

Xujie Lin

Shien-Ming Wu School of Intelligent Engineering, South  
China University of Technology, Guangzhou, China  
lxj1435359352@gmail.com

Patrick P. K. Chan

Shien-Ming Wu School of Intelligent Engineering, South  
China University of Technology, Guangzhou, China  
patrickchan@ieee.org

Siqi Cai

Department of Electrical and Computer Engineering,  
National University of Singapore, Singapore  
elesiqi@nus.edu.sg

Longhan Xie\*

Shien-Ming Wu School of Intelligent Engineering, South  
China University of Technology, Guangzhou, China  
melhxie@scut.edu.cn

## ABSTRACT

Highly engaging rehabilitation promotes functional reorganization of the brain in stroke patients. Engagement detection in game-based rehabilitation can help rehabilitation practitioners get real-time feedback, and then provide patients with appropriate training programs. Previous research on engagement detection has focused on wearable devices, and the complicated laboratory setup makes them unsuitable for use in clinics and homes. In this work, we propose a method to automatically extract facial and posture features from camera-captured videos. Then we design an automatic engagement detection model using the facial and posture features as the input. In the dataset of engagement in virtual game rehabilitation scenarios, our model detects engagement levels with an average accuracy of 96.85%, achieving remarkable performance. This study sheds new light on engagement detection for stroke patients in clinical applications.

## CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **HCI design and evaluation methods**; • **User studies**;

## KEYWORDS

Engagement detection, Stroke rehabilitation, Face and pose tracking, Ensemble model

### ACM Reference Format:

Xujie Lin, Siqi Cai, Patrick P. K. Chan, and Longhan Xie. 2023. An Ensemble Model using Face and Pose Tracking for Engagement Detection in Game-based Rehabilitation. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590085>

\*corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590085>

## 1 INTRODUCTION

Rehabilitation engagement is defined as the state of being motivated and making an active effort to engage in rehabilitation training [1]. It can reflect the patient's attitude towards rehabilitation, understanding of task requirements, and the effectiveness of the entire training process. Previous studies have shown that the high-level engagement of patients is an important factor in promoting nerve reorganization [2]. Even if patients do not have the ability to complete movements, the willingness to active exercise is also necessary for rehabilitation [3]. However, the repetitive nature of rehabilitation exercises often makes patients feel bored and fatigue easily. In fact, patients usually show low-level engagement behaviors during rehabilitation training, which leads to inefficient rehabilitation performance and safety hazards. In traditional rehabilitation training, the supervision and correction of therapists can reduce the impact of patients' low-level engagement. However, there is a shortage of rehabilitation physicians, and patients often perform training tasks without supervision and correction. Therefore, it is important to assess the patient's engagement in the rehabilitation training, which facilitates the assessment of rehabilitation outcomes and the adjustment of training tasks.

This paper presents a vision-based approach for engagement detection for stroke patients using facial and posture features. We first construct a dataset on engagement detection for stroke patients in virtual game rehabilitation scenarios. A sequence of upper body postures and facial images are then automatically extracted from videos captured by cameras. Finally, we propose a face and pose hybrid network, refer to as FPNnet, for learning spatio-temporal information from facial and posture features to detect engagement levels.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of related work on engagement detection. In Section 3, we describe the proposed detection method, including posture tracking method and engagement detection framework. In Section 4, we evaluate the performance of the proposed detection method through extensive experiments. Finally, Section 5 concludes this paper.

## 2 RELATED WORK

Many approaches have been developed to evaluate patients' engagement, mainly including rating-scale based and physiological signals-based methods. The evaluation method using scales requires

rehabilitation physicians to observe and score according to the engagement indicators of patients in the rehabilitation training process. This method inevitably introduces the subjective judgment of physicians and increases the workload of rehabilitation physicians [4]. Physiological signals generated during rehabilitation training are also often used to assess patient engagement. Zimmerli et al. [5] used electromyography (EMG) to describe the state of motor engagement and applied it to the assessment of participation in gait rehabilitation training in patients with spinal cord injury. Li et al. [6] used electroencephalogram (EEG) to characterize the state of cognitive engagement and applied it to assess engagement in upper limb rehabilitation training in stroke patients. However, the acquisition of these physiological signals relies on wearable sensors, which are inconvenient and likely cause discomfort to patients during rehabilitation exercises.

Low engagement behaviors of stroke patients during virtual game rehabilitation training are often accompanied by many facial and posture behaviors [7]. Inspired by this finding, we would like to use vision technology to capture changes in facial behavior, such as eyelid movement, pupil movement, degree of eye openness, head posture, and facial fatigue expression, to intuitively detect the patient's subjective engagement. Compared with physiological signal-based methods, vision-based methods can be assessed in a non-invasive manner and will not affect the patient's rehabilitation training process.

Recently, vision-based methods have been used to detect behavioral engagement in healthy individuals. Geng et al. [8] built an end-to-end video classifier using C3D [9] and implemented it in detecting student engagement levels. Huynh et al. [10] used the OpenFace to extract Gaze-AU-Pose (GAP) features and built an LSTM network [11] to classify student engagement. Nezami et al. [12] proposed a method to detect student engagement in a small-sample dataset using a pre-trained CNN model. Moreover, Schulc et al. [13] proposed a CNN-LSTM method to detect user engagement in the process of watching advertisements based on a large amount of video data. However, to the best of our knowledge, there is no related study applying vision technology to the engagement detection for stroke patients.

### 3 METHODS

The proposed FPnet includes pose feature points localization, face feature extraction, data preprocessing, and model building for classifying the engagement of stroke patients in rehabilitation training, as shown in Figure 1.

#### 3.1 Posture tracking method

Our engagement detection system needs to obtain facial information and upper body pose information from image sequences. First, we use OpenPose [14] to extract upper body feature points in the image sequence, and select the left shoulder point, right shoulder point, torso vertex, nose tip point and two eye positioning points as the pose feature points. Then, with the nose tip point as the center, we cut out the facial region from the image sequence according to the size pre-corrected using a Multi-Task Convolutional Neural Network (MTCNN) [15]. Considering that when the patient's head is turned close to 90 degrees or the face is occluded, MTCNN often

has a low success rate of face detection due to insufficient information. Therefore, we did not directly use MTCNN to participate in the whole process of face detection, and the face region extraction method centered on the nose tip can ensure the robustness of our system to extreme angles and occlusions.

Each face image is first resized to  $224 \times 224$ ; we then subtract the coordinate values of the torso vertices from the coordinate values of each pose feature point to obtain the relative coordinates. The relative xy-coordinates of the pose feature points are connected as the pose feature vector and normalized. The specific processing method is as follows:

$$\begin{aligned}\mu &= \frac{\sum_{k=1}^N P_k}{N} \\ \sigma &= \sqrt{\frac{1}{N} \sum_{k=1}^N (P_k - \mu)^2} \\ P'_k &= \frac{P_k - \mu}{\sigma + C}\end{aligned}$$

where  $\mu$  and  $\sigma$  are the mean and standard deviation of the data, respectively;  $P_k$  are the original input vectors; and  $P'_k$  are the normalized input vectors.

#### 3.2 Engagement detection framework

According to the engagement characteristics of subjects during game training, the main task of the model is to detect three levels of engagement: low engagement (LE), medium engagement (ME) and high engagement (HE). As shown in Figure 2, the performance of LE is that the subject blinks frequently or closes his/her eyes for a long time; the performance of ME is that the subject turns his/her head and does not focus on the screen; the performance of HE is that the subject looks ahead and actively participates in the game training.

In our FPnet model, we use DenseNet [16] to extract high-level facial features from face image sequence, and then use TCN [17] to fuse temporal information from facial and posture features and detect engagement levels. Facial behavior features can intuitively reflect the subject's subjective engagement. In our engagement detection system, in order to obtain features that are more robust to adverse factors such as extreme angles and occlusions, we use a DenseNet pretrained on the ImageNet-1K [18] dataset and fine-tuned on our engagement dataset. DenseNet achieves feature reuse by using a dense connection mechanism to connect feature maps of different layers, which significantly improves the feature representation ability of the model. For each face frame, 128-dimensional features from the last pooling layer of DenseNet are used to represent the facial spatial features. The facial spatial feature sequence extracted from an image sequence of  $N$  frames can be expressed as:

$$F = \{f_1, f_2, \dots, f_N\}, f_n \in R^D$$

where  $f_n$  represents the facial spatial features of the  $n$ -th frame, and  $D$  is the feature dimension.

For engagement behaviors, facial behavioral features and posture features are difficult to separate in image sequences of several frames, such as eye closure versus blinking, and head-turning versus head-tilting. Therefore, in order to obtain sufficient distinguishing

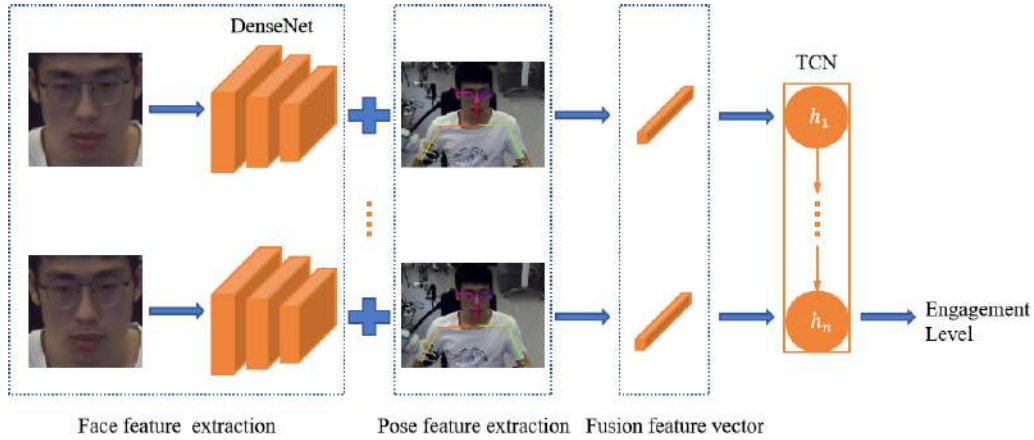


Figure 1: The proposed engagement detection framework.

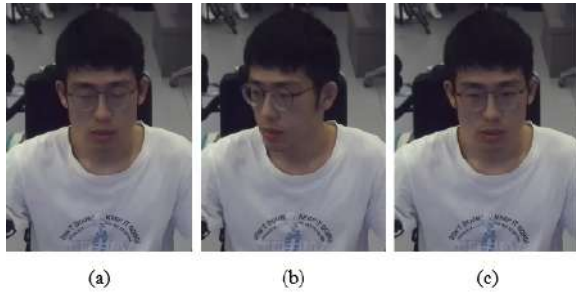


Figure 2: (a) Low engagement (LE): blinks frequently or closes eyes for a long time. (b) Medium engagement (LE): turns head and does not focus on the screen. (c) High engagement (HE): looks ahead and actively participates in the game training.

features, we choose an image sequence length of 2s, sampling at a frequency of 15 frames per second.

Temporal features between consecutive image sequences are crucial information for accurately detecting engagement. TCN is the most common architecture used to solve timing problems. Due to the ingenious design of causal convolution, dilated convolution, and residual connection, TCN can efficiently capture the contextual information of time series. TCN has excellent classification performance in tasks such as video classification and emotion recognition. Compared with general time series architectures such as LSTM [11], TCN performs better in extracting longer time series information. Therefore, we use TCN as a temporal information extraction model for engagement.

We concatenate the facial and posture features to form a fused feature vector of each image frame. The fused feature vectors of an image sequence of  $N$  frames are used as the input of the time step of TCN to extract temporal information. To detect patient’s engagement from the input video, the output of the last time step of TCN is passed into a fully connected layer and a softmax function.

## 4 EXPERIMENTS

In this section, we describe the dataset for engagement detection, the validation strategy, and the detection performance of our model.

### 4.1 Dataset

We recruited 12 patients with stroke ( $51.6 \pm 10.4$  years old, 7 males, 5 females) from the Third Affiliated Hospital at Sun Yat-sen University to complete the experiment. All patients gave written informed consent and met the following selection criteria: 1) good cognitive and comprehension skills, 2) able to sit alone in a chair, and 3) able to actively complete virtual game rehabilitation tasks.

The subjects with stroke performed 20 minutes of virtual game rehabilitation training under the supervision of a rehabilitation therapist, and a 6-degree-of-freedom UR5 robotic arm was used to interact with the whack-a-mole game. A high-definition camera (with a resolution of  $1920 \times 1080$ , a sampling rate of 30fps, and a focal length of 3.6mm) was placed directly in front of the subject to collect videos that obtain engagement behavior. Each subject’s engagement behavior, including head rotation, facial expressions, and sleepiness due to fatigue, was spontaneous.

We collected engagement data in game training from 20 healthy subjects ( $24.1 \pm 4.2$  years old, 15 males, 5 females) to expand the training dataset. To keep the engagement samples of healthy and patients similar, we design three scenarios for the experiment of healthy subjects: high engagement, attention disturbance and sleep deprivation. The rest of the experimental methods are the same as those in the patient dataset.

For each video clip of patients and healthy subjects, the engagement labels are annotated by 5 rehabilitation therapists, and the Dawid-Skene vote aggregation strategy is adopted to get the final labels [19]. Finally, we obtain 2537 patient samples and 5080 healthy person samples as shown in Table 1.

It is necessary to have a sufficient amount of training data covering various situations to train our engagement detection model and accurately classify samples without overfitting. In this paper, we perform dataset augmentation using horizontal picture flips and rotations of 5 and -5 degrees, which can make the model robust to

**Table 1: Information on datasets**

Engagement level	LE	ME	HE	Total
Stroke subjects	397	282	1858	2537
Healthy subjects	814	572	3694	5080

slight in-plane head rotations of subjects. Finally, we obtain 6 times more data.

## 4.2 Validation Strategy

Our experiments employ a patient-based leave-one-subject-out (LOSO) cross-validation [20] strategy to evaluate classification performance. With LOSO cross-validation, the data of one patient is selected as the test dataset each time, and the data of the remaining patients and healthy people form the train dataset. The data from each patient should be used as a test dataset, and the average detection rate is used to evaluate the model’s performance.

To graphically present information about the sample distribution of our model’s actual and predicted classes, we utilize a confusion matrix. True positive (TP), true negative (TN), false positive (FP), and false negative (FN) can reflect the actual classification effect of the classifier from multiple perspectives. For the requirements of different scenarios on the model, the precision rate can be used to describe the proportion of the actual positive samples to all the predicted positive samples, and the recall rate can be used to describe the degree to which the target object is detected without being missed. The  $F_\beta$  score can comprehensively evaluate the two scores of precision and recall according to their importance [21]. In our engagement detection task, our purpose is to detect low-engagement behaviors of subjects, and the proportion of low-engagement samples is small, so we believe that the recall rate is more important than the precision rate. We use the  $F_2$  score to evaluate the classification performance of the model. Precision, recall and score can be calculated with the following formulas:

$$precision = TP / (TP + FP)$$

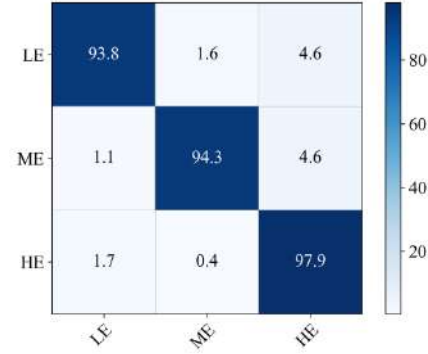
$$recall = TP / (TP + FN)$$

$$F_2 = (5 * precision * recall) / (4 * precision + recall)$$

## 4.3 Detection Performance

The confusion matrix shows the overall distribution of true and predicted values in our test results. As shown in Figure 3, the prediction results of most samples are concentrated on the diagonal position, and each category has achieved a high accuracy rate. The average accuracy of engagement detection was 96.85%, with HE achieving the highest accuracy (97.88%), followed by ME (94.33%) and LE (93.79%).

As can be seen in Table 2, our classifier achieves very high recall and scores in each category. This is mainly because our method combining facial features and head poses can well distinguish features between categories, and LE and ME samples can be better distinguished. For LE and ME, ME performs better in detection, while LE has low precision and recall. We think this is mainly because some samples of LE and HE are not clearly distinguishable in terms of facial features and posture features. We believe that

**Figure 3: The confusion matrix of the proposed model.**

by increasing the resolution and sampling rate of the camera, the detection accuracy of LE can be effectively improved.

We compare the performance of the proposed method with two other methods: DensNet+LSTM, C3D [9]. As can be seen in Table 3, our FPnet achieves the best classification performance (96.85%). The experimental results show that the proposed method is superior to the other two methods in extracting spatio-temporal information of engagement.

## 5 CONCLUSION

In this paper, we proposed a reliable method for detecting engagement in game rehabilitation training based on facial features and head poses, which can be used to help rehabilitation physicians obtain timely feedback and provide patients with appropriate training prescriptions. On our collected engagement dataset of game rehabilitation training for stroke patients, the average accuracy of the proposed method is 96.85%. Experimental results show that facial features and head poses extracted from image sequences can complement each other to correctly reflect the patient’s engagement level. This fast, stable, and low-cost method is expected to realize automatic detection of engagement in game rehabilitation training scenarios, complementing the engagement detection method of wearable devices, and helping patients complete rehabilitation training more safely and reliably.

However, the limitation of the proposed method is that the position of the camera that captures the image sequence is relatively fixed, and the camera position and angle will affect the accuracy of engagement detection. Methods that acquire image sequences from multiple angles and combine other useful information (such as pupil motion) with the features extracted in this paper may yield better performance. This will be studied in our future work.

**Table 2: Information on subjects**

Engagement level	Precision	Recall	F <sub>2</sub> -score
LE	91.41	93.78	93.30
ME	95.06	94.32	94.47
HE	98.32	97.88	97.97
Average	96.85	96.85	96.85

**Table 3: Comparison the results using the proposed algorithm and other algorithms**

Model	LE	ME	HE	Average
FPnet	93.78	94.32	97.88	96.85
DensNet+LSTM	91.32	91.52	95.87	94.68
C3D	89.34	86.15	90.74	90.01

## ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China (Grant No. 52075177), Joint Fund of the Ministry of Education for Equipment Pre-Research (Grant No. 6141A02033124), Guangzhou Research Foundation (Grant No. 202002030324 and 201903010028), Zhongshan Research Foundation (Grant No.2020B2020), and Shenzhen Institute of Artificial Intelligence and Robotics for Society (Grant No. AC01202005011).

## REFERENCES

- [1] Lequerica, Anthony H., and Kathleen Kortte. "Therapeutic engagement: a proposed model of engagement in medical rehabilitation." *American journal of physical medicine & rehabilitation* 89.5 (2010): 415-422.
- [2] Langhorne, Peter, Julie Bernhardt, and Gert Kwakkel. "Stroke rehabilitation." *The Lancet* 377.9778 (2011): 1693-1702.
- [3] Kwakkel, Gert, Boudewijn J. Kollen, and Hermano I. Krebs. "Effects of robot-assisted therapy on upper limb recovery after stroke: a systematic review." *Neurorehabilitation and neural repair* 22.2 (2008): 111-121.
- [4] Kortte, Kathleen B., *et al.* "The Hopkins rehabilitation engagement rating scale: development and psychometric properties." *Archives of physical medicine and rehabilitation* 88.7 (2007): 877-884.
- [5] Zimmerli, Lukas, *et al.* "Increasing patient engagement during virtual reality-based motor rehabilitation." *Archives of physical medicine and rehabilitation* 94.9 (2013): 1737-1746.
- [6] Li, Chong, *et al.* "Development of engagement evaluation method and learning mechanism in an engagement enhancing rehabilitation system." *Engineering Applications of Artificial Intelligence* 51 (2016): 182-190.
- [7] Li, Chong, *et al.* "Implementation and validation of engagement monitoring in an engagement enhancing rehabilitation system." *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.6 (2016): 726-738.
- [8] Geng, Lin, *et al.* "Learning deep spatiotemporal feature for engagement recognition of online courses." 2019 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2019.
- [9] Tran, Du, *et al.* "Learning spatiotemporal features with 3d convolutional networks." *Proceedings of the IEEE international conference on computer vision*. 2015.
- [10] Thong Huynh, Van, *et al.* "Engagement intensity prediction with facial behavior features." 2019 International Conference on Multimodal Interaction. 2019.
- [11] Graves, Alex. "Long short-term memory." *Supervised sequence labelling with recurrent neural networks* (2012): 37-45.
- [12] Mohamad Nezami, Omid, *et al.* "Automatic recognition of student engagement using deep learning and facial expression." *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Cham, 2020.
- [13] Schulc, Attila, *et al.* "Automatic measurement of visual attention to video content using deep learning." 2019 16th International Conference on Machine Vision Applications (MVA). IEEE, 2019.
- [14] Cao, Zhe, *et al.* "Realtime multi-person 2d pose estimation using part affinity fields." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [15] Zhang, Kaipeng, *et al.* "Joint face detection and alignment using multitask cascaded convolutional networks." *IEEE signal processing letters* 23.10 (2016): 1499-1503.
- [16] Huang, Gao, *et al.* "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [17] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." *arXiv preprint arXiv:1803.01271* (2018).
- [18] Russakovsky, Olga, *et al.* "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115 (2015): 211-252.
- [19] Dawid, Alexander Philip, and Allan M. Skene. "Maximum likelihood estimation of observer error-rates using the EM algorithm." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28.1 (1979): 20-28.
- [20] Cai, Siqu, *et al.* "Real-time detection of compensatory patterns in patients with stroke to reduce compensation during robotic rehabilitation therapy." *IEEE Journal of Biomedical and Health Informatics* 24.9 (2020): 2630-2638.
- [21] Cai, Siqu, *et al.* "Online compensation detecting for real-time reduction of compensatory motions during reaching: a pilot study with stroke survivors." *Journal of neuroengineering and rehabilitation* 17.1 (2020): 1-11.

# A U-Net based Self-Supervised Image Generation Model Applying PCA using Small Datasets

Sang Hun Han

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
shhan@vim.cau.ac.kr

Asim Niaz

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
asim@vim.cau.ac.kr

Kwang Nam Choi\*

Department of Computer Science &  
Engineering, Chung-Ang University,  
Seoul (06974), South Korea  
knchoi@cau.ac.kr

## ABSTRACT

Generative Adversarial Networks (GAN) is a research-based on deep learning technology that synthetically generates, combines, and transforms images similar to the original images. The main focus of GAN existing work has been to improve the quality of generated images and to generate high-resolution images by changing the training scheme or devising more complex models. However, these models require a large amount of data and are not suitable for training with a small amount of data. To address these challenges, this paper aims to improve the quality of images and the stability of training with a small dataset by proposing a novel training method for generating real-world images by using PCA and Self-Supervised GAN. Previously, PCA was applied to DCGAN to generate images with a small dataset, but some images showed poor results. By preparing quantitatively different datasets, we show that the quality of generated image with a small dataset is equivalent, or even better when compared to the quality of the image generated with a large dataset.

## CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Computer vision; Computer vision problems; Reconstruction.

## KEYWORDS

Generative Adversarial Network, Principal Component Analysis, Self-Supervised Learning, U-Net

### ACM Reference Format:

Sang Hun Han, Asim Niaz, and Kwang Nam Choi. 2023. A U-Net based Self-Supervised Image Generation Model Applying PCA using Small Datasets. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590086>

\*Corresponding author: Kwang Nam Choi (knchoi@cau.ac.kr)

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590086>

## 1 INTRODUCTION

The need for more data is the most significant restriction in introducing artificial intelligence to practical usage. A large amount of refined and densely annotated data is required to train an artificial intelligence model. Training with a small dataset takes less time, but the prediction accuracy quality of results is compromised. The main focus of existing work in the GAN [1] domain has been to improve result image quality by changing the training scheme or devising more complex models. However, these models require large amounts of data and are unsuitable for training with small datasets. This research attempts to solve this problem by applying Principal Component Analysis (PCA) to DCGAN [2].

Our approach focuses on the Generator part of GAN, responsible for producing the image. Images are generated from the image generation model as input from the convolutional neural network by synthesizing the Latent Space. That is, the Latent Space and the Convolutional Neural Network determine the main features of the image. In this paper, we devise a method to improve the quality of images even with a small dataset by modifying Latent Space and Convolutional Neural Network that majorly contributes to image generation.

Our work has the following three-fold contributions:

- As confirmed in the experiments section, the average error rate of the generator reduces compared to the existing model.
- The proposed model produces quality images, despite being trained on a small dataset compared to the previous model [3], which is trained on large datasets.
- The proposed method sets the future research direction for image data generation models using smaller datasets.

## 2 RELATED WORK

### 2.1 PCA (Principal Component Analysis)

In machine learning, PCA [4] is one of the dimensional reduction methods of projecting high-dimensional space data into low-dimensional space to find the primary component of the data. In machine learning, the number of dimensions also increases as the features of the dataset increase. PCA technique is mainly applied to high-dimensional data such as face datasets and is used in face recognition algorithms. Furthermore, previous work has confirmed that PCA can be applied to GAN to improve the quality of the image with a small dataset.

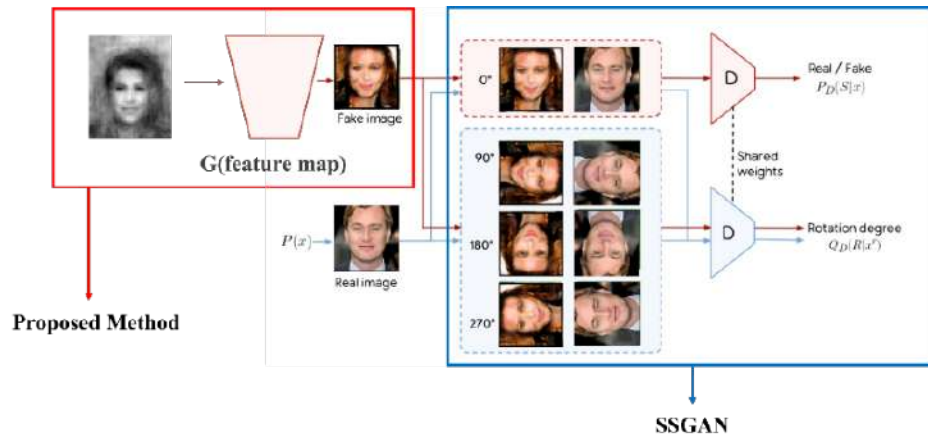


Figure 1: Proposed Method Architecture

## 2.2 SS-GAN (Self-Supervised Generative Adversarial Network)

Self-supervised [5] learning is a technology that enables deep learning models to understand the dataset by learning with unlabeled data. And it's a technology that allows you to learn with less data. GAN studies using self-supervised learning such as Match GAN[6] and OP-GAN[7] are also being conducted, and these studies improve performance using image quality and unpurified datasets. This paper adopted the SSGAN [8] model as its backbone model. SSGAN is a model that applies self-supervised learning to GAN and applies Image Rotation, one of the Self-Supervision methods, to GAN.

## 2.3 U-Net

U-Net [9] is an image segmentation model proposed originally for biomedical images, which can obtain overall contextual information about the image. U-Net demonstrates superior performance despite being trained on small data sets. Through this study, it is possible to generate an image without losing the characteristics of the image in the Expansive Path stage. Taking advantage of these points, they propose a new training scheme of the generator through the structure of U-Net such as U-Net GAN [10] and CycleGAN [11], and we can generate new images.

# 3 PROPOSED METHODS

## 3.1 Feature Extractor

In this paper, a feature map is extracted by analyzing the principal components of the face image. We obtain 2,000 images from CelebA [12] dataset and analyze the principal components of the data to output an average face (Mean Face) for 2,000 images. The average face image size is output as  $[64, 64, 1]$  to match the structure of the proposed model. The characteristics in which the face image is expressed are different according to the main component  $k$ . As a result of representing a person's face using one to 800 principal components in a previous study, the higher the number of principal components, the more detailed personal characteristics and noise

appear. In this case, it is said that if an image is created, it can be learned to create only personal features. When the number of principal components is 1, only the approximate features of the human face are shown, so one is used as the number of principal components to generate various images..

## 3.2 Model

Instead of randomly generating a noise image in the latent space using a feature map of a face image output through PCA, an image was generated from a feature map of a face image. The Latent Space of the existing model generated images from vectors  $[1, 1, 100]$ . However, the proposed method changed the structure of the generator model because it generates images from feature maps of facial images  $[64, 64, 1]$ . Figure 1 shows the overall structure of the proposed method for applying PCA to SSGAN.

The difference from the existing SSGAN model is that it does not enter the generator as an input from the noise vector  $z$  to generate an image, but rather that the feature map obtained through PCA becomes the input of the generator model. However, since the existing SSGAN generator model structure is DCGAN, the network consists of five layers of the Transpose Convolution Layer, starting with Noise Vector  $z$   $[1, 1, 100]$ . The proposed model's feature Map  $[64, 64, 1]$  could not be received as input. In addition, we implemented a model consisting of a transpose convolution layer that can receive a feature map  $[64, 64, 1]$  as an input. However, the size of the model increased, and the number of parameters increased, resulting in many artifacts. To address this problem, in this paper, U-Net is applied to the structure of the generator model. The generator structure of the proposed method is shown in Figure 2.

The generator structure is similar to that of U-Net, and the model was modified to output the resulting images  $[64, 64, 3]$  by receiving the Feature Map  $[64, 64, 1]$  as input. First, the feature map  $[64, 64, 1]$  was received as input from the extracting path, and the overall information of the image was obtained by reducing the size of the image through the convolutional neural network and the max

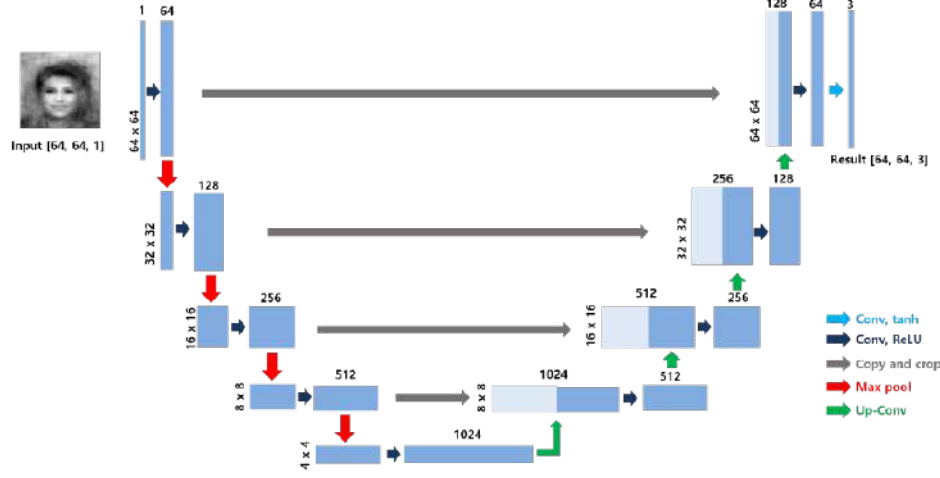


Figure 2: Generator Structure of Proposed Method

pooling layer. The Transpose Convolution Layer repeats the Up-Sampling process several times, halving the number of channels applied each time an Up-Sampling is performed. If you look at the gray arrow in Figure 2, it is combined with the Feature Map on the same layer in the contraction path on the opposite side. Finally, by iterating through the convolutional neural network, we generate the image by applying the tanh activation function to the last convolutional neural network.

### 3.3 Loss Function

The objective function used in the proposed method uses the objective function of the traditional GAN. (Equation 1)

$$\text{MinMaxV}(G, D) = E_{x \sim P_{\text{data}}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

$$L_G = -V(G, D) - \alpha E_{x \sim P_G} E_{r \sim R} [\log Q_D(R = r|x^r)] \quad (2)$$

$$L_D = V(G, D) - \beta E_{x \sim \text{data}} E_{r \sim R} [\log Q_D(R = r|x^r)] \quad (3)$$

The objective function of the proposed model uses the objective function of the existing SSGAN, and the objective function of SSGAN is the addition of Rotation-based Loss to Equation 1. The objective function of the constructor G is Equation 2, and the objective function of the discriminator D is Equation 3. (Equation 2, 3)

In Equations 2 and 3,  $V(G, D)$  is the value of Equation 1, and  $r \in R$  represents a set of possible rotations. Use  $R = (0^\circ, 90^\circ, 180^\circ, 270^\circ)$  as possible rotation. An image  $x$  rotated at an angle of  $r$  is represented by  $x^r$ , where  $Q(R|x^r)$  is the predicted distribution of the discriminator D with respect to the rotation angle of the data. By rotating the real image and the generated image at four angles, the discriminator's goal for the unrotated image is to determine whether the input is genuine or fake. And the goal is to determine the rotation angle in the rotated real image. In actual code, there are four types of classes (angle), and when data belongs to a certain angle, it is classified. Therefore, Cross Entropy Loss was used.

## 4 EXPERIMENTS

In this paper, we experiment with a model by preparing two quantitatively different datasets for comparison. CelebA dataset is a large face dataset. It consists of more than 200,000 facial images. The Large Age-Gap (LAG)[13] dataset contains images of persons of various ages, including children/adolescents and adults/seniors. The dataset consists of 3,828 celebrity images of 1,010.

Visually from a human evaluation perspective, Figure 3 shows that the resulting images of the proposed model produce better results than those generated by the previous models. (c) Resulting images of the proposed model. (b) Compared to the result image, a lot of noise was removed, and a clearer image was generated. Figure 4 is a result image generated using the existing SSGAN model. (a) is the result of using CelebA datasets with large amounts of data, and (b) is the result of using LAG datasets with small amounts of data, resulting in noise and poor-quality images.

Through Figure 3 and Figure 4, we demonstrate that the method proposed in this paper produces images better than previous and existing models in terms of human evaluation. The proposed method applied the U-Net model to put the Feature Map obtained through PCA and PCA into the existing SSGAN model as an input. Using the CelebA dataset in Figure 4 (a), we generate similar or good-quality images compared to the images generated by existing SSGAN models. The proposed method allows us to produce similar or clearer images than the results of data-rich datasets, even with small datasets.

As shown in Figure 5, the error graph of the proposed method has less vibration width than the error graph of the existing model, indicating that the learning is more stable. Table 1 shows that stable learning is conducted with the lowest average error value of the proposed method.

## 5 CONCLUSION

The main focus of existing GAN research has been to improve the quality of generated images by changing the training scheme or devising more complex models. However, a large amount of

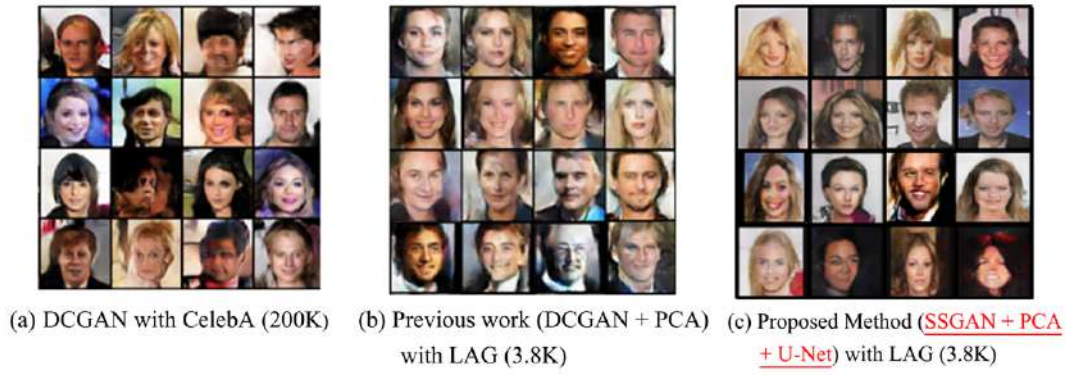


Figure 3: Image of Comparison with Previous Model

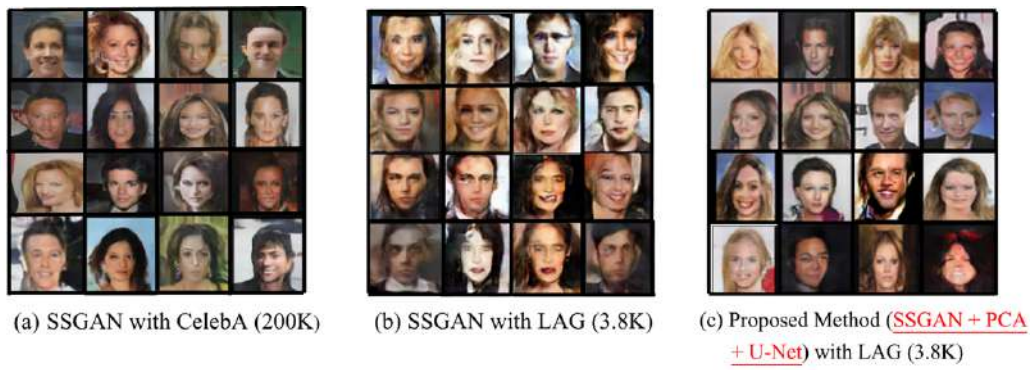


Figure 4: Image of Comparison with Existing Model

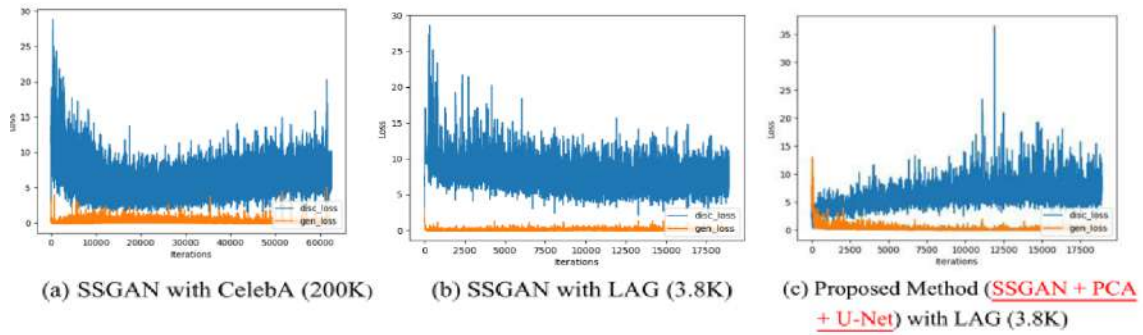


Figure 5: Comparison of Loss Graph between Generator and Discriminator

Table 1: Comparison of Mean Error between Generator and Discriminator

	<i>SSGAN(CelebA)</i>	<i>SSGAN(LAG)</i>	<i>Proposed Method(LAG)</i>
$G_{Loss}$	8.072	7.894	6.606
$D_{Loss}$	0.272	0.233	0.232

data is required to produce an image similar to a real image. This paper uses the SSGAN model as a backbone to achieve effective performance with small datasets. The proposed model does not require large datasets, and by proposing a new training method for generators, the image quality is improved with a small dataset, and the average error rate is reduced.

In previous work, PCA was applied to the DCGAN model to improve the quality of the image with some small data sets. However, unstable training of the DCGAN model itself still existed and produced an image that needed to be clarified. Although we have applied the proposed U-Net structure in the initial experimental phase to previous studies, the image quality has mostly stayed the same.

In this paper, the main objective is the generation of facial images, and PCA and U-Net structures are applied to save the overall features of facial images. It has been demonstrated in previous studies that putting a feature map of the face image as an input to the generator model via PCA helps to learn more than creating images in the existing latent space. In this paper, we propose a new training method by modifying the generator model to the U-Net structure to put the Feature Map as input to the generator model. The image quality is improved compared to the previous work, and the U-Net structure is clearer when using the SSGAN model than the DCGAN model. When the existing SSGAN model was learned with CelebA dataset and the proposed method was learned with the LAG dataset, the results showed that the images were similar in quality. The vibration width was reduced through the compared error graph and learning timetable during learning, and the average error rate decreased by 1.46 compared to the existing models. It is confirmed that the learning time was also the fastest to generate an image using the U-Net structure.

This study experimented on facial images; however, in the future, we plan to conduct additional experiments to generate high-resolution images of other domains and face images. For example, it plans to apply it to medical fields that lack data. In addition, we plan

to study a training scheme for generator models that can generate images with small datasets.

## ACKNOWLEDGMENTS

This research is a contribution to the "HPC Support" Project, supported by the 'Ministry of Science and ICT' and NIPA.

## REFERENCES

- [1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, 63(11), 139-144.
- [2] Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- [3] Cha, G. S., Asim, U., Song, M. K., Niaz, A., & Choi, K. N. (2022, August). Image Generation Network Model based on Principal Component Analysis. In 2022 Asia Conference on Advanced Robotics, Automation, and Control Engineering (ARACE) (pp. 76-80). IEEE.
- [4] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.
- [5] Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.
- [6] Sun, J., Bhattarai, B., & Kim, T. K. (2020). MatchGAN: a self-supervised semi-supervised conditional generative adversarial network. In *Proceedings of the Asian Conference on Computer Vision*.
- [7] Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., & Zheng, Y. (2020, August). Self-supervised cyclegan for object-preserving image-to-image domain adaptation. In *European Conference on Computer Vision* (pp. 498-513). Springer, Cham.
- [8] Chen, T., Zhai, X., Ritter, M., Lucic, M., & Hounsby, N. (2019). Self-supervised gans via auxiliary rotation loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12154-12163).
- [9] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.
- [10] Schonfeld, E., Schiele, B., & Khoreva, A. (2020). A u-net based discriminator for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8207-8216).
- [11] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).
- [12] Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15(2018), 11.
- [13] Bianco, S. (2017). Large age-gap face verification by Feature injection in deep networks. *Pattern Recognition Letters*, 90, 36-42.

# An Unmanned Lane Detection Algorithm Using Deep Learning and Ordered Test Sets Strategy

Zhang Shenwei  
North China University of  
Technology, School of Information  
Science and Technology  
2778922806@qq.com

Lin Xiaoyan  
Wuyi University  
514653742@qq.com

Zhang Mingwei  
Wuyi University  
1067680938@qq.com

Zhang Zhen  
Wuyi University  
1206783534@qq.com

Hou Yun  
Hang Seng University of Hong Kong,  
Department of Computing  
aileenhou@hsu.edu.hk

Ning Honglong  
South China University of Technology  
ninghl@scut.edu.cn

Qiu Tian\*  
Wuyi University  
timeqiu@hotmail.com

## ABSTRACT

The traditional method of automatic lane detection is mostly based on Hough detection. However, this category of methods has low robustness and is vulnerable to interference. In order to improve the accuracy of lane detection, the presented paper compares and analyzes the end-to-end lane line detection network based on deep learning, including Unet-base and Deeplabv3+, in view of gradient explosion and slow running speed during model training, solutions are also given. Ordered test sets are used to speed up the training processing and validate the deep learning algorithm, in the case of different image resolutions, uses Unet-base and Deeplabv3+ to perform experiments respectively. Experiments show that under the same resolution, the Unet-base model with FCN network structure incorporating a better training strategy outperforms the Deeplabv3+ algorithm model that uses a classical ASSP module to solve the downsampling layer problem in terms of model generalization capability. And the MIOU of improved Unet-base is higher than Deeplabv3+. Therefore, compared to Deeplabv3+, the improved Unet-base model is more generalized.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Object detection;

## KEYWORDS

Lane Detection, Deep Learning, Unet, MIOU, Deeplabv3+

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590087>

## ACM Reference Format:

Zhang Shenwei, Lin Xiaoyan, Zhang Mingwei, Zhang Zhen, Hou Yun, Ning Honglong, and Qiu Tian. 2023. An Unmanned Lane Detection Algorithm Using Deep Learning and Ordered Test Sets Strategy. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590087>

## 1 INTRODUCTION

Traditional lane detection methods usually use Hough transform and other methods to extract lane features. Research on end-to-end lane detection method has steadily evolved into a novel technique in the field of autonomous driving due to the rapid development of deep learning and object identification technologies [1]. For the problem of lane detection, domestic and foreign scholars have carried out relevant research in this field and achieved certain research results. Wu Lingling et al. proposed a driverless road detection method based on traditional image processing to detect the contour of lane lines [2]. Shi Linjun et al. collected the information of all colour channels in RGB images, converted them into grayscale, and then detected the lane lines by Hough transform method under multiple constraints [3]. However, the traditional road lane detection method has low robustness and is easy to be interfered with. The unmanned road detection algorithm based on deep learning learns the characteristics and distinction of the lane through the convolutional neural network, and finally detects it. Girshick R and other foreign scholars have used convolutional neural network technology to solve a series of problems such as semantic segmentation, and gradually speed up the detection speed and improve the accuracy [4]. Redmon J and other foreign scholars proposed the YOLO (You Only Look Once) algorithm, which is an end-to-end algorithm to transfer the target detection strategy to the regression strategy, and further classify the detected target well [5].

This paper takes the lane line as the research object. Firstly, the road is detected by the traditional detection method, and the deeplabv3+ object detection algorithm model is provided based on the paddle to design a more accurate algorithm. A variety of

algorithms are comprehensively compared, and the feasibility of the algorithm model is verified through more data training, and the generalization ability of the lane detection algorithm model is constantly improved.

## 2 ROAD DETECTION ALGORITHM FOR UNMANNED DRIVING

In the development history of image segmentation, it is roughly divided into two stages. The first stage is based on the traditional manual feature algorithm for detection, and the second stage is based on the deep learning algorithm for detection [6]. The deep learning model approach can be trained to segment pixel lanes and is computationally less expensive than traditional lane detection methods that rely on a combination of hand-crafted features and heuristics [7]. This paper will first describe the algorithm based on Hough transform method and then focus on the lane line detection algorithm based on deep learning, including three important convolutional neural network models: fully connected convolutional neural network FCN (Fully Convolutional Networks for Semantic Segmentation), deep residual network ResNet, and Deeplabv3+.

### 2.1 Algorithms using Hough Transform method as a technical mean

The process of the algorithm is shown in Figure 1. Firstly, Gaussian filtering is used to remove the noise, and then Canny edge detection is used to roughly detect the edge of the object in the image [8]. The detection results are shown in Figure 2. The lane lines on both sides of the highway are roughly fixed in a range, so we can draw an ROI (region of interest), filter out the edges outside the ROI [9], and the detection results are shown in Figure 3. In the image that basically extracts the lane lines on both sides of the highway, there are still multiple lane lines on both sides, so we need to perform Hough transform to detect two clear lane lines, and finally determine and redraw the lane lines. The following steps are used: 1) distinguish the lane lines on both sides of the highway by using different slopes; 2) remove outliers from the lane lines on both sides of the highway in turn: Calculate the slope and the mean of the slope of each line, repeatedly find out the difference between the two, and remove the lines that do not conform to the rules one by one 3) Perform algebraic operations on the vertex set of the lane lines on both sides of the highway in turn to obtain the lane we need 4) Superimpose the result with the original image, and finally output the image (as shown in Figure 4)

### 2.2 Detection algorithm based on deep learning

**2.2.1 FCN.** When semantic segmentation is performed using deep learning, local location information is usually ignored in the process of high-level feature extraction. To avoid this problem, we propose a graph model initialized by a fully convolutional network named FCN for image semantic segmentation [10]. FCN is a very famous network, which has had a profound impact on the development of image segmentation. It is the first network to realize end-to-end semantic segmentation. Usually, after completing convolution, the CNN (Convolutional Neural Network) network is connected with many Affine layers, and the semantic information map after convolution operation is projected to an output vector of exact length

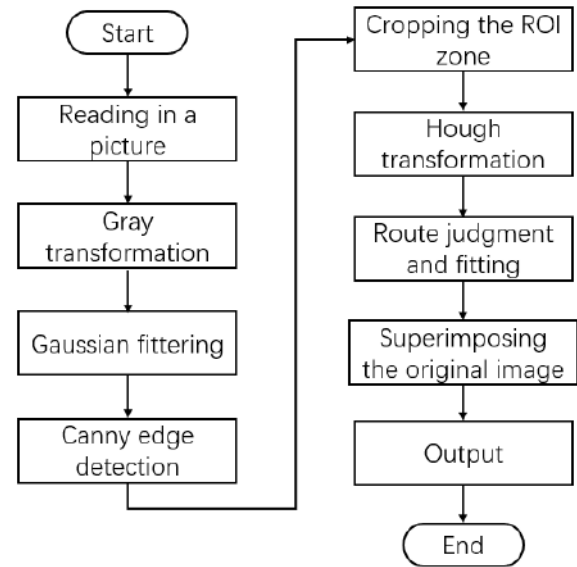


Figure 1: Hough transform method algorithm flow

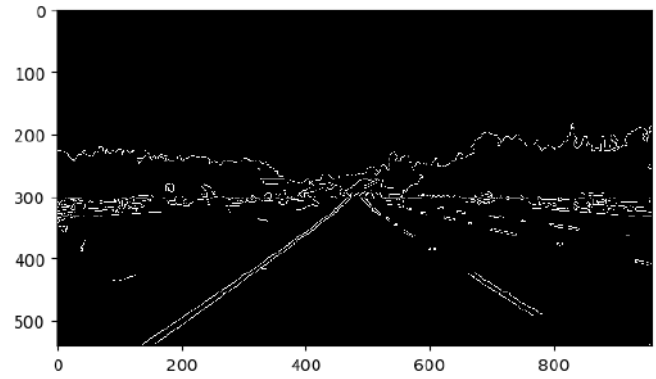


Figure 2: Canny edge detection

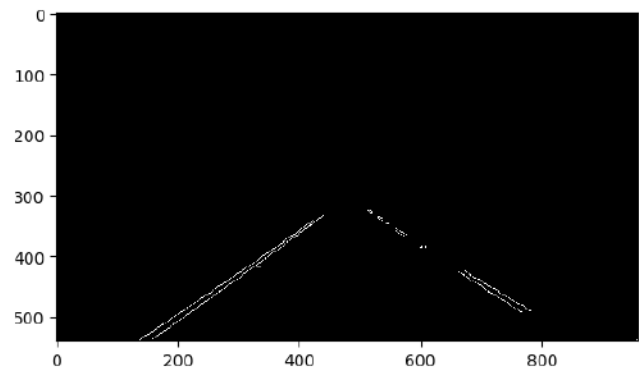


Figure 3: based on ROI edge filtering

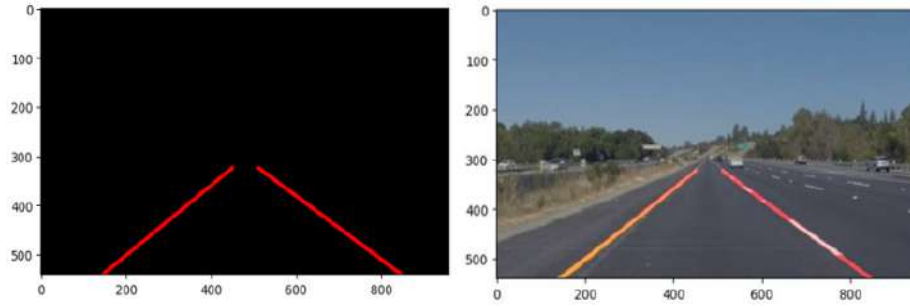


Figure 4: Route judgment and redrawing

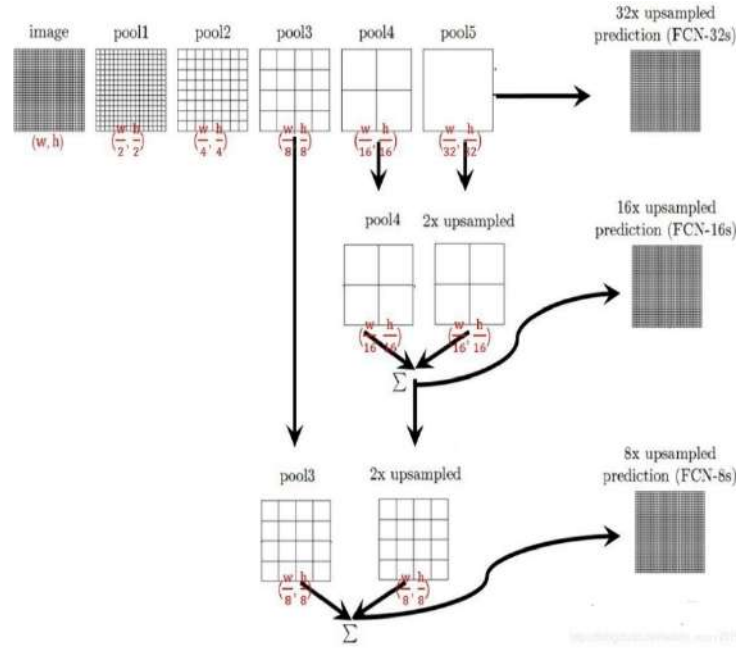


Figure 5: FCN-8 network structure

[11]. Finally, a numerical value is used to express the probability of an input image. FCN-8s combines the results of the previous output twice, and then outputs the 8 times enlarged semantic information map by deconvolution. Take FCN-8s as an example in Figure 5, that is, the feature map of pool5 is enlarged by 2 times, then fused with the feature map of pool4, expanded by 2 times by deconvolution, fused with the feature map of pool3, and finally enlarged by 8 times by deconvolution to output an image that remains the same size as the original image. As shown in Figure 5, we can see intuitively that the semantic information of FCN segmentation is continuously improved as the number of fused layers increases [12].

**2.2.2 ResNet (Deep residual network).** For traditional CNN networks, arbitrarily increasing the number of network layers will easily lead to gradient disappearance and explosion. Usually, regularization is used to solve this problem. However, when the number of network layers is deepened, the error rate of training and testing on the data set also increases. Here, ResNet solves the degradation

problem very well. ResNet uses the residual learning structure, and residual learning is popularly understood as the difference between two numbers. The learning of a network is equivalent to a function fitting. It is difficult to learn  $H(x)$  directly, so we first learn  $F(x)$  indirectly. ResNet gives two ideas to avoid degeneracy problems: identity mapping and residual mapping. The identity map refers to the "curved" part of the graph, and the residual map refers to the remaining part of the graph that is not "curved".

**2.2.3 Deeplabv3+.** DeepLabv3+ is a dilated fully convolutional network, which can reduce the down-sampling rate while stabilizing the receptive field. The final output image contains a large amount of semantic information and then restores the image as large as the original input image through technical means such as bilinear interpolation. Its network structure is shown in Figure 6. Its core framework is ResNet residual learning network and spatial pyramid pooling module, which is conducive to mining multi-dimensional

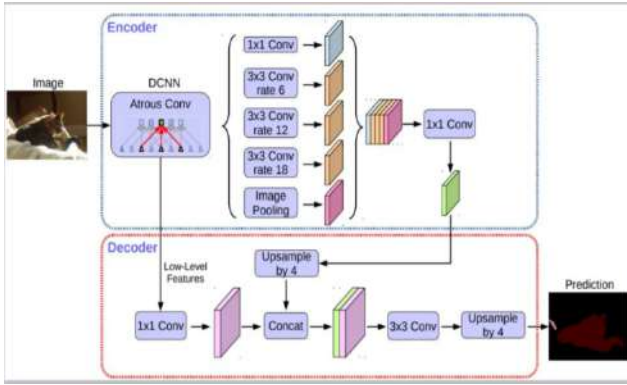


Figure 6: DeepLabv3 network structure

semantic information. The encoder-decoder architecture is an effective architecture in semantic segmentation. This is because the encoder network helps to obtain various scale features from the deep convolutional layers, while the decoder network often helps to recover the details of spatial resolution and location information [13]. DeepLabv3+ introduces an encoder-decoder module on the premise of the dilated full convolutional network, which is to better integrate high-dimensional and low-dimensional semantic information maps into one and improve the accuracy of semantic segmentation.

For DeepLabv3, the resolution ratio of the semantic information map output by the ASPP (Atrous Spatial Pyramid Pooling) module is 8 or 16, and then the  $1 \times 1$  classification is adopted, and finally the interpolation method is used to restore the original shape. There are certain defects in this practice, especially the resolution ratio of 16, which does not achieve an accurate segmentation effect. So DeepLabv3+ emulated the encoder-decoder module and re-used the new decoder structure. As shown in Figure 6, the encoder output is first restored to a 4x effect using a linear interpolation technique, which is then merged with the matching low-level semantic map in the encoder. Because the encoder only outputs 256 kinds of semantic information maps, and the low-level semantic information map has a high dimension, in order to avoid weakening the high-level semantic information map output by the encoder, the  $1 \times 1$  convolution is first used to reduce the dimension of the low-level semantic information map. After the two semantic information maps are merged, the fusion effect is strengthened by  $3 \times 3$  convolution. Then the bilinear interpolation technique is used to recover the predicted image with the same size as the original shape.

**2.2.4 Dilated Convolution.** On the premise of stabilizing the semantic information map and managing the receptive field well, dilated convolution can easily extract multi-dimensional semantic information. As shown in Figure 7, rate ( $r$ ) manages the receptive field size, and the receptive field grows as  $r$  increases. The output\_stride represents the resolution ratio of the input image to the output semantic information map. Generally, the resolution ratio of the two for the CNN classification network is 32. Only by changing the stride of the final downsampling layer to 1 and the rate of all the later convolutional layers to 2, the receptive field will remain unchanged, and the resolution ratio of the two of the

dilated full convolutional neural network is 16. In the same way, only by changing the stride of the last two downsampling layers to 1 and the rate of the later paired convolutional layers to 2 and 4, the ratio of the two resolutions of the dilated FCN becomes 8.

### 3 EXPERIMENT

#### 3.1 Experimental Instructions

This paper uses python as the development language; Software environment: cuda9, cudnn-9.0, relying on Python3.6, opencv-python, paddlepaddle-gpu, imgaug; The hardware environment is quad-core i5- 6300HQ\_@\_2.30GHz, 8G RAM, NVIDIA GeForce 950M [14]. The data set adopted in this paper comes from the pictures of Baidu Apollo lane line segmentation competition. These pictures are all traffic road pictures of different scenarios taken by cars equipped with laser radars and cameras on different urban roads. In the experiment, these images were divided into more than 50000 training set images and more than 10000 test set images, and the resolution of each image is  $3384 \times 1710$ .

#### 3.2 Training Strategy

Because the resolution of the original image is very large, it takes up a lot of graphics card resources. In order to save resources, the input image is processed. As shown in Figure 8, the image label classification is analyzed first, and the  $3384 \times 690$  part of the top of the image is removed (the experimental results show that the lane detection does not affect the lane after removal). Some types of data in the test set are preprocessed, and the data enhancement is completed from multiple angles such as brightness, saturation, noise and contrast in more than 10,000 images. All the images in Road 2 and road 4 scenes were verified, a large amount of erroneous data were removed, and the results were filtered one by one. Only nearly 60,000 images met the training requirements. In order to ensure that the size of the images does not change, three types of images with different resolutions ( $768 \times 256$ ,  $1024 \times 384$ ,  $1536 \times 512$ ) are trained. We set the image size as a parameter to decide which category of images should be trained first. The experimental results show that training from small-resolution images can not only efficiently verify the feasibility of the algorithm model but also give a good effect and distribution. At the same time, it solves the problem that large-resolution images are difficult to train due to the field.

#### 3.3 Evaluation metrics

Here, we choose the size of the average intersection to measure the accuracy of the lane detection algorithm model, and the calculation formula of the average intersection ratio is as follows:

$$IoUc = \frac{TP}{TP + FP + FN} \quad (1)$$

$$MIoU = \frac{1}{k+1} \sum_i^k IoUc \quad (2)$$

Among them, the TP represents the classifier considers the sample as a positive sample; the FP represents the classifier considers the sample as a positive sample, but in fact the sample is a negative

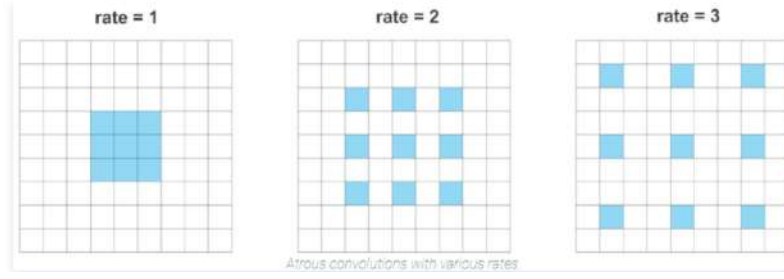


Figure 7: Dilated convolution

#	name	id	trainId	category	catId	hasInstances	ignoreInEval	color
Label(	'void'	0	0	'void'	0	False	False	( 0, 0, 0 )
Label(	's_w_d'	200	1	'dividing'	1	False	False	( 70, 150, 180 )
Label(	's_y_d'	204	1	'dividing'	1	False	False	( 220, 30, 60 )
Label(	'ds_w_d'	218	1	'dividing'	1	False	True	( 128, 0, 128 )
Label(	'ds_y_d'	203	1	'dividing'	1	False	False	( 255, 0, 0 )
Label(	'sb_w_d'	206	1	'dividing'	1	False	True	( 0, 0, 60 )
Label(	'sb_y_d'	207	1	'dividing'	1	False	True	( 0, 60, 100 )
Label(	'b_w_g'	201	2	'guiding'	2	False	False	( 0, 0, 142 )
Label(	'b_y_g'	203	2	'guiding'	2	False	False	( 119, 11, 32 )
Label(	'db_w_g'	211	2	'guiding'	2	False	True	( 244, 35, 232 )
Label(	'db_y_g'	208	2	'guiding'	2	False	True	( 0, 0, 160 )
Label(	'ds_w_s'	214	3	'stopping'	3	False	True	( 159, 159, 153 )
Label(	's_w_s'	217	3	'stopping'	3	False	False	( 220, 220, 0 )
Label(	'ds_w_s'	213	3	'stopping'	3	False	True	( 250, 170, 30 )
Label(	's_w_c'	215	4	'chevron'	4	False	True	( 102, 102, 156 )
Label(	's_y_c'	216	4	'chevron'	4	False	True	( 128, 0, 0 )
Label(	's_w_p'	210	5	'parking'	5	False	False	( 128, 64, 128 )
Label(	's_y_p'	222	5	'parking'	5	False	True	( 238, 232, 170 )
Label(	'c_wy_s'	214	6	'zebra'	6	False	False	( 190, 153, 153 )
Label(	'a_w_u'	220	7	'thru/turn'	7	False	True	( 0, 0, 230 )
Label(	'a_w_t'	220	7	'thru/turn'	7	False	False	( 128, 128, 0 )
Label(	'a_w_tl'	221	7	'thru/turn'	7	False	False	( 128, 78, 160 )
Label(	'a_w_tr'	222	7	'thru/turn'	7	False	False	( 150, 100, 100 )
Label(	'a_w_tlr'	223	7	'thru/turn'	7	False	True	( 255, 165, 0 )
Label(	'a_w_l'	224	7	'thru/turn'	7	False	False	( 180, 165, 180 )
Label(	'a_w_r'	225	7	'thru/turn'	7	False	False	( 107, 142, 35 )
Label(	'a_w_lr'	226	7	'thru/turn'	7	False	False	( 201, 255, 229 )
Label(	'a_w_lu'	230	7	'thru/turn'	7	False	True	( 0, 191, 255 )
Label(	'a_w_tr'	228	7	'thru/turn'	7	False	True	( 51, 255, 51 )
Label(	'a_w_m'	229	7	'thru/turn'	7	False	True	( 250, 128, 114 )
Label(	'a_y_t'	223	7	'thru/turn'	7	False	True	( 127, 255, 0 )
Label(	'b_n_sp'	205	8	'reduction'	8	False	False	( 255, 128, 0 )
Label(	'd_wy_sp'	212	8	'attention'	8	False	True	( 0, 255, 255 )
Label(	'r_wy_sp'	227	8	'no parking'	8	False	False	( 178, 132, 190 )
Label(	'vom_wy_n'	223	8	'others'	8	False	True	( 128, 128, 64 )
Label(	'om_n_n'	250	8	'others'	8	False	False	( 102, 0, 204 )
Label(	'noise'	249	0	'ignored'	0	False	True	( 0, 153, 153 )
Label(	'ignored'	255	0	'ignored'	0	False	True	( 255, 255, 255 )

Figure 8: classification of the labels

Table 1: Unet-base Different resolution experiments and Deeplabv3+ comparison results

Models		Batch Size	Resolution	MIoU
(Train models)	Base LR	(Sample size)	(image resolution)	(Mean Intersection over union)
	(Learning rate)			
Unet-base	0.001	8	768 * 256	0.522
Unet-base	0.001	4	1024 * 384	0.551
Unet-base	0.001	2	1536 * 512	0.605
Deeplabv3+	0.001	2	1536 * 512	0.599

sample; the FN represents the samples as negative samples, but actually, it is a positive sample;

### 3.4 Experimental results

It can be seen from Table 1, the MIoU of the Unet-base algorithm model reached the highest value under the premise of Base

LR=0.001, Batch Size=2, and Resolution of 1536\*512. Four experiment models are shown in Figure 9, Figure 10, Figure 11, Figure 12.

Although the Deeplabv3+ algorithm model uses the classical ASPP module to solve the problem of downsampling layer in the previous introduction, in the lane line detection task, Unet - base adopts

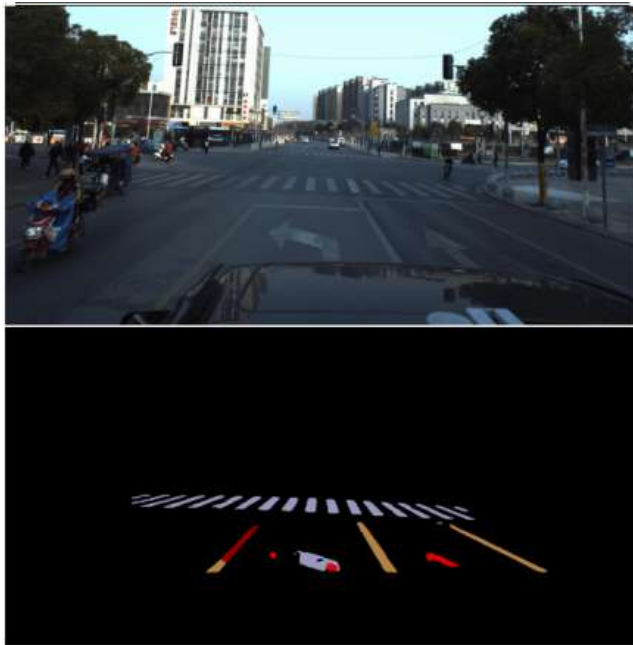


Figure 9: Unet-base first set of test results

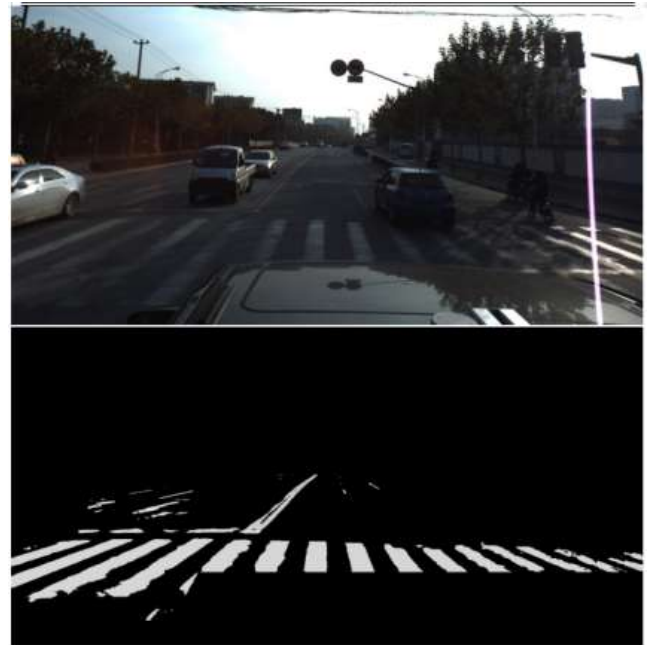


Figure 11: Deeplabv3+ first set of test results

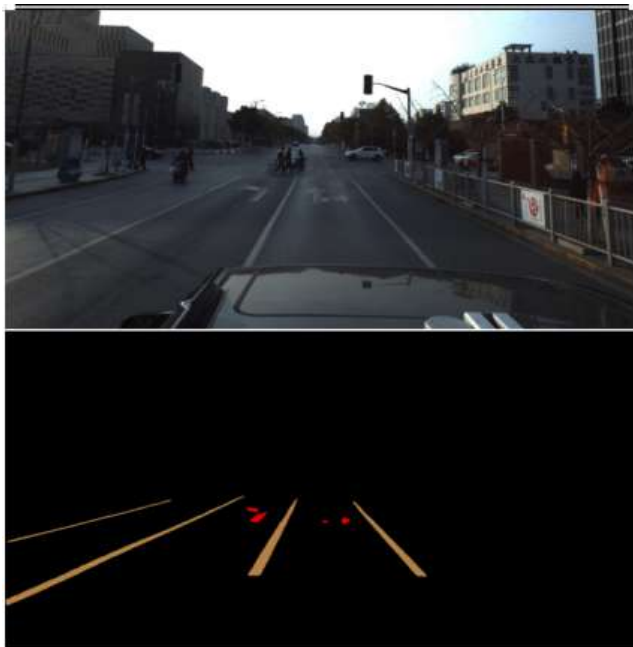


Figure 10: Unet-base second set of test results

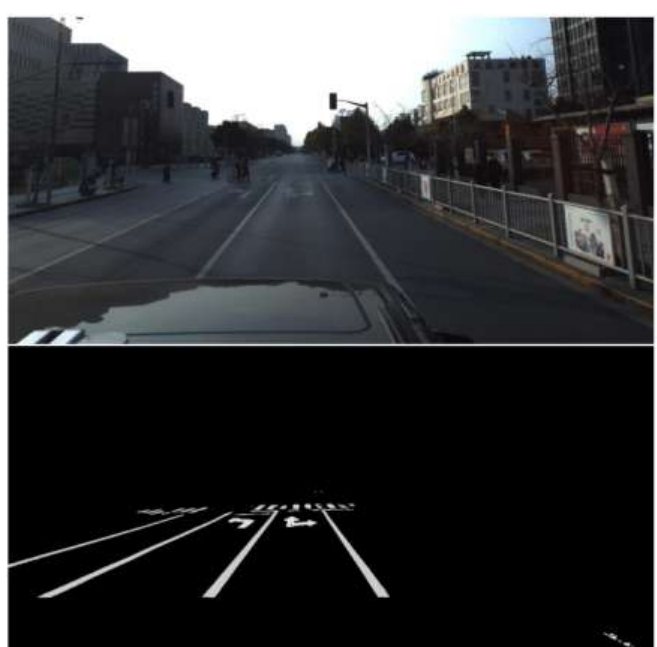


Figure 12: Deeplabv3+ second set of test results

the structure of FCN network, combined with a better training strategy, its model generalization ability is better than deeplabv3+.

#### 4 CONCLUSIONS

This paper first introduces the method based on Hough detection, but the traditional method has been difficult to adapt to complex

road conditions, so compares the lane line detection network of Unet-base and Deeplabv3+. The experimental results show that under the same model, MIOU increases with the increase of resolution, and under the same resolution, the MIOU of Unet-base model with FCN network structure is higher. At the same time, the method and

precautions of model training are given, which shows that a good algorithm model must be combined with a better training strategy in order to produce good results in practical problems. This paper provides a new idea for road detection and has important reference value.

## REFERENCES

- [1] TANG Zhiwei. Survey on Vision-based Autonomous Vehicles [J]. *Manufacturing Automation*, 2016, 38(8): 134-136.
- [2] Wu Lingling, Lin Zhixian, Guo Tailiang. Hough Transform lane Detection Based on Superposition Constraint [J]. *Cable TV Technology*, 2019, 2019(3): 44-49.
- [3] SHI L J, YU Su. Lane detection method based on Hough transform under multiple constraints [J]. *Computer Measurement and Control*, 2018, 26(9): 9-12.
- [4] Girshick R, Donahue J, Darrell T, *et al*. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, 2014: 580-587.
- [5] Redmon J, Divvala S, Girshick R, *et al*. You Only Look Once: Unified, Real-time Object Detection[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016: 779-788. SSS
- [6] Qin Z, Wang H, Li X. Ultrafast structure-aware deep lane detection[C]//*European Conference on Computer Vision*. Springer, Cham, 2020: 276-291.
- [7] Neven D, De Brabandere B, Georgoulis S, *et al*. Towards end-to-end lane detection: an instance segmentation approach[C]//*2018 IEEE intelligent vehicles symposium (IV)*. IEEE, 2018: 286-291.
- [8] Li Yadi, Huang Haibo, Li Xiangpeng. Night lane detection Based on Canny operator and Hough Transform [J]. *Science Technology and Engineering*, 2016, 16(31): 234-237, 242
- [9] Wang Xin, Liu Yuchao, Hai Dan. Lane detection based on Double ROI and variable interval scanning [J]. *Journal of Command and Control*, 2017, 3(2) :154-159.
- [10] Lu, Y., Chen, Y., Zhao, D., Chen, J. (2019). Graph-FCN for Image Semantic Segmentation. In: Lu, H., Tang, H., Wang, Z. (eds) *Advances in Neural Networks – ISNN 2019*. ISNN 2019. *Lecture Notes in Computer Science*(), vol 11554. Springer, Cham. [https://doi.org/10.1007/978-3-030-22796-8\\_11](https://doi.org/10.1007/978-3-030-22796-8_11)
- [11] Biswas Sanad, Chambers Destini, Hairston W David, Bhattacharya Sylvia. Head pose classification for passenger with CNN[J]. *Transportation Engineering*, 2023, 11.
- [12] Kamal Shoaib, Shende Vaishali, Gajendra, Swaroopa Korla, Bindhu Madhavi P., Akram Patan Saleem, Pant Kumud, Patil Shantala Devi, Sahile Kibebe. FCN Network-Based Weed and Crop Segmentation for IoT-Aided Agriculture Applications[J]. *Wireless Communications and Mobile Computing*, 2022, 2022.
- [13] Liu J, Wang Z, Cheng K. An improved algorithm for semantic segmentation of remote sensing images based on DeepLabv3+[C]//*Proceedings of the 5th international conference on communication and information processing*. 2019: 124-128.
- [14] Li Meimei, Hu Chunhai, Long Ping, Liu Shaonan. Lane Detection Algorithm Based on MultiRes+UNet Network [J]. *Journal of Electronic Measurement and Instrument*, 2020, 34(9): 1-6 [10] KIM J, LEE M. Robust lane detection based on convolutional neural network and random sample consensus [C]. *ICONIP*, 2014: 454-461.

# Intelligent perception recognition and positioning method of distribution network drainage line

Shuzhou Xiao\*

Qiuyan Zhang\*

irina@stu.csust.edu.cn

tyx002@protonmail.com

Electric Power Research Institute of Guizhou Power Grid  
Co., LTD, Electric Power  
Guiyang, Guizhou, China

Jianrong Wu

Electric Power Research Institute of Guizhou Power Grid  
Co., LTD, Electric Power  
Guiyang, China

Qiang Fan

Electric Power Research Institute of Guizhou Power Grid  
Co., LTD, Electric Power  
Guiyang, China  
larst@affiliation.org

Chao Zhao

Electric Power Research Institute of Guizhou Power Grid  
Co., LTD, Electric Power  
Guiyang, China

## ABSTRACT

Due to the serious interference of illumination and background on the camera during the live operation of the distribution network robot, it is difficult to match, identify, and locate the feature points of the target image, such as the drainage line. This paper proposes the intelligent perception recognition and positioning method of the distribution network drainage line. First, YOLOv4 is used to identify and classify the typical parts of the distribution network and determine the two-dimensional position of the operation point. Subsequently, the Res-UNet segmentation network was improved to perform image segmentation of drainage lines and wires to avoid complex background interference. Finally, binocular vision is used to extract the center line of the wire through the image geometric moment and determine the image line of the wire and the center of the double eyes. The intersection line of the wire is the spatial three-dimensional coordinates of the wire. After the target detection, wire segmentation, and operation point positioning experiments, this method can achieve a positioning accuracy of 1 mm in the x and y directions and 3 mm in the z direction under the camera coordinate system, which provides a guarantee for accurate perception and recognition and reliable operation control of the power distribution robot operation.

## CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590088>

## KEYWORDS

Yolov4, U-Net, binocular vision location, Distribution network drainage line

### ACM Reference Format:

Shuzhou Xiao, Qiuyan Zhang, Qiang Fan, Jianrong Wu, and Chao Zhao. 2023. Intelligent perception recognition and positioning method of distribution network drainage line. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590088>

## 1 INTRODUCTION

In the electric power industry, as China's demand has increased in recent decades, the distribution network equipment has had to maintain a high-intensity working state for a longer period of time, with fatigue damage risk increasing year by year and distribution network maintenance operation numbers growing [10]. In order not to affect people's daily lives and maintain a stable supply, live operations have become the main distribution network maintenance operation. When performing manual live maintenance operations, it usually faces a high voltage of 10 kV or even higher. Maintenance personnel needs to wear heavy shielding clothing, which makes the safety risk of a long-term operation high. The robot can replace the operator to perform tasks remotely, so the distribution network's maintenance robot is widely concerned about [6]. One of the biggest challenges of outdoor work robots is the accurate identification and positioning of outdoor light operation targets under interference.

Traditional distribution network equipment detection technology mainly extracts image features through manually designed feature descriptors, such as SVM [8], DPM [4], and HOG algorithm [3]. Zhao Junmei et al. [5] proposed an algorithm for binary segmentation of distribution network equipment images using the pixel statistics method, and Qi et al. [9] proposed an insulator detection algorithm based on contour segmentation. Since CNN's proposal, it has quickly replaced the traditional algorithm in the field of image processing [1]. The model based on deep learning is gradually applied to the object detection and defect diagnosis of aerial inspection images of the power grid. Li Junfeng et al. [13] combined the

random forest with CNN, first using AlexNet to extract the features of the power equipment images, and then using the trained random forest algorithm to classify the extracted features of the network, completing the detection of multiple types of power equipment. Cheng Haiyan et al. [7] directly used the Faster R-CNN model to train and detect the aerial insulator image data, but this algorithm has problems such as missing detection on smaller insulator targets. For insulator occlusion problems, Zhao Zhenbing and others [15] added an ASDN to the R-CNN model to generate a mask on the feature map part and obtain more obscured samples, improving the detection effect of the model on occluded samples. Wang et al. [12] used a monocular SSD algorithm based on candidate regions to locate and identify insulator defects. CHEN and others [2] use the SOFCN to conduct fault detection on insulators.

At present, the mainstream robot operation target identification and positioning methods can be divided into three kinds: lidar, structured light cameras, and binocular cameras, according to the different sensors. Zhang Jing [14] uses lidar to identify large scene operation targets, but it generates sparse point clouds for weakly textured thin wires with poor quality. Sarbolandi H [11] combined with the active light emitting principle and experimental verification discussed the characteristics of structured light cameras that make them vulnerable to sunlight interference and poor outdoor effects.

For the complex lighting and background environment of the distribution network, this paper proposes the identification and positioning method of the distribution network drainage line, which can be divided into three steps: target detection, image segmentation, and binocular visual positioning. First, the YOLOv4 network is improved to identify the insulator, drainage line clip, and grounding ring of the distribution network, and then Res-Unet is used to divide the drainage line and wire of the distribution network. Finally, the spatial position of the drainage line and wire of the distribution network is obtained based on the binocular vision algorithm.

## 2 IDENTIFICATION METHOD OF DISTRIBUTION NETWORK DRAINAGE LINE BASED ON YOLOV4

YOLOv4 The detection process for the distribution network equipment is as follows:

Enter a grounding ring picture of any size and adjust the size of the picture to 416×416 (the excess part with a pixel value of 0 while keeping the length and width ratio unchanged). The newly obtained images serve as input for the network. The YOLOv4 algorithm directly divides the pictures into SS grids. Since YOLO v4 makes a multi-scale prediction, as shown in Figure 2, the predictions are made on three different layers, and the SS grids in each layer are 13×13, 26×26, and 52×52 respectively. This project needs to predict four cases of a grounding ring, insulator, bolt, and drain clamp, so the category M is 4, and the shape of the output layer is (13,13,27), (26,26,27), (52,52,27,27) in the last dimension of 3 (M + 4 + 1), where 4 is 4 coordinate information, including the center coordinates of the prediction box (x, y) and the input image (w, h), and 1 is a Confidence value. Confidence expression is as described in equation (1):

$$confidence = Pr(object) * IOU_{pred}^{truth} \quad (1)$$

When there is an object in the grid, it is indicated by 1; otherwise 0. It indicates the intersection of the annotation box and the prediction box, so confidence reflects the confidence that the prediction box contains the object. product of prediction confidence and predicted analogy probabilities.

Take out the box with a score greater than the threshold, and finally select the final prediction box by non-maximum suppression (NMS) using the position and score of the box.

## 3 DRAINAGE LINE SEGMENTATION METHOD BASED ON RES-UNET

According to the insufficient anti-interference resistance and poor robustness of traditional methods, this project improves the U-Net semantic segmentation algorithm when dividing wire and drainage line. First, the backbone network VGG 16 used during the down-sampling process was changed to Resnet101. Secondly, this project improves the four downsampling process of the low-level network in the original U-Net network and the four high-level network. Four downsampling processes are added in the low-level network, and the two sampled feature maps are merged and superimposed to the last feature graph of the layer network. The improved U-Net network strengthens the extraction of low-level feature information, controlling the receptive fields of neurons at different levels at a more reasonable level. Because the gradient dispersion phenomenon is alleviated to some extent, the depth of the network can be increased to improve the image segmentation effect. The structure of the Res-Unet network is shown in Fig.1.

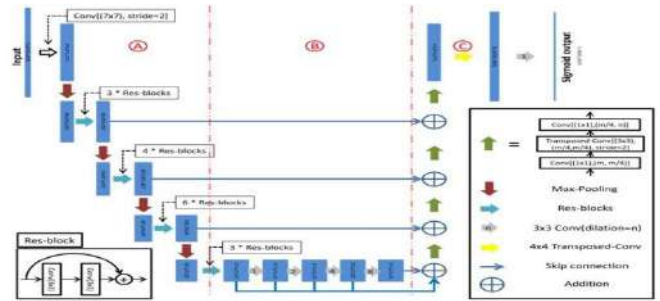


Figure 1: Res-Unet structure diagram.

## 4 A WIRE POSITIONING METHOD BASED ON BINOCULAR VISION

After the robot is driving along the road to the designated operation position, stop the movement and lock the robot's body moving mechanism and the distribution network wire. To control the specified position of the wire of the operating arm, a parallel binocular visual measurement system is constructed. Combined with the internal and external parameter transformation matrix of the camera, the 3-D coordinates of the specified point in the robot coordinate system and the relative position with the end effector are calculated.

### 4.1 Lead centerline positioning

The extraction of the wire centerline directly affects the subsequent image matching and the accuracy of point cloud computing. To

reduce the zigzag noise interference at the edge of the wire image segmentation, the image geometric moment is used to extract the wire centerline. Image geometric moment is the transformation of image pixel integration, with transformation invariances such as rotation, translation, and scale, and is not sensitive to random noise. The image geometric moment is selected as the image feature extraction tool to enhance the robustness of image matching.

The above segmented wire image is regarded as a two-dimensional distribution  $f(u, v)$ , and  $f(x, y)$  is the gray value at the coordinates of the pixel coordinate system  $(x, y)$ . At this case, the  $(p + q)$  order geometry of the wire is:

$$\mu_{pq} = \sum_{x=0}^m \sum_{y=0}^n x^p y^q f(x, y) \quad (2)$$

Then there are the center coordinates of the wire image  $(x_c, y_c)$  is:  $x_c = \frac{\mu_{10}}{\mu_{00}}, y_c = \frac{\mu_{01}}{\mu_{00}}$

The direction angle  $\theta$  of the wire is:

$$\theta = \frac{1}{2} \arctan\left(\frac{2b}{a-c}\right), \theta \in [-90, 90] \quad (3)$$

where  $a = \frac{\mu_{20}}{\mu_{00}} - x_c^2, b = \frac{\mu_{11}}{\mu_{00}} - x_c y_c, c = \frac{\mu_{02}}{\mu_{00}} - y_c^2$

Then it can be deduced that the center line of the wire in the pixel coordinate frame is:

$$y = \tan(\theta) * (x - x_c) + y_c \quad (4)$$

Similarly, the centerline equation of the left and right binocular wires is deduced, respectively:

$$\begin{cases} y_l = \tan(\theta_l) * (x_l - x_{cl}) + y_{cl} \\ y_r = \tan(\theta_r) * (x_r - x_{cr}) + y_{cr} \end{cases} \quad (5)$$

## 4.2 Lead depth calculation

As shown in Fig.2, the center of the cochlear wheel of the working arm connected to the body moving mechanism is used as the origin of the robot coordinate system  $O_w - x_w y_w z_w$  of the binocular vision system,  $O_{cl} - x_{cl} y_{cl} z_{cl}, O_{cr} - x_{cr} y_{cr} z_{cr}$  are the left and right camera coordinate system,  $O_l - x_l y_l, O_r - x_r y_r$  are the left and right image coordinate systems after focal length normalization.

The space point P, in the line of the center point  $P_{1cl}$  and the optical axis of the camera  $c_l$ , which is consistent to the equation (6):

$$\begin{cases} x = x_{1cl} t_l \\ y = y_{1cl} t_l \\ z = t_l \end{cases} \quad (6)$$

Convert the coordinates of the point  $P_{1cl}$  in the camera coordinate  $c_r$  frame to coordinates in the camera coordinate  $c_l$  system  $P_{1cl}$ :

$$\begin{bmatrix} x_{1cl} & y_{1cl} & z_{1cl} & 1 \end{bmatrix}^T = {}^{c_l} M_{c_r} \begin{bmatrix} x_{1cr} & y_{1cr} & 1 & 1 \end{bmatrix}^T \quad (7)$$

${}^{c_l} M_{c_r}$  is the position vector of the central point of the optical axis of the right camera in the coordinate system of the left camera. The position vector of the central point of the optical axis of the right camera in the left camera frame, known from the geometric relationship of binocular vision in Figure 2, only the coordinate value

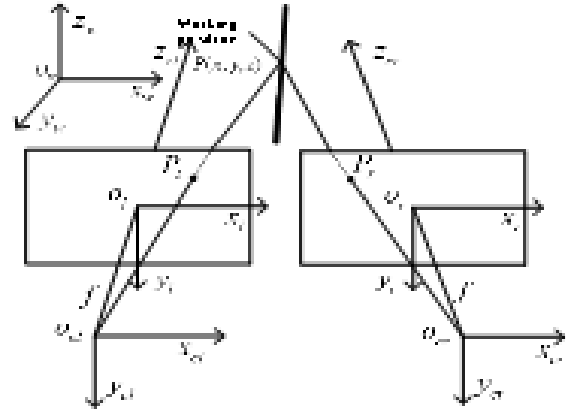


Figure 2: Binocular visual system ranging model.

$x_{1cr}$  becomes  $x_{1cl}$ . Therefore, the line equation can be expressed as equation (8):

$$\begin{cases} x = p_x + (x_{1cl} - p_x) t_r \\ y = p_y + (y_{1cl} - p_y) t_r \\ z = p_z + (z_{1cl} - p_z) t_r \end{cases} \quad (8)$$

The intersection point of the above two straight lines is the point of space P. If equation (6) and (7) stand together, you can solve the three-dimensional coordinates of the space point P in the camera coordinate system  $c_l$ . Using the external parameters between the camera coordinate system and the robot coordinate system, the 3D coordinate of the point P in the robot coordinate system  $c_l$  can be calculated from the matrix transformation of the camera coordinate system.

$$\begin{bmatrix} x_w & y_w & z_w & 1 \end{bmatrix}^T = \begin{bmatrix} R & P \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} x & y & z & 1 \end{bmatrix}^T \quad (9)$$

Because the wire center line, the image center line, the camera light center, the camera light center O and the wire image center line  $\vec{l}$  can determine a set of image surfaces, the intersection of the two sets of images of the binocular camera is the position of the real wire. After the image is corrected, the camera imaging model can be approximated to the ideal small-hole imaging model. The coordinates of the wire in the image coordinate system are  $P_{pixel}(x, y)$ , meters. The coordinates in the pixel coordinate system are given as follows  $P_{pixel}(u, v)$ , the number of pixels and the coordinate of the origin of the image coordinate system in the pixel coordinate system is  $(C_x, C_y)$ . At this time, the wire equation in the pixel coordinate frame is converted to the image coordinate frame wire equation, namely:

$$\begin{cases} x = (u_c + t * \cos\theta - C_x) * dx \\ y = (v_c + t * \sin\theta - C_y) * dy \end{cases} \quad (10)$$

For the wires in the pixel coordinate system  $(u_c, v_c)$  and the image coordinate system  $(x, y)$ ,  $dx$  is the number of pixels corresponding to 1 meter in the x direction in the pixel coordinate system, and similarly  $dy$  is the number of pixels corresponding to 1 meter in the y direction. There are also conductors, wire images, and camera

centers, namely the object image surface. At this time, the vector  $\vec{OC}$  and direction vector of the wire  $\vec{l}$  determined by the conductor center of gravity C and the camera center O are known, and the normal line  $\vec{n}$  of the image surface can be determined as:

$$\vec{n} = \vec{l} \times \vec{OC} = \begin{bmatrix} \vec{i} & \vec{j} & \vec{k} \\ \cos\theta & \sin\theta & 0 \\ x_c & y_c & f \end{bmatrix} = \{f * \sin\theta \quad f * \cos\theta \quad y_c * \cos\theta - x_c\} \quad (11)$$

Then the two groups of two objects are:

$$\begin{cases} x_l * f * \sin\theta_l + y_l * f * \cos\theta_l + z * y_{cl} * \cos\theta_l - x_{cl} * \sin\theta_l = 0 \\ x_r * f * \sin\theta_r + y_r * f * \cos\theta_r + z * y_{cr} * \cos\theta_r - x_{cr} * \sin\theta_r = 0 \end{cases} \quad (12)$$

## 5 EXPERIMENTAL VERIFICATION

To verify the feasibility of this method, the network equipment identification experiment, network wire segmentation experiment, and wire operation point location experiment are conducted in this section.

### 5.1 Identification and detection experiments of distribution network equipment

**5.1.1 Dataset building.** Using the annotation tool labellmg, the distribution network equipment in the data set: the insulator, grounding ring, wire clip, and bolt (including bolt corrosion) are marked to establish the data set of distribution network transmission line components and collected in the 10KV distribution network line pilot, to enrich the diversity of samples. Ground ring samples of different colors were collected at different angles, different distances, and different times, and 5000 data sets were made. Training using the YOLOv4 model, 4500 were randomly selected for training and 500 remaining for outcome testing.

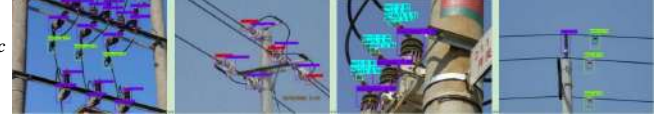
**5.1.2 YOLOv4Model output results.** The training parameters are set as follows: learning rate is set to 0.001, the number of Batchsize is set to 32, and the training is stopped after 800 epochs. The training results are shown in Table1 below.

**Table 1: Sample test accuracy table**

Classification	accuracy	recall
Ground ring	94.51%	94.91%
insulator	90.72%	90.26%
clamp	92.29%	92.91%
bolt	92.29%	92.91%

Accuracy refers to the proportion of all targets that contain the true target detected in the test results. Recall is the proportion of the detected true target up to all the true targets. It can be seen from the test accuracy table that the detection accuracy of the algorithm on the picture accuracy of the typical parts of the distribution network is higher than 85%, which meets the requirements of the robot identification and positioning during the installation of the distribution network grounding line. The weight files generated after

training and the configuration files used for training are extracted, and the trained model is loaded in the control program. The final network identification training results are shown in Figure 3 below:



**Figure 3: Inspection effect diagram of typical components of distribution network lines.**

### 5.2 Lead segmentation experiment of the distribution network

In this project, the Res-Unet model has been used as the basis to segment the drainage line of the distribution wire. The 5000 images after target detection are taken as the data set, and the data set is manually annotated with the annotation tool Labelme. The data sets in PNG format were converted into. After json file, 90% of images are taken as training sets and 10% of images as test set; load training set into detection model training and select 400 images for input to Res-Unet network for testing. The segmented images of the distribution wire and drainage line are shown in Figure 4.



**Figure 4: Separation result diagram of distribution network wire and drainage line.**

### 5.3 Lead operation point positioning experiment

This project simulates the measurement of the outdoor straight wire by measuring the black round pipe with the same diameter as the real 10 kV wire. The experiment transforms the measurement of the absolute position of the wire operation point into the distance between the wire operation point and the specified typical part (insulator). The experiment uses the coordinate value of the insulator and the coordinate system of the camera to calculate the position of the insulator and compare the accuracy of the algorithm. The results are shown in Table2, the method has achieved accuracy within 1mm in the x and y directions, and the accuracy is slightly worse within 3mm in the z-direction.

## 6 CONCLUSION

A distribution network drainage line positioning method is presented in this paper based on neural network and binocular vision. The improved YOLOv4 target detection network is used to realize the recognition rate of more than 85% of typical components of the distribution network; Res-Unet segmentation network is used to avoid background interference and improve the robustness of the

**Table 2: Lead work point location result**

Order	Measured moving values( $\Delta x, \Delta y, \Delta z$ )/mm	Measured distance values/cm	Actual distance value/cm	Error/cm
1	(-9.66, 2.55, 0.56)	9.77	10	0.23
2	(-19.63, 2.18, 0.12)	19.72	20	0.28
3	(29.68, -4.28, 1.49)	30.02	30	0.02
4	(1.01, 3.24, 40.79)	40.97	40	0.97
5	(0.58, -3.51, 50.20)	50.38	50	0.38
6	(0.10, -4.53, 60.77)	60.33	60	0.33
7	(3.62, -69.64, 0.01)	69.86	70	0.14
8	(2.78, -79.63, 0.25)	79.76	80	0.46

wire recognition and positioning algorithm; by fitting the image geometric moment to the wire pixels, the positioning accuracy can reach 1mm and 3mm in the xy direction, and realize the operation accuracy within 1cm in combination with the control error of the mechanical arm.

In general, the method proposed in this paper realizes high-precision wire identification and positioning, which can provide high-precision visual positioning for the live operation robot of the distribution network, and can also provide visual guidance for other outdoor wire operation target robots.

## REFERENCES

- [1] Azzedine Boukerche and Zhijun Hou. 2021. Object detection using deep learning methods in traffic scenarios. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–35.
- [2] Jingwen Chen, Xin Xu, and Hongshe Dang. 2019. Fault detection of insulators using second-order fully convolutional network model. *Mathematical Problems in Engineering* 2019 (2019).
- [3] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 1. Ieee, 886–893.
- [4] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* 32, 9 (2009), 1627–1645.
- [5] Amr Ibrahim, Ahmad Dalbah, Ayaat Abualsaud, Usman Tariq, and Ayman El-Hag. 2019. Application of machine learning to evaluate insulator surface erosion. *IEEE Transactions on Instrumentation and Measurement* 69, 2 (2019), 314–316.
- [6] W. Jiang, G. C. Ye, D. Zou, A. Zhang, G. Zuo, Y. Yan, and H. Li. 2021. Dynamic model based energy consumption optimal motion planning for high-voltage transmission line mobile robot manipulator:. *Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics* 235, 1 (2021), 93–105.
- [7] Xinyu Liu, Hao Jiang, Jing Chen, Junjie Chen, Shengbin Zhuang, and Xiren Miao. 2018. Insulator detection in aerial images based on faster regions with convolutional neural network. In *2018 IEEE 14th international conference on control and automation (ICCA)*. IEEE, 1082–1086.
- [8] Navin Prakash and Yashpal Singh. 2015. Fuzzy support vector machines for face recognition: A review. *International Journal of Computer Applications* 131, 3 (2015), 24–26.
- [9] Yincheng Qi, Lei Xu, Zhenbing Zhao, and Yiping Cai. 2015. A Cosegmentation Method for Aerial Insulator Images. In *Advances in Image and Graphics Technologies: 10th Chinese Conference, IGTA 2015, Beijing, China, June 19-20, 2015, Proceedings 10*. Springer, 113–122.
- [10] Zhi Ming Qiu, M. A. Yan, Xiang Yao Meng, Jian Hua Chen, and Wei Feng. 2023. Analysis on the Development Trend and Key Technologies of Unmanned Underwater Equipment. *journal of unmanned undersea systems* 31, 1 (2023), 1–9.
- [11] Hamed Sarbolandi, Damien Lefloch, and Andreas Kolb. 2015. Kinect range sensing: Structured-light versus Time-of-Flight Kinect. *Computer vision and image understanding* 139 (2015), 1–20.
- [12] Wang Wanguo, Wang Zhenli, Liu Bin, Yang Yuechen, and Sun Xiaobin. 2019. Typical defect detection technology of transmission line based on deep learning. In *2019 Chinese Automation Congress (CAC)*. IEEE, 1185–1189.
- [13] Guobing Yan, Qiang Sun, Jianying Huang, and Yonghong Chen. 2021. Helmet detection based on deep learning and random forest on UAV for power construction safety. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 25, 1 (2021), 40–49.
- [14] Jing Zhang, Guofang Huang, Xiaoming Liu, Chao Shan, Wenzheng Wang, Yuhui Tong, Xin Ning, and Congli Li. 2021. Intelligent Perception Method for Distribution Network Live Working Robot. In *2021 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia)*. IEEE, 1457–1462.
- [15] Zhenbing Zhao, Zhen Zhen, Lei Zhang, Yincheng Qi, Yinghui Kong, and Ke Zhang. 2019. Insulator detection method in inspection image based on improved faster R-CNN. *Energies* 12, 7 (2019), 1204.

# KRE: A Key-retained Random Erasing Method for Occluded Person Re-identification

HongXia Wang  
whx\_green@whut.edu.cn  
School of Computer Science and  
Technology, Wuhan University of  
Technology  
Wuhan, China

Yao Ma  
School of Computer Science and  
Technology, Wuhan University of  
Technology  
Wuhan, China  
1074759965@qq.com

Xiang Chen  
School of Computer Science and  
Technology, Wuhan University of  
Technology  
Wuhan, China  
417545906@qq.com

## ABSTRACT

Occluded person re-identification (ReID) is a challenging task in the field of computer vision, facing the problem that the target pedestrians in probe images are obscured by various occlusions. Random Erasing in data augmentation techniques is one of the effective methods used to deal with the occlusion problem, but it may introduce noise into the training process, which affects the training of the model. In order to solve this problem, we propose an novel data augmentation method named Key-retained Random Erasing (KRE) which preserves the critical parts in images for occluded person ReID. Based on the regular Random Erasing, we utilize the naturally generated attention map in Vision Transformers and introduce an adaptive threshold selection method to detect the key areas of the image to be augmented. The complexity of the training samples can be improved without losing the key information of the images by reserving the key areas in Random Erasing process, which can finally alleviate the occluded person ReID problem. Validating the proposed method on occluded, partial and holistic ReID datasets, extensive experimental results demonstrate that our method performs favorably against state-of-the-art methods on ViT-based models.

## CCS CONCEPTS

• Computing methodologies → Object identification.

## KEYWORDS

occluded person re-identification, data augmentation, vision transformer, self attention

## ACM Reference Format:

HongXia Wang, Yao Ma, and Xiang Chen. 2023. KRE: A Key-retained Random Erasing Method for Occluded Person Re-identification. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590089>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China  
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590089>

## 1 INTRODUCTION

Person re-identification (ReID) [34, 37] is to retrieve a probe pedestrian of interest from multiple non-overlapping camera views. It is an important subject in the field of computer vision, which combined with pedestrian detection and tracking for a wide range of applications, such as video surveillance, intelligent security and criminal investigation.

In recent years, person ReID have attracted more and more attention from the academic community, and with it, a series of methods have been proposed [14, 18, 20, 27, 31, 39]. However, most of these methods assume that the target pedestrians are holistic and unobstructed. They only perform well on holistic datasets, but not when dealing with occluded datasets. This is because the assumption cannot be satisfied in realistic scenarios, such as airports, hospitals and shopping malls, where pedestrians may be occluded by objects such as trees, cars, barricades or even occluded by other pedestrians, as shown in Figure 1. Therefore, it is necessary to study the occluded ReID problem.

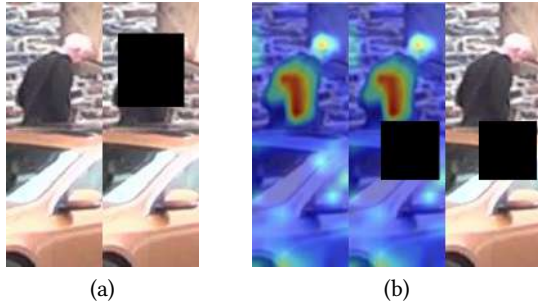


**Figure 1: Illustration of occluded person ReID. The red bounding box indicates the target person occluded by various obstacles or other pedestrians.**

A simple solution is to apply data augmentation during the training stage, of which Random Erasing [40] has proven to be useful for the occluded problem. Random Erasing takes as input an original image, the erasing probability, the erasing area ratio range, and the erasing aspect ratio range, and as output an erased image. After Random Erasing, part of the original image is covered by a rectangular box with random values to indicate that the image is obscured.

Although Random Erasing specifies the size and shape of the erased rectangular area, it does not specify the location of the erased area in the image. It may result in erasing key areas in the image, where key areas refer to key parts of the target pedestrian, such as the head, torso, etc. If the target pedestrian in the image is erased

without the corresponding labels being changed, noise is introduced into the training set, affecting model learning, as shown in Figure 2(a).



**Figure 2: Examples of two Random Erasing methods. (a) shows the drawback of the regular Random Erasing, i.e. the tendency to erase key areas of the image. (b) shows Key-retained Random Erasing, which uses the attention map in Vision Transformers for key area protection and can effectively reduce noise enhancement samples.**

In this paper, we investigate how to avoid erasing critical regions in an image, which requires calculating a keyness score for each region in the image. We find that the naturally generated attention map in Vision Transformers is perfect for this task. Simply using a sliding window approach, the sum of the weights of all pixels in each region of the attention map is calculated as the keyness score. Next, we adaptively select a key area threshold by sorting the keyness scores. Then, regions with a keyness score larger than the threshold are called key areas and are prohibited from being erased.

In conclusion, we propose the Key-retained Random Erasing (KRE) data augmentation method for occluded ReID tasks. Extensive experimental results on occluded datasets demonstrate that the proposed method is effective and performs favorably against the state-of-the-art methods. In addition, KRE also achieves competitive performance on holistic and partial datasets.

## 2 RELATED WORK

### 2.1 Occluded Person Re-identification

Compared to holistic person ReID, occluded person ReID is more challenging and relevant due to the lack of target information and the introduction of additional interference. Existing methods can be broadly divided into methods based on hand-craft splitting, methods introducing semantic information, and methods based on transformer. Hand-craft splitting based methods can learn discriminative features in a straightforward way. Sun et al. [27] propose a network named Part-based Convolution Baseline (PCB) which partition feature maps into 6 horizontal stripe to learn local features. Fan et al. [3] use a Spatial Channel Parallelism Network (SCPNet) to encode local information into global features by designing the SCP loss function. Extra semantic information generally refers to pose information. Miao et al. [23] introduce Pose-Guided Feature Alignment (PGFA) that utilizes pose landmarks to learn local features not disturbed by occlusion noise. Gao et al. [4] learn discriminative part

features in an end-to-end framework by combining pose-guided attentions and the self-mining part visibility. Wang et al. [30] firstly introduce high-order information into occluded person ReID, making full use of high-order relation and human-topology information to learn robust features. Methods based on transformer have recently received increasing attention from researchers. Li et al. [19] first propose Part-Aware Transformer (PAT) for occluded person ReID, which exploits the transformer encoder-decoder architecture to achieve diverse part discovery in a unified deep model.

### 2.2 Vision Transformers

Transformer [29] has been widely used in the field of natural language processing. Inspired by its powerful representation capabilities, researchers are looking at ways to apply Transformer to computer vision tasks. Dosovitskiy et al. [2] propose Vision Transformer (ViT) which reshapes a image into a sequence of flattened 2D patches as the input of the Transformer encoder. ViT achieves state-of-the-art performance on multiple image recognition benchmarks. However, ViT lacks the inductive bias of CNNs so that it requires a large amount of data for training. To address this problem, Touvron et al. [28] introduce a new token-based distillation procedure for training. Due to the excellent performance of ViT, more and more computer vision tasks, such as object detection, image segmentation, are using transformer-based models. For example, He et al. [13] is the first one to propose a transformer-based model named TransReID for object ReID. They design the SIE module and JPM module to encode various kinds of side information and learn more robust feature representation, respectively.

### 2.3 Data Augmentation

Data augmentation is a technique of increasing the amount of training data by adding slightly modified copies of already existing data. Although data augmentation can solve the problem of insufficient data, there is a risk of introducing noise and ambiguity when left uncontrolled [7, 33]. Chen et al. [1] propose TransMix to bridge the gap between the input and label spaces of mixup-based augmentation methods. Gong et al. [6] propose KeepAugment for both Cutout and AutoAugment, which can detect important regions through saliency map and preserve these regions during augmentation to increase the fidelity of data augmentation. Inspired by this, we propose a novel attention-based data augmentation method KRE, which differs from KeepAugment in that it is based on random erasing commonly used in occluded person ReID and we determine the key areas according to the attention map naturally generated in ViTs and the adaptive threshold selection method proposed in this paper. It is noteworthy that since the attention map is naturally generated in ViT-based models, our method can be merged into the training pipeline with no additional parameters and minimal computation overhead.

## 3 THE PROPOSED METHOD

### 3.1 Background

**Random Erasing** Random Erasing [40] is a simple and effective data augmentation method. For an image  $I$  with area  $S = H \times W$ , the probability of it undergoing Random Erasing is set to  $p$ , and the probability of it being kept unchanged is  $1 - p$ . Assume that the area

of the randomly initialized erasing rectangle region  $I_e$  is  $S_e$ , where  $\frac{S_e}{S}$  is in range specified by minimum  $s_l$  and maximum  $s_h$ , and the aspect ratio of  $I_e$  which set to  $r_e$  is between  $r_1$  and  $r_2$ . Then the size of  $I_e$  is  $H_e = \sqrt{S_e \times r_e}$  and  $W_e = \sqrt{\frac{S_e}{r_e}}$ . For the point  $P = (x_e, y_e)$  randomly initialized in  $I$ , if  $x_e + W_e \leq W$  and  $y_e + H_e \leq H$ , we set the region,  $I_e = (x_e, y_e, x_e + W_e, y_e + H_e)$ , as the selected rectangle region. Otherwise repeat the above process until an appropriate  $I_e$  is selected. The Random Erasing process is completed by assigning each pixel in  $I_e$  to a random value in  $[0, 255]$ , respectively.

**Multi-head Self-attention** Self-attention [29] exploits attention mechanisms to dynamically generate different connection weights in order to process long sequences. Specifically, given an input matrix  $\mathbf{x} \in \mathbb{R}^{N \times d}$ , the query matrix  $Q$ , the key matrix  $K$ , and the value matrix  $V$  are obtained by linearly projecting  $\mathbf{x}$ , which can be expressed by:

$$\begin{aligned} Q &= W^Q \mathbf{x} \\ K &= W^K \mathbf{x} \\ V &= W^V \mathbf{x} \end{aligned} \quad (1)$$

Where  $N$  is the number of tokens,  $d$  is dimension of each token, the projections are parameter matrices  $W^Q \in \mathbb{R}^{d \times d_k}$ ,  $W^K \in \mathbb{R}^{d \times d_k}$ , and  $W^V \in \mathbb{R}^{d \times d_v}$ . Then, we scale the dot products of  $Q$  and  $K$  and apply the softmax function to obtain the attention map, which can be formulated as below:

$$A(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (2)$$

If the attention map is then multiplied by the corresponding value matrix as weights, the weighted sum is the output of the self-attention operation. Formally,

$$\text{Attention}(Q, K, V) = A(Q, K)V \quad (3)$$

In order to allow the model to jointly attend to information from different representation subspaces at different positions, single-head self-attention can be extended to multi-head self-attention. Given a Transformer with  $h$  attention heads, the dimension of each head is  $\frac{d}{h}$ .

### 3.2 Key-retained Random Erasing

**Random Erasing Based on Key Area Preservation** We propose Key-retained Random Erasing (KRE) which is a novel Random Erasing data augmentation method with key area preservation based on attention map. By detecting key areas in images before Random Erasing using the attention map automatically generated in ViTs, we ensure that the critical regions of the images are not erased.

Specifically, for each image in training, we simply extract the attention map in the last layer of the model and average across all attention heads. Then, a randomly selected rectangular region in the attention map with the required area size is used to calculate the keyness score  $keyScore$  of the region as follows:

$$keyScore = \text{sum}(A[x : x + H_e, y : y + W_e]) \quad (4)$$

We use  $keyScoreList$  to count the keyness scores of all regions with required size in the attention map, and sort these keyness

scores from smallest to largest to adaptively select the appropriate key area threshold.

With the key area threshold constraint, it can be ensured that no key areas are selected in the process of randomly selecting erasure areas. For example, if the keyness score of the selected region to be erased is greater than the key area threshold, which means the region is the key area of the image, we will discard the region and search for the next region until a region with a keyness score less than or equal to key area threshold is found. With the selected non-critical region, we assign each pixel in the region to a random value in  $[0, 255]$ , respectively, and return the erased image to the model for training. Figure 2(b) visualises the algorithmic idea of our method.

**Adaptive Threshold Selection** We also propose an adaptive selection method for key area threshold. Unlike the direct determination of the threshold as a specific value, we introduce the concept of key area fraction  $\lambda$ , which can be used to determine the index of key area threshold by dividing the length of  $keyScoreList$ . Formally,

$$keyThreshold = keyScoreList[\text{len}(keyScoreList) \cdot \lambda] \quad (5)$$

Specifically, we set the key area fraction  $\lambda$  to 0.6, i.e. the sorted keyness score list is divided in a ratio of 3:2. Then, the index of the list at the division is the index of key area threshold. Therefore, we can determine different thresholds based on different indexes, as different images have different  $keyScoreList$ .

The adaptive threshold selection method is able to select the most suitable key area threshold for each image, in terms of the ratio of non-critical regions to critical regions. Thus, instead of setting the threshold for all images to a uniform value, each image has its own key area threshold, which can effectively use the context information to identify the key areas in the image, and is more flexible.

### 3.3 Pseudo-code

Algorithm 1 shows in detail the whole procedure of our method.

## 4 EXPERIMENTS

### 4.1 Datasets and Settings

**Datasets** To demonstrate the efficacy of our method on the occlusion problem, we evaluate our proposed method on two occluded datasets: Occluded-Duke [23], Occluded-REID [41], one partial dataset: Partial-REID [38], and two holistic datasets Market-1501 [36], DukeMTMC-reID [24].

1) Occluded-Duke [23] is a subset of DukeMTMC-reID [24] dataset, containing 15618 training images, 2210 occluded query images and 17661 gallery images.

2) Occluded-REID [41] is captured by mobile camera on campus, which contains 2000 images of 200 pedestrians, each of which has 5 images of the holistic body and 5 images of varying degrees of occlusion, respectively.

3) Partial-REID [38] contains 600 images of 60 pedestrians, each with 5 holistic images and 5 partial images respectively, and is the first dataset for partial ReID.

4) Market-1501 [36] contains 32668 holistic images of 1501 pedestrians that were collected from six camera views, with 12936 images

**Algorithm 1** Key-retained Random Erasing Procedure

---

**Input:** input image  $I$   
size of the input image  $H$  and  $W$   
erasing probability  $p$   
erasing area ratio range  $s_l$  and  $s_h$   
erasing aspect ratio range  $r_1$  and  $r_2$   
key area fraction  $\lambda$

**Output:** erased image after key areas are retained  $I^*$

```

1: Initialization:  $p_1 \leftarrow \text{Rand}(0, 1)$ .
2: if  $p_1 \geq p$  then
3:    $I^* \leftarrow I$ ;
4:   return  $I^*$ .
5: else
6:   while True do
7:      $S_e \leftarrow \text{Rand}(s_l, s_h) \times H \times W$ ,  $r_e \leftarrow \text{Rand}(r_1, r_2)$ ;
8:      $H_e \leftarrow \sqrt{S_e \times r_e}$ ,  $W_e \leftarrow \sqrt{\frac{S_e}{r_e}}$ ;
9:      $A = \text{model}(I)$ ;
10:    for  $x = 0$  to  $H - H_e$ 
11:      for  $y = 0$  to  $W - W_e$ 
12:         $\text{keyScoreList}[i++] \leftarrow$ 
13:         $\text{sum}(A[x : x + H_e, y : y + W_e])$ ;
14:      end for
15:    end for
16:     $\text{sort}(\text{keyScoreList})$ ;
17:     $\text{keyThreshold} \leftarrow \text{keyScoreList}[\text{len}(\text{keyScoreList}) \cdot \lambda]$ ;
18:    while True do
19:       $x_e \leftarrow \text{Rand}(0, H - H_e)$ ,  $y_e \leftarrow \text{Rand}(0, W - W_e)$ ;
20:       $\text{keyScore} \leftarrow \text{sum}(A[x_e : x_e + H_e, y_e : y_e + W_e])$ ;
21:      if  $\text{keyScore} > \text{keyThreshold}$  then
22:        continue.
23:      else
24:         $I_e \leftarrow (x_e, y_e, x_e + H_e, y_e + W_e)$ ;
25:         $I(I_e) \leftarrow \text{Rand}(0, 255)$ ;
26:         $I^* \leftarrow I$ ;
27:        return  $I^*$ .
28:      end while
29:    end while
30:  end if

```

---

of 751 pedestrians used for training and 19732 images of 750 pedestrians used for testing.

5) DukeMTMC-reID [24] contains 36411 images of 1404 pedestrians, which were captured in 8 camera views. Of these, 16522 images were used as training images, 17661 images were used as gallery images, and 2228 images were used as query images.

**Evaluation Metrics** We adopt Cumulative Matching Characteristic (CMC) [8] curves and Mean average precision (mAP) [36], commonly used in the field of person ReID, to evaluate the quality of different ReID models. All the experiments results are performed in a single query setting.

**Implementation Details** As the attention map is naturally generated in the ViT-based models, we adopt TransReID [13] as the baseline for our approach. Our experiments are conducted on NVIDIA 2080Ti GPU environment and Pytorch platform. As with the setting in TransReID, the data augmentation operations

of random horizontal flipping, random cropping, random erasing, and padding 10 pixels were used on the training set. The input images are resized to  $256 \times 128$ , and batch size is set to 32. SGD optimizer is employed with the weight decay of  $1e-4$ . The training epoch number is set to 120 with the initial learning rate 0.008 and use the cosine learning rate decay strategy.

## 4.2 Experimental Results

We compare our method with other state-of-the-art methods on occluded datasets, partial datasets and holistic datasets.

**Results on Occluded Datasets** We conduct experiments on two occluded datasets: Occluded-Duke and Occlude-REID. The experimental results are shown in Table 1. We list the results of three mainstream methods and the work of our method for comparison. The methods in the first group are proposed for the holistic person ReID problem, including Part Aligned [35], Adver Occluded [15] and PCB [27]. The models in the second group are designed for occluded ReID, including Part Bilinear [25], FD-GAN [5], PGFA [23], and HOREID [30], Mos [16] and SRNet [32]. The approaches in the third group are transformer based person ReID methods, including PAT [19] and TransReID [13]. As can be seen from the data in the table, the holistic person ReID methods do not perform very well on the two occluded datasets, for example, PCB has a 26.5% difference in Rank-1 and 27.2% difference in mAP from our method on Occluded-Duke dataset. The performance of occluded person ReID methods was similar to that of the holistic person ReID methods when they were first proposed, and has steadily improved in recent years as more and more methods have been proposed, but is still not as superior as Transformer-based methods. Our method achieves 69.1% Rank-1 and 60.9% mAP on Occluded-Duke dataset, outperforming the other methods on the table.

**Table 1: Comparison with state-of-the-arts on two occluded datasets: Occluded-Duke and Occluded-REID.**

Methods	Occluded-Duke		Occluded-REID	
	Rank-1	mAP	Rank-1	mAP
Part Aligned [35]	28.8	20.2	-	-
Adver Occluded [15]	44.5	32.2	-	-
PCB [27]	42.6	33.7	41.3	38.9
Part Bilinear [25]	36.9	-	-	-
FD-GAN [5]	40.8	-	-	-
PGFA [23]	51.4	37.3	-	-
HOREID [30]	55.1	43.8	80.3	70.2
Mos [16]	61.0	49.2	-	-
SRNet [32]	65.5	52.7	80.6	72.4
PAT [19]	64.5	53.6	<b>81.6</b>	72.1
TransReID [13]	66.4	59.2	-	-
KRE(Ours)	<b>69.1</b>	<b>60.9</b>	73.3	<b>72.4</b>

**Results on Partial Datasets** To evaluate our method even further, we compared it with other methods [4, 10–12, 21, 23, 26, 38] on one partial dataset: Partial-REID. The experimental results are shown in Table 2. Partial datasets differs from occluded datasets in that its images are at different scales due to cropping, which may

**Table 2: Comparison with state-of-the-arts on one partial datasets: Partial-REID.**

Methods	Partial-REID	
	Rank-1	Rank-3
MTRC [21]	23.7	27.3
AMC+SWM [38]	37.3	46.0
DSR [10]	50.7	70.0
SFR [11]	56.9	78.5
VPM [26]	67.7	81.9
PGFA [23]	68.0	80.0
PVPM [4]	78.3	<b>87.7</b>
FPR [12]	<b>81.0</b>	-
KRE(Ours)	80.3	85.0

lead to misalignment. And partial datasets contains little information about the occlusion. As a result, our method did not perform as well on Partial-REID dataset compared to the previous two occluded datasets. But our method also achieved similar performance compared to other mainstream methods. Specifically, Rank-1 and Rank-3 of our method on Partial-REID is 80.3% and 85.0%, respectively. Compared to PVPM [4], with Rank-1 outperforming by 2.0%, although Rank-3 differed by 2.7%.

**Results on Holistic Datasets** We hope that our proposed method will not only perform well on occluded problem, but also show generalizability and robustness on the holistic datasets. Therefore, we conduct experiments on two classical holistic datasets: Market-1501 and DukeMTMC-reID, as shown in Table 3.

We divide the compared methods into three categories. The methods in first group are hand-crafted splitting based models, including PCB [27], BOT [22] and MGN [31]; the methods in second group are based on extra semantic, including SPReID [17],  $P^2$ Net [9], PGFA [23] and HOREID [30]; the methods in third group are transformer based models, including PAT [19] and TransReID [13]. According to the experimental data in the table, we can see that in scenes without extensive occlusion, the performance of the extra semantic based methods is similar to that of the hand-crafted splitting based models, due to the fact that the former relies heavily on extra cues for feature alignment, but does not play a significant role in the holistic datasets. The transformer based methods have better performance than either of the previous two kinds of methods. Because they are able to use attention to capture global contextual information, thereby establishing long-range dependence on the target and extracting more robust features. Whereas our method outperforms all the other methods in the table. Specifically, the Rank-1/mAP of our method achieves 95.4%/89.4% and 90.6%/82.1% on Market-1501 and DukeMTMC-reID datasets, respectively, reaching state-of-the-art performance and demonstrating the continued validity of our method on the holistic datasets.

### 4.3 Analysis of Threshold Selection Methods

In this section, we analyze the impact of the threshold selection methods on our approach. The experimental results are shown in Table 4. We compare the impact on experimental results produced by thresholds with fixed values and the threshold selected by the adaptive threshold selection method on Occluded-Duke dataset,

**Table 3: Comparison with state-of-the-arts on two holistic datasets: Market-1501 and DukeMTMC-reID.**

Methods	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
PCB [27]	92.3	77.4	81.8	66.1
BOT [22]	94.1	85.7	86.4	76.4
MGN [31]	95.7	86.9	88.7	78.4
SPReID [17]	92.5	81.3	84.4	71.0
$P^2$ Net [9]	95.2	85.6	86.5	73.1
PGFA [23]	91.2	76.8	82.6	65.5
HOREID [30]	94.2	84.9	86.9	75.6
PAT [19]	95.4	88.0	88.8	78.2
TransReID [13]	95.2	88.9	<b>90.7</b>	82.0
KRE(Ours)	<b>95.4</b>	<b>89.4</b>	90.6	<b>82.1</b>

and it can be intuitively seen that the latter gives a better performance of our method. Compared to the optimal performance of 0.6 for fixed values, the threshold selected by the adaptive threshold selection method achieves a 0.5% improvement on mAP and a 1.3% improvement on Rank-1, which leads to the conclusion that selecting different thresholds for different images better preserves information in key areas.

**Table 4: The results of the impact of the threshold selection methods on our approach.**

Threshold	Occluded-Duke	
	mAP	Rank-1
0.4	58.9	65.5
0.6	60.4	67.8
0.8	59.3	65.9
KRE(Ours)	<b>60.9</b>	<b>69.1</b>

### 4.4 Ablation Study

In this section, we analyze the effectiveness of the proposed method through ablation experiments. The ablation study results are shown in Table 5. As can be seen from the data in the table, the traditional Random Erasing is able to improve the performance of the TransReID baseline on Occluded-Duke dataset. Specifically, mAP improves by 2.8%, but Rank-1 decreases by 0.6%. In contrast, KRE outperforms the regular Random Erasing, with not only a 4.5% improvement in mAP but also a 2.1% improvement in Rank-1, indicating that the method based on key area preservation facilitates

**Table 5: Results of ablation study for KRE on Occluded-Duke dataset.**

Methods	Occluded-Duke	
	mAP	Rank-1
Baseline	56.4	67.0
Baseline + REA	59.2	66.4
Baseline + KREA	<b>60.9</b>	<b>69.1</b>

the model to identify easy samples in the gallery images, resulting in an improvement in both metrics. This shows that our proposed method is important for the performance improvement of occluded person ReID and proves the feasibility and validity of the research in this paper.

## 5 CONCLUSION

In this paper, we propose a simple yet effective data augmentation method named Key-retained Random Erasing (KRE) to address the problem that invalid samples generated by traditional Random Erasing are detrimental to model training for occluded ReID. We utilize the attention map naturally generated in Vision Transformers to calculate keyness score of each region in an image, and then set the key area threshold to retain the information of critical regions by the adaptive threshold selection method. The 1.7% improvement in mAP on the Occluded-Duke dataset verifies the effectiveness of KRE. Moreover, the experimental results on the partial and holistic datasets also demonstrate the robustness and generalizability of our proposed method. However, there is still a certain gap between the method and the actual landing application. Therefore, how to transform the existing theoretical methods into practical applications is still a subject to be studied at present.

## REFERENCES

- [1] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. 2022. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12135–12144.
- [2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [3] Xing Fan, Hao Luo, Xuan Zhang, Lingxiao He, Chi Zhang, and Wei Jiang. 2019. Sepnet: Spatial-channel parallelism network for joint holistic and partial person re-identification. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II*. Springer, 19–34.
- [4] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. 2020. Pose-guided visible part matching for occluded person ReID. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11744–11752.
- [5] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. 2018. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. *Advances in neural information processing systems* 31 (2018).
- [6] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. 2021. Keppaugmant: A simple information-preserving data augmentation approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1055–1064.
- [7] Raphael Gontijo-Lopes, Sylvia J Smullin, Ekin D Cubuk, and Ethan Dyer. 2020. Affinity and diversity: Quantifying mechanisms of data augmentation. *arXiv preprint arXiv:2002.08973* (2020).
- [8] Douglas Gray, Shane Brennan, and Hai Tao. 2007. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, Vol. 3. 1–7.
- [9] Jianyuan Guo, Yuhui Yuan, Lang Huang, Chao Zhang, Jin-Ge Yao, and Kai Han. 2019. Beyond human parts: Dual part-aligned representations for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3642–3651.
- [10] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. 2018. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7073–7082.
- [11] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. 2018. Recognizing partial biometric patterns. *arXiv preprint arXiv:1810.07399* (2018).
- [12] Lingxiao He, Yinggang Wang, Wu Liu, He Zhao, Zhenan Sun, and Jiashi Feng. 2019. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8450–8459.
- [13] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 15013–15022.
- [14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).
- [15] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. 2018. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5098–5107.
- [16] Mengxi Jia, Xinhua Cheng, Yungpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. 2021. Matching on sets: Conquer occluded person re-identification without alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1673–1681.
- [17] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. 2018. Human semantic parsing for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1062–1071.
- [18] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2285–2294.
- [19] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. 2021. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2898–2907.
- [20] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. 2015. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2197–2206.
- [21] Shengcai Liao, Anil K Jain, and Stan Z Li. 2012. Partial face recognition: Alignment-free approach. *IEEE Transactions on pattern analysis and machine intelligence* 35, 5 (2012), 1193–1205.
- [22] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 0–0.
- [23] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. 2019. Pose-guided feature alignment for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 542–551.
- [24] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II*. Springer, 17–35.
- [25] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*. 402–419.
- [26] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. 2019. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 393–402.
- [27] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*. 480–496.
- [28] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*. PMLR, 10347–10357.
- [29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [30] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6449–6458.
- [31] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*. 274–282.
- [32] HongXia Wang, Xiang Chen, and Chun Liu. 2021. Pose-guided part matching network via shrinking and reweighting for occluded person re-identification. *Image and Vision Computing* 111 (2021), 104186.
- [33] Longhui Wei, An Xiao, Lingxi Xie, Xiaopeng Zhang, Xin Chen, and Qi Tian. 2020. Circumventing outliers of autoaugment with knowledge distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*. Springer, 608–625.
- [34] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence* 44, 6 (2021), 2872–2893.
- [35] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. 2017. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 3219–3228.

- [36] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.
- [37] Liang Zheng, Yi Yang, and Alexander G Hauptmann. 2016. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984* (2016).
- [38] Wei-Shi Zheng, Xiang Li, Tao Xiang, Shengcai Liao, Jianhuang Lai, and Shao-gang Gong. 2015. Partial person re-identification. In *Proceedings of the IEEE international conference on computer vision*. 4678–4686.
- [39] Zhedong Zheng, Liang Zheng, and Yi Yang. 2018. Pedestrian alignment network for large-scale person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 10 (2018), 3037–3045.
- [40] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13001–13008.
- [41] Jiaxuan Zhuo, Zeyu Chen, Jianhuang Lai, and Guangcong Wang. 2018. Occluded person re-identification. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

# Digital image denoising by partial differential equation based on P-M model and its fuzzy evaluation method system

Jingying, L, and Liu\*

The Fifth Institute of Electronics, Ministry of Industry and  
Information Technology  
1099098464@qq.com

Yang, H, and Hu

The Fifth Institute of Electronics, Ministry of Industry and  
Information Technology  
274830594@qq.com

## ABSTRACT

Aiming at the problems of storage, batch migration and centralized processing of visual digital images of infrared imaging products, this paper takes digital image noise reduction as the main research object and starts with the concept of image partial differential equation processing. Based on the development history, advantages, practicability and operability of digital image processing by partial differential equation, it is concluded that digital image processing technology based on P-M model method is more suitable for modern image processing, and also broadens and improves the basic algorithm of digital image processing in the past. On this basis, the image quality is evaluated by using the fuzzy comprehensive evaluation theory based on analytic hierarchy process. The results show that the optimized processing system can screen the advantages and disadvantages of visual digital images of infrared imaging products and provide technical support.

## CCS CONCEPTS

• Picture Processing, Fuzzy comprehensive evaluation;

## KEYWORDS

Digital image, Partial differential equation, Image denoising, Fuzzy comprehensive evaluation, Hierarchical analysis

### ACM Reference Format:

Jingying, L, and Liu and Yang, H, and Hu. 2023. Digital image denoising by partial differential equation based on P-M model and its fuzzy evaluation method system. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590091>

## 1 INTRODUCTION

Image is the general term of two-dimensional image and three-dimensional image, which is a material representation of human visual perception [1]. Image has become an indispensable way to obtain, interpret and re-develop information in modern life. In the

\*Jingying Liu(1996-), female, assistant engineer, Master degree in reliability, engaged in reliability and systems engineering research. Yang Hu (1990-), male, engineer, Master degree in reliability, engaged in reliability and systems engineering research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590091>

course of the continuous development of modern society, images can be widely obtained from various optical equipment. Among them, digital image plays a particularly important role in the field of computer science, optics and flight driving because of its unique reliability, authenticity and high color saturation. At present, image processing technology is becoming more mature and widely used in all kinds of daily life.

With the continuous development of information technology, in order to enhance the practicability of image information processing, digital image processing is formed. Through computer or other hardware digital processing of image information, including image cutting, noise reduction, enhancement, detection and many other aspects.

The development history of digital image processing in China can be traced back to the 1960s, when medium and large equipment was used to carry out prime raster scanning display of images. At this stage, the cost of image storage was high and the application scope was narrow. In 1970, a large number of small and medium machines were used for batch processing of images, which was changed to raster scanning mode, which reduced the problem of high cost in the past and also improved the processing rate. After 1980, with the appearance of satellite remote sensing technology, image information processing by microcomputer using ultra-large integrated circuit is gradually popularized, which greatly promotes the innovation of digital image processing. Since 1990, with different application fields, image processing has gradually divided into different architectures and introduced new theories and algorithms. In recent years, artificial neural networks, fractal and other methods have become a research hotspot in digital image processing [2].

The partial differential equation method studied in this paper is a common method of digital image processing, which has experienced the development of linear diffusion, nonlinear diffusion, and then anisotropic nonlinear diffusion. After comparing various traditional image denoising processing methods, this paper focuses on the image denoising based on P-M model. The essence of the digital image processing method is to process discrete data into a continuous mathematical model that can be concluded. By analyzing and comparing a large number of data, the optimal solution is iteratively calculated and the processing results are obtained to ensure that the image quality reaches the optimal value.

## 2 APPLICATION OF PARTIAL DIFFERENTIAL EQUATION IN DIGITAL IMAGE DENOISING

This chapter introduces the noise model of digital image and the use model of partial differential equation to remove noise.



Figure 1: Finite difference method number line

## 2.1 Digital image noise model

The difference between the image and the real image is called noise. The sources of digital image noise are diversified and unavoidable, which can only be improved by technical means. Among them, military airborne products infrared imaging visualization digital image common noise sources such as temperature, heat, turbulence, etc [3].

Noise is an unpredictable random signal, which can be considered as a random variable represented by probability density function. Therefore, digital image noise can generally be modeled as an additive noise term after the image is processed by a transmission function. Let's say the original graph is  $w(x,y)$ , then the image  $W$  to be processed should be:

$$W(x, y) = h(x, y) * w(x, y) + n(x, y)$$

It convolves the original image with the transfer function  $h(x,y)$ , and adds an  $n(x,y)$  noise term to it [4].

## 2.2 Model construction and discretization of partial differential equation in digital image denoising

An equation containing the partial derivative of an unknown variable of a function is called a partial differential equation. In many engineering problems, it is usually converted into a partial differential equation for the convenience of solving the quantity value changing with space or time.

Partial differential equations in the field of image processing mainly involve diffusion equations and heat conduction equations, and their general models are as follows:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x_1} \left( c_1 \frac{\partial u}{\partial x_1} \right) + \dots + \frac{\partial}{\partial x_n} \left( c_n \frac{\partial u}{\partial x_n} \right) + F(x, t)$$

Where,  $t$  represents the time variable,  $x = (x_1, x_2, \dots, x_n)$  is the space variable,  $u = u(x, t)$  is the size of the gray value in the diffusion process, and coefficient  $C_i$  is the diffusion coefficient. If  $c_1 = c_2 = \dots = c_n = A$ , the equation is: if  $i$  is 1, then the diffusion equation is a one-dimensional partial differential equation [5].

If the initial conditions  $u(x,0) = u_0(x)$  are given, the initial boundary value problem is formed. The equation is as follows:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}$$

In this case, finite difference method can be used to discretize the equation.

The equation approximation can be derived from the grid points:

$$\frac{\partial u(x, t)}{\partial t} = \lim_{\Delta t \rightarrow 0} \frac{u(x, t + \Delta t) - u(x, t)}{\Delta t}$$

Appropriate selection can get a suitable approximation solution, and the variable at the time of  $(n+1)$  can be obtained through iteration.

1	0	0	1
0	-1	-1	0

Figure 2: Roberts operator

For such two-dimensional partial differential processing method, small-region convolution is usually used for calculation in practical application. For two-dimensional two directions, a template [6] is used to calculate the partial derivative values in the two directions respectively. The template is used for derivative calculation, and the partial derivative values of corresponding pixels can be calculated at the same time of sliding. For example, the template of Roberts gradient operator is as follows.

In the process of image processing, after modeling an image processing method into a partial differential equation model, the calculation of partial differential equation is realized in the discrete case. Based on the discretization of partial differential equation, this paper proposes a better method.

## 3 RESEARCH ON IMAGE DENOISING THEORY OF PDE BASED ON P-M MODEL

### 3.1 Theory of partial differential equations

In recent years, there are a variety of image denoising processing methods based on partial differential. Compared with traditional methods, this paper focuses on the denoising of partial differential equation based on P-M model, and improves it according to its characteristics, so that the denoising of such partial differential equation can retain the edge texture features to a greater extent.

Perona and Malik proposed that the use of directional diffusion equations that preserve edges can be better than other methods in image denoising, namely anisotropic nonlinear diffusion equations. This method can effectively avoid the situation of boundary blurring when using the isotropic diffusion equation. By choosing the diffusion coefficient, the diffusion degree can be solved in all directions. The image gradient value is considered as the edge detection operator, and the diffusion coefficient is taken as the image gradient function, which is opposite to the gradient value. In this way, the noise can be smoothed and the edge features can be preserved.

The classic P-M model algorithm formula is as follows:

$$\begin{cases} \frac{\partial u}{\partial t} = \text{div}(c(|\nabla u|) \nabla u) \\ u_0 = u(x, y, 0) \end{cases}$$

Where,  $u_0$  is the original image,  $c(|\nabla u|)$  is the diffusion coefficient,  $\text{meet}(0)=1$ , as the edge detection operator, P-M proposed a diffusion coefficient function integrating noise removal and edge enhancement:

$$C1(|\nabla u|) = \frac{1}{1 + |\nabla u|^2 / k^2}$$

$C1$  here meets all the above characteristics of the diffusion coefficient. According to the calculation and iteration, the diffusion coefficient increases with the increase of parameter  $k$ . For pixel points with the same gradient, the larger parameter  $k$  is, the faster diffusion speed will be. If  $k$  is too small, the number of iterations is greatly increased.

$C1$  is smoothed within the selection range of the entire image, which can smooth all contrast image regions to a certain extent, satisfying the anisotropic nonlinear theory.

According to the above model algorithm formula, if  $\xi$  is assumed to be the unit vector along the gradient direction and the unit vector perpendicular to it  $\eta$ . Then there is:  $\xi = \frac{1}{\sqrt{u_x^2 + u_y^2}} \begin{pmatrix} u_x \\ u_y \end{pmatrix}$ ,  $\eta = \frac{1}{\sqrt{u_x^2 + u_y^2}} \begin{pmatrix} -u_y \\ u_x \end{pmatrix}$ .

By solving for the partial derivative of  $u$  with respect to  $\xi$ , then  $u_\xi = u_x \cos \alpha + u_y \cos \beta$ ,  $\cos \alpha$ ,  $\cos \beta$  cosine in the  $\xi$  direction. It can be concluded that  $u_\xi = \frac{u_x^2 + u_y^2}{\sqrt{u_x^2 + u_y^2}}$ . And the second derivative is calculated the same way [7]:

$$u_{\xi\xi} = \frac{u_{xx}u_x^2 + 2u_xu_yu_{xy} + u_yy u_y^2}{u_x^2 + u_y^2}$$

$$u_{\eta\eta} = \frac{u_{xx}u_y^2 + 2u_xu_yu_{xy} + u_yy u_x^2}{u_x^2 + u_y^2}$$

Add left and right to get:  $u_{\xi\xi} + u_{\eta\eta} = u_{xx} + u_{yy}$ . By substituting the diffusion coefficient, we obtain:

$$\frac{\partial u}{\partial t} = \frac{k^2}{k^2 + |\nabla u|^2} u_{\eta\eta} + \frac{k^2 (k^2 - |\nabla u|^2)}{(k^2 + |\nabla u|^2)^2} u_{\xi\xi}$$

From the relationship between the values of  $C_\eta(|\nabla u|)$ ,  $C_\xi(|\nabla u|)$  and  $k$  and the gradient, it can be seen that the diffusion coefficient along the gradient direction has a larger value in the flat region and has a strong diffusion ability, while the diffusion coefficient perpendicular to the gradient direction has a diffusion effect in the whole gradient range. By combining the two as the coefficient of P-M diffusion equation, the image edge can be maintained and the noise can be effectively removed. The anisotropic nonlinear diffusion process is realized.

### 3.2 Case analysis

In order to verify the feasibility of the above method and its superiority over other kinds of denoising methods, we compare the original image, the image with noise and the image after denoising by using the method studied in this paper through simulation analysis, which further proves the practicability of this method. Aiming at the above P-M model algorithm, this paper intends to use this method to analyze the original image of a color image, the noise image after adding Gaussian white noise with variance of 0.01, the P-M diffusion denoising image and the denoising image when adding the diffusion coefficient  $C1$ .



Figure 3: Original image (1), noise image (2)

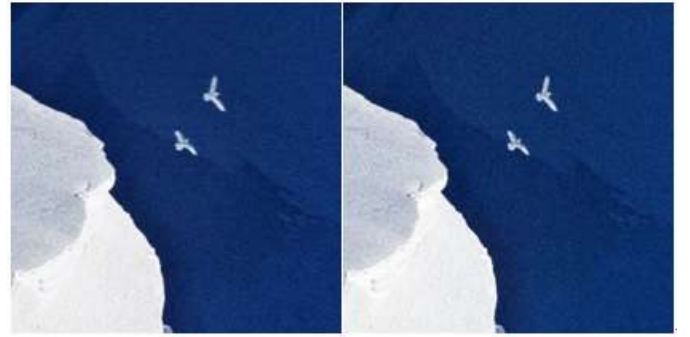


Figure 4: P-M diffusion denoising image (1), denoising image when adding diffusion coefficient  $C1$  (2)

It can be seen that the denoising ability of P-M diffusion model is strong, but its denoising effect is closely related to the number of iterations. If the number of iterations is too much, the noise removal will deepen the blur degree of image edge.

The denoising image of the P-M diffusion model with the addition of the diffusion coefficient  $C1$  can make the gray value of the image excessively smooth [8], and the denoising effect is better than that of the basic P-M diffusion model to achieve the purpose of anisotropic diffusion. However, according to the result analysis, there is still the situation of image edge blur, which is also a problem to be solved in the future.

## 4 FUZZY COMPREHENSIVE EVALUATION OF IMAGE QUALITY BASED ON ANALYTIC HIERARCHY PROCESS

### 4.1 Introduction of comprehensive evaluation

Evaluation refers to the conclusion after the judgment and analysis of something. In the traditional sense, evaluation can only be good or bad, and the evaluation of complex systems or things cannot rely solely on the unique element index to make decisions, which will not only make the evaluation results unconvincing, but also lead to errors caused by incomplete evaluation [9].

Comprehensive evaluation, developed from traditional evaluation, is characterized by the introduction of a new evaluation

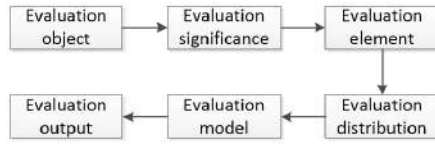


Figure 5: Key elements of comprehensive evaluation

method when the evaluation object has a large number of connotation system levels, all kinds of factors have an impact on it, and a single factor cannot describe the degree in detail. Among them, the evaluation object, significance, elements, weight distribution, model building, output and other aspects constitute the comprehensive evaluation method.

Starting from the digital image of airborne infrared imaging products, the research object of this paper, when evaluating the quality of the generated images, the comprehensive evaluation method can select the key elements of image information through the comprehensive evaluation method, integrate subjective and objective, achieve multi-dimensional evaluation, and finally achieve the purpose of distinguishing and evaluating the image quality of infrared imaging products of different manufacturers.

## 4.2 Fuzzy evaluation and application

Based on the investigation and analysis of infrared imaging products, in view of their own characteristics and the characteristics of image information generation, according to the different characteristics of their images and considering the comparability, observability and independence of each factor, this paper divides the evaluation factors into the following five points. They are double-stimulus damage classification, double-stimulus continuous mass classification, single-stimulus continuous mass classification, mean square error, peak signal-to-noise ratio. Its evaluation model is as follows:

Among them, double stimulus damage grading, double stimulus continuous quality grading and single stimulus continuous quality grading are subjective evaluation methods, and mean square error and peak signal-to-noise ratio are objective evaluation methods. These five items constitute the first-level factors of image quality evaluation. Combining the two major types of evaluation methods can effectively avoid the unicity of evaluation and improve the evaluation level [10].

Because different factors have different importance degrees to the evaluation of the whole system, the distribution of weight plays a decisive role in the final evaluation. On the basis of this study,

it is necessary to further determine the severity of the influence between factors on the results, and then assign normalized values [11].

It should be noted here that  $w=(w_1, w_2, w_3, w_4)$  is the membership degree of target object  $k$  in  $L_j(j=1,2,3,4)$ , and is expressed as the probability of a certain work at this decision level. Normalization processing is required, so that different levels in the table can be output into a vector. After determining the weight of the index, the overall component factors are disassembled through the analytic hierarchy process, and different levels are aggregated according to the membership and correlation degree between the factors, and the advantages and disadvantages are compared in pairs for scheduling or relatively important weight determination. The judgment factor domain  $M$  collects all evaluation factors of the object to be evaluated, and all decision levels are stored in the evaluation level domain  $N$ . The finer the degree division, the higher the accuracy of the comprehensive evaluation results.

## 4.3 Analysis of fuzzy evaluation mathematical model

Fuzzy comprehensive evaluation can be divided into single-level evaluation and multilevel evaluation in engineering practice. According to the characteristics of infrared product imaging, this paper selects single-level evaluation for research. Let  $U = \{u_1, u_2, \dots, u_m\}$  be  $m$  evaluation indexes of the evaluated object;  $V = \{v_1, v_2, \dots, v_n\}$  is the  $n$  evaluation levels that describe the state of each indicator.  $m$  is the number of evaluation indicators,  $n$  is the number of comments.

First, for the single index in the index set, ( $i=1,2,\dots,m$ ) to make a single index evaluation, from the index of the object's evaluation level ( $j=1,2,\dots,n$ ), and then the single index evaluation set of the  $i$ th index is obtained:  $u_i$ :

<?TeX

$$r_i = (r_{i1}, r_{i2}, \dots, r_{in})$$

?>

A total evaluation matrix  $R$  is constructed from the evaluation set of  $m$  focus indicators [12], that is, each evaluated object determines the relation  $R$  from  $U$  to  $V$ , which is a matrix:

<?TeX

$$R = (r_{ij})_{m \times n} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}$$

?>

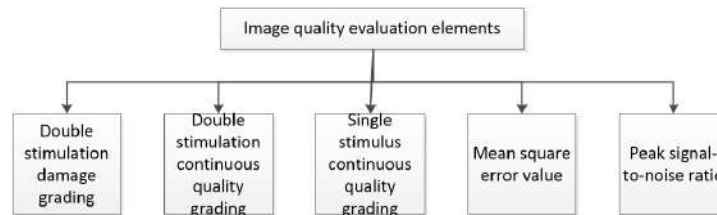


Figure 6: Primary image quality evaluation elements

**Table 1: Classification of conformity grade**

Grade classification of qualification degree L	corresponding value range k	corresponding coefficient
Highly qualified	(a1,a2)	W1
Good pass	(b1,b2)	W2
Basically qualified	(c1,c2)	W3
Not qualified	(d1,d2)	W4

**Table 2: A case study of the coincidence evaluation grade of the damage grade of double stimulation**

First-order index	Compliance evaluation grade V			
	High compliance	Good coincidence	Basic agreement	Do not conform to
Double stimulation damage grading	0.70	0.15	0.15	0

Where  $r_{ij}$  represents the membership degree that the evaluation object can be rated from the perspective of indicator  $u_i$ , namely the frequency distribution of the  $i$ th indicator  $u_i$  on the  $j$ th comment  $v_j$ , which is generally normalized to meet  $\sum r_{ij} = 1$ . In this way, the R matrix itself is dimensionless and requires no special treatment.

Such membership matrix is not enough to evaluate things. Each index in the evaluation index set has different status and role in the "evaluation objective", that is, each evaluation object occupies different proportion in the comprehensive evaluation. Introducing a subset of U, called the weight or weight assignment set,  $A = (a_1, a_2, \dots, a_m)$ . Where  $a_i > 0$ , and  $\sum a_i = 1$ , it reflects a trade-off between these indices.

There are two kinds of fuzzy sets. One is the quantity that marks the importance degree of each element in the index set U in people's mind, which is represented by the weight vector on the index set U  $A = (a_1, a_2, \dots, a_m)$ , another is  $U \times V$ , It is represented by the membership matrix R.

Let's introduce some subset  $B = (b_1, b_2, \dots, b_n)$  on V, let's say  $B = A \bullet R$ .

Different rows in R reflect the degree of membership of a certain evaluated thing to the subset of the evaluation level from different single indicators. Weight vector A is used to synthesize different rows, and then the degree of membership of the evaluated thing to the subset of the evaluation level can be obtained in general, namely the evaluation result vector B, so B is also known as the decision set.

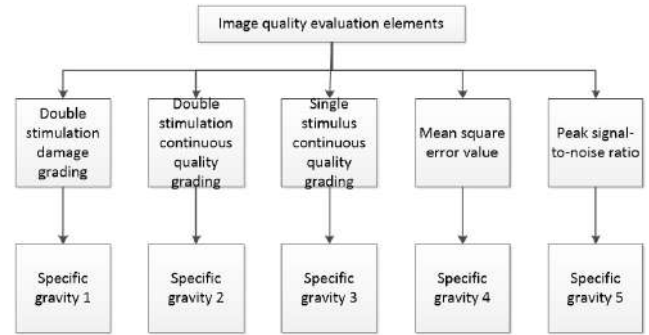
Taking the damage classification of double stimulation as an example, the following data were obtained after preliminary processing with the above coincidence table.

Then the evaluation matrix of first-order index double-stimulus damage classification is:

$$R = [0.70, 0.15, 0.15, 0]$$

#### 4.4 Application of Analytic Hierarchy process

In the face of a large number of different indicators in the comprehensive evaluation, from an objective point of view, the impact significance of each item is quite different, so the analytic hierarchy process is introduced.

**Figure 7: Weight distribution of first level evaluation index**

According to the top-down study of objective layer, criterion layer and scheme layer, starting from the overall level of image quality and combining with expert scoring method, the matrix S can be generated by comparing the factors of index layer in pairs:

$$S = (s_{ij})_{1 \times 1} = \begin{bmatrix} S_{11} & S_{12} & \dots & S_{1l} \\ S_{21} & S_{22} & \dots & S_{2l} \\ \dots & \dots & \dots & \dots \\ S_{l1} & S_{l2} & \dots & S_{ll} \end{bmatrix}$$

Take the five factors of double stimulus damage classification, double stimulus continuous mass classification, single stimulus continuous mass classification, mean square error and peak signal-to-noise ratio as an example. The five factors are arranged in horizontal and vertical order. If the five items are  $O_1, O_2, O_3, O_4, O_5$ , and the importance quotient is used [13], then:

$$S = (s_{ij})_{1 \times 1}, s_{ij}, 0, s_{ij} = \frac{1}{s_{ji}}$$

After solving the maximum characteristic root of this matrix, the eigenvector corresponding to the maximum characteristic root is the required weight vector after normalization.

Suppose the weight  $A_i = (a_{i1}, a_{i2}, \dots, a_{in})$  of the first-level index is finally obtained, and each  $R_i$  is a single index evaluation matrix, then the first-level evaluation vector is obtained:  $B_i = A_i \bullet R_i = (b_{i1}, b_{i2}, \dots, b_{im}), i = 1, 2, \dots, s$ . According to the

principle of maximum membership degree, we can judge which level of conformity the current image quality belongs to [14]. Take the double stimulus continuous quality classification as an example, if the calculation results are as follows:  $B = A \bullet R$ .

B is the final evaluation result of the audit data, which is the grade description of the comprehensive status of the audited object, and represents the probability that the audited object belongs to each evaluation level. Here, B represents the probability that the audited object belongs to excellent, good [15], medium and unqualified. Here, the final quality level of the image can be determined by the maximum membership degree method.

## 5 CONCLUSION

Based on the research and analysis of airborne infrared imaging products, this paper introduces the development history of digital image denoising in recent years. Considering the combination of modern mature denoising technology and effective image denoising ability, the paper analyzes the partial differential equation algorithm based on P-M model, and verifies the conjecture by using the algorithm calculation and actual simulation. On this basis, the study of image evaluation method is introduced, and the fuzzy evaluation theory based on analytic hierarchy process is used to systematically evaluate the quality of airborne infrared imaging products. This method can effectively solve the deviation caused by single factor evaluation and subjective assumptions caused by human factors, and combine image denoising with image quality evaluation. It can form a complete set of systematic process for screening the image quality level of different infrared imaging products. The process can provide screening means for users who need to buy infrared imaging products, but also conducive to the development of infrared imaging products personnel to find their own problems, to achieve the purpose of comprehensively improving the image level of domestic infrared imaging products.

## REFERENCES

- [1] Application mode of partial differential equation in digital image processing [J]. Journal of Beijing Institute of Printing and Technology, 2017, 25(7): 175-177. (in Chinese) DOI: 10.3969/j.issn.1004-8626.2017.07.060.
- [2] An C S. Analysis on the status quo and development of image processing technology [J]. Science and Technology Information, 2018, 16(25): 72-73. (in Chinese) DOI: 10.16661/j.cnki.1672-3791.2018.25.072.
- [3] JIA Shuxiang. Research on digital image processing based on Partial Differential Equation [J]. Electronic Technology and Software Engineering, 2020, No. 184(14): 143-144.
- [4] Zhang Mingwu, Chen Sheng. Denoising of partial differential equation model in digital image processing research [J]. Journal of information technology, 2016 (11) : 143-146 + 151. DOI: 10.13274/j.carol carroll nki HDZJ. 2016.11.036.
- [5] JIA Chaoxian. Comparative analysis of Four Kinds of Partial Differential Equation image Denoising Techniques [J]. Journal of Changchun Normal University, 2022, 41(12): 41-47.
- [6] YanSuYa. Image denoising method based on partial differential equation of research [D]. Xinjiang normal university, 2021. The DOI: 10.27432/, dc nki. Gxsfu. 2021.000103.
- [7] WU Denghui. Research on a Class of image Denoising Methods Based on Partial Differential Equation [D]. Nanjing University of Information Science and Technology, 2017.
- [8] JIANG Bin. Research on Partial Differential Equation Method for Image Processing [D]. Jilin University, 2009.
- [9] Ye Xiaojing, Fang Guohua, Liao Tao, *et al.* Risk assessment of river shoreline development and utilization based on improved grey fuzzy evaluation [J/OL]. Yangtze river proceedings of the national academy of sciences: 1-12 [2023-02-13]. <http://kns.cnki.net/kcms/detail/42.1171.TV.20230209.1323.002.html>.
- [10] Huang Yingji, clock fierce, family Wei, *etc.* Image quality evaluation method based on objective evaluation review [J]. Computer knowledge and technology, 2021 (28) : 92-94. The DOI: 10.14004 / j.carol carroll nki CKT. 2021.3006.
- [11] Zhang Dandan, Zhao Yinghui. Natural image quality evaluation method review [J]. Computer knowledge and technology, 2020 (9) : 203-205. The DOI: 10.14004 / j.carol carroll nki CKT. 2020.1072.
- [12] Huang Min, Zhou Dehong, Fan Xuyan *et al.* Comparison of safety risk assessment of long distance natural gas pipelines based on two combination weighting methods [J]. Industrial Safety and Environmental Protection, 2023, 49(02): 42-47.
- [13] tone. Index weight scheme design based on improved analytic hierarchy process [J]. China's new technology and new products, 2022, No. 472 (18) : 139-141. The DOI: 10.13612 / j.carol carroll nki CNT. 2022.18.045.
- [14] GE Lulu. Weight Determination of Railway Informatization Evaluation Index System Based on Group Decision Analytic Hierarchy Process [J]. Railway Computer Applications, 2017, 26(04): 6-9.
- [15] Zhu Xiaofei, Wang Yongjun, Li Dajun. Validity test of maximum membership principle in fuzzy evaluation [J]. Surveying, Mapping and Spatial Geographic Information, 2016, 39(05): 135-137+143.

# Deep Vision Network Based CT Image Detection for Aiding Lumbar Herniated Disc Diagnosis

Wenzhe Xie  
Zhuoyue Honors College, Hangzhou  
Dianzi University  
xwz13@hdu.edu.cn

Feiwei Qin  
School of Computer Science and  
Technology, Hangzhou Dianzi  
University  
qinfeiwei@hdu.edu.cn

Yanli Shao  
School of Computer Science and  
Technology, Hangzhou Dianzi  
University  
shaoyanli@hdu.edu.cn

## ABSTRACT

Recently, artificial intelligence (AI) technologies have applied in the field of clinical medicine widely. And some of researches try to use AI to assist the diagnosis of spinal disease. In this study, Herniation-Det: an automatic lumbar disc herniation detection method based on two stage detection framework (e.g. R-CNN, Fast R-CNN, Faster R-CNN, etc.) is presented. Firstly, after comparing the performance of various backbone networks such as VGG, ResNet, EfficientNet, etc., a feature extractor based on VGG16 is constructed to automatically and efficiently extract the necessary feature information from medical images. Secondly, we use region proposal network (RPN) to generate region proposals and provide them to the part of Fast R-CNN for classification and regression. After precisely studying the image of disc herniation, we adjust the scale and radio of the anchor, to make them more in line with the characteristics of the lumbar disc image dataset. Finally, the object detection algorithm is first used on CT images which achieved 89.50% mAP, and then applied to MR images of the lumbar disc to achieve the goal of automatically identifying lumbar disc herniation with or without calcification. Hence, artificial intelligence assisted diagnosis of calcified lumbar disc herniation on MR images can be achieved with 81.24% mAP, by further using a multi-modal learning strategy.

## CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning approaches.

## KEYWORDS

Deep learning, Lumbar disc herniation, Object detection, Multi-modal learning

## ACM Reference Format:

Wenzhe Xie, Feiwei Qin, and Yanli Shao. 2023. Deep Vision Network Based CT Image Detection for Aiding Lumbar Herniated Disc Diagnosis. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590092>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590092>

## 1 INTRODUCTION

Intervertebral disc herniation is one of the common spinal degenerative diseases in clinical practice, which may cause severe pain and even paralysis of the lower limbs. Among the patients with lumbar disc herniation, some patients have calcified lumbar disc herniation, which is caused by the deposit of calcium. The treatment of lumbar disc herniation varies with or without calcification. The treatment methods are different for whether the lumbar disc herniation is combined with calcification. For a single segment of the lumbar disc herniation, foraminoscopic surgery can be chosen for treatment, which is a minimally invasive lumbar spine surgery. However, calcified lumbar disc herniation is difficult to be treated by foraminoscopic surgery, because the surface of severely calcified lumbar disc tissue is often uneven and the calcified tissue is hard, inelastic, and closely adhered to surrounding tissues such as nerves, which increases the difficulty of the operation. Hence, it is very important to estimate whether the herniated disc is calcified or not before we take the operation.

Currently, CT scanning is mainly used to evaluate whether lumbar disc herniation is accompanied by calcification or not. However, CT scan still has many problems. Firstly, CT scan is less effective than MR in soft tissue imaging, while whether the lumbar disc is diseased can be fully observed from MR images. So sometimes, we have to perform both CT and MR examinations for one patient with a herniated disc, which is a waste of medical resources. In addition, CT scan has a certain degree of radiation, repeated CT scan is not good for patients.

Localizing the position of the intervertebral disc in images is commonly the first step to detect the calcified lumbar disc herniation. Schmidt et al. [1] proposed a classification tree algorithm in 2007. This algorithm first generated the probability distribution of centroid of each intervertebral disc in MR images, and then used the model of probability map to infer the most likely position. The final result was an average localization error of 6.2 mm with respect to the human experts. Oktay et al. [2] achieved the positioning of intervertebral disc based on the Histogram of Oriented Gradient (HOG) using the Support Vector Machine (SVM) in 2011, with an average positioning error of the algorithm between 2.6 mm and 3.6 mm. Glocker et al. [3] used the random forest to determine the position of the centroid of vertebral bodies, which achieved average errors between 6 mm and 8.5 mm.

Computer-aided diagnosis (CAD) refers to the use of imaging, medical image processing technology, which combines with computer analysis and calculation, to improve the accuracy of diagnosis. Bounds et al. [4] built a multilayer perceptron model to diagnose low back pain and sciatic nerve pain in 1988. The accuracy of the

network was between 77% and 88%. In 2011, Koopman et al. [5] achieved an accuracy rate of 94% using SVM on the MR image dataset. Hao et al. [6] proposed a new algorithm based on SVM to further study the shape of Preprint submitted to Artificial Intelligence in Medicine February 12, 2021 the intervertebral disc. And the accuracy was 91.6%. Silvia et al. [7] and Isaac et al. [8] adopted the research plan proposed by Pfirrmann et al. [9], which divided disc degeneration into five categories. In addition to spinal degenerative lesions, machine learning has also been applied to the study of spinal deformities. Machine learning has a quite good advantage in estimating the degree of adult congenital scoliosis. Ramirez et al. [10] divided the degree of scoliosis from normal to severe into 3 levels, by using SVM, decision tree, statistical model, linear discriminant analysis, etc. Finally, the SVM reached the highest accuracy of 85%. Seoud et al. [11] trained the SVM on a dataset consisting of 97 patient samples, and the accuracy reached 72.2%. Yu et al. [12] provided the edgebased active contour model (ACM) to overcome the problem of low contrast, complex noises, and intensity inhomogeneity in segment medical images.

With the development of AI, deep learning was also applied to positioning the spine structure. Chen et al. [13] proposed a hybrid algorithm combining random forest and convolutional neural network (CNN). This algorithm used the random forest to get the preliminary position and input it to the CNN, which achieved average errors between 1.6 mm and 2 mm. Suzani et al. [14] regraded the positioning as a regression task, using a neural network with the depth of 6 layers. In 2017, Yang et al. [15] achieved average errors between 6.9 mm and 9 mm in the positioning of cone centroid. Each sample in the CT dataset not only suffered from different diseases, but also was taken with different image resolutions. After the success of AlexNet [16], many famous convolutional neural networks were proposed, such as ZFNet [17], VGGNet [18], GoogLeNet [19], ResNet [20], and DenseNet [21]. Chen et al. [22] proposed a deep convolutional neural network with 3D convolutional layers for intervertebral disc localization and segmentation. The network could generate the probability of classification, and then used the threshold segmentation and smoothing to adjust the segmentation. Lessman et al. [23] proposed a 3D convolutional network with memory components that segment and label vertebrae. The network analyzed images using information from images and memory to search the next vertebra. The average DSC of this method reached 94% and the MSD was 0.2 mm.

As one of the basic problems of computer vision, object detection is the basis of many other computer vision tasks, such as instance segmentation, image captioning, object tracking, etc. Girshick et al. [24] first proposed R-CNN in 2014, which was a pioneering work of applying deep learning into the field of object detection. Since then, object detection has evolved at a rather amazing rate. Different from R-CNN, Fast R-CNN [25] extract feature from a picture only once to generate RoIs, and then use the RoIs for classification and regression. So far, many representative object detection algorithms were proposed. They all have good performance in natural scenes. Among them, Faster R-CNN [26] is very famous. It is a two-stage detection network, the object in the image is roughly located and then finetune, which is more accurate than the one-stage detection algorithms [27, 28].

In the past, the diagnosis of calcified lumbar disc herniation was done by the experienced experts who basically rely on the spine CT scan to complete. Manual recognition has the problems of high labor consumption, high image dependency and high error rate of human eye recognition. Therefore, it will be a significant breakthrough in clinical imaging technology, if we can automate this task and realize this diagnosis of calcified lumbar disc herniation with high-precision on CT images.

In this paper, aiming at the problem of recognizing calcified lumbar disc herniation on CT images, an object detector called HerniationDet based on two stage detection framework is proposed. First, CT images are used for training the neural network model. And then based on the idea of multi-modal learning, MR images will be used for training the neural network model. Our main contributions can be summarized as follows:

- Propose HerniationDet which is an end-to-end trainable network based on two stage detection framework, the network achieves state-of-the-art results on the task of calcified lumbar disc herniation detection.
- Take VGGNet as a feature extractor compared with other backbone networks in practice.
- A multi-modal feature fusion strategy is used in the network, which achieves better performance than without it.
- Experimental results show that it is more advantageous to use the proposed method to detect calcified lumbar disc herniation than the orthopedists.

## 2 MATERIALS

### 2.1 Data

Because of the difficulty in collecting and labeling medical images of the lumbar disc, until now there is no available public dataset of lumbar disc. In this paper, we built our own dataset of the lumbar disc on CT and MR images as the basis of the study. All the cross-sectional images of the lumbar disc were collected from the local hospitals. As is shown in the Table 1, this dataset includes 428 CT scan images and 57 MR images from 191 patients. In detail, there are 231 normal, 105 herniation without calcification and 92 herniation with calcification cross-sectional images of the lumbar disc in CT scan and 33 normal, 12 herniation without calcification, and 12 herniation with calcification cross-sectional images of the lumbar disc in MR images..

**Table 1: Sample image dataset statistics**

	normal	herniation	calcification	total
CT	231	105	92	428
MR	33	12	12	57

In this paper, we used the labelImg [29], an open labeling tool on GitHub, to label CT images. Fig. 1 shows a part of the CT images which are manually labeled. We divided the whole images of the lumbar disc into a training set, a verification set and a test set according to the ratio of 6:2:2. The training set was used to train network, learn the features of the image, the verification set was used to evaluate the performance of current model and adjust the

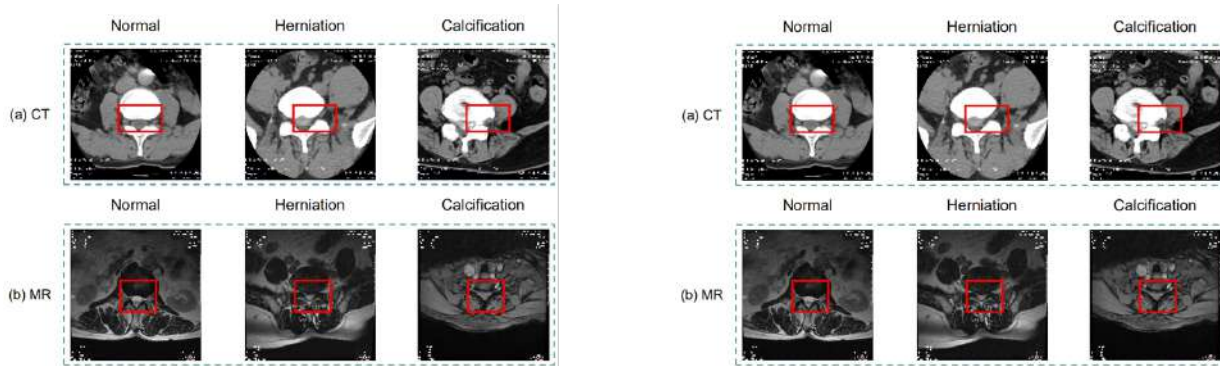


Figure 1: (a)Three kinds of images in CT dataset. (b)Three kinds of images in MR dataset.

parameters, and the test set was used to evaluate the performance of optimal model.

## 2.2 Data pre-processing

Before training the model, we pre-process the data to make it more suitable for detection.

1) Data normalization: To avoid vanishing gradient problem, we divide the image data by 255 to normalize the data.

2) Data augmentation: To avoid overfitting and underfitting problem, the original dataset is manually added with a corresponding transformation or interference to increase the dataset. We expand the dataset of the lumbar disc by rotating, cropping randomly, and scaling on the original images

## 3 METHODOLOGY

As is shown in the Fig. 2, the proposed HerniationDet (Herniation Detection) based on two stage detector framework consists of three main parts: Backbone Network, Region Proposal Network, and RoIHead (part of Fast R-CNN), each part will be described in more detail within following subsections.

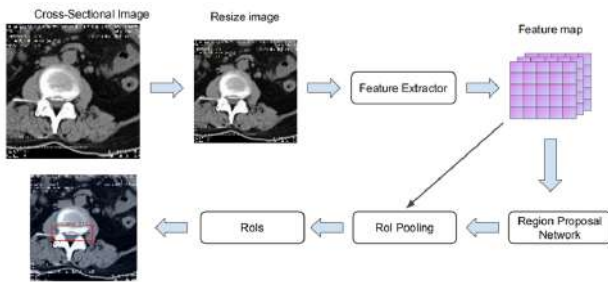


Figure 2: The overall architecture of the HerniationDet model. The input image, which is resized to the fixed size, is fed into the Feature Extractor to generate the feature map; then, the region proposals are generated by the Region Proposal Network on the basis of the feature map. Since the region proposals are numerous, the NMS algorithm is used for selecting RoIs, which are put into the RoIHead for further classification and positioning.

## 3.1 Backbone Network

Convolutional neural networks (CNNs) are very good at learning complex features from raw data. There are some representative CNN architectures, such as AlexNet [16], VGGNet [18], ResNet [20], DenseNet [21], Xception [30] etc. At first, these networks are used to conduct the image classification. With the widespread application of transfer learning, the above networks can be as the feature extractor for object detection, called backbone networks. Different backbone networks can achieve different effects.

Only when the backbone network and RoIHead match can the expression ability of the entire network be maximized. If the extractive ability of the backbone network is very weak, and the extractive ability of the RoIHead is very strong, the performance of the overall detection model will be limited by the backbone network. Hence, we perform experiments on different backbones. As shown in Section 4.1, we find that VGG16 works well in practice.

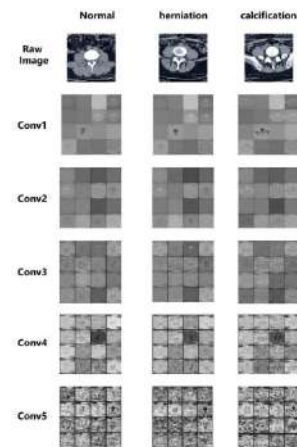


Figure 3: The features extracted by VGGNet.

Fig. 4 shows the architecture of the VGG16. However, instead of using the entire VGG16 Network as a feature extractor, we need to choose and adapt the network to the object detection tasks. We only need its convolution part and delete the pooling layer after the fifth

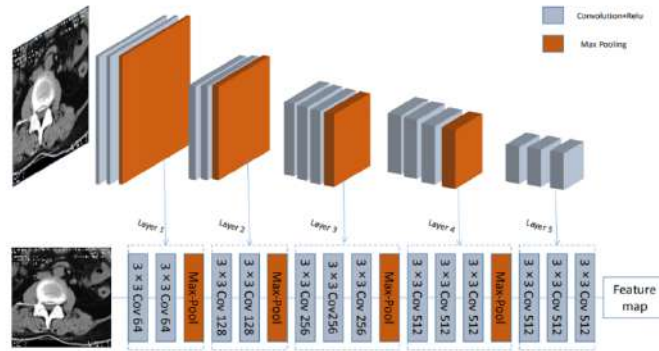
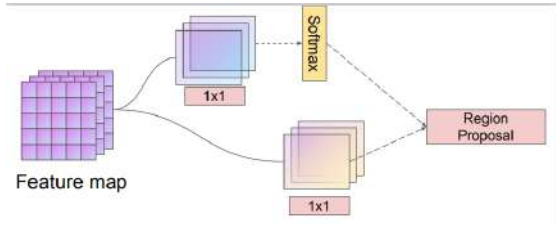


Figure 4: The feature extractor based on VGGNet.

convolutional layer. Hence, there are 13 convolutional layers and 4 pooling layers. Fig. 3 visualizes the features extracted by VGG16 from different categories.

### 3.2 Region Proposal Network

Classic object detection algorithms are very time-consuming (e.g. AdaBoost in OpenCV, Selective Search in R-CNN). HerniationDet abandons the traditional method and uses the Region Proposal Network (RPN) to generate the region proposals. Because the RPN and the RoIHead networks share the feature maps which are generated by VGG16, it can greatly improve the speed of generating region proposals. Actually, the RPN is a Fully Convolution Network (FCN), which can accept images of any size. The RPN will generate a series of rectangular region proposals with high-quality and corresponding probabilities. Fig. 5 shows the architecture of RPN.



**Figure 5: The architecture of RPN. Feature maps are input into two parallel 1×1 conv layers for classification and regression, respectively.**

**3.2.1 Anchor.** To achieve translation invariant, a set of the bounding boxes of the specific size and aspect ratio centered on the anchor is applied to generate region proposals. In HerniationDet, there are two parameters related to the generation of anchor boxes, namely ratio and scale. (1) Ratio: It is the aspect ratio of the anchor box. (2) Scale: The area of the anchor box. Supposing the height of the anchor box is  $h$ , the width is  $w$ , the base area of the anchor box is  $S$ , and the aspect ratio is  $r$ , the conversion formula is defined as:

$$\begin{cases} S = w \times h \\ h = \sqrt{S/r} \\ w = r \times h \end{cases} \quad (1)$$

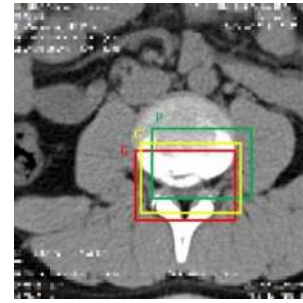


Figure 6: Bounding box regression.

Because it is uncertain whether the anchor box contains the object that we want to detect and locate the exact position. Therefore, we need to conduct the anchor box classification and bounding box regression. On the feature map obtained by the backbone network, RPN adds a convolutional layer with a 3×3 convolutional kernel to conduct semantic space conversion, and activates the output using ReLU algorithm. The features will be input into the two parallel convolutional layers to achieve classification and bounding box regression.

**3.2.2 Classification and Regression Layer.** In the classification layer, we used the 1x1 conv layer followed by the softmax function to estimate the probability of object/nonobject for each proposal. Because there are 9 shapes of anchor boxes and the number of the categories is 2, the number of channels is 18. In the regression layer, HerniationDet uses bounding box regression to fine-tune the position of the anchor box, so that the predicted border is closer to the true position. During the training, the input of regression is a pair of anchor box and ground truth box, the output is the transformation between the anchor box and ground-truth box. To obtain the position of the region proposals, the calculation formula is defined as:

$$\begin{cases} \hat{G} = P_w \times d_x(P) + P_x \\ \hat{G} = P_h \times d_y(P) + P_y \\ \hat{G} = P_w \times \exp(d_w(P)) \\ \hat{G} = P_h \times \exp(d_h(P)) \end{cases} \quad (2)$$

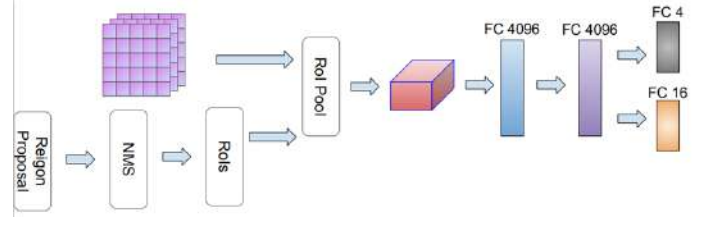


Figure 7: The structure of RoIHead network.

We define the anchor box of center coordinate, width and height as the  $P = (P_x, P_y, P_w, P_h)$ , the ground truth box of center coordinate, width and height as  $G = (G_x, G_y, G_w, G_h)$ , the aim of the algorithm is to learn the transformation that maps  $P$  to  $G$ . This transformation is defined as  $d_x(P), d_y(P), d_w(P), d_h(P)$ , the first two are the translation for the center coordinate of the anchor boxes, the latter two are the zoom for width and height of the anchor boxes.

Fig. 6 shows the result of bounding box regression. But not all anchor boxes are involved. Firstly, select 2,000 RoIs (Region of Interests) from more than 20,000 anchor boxes. Next, select 128 from 2,000 RoIs. In order to select the anchor boxes reasonably, HerniationDet uses Intersection over Union (IoU) as the evaluating indicator. The IoU is defined as:

$$IoU = \frac{\text{Area of overlap}}{\text{Area of Union}} \quad (3)$$

To select positive and negative samples from anchor boxes, we assign a binary class label to each anchor. In this paper, the positive label is assigned in two ways: (1) the anchor boxes which have the highest overlap with ground-truth boxes. (2) the IoU of anchor boxes higher than 0.7 with ground-truth boxes. And we assign a negative label if the IoU of anchor boxes lower than 0.7 with any ground-truth boxes. During the first selection, we use the Non-Maximum Suppression (NMS) algorithm to select RoIs from the more than 20,000 region proposals. In fact, NMS is a process of finding a locally optimal solution:

- 1) Select one with the highest probability of pre-detected objects in all region proposals of an image, record it as box best and keep it.
- 2) Calculate the IoU of box best with the remaining region proposals, and set a threshold.
- 3) If the IoU is bigger than the threshold, delete the region proposal.
- 4) Continue to select the one with the highest score from the unprocessed region proposals and repeat the above process.

### 3.3 RoIHead

Based on the RPN, the RoIHead continues to conduct the classification and regression. Fig. 7 shows the structure of RoIHead network. To fix the size of the input, He et al. [31] proposed the Spatial Pyramid Pooling (SPP). In SPP, the convolutional feature map is transformed into different sizes. Each size extracts a feature of fixed dimension, and finally obtains the fixed-size output after maximum pooling. HerniationDet uses RoI Pooling to fix size. After RoI Pooling layer, there are two fully connected layers whose parameters were initialized by the weights of the VGG16 pre-trained model. Finally, HerniationDet uses two sibling fully connected layers to

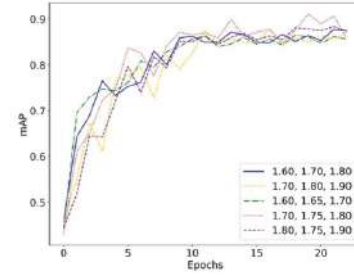


Figure 8: Comparison of network accuracy under different ratio settings.

classify RoIs and conduct bounding box regression. The fully connected layer used for the classification and regression is composed of 4,096 neurons respectively. The output length of classification is 4, namely there are three categories and back-ground. And the output length of regression vector is 12, because each category has four positional parameters.

## 4 RESULTS

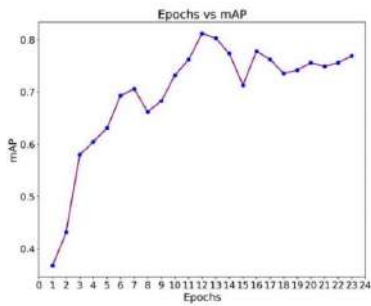
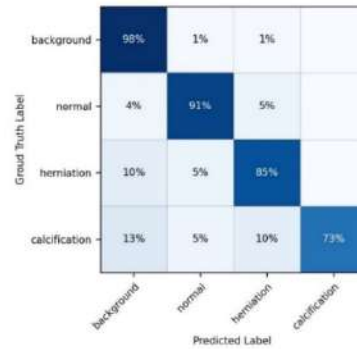
In this section, the performance of the HerniationDet model is evaluated on the CT dataset. To perform a better comparison, we implemented the model by Python. The IDE is PyCharm, and the PyTorch framework is used. The experiments are conducted on a single GPU (GeForce RTX 2070). The operating system is Ubuntu 16.04. After running for 87 minutes, with 25 epochs, the optimization completed. The experiments consist of two parts. In the first part, we perform the experiment to choose the better backbone network and ratio settings. Then we conduct the multi-modal learning to achieve a better performance on the MR dataset.

### 4.1 Experiment

In this subsection, in order to achieve the best performance of the model, we evaluate different backbone networks and ratios of the anchor. All networks are training from scratch with the same data argumentation. For this experiment, the CT dataset is used for training and evaluation. The detectors are optimized by SGD with a weight decay of 0.0005. And the learning rate is set to 0.0001 at begin of training. From Table 2, we can see the different performances of each backbone network, and the model with VGG16 has the best performance. As is shown in Fig. 8, the ratio setting of [1.70, 1.75, 1.80] has the best performance.

**Table 2: The comparison of different backbones on CT dataset and the different performance with multi-modal learning.**

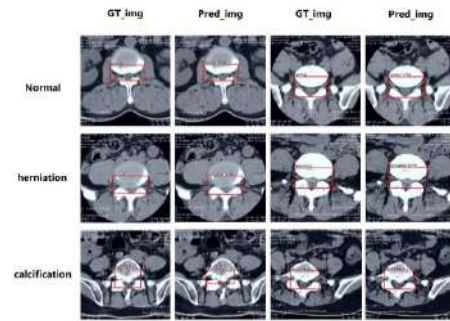
Model	Dataset		Multi-model learning	mAP
	CT	MR		
VGG16	✓			89.50%
VGG19	✓			87.45%
ResNet50	✓			47.58%
ResNet101	✓			86.81%
ResNetXt50_32×4d	✓			42.51%
ResNetXt101_32×8d	✓			72.77%
EfficientNetB5	✓			80.48%
VGG16	✓	✓		70.34%
VGG16	✓	✓	✓	81.24%

**Figure 9: The training curve of mAP on MR dataset with VGG16.****Figure 10: Confusion matrix of MR image dataset.**

## 4.2 Multi-Modal Learning

Modality means the specific way in which something happens, or we receive information. We observe objects with the eyes, hear sound with our ears, and feel the texture with our hands, all of which are different modes. Multi-Modal Machine Learning (MMML) aims to build models that can handle multiple modal information. Multi-modal fusion is one of the earliest research topics in MMML. It means the fusion of information from multiple modalities to complete a certain task. The reason why multi-modal fusion is so attractive is that it has many advantages. First, it will be more robust (e.g. audio-visual speech recognition). Second, there is complementary between multimodal information, it may be possible to find information that is difficult to observe on a single modality after fusion. Finally, a model that uses multi-modal information can still work when one of the modal information is missing.

Multi-modal medical image fusion [32] is one of the application fields of multi-modal fusion, mainly for MR, CT, PET, and SPECT images. As a three-dimensional image technology, CT scan has the advantages of short scanning time and high image resolution. However, the characterization of CT scan is limited, and it is difficult for CT equipment to reverse the slice of the image into one image in a short scanning time, while MR provides better clarity of soft tissue and higher spatial resolution. If the feature of CT and MR images can be multi-modally fused, perhaps the object detection of a lumbar disc can achieve better accuracy.

**Figure 11: The detection results of some MR images.**

To evaluate the performance with multi-modal learning, we did the experiment. From Table 2, we can see that multi-modal fusion achieved the better performance with the same network and the process of training can be seen from Fig. 9. The confusion matrix of the MR dataset is shown in Fig. 10, the reason why the accuracy of the calcified category is lower than others is that the sample of calcified lumbar disc is relatively small. And the detection results can be seen in Fig. 11

In order to test the performance of HerniationDet, we invited several experienced orthopedic experts to diagnose the MR images in the test set. From Table 3 and Table 4, we can see that the network

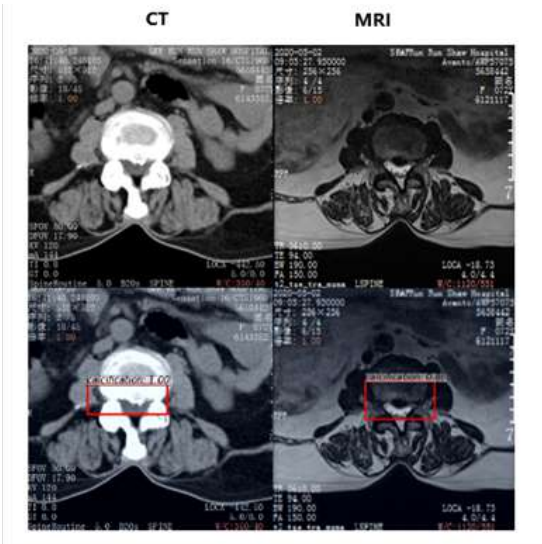


Figure 12: CT and MR test results of the same patient.

has a significant effect on the detection of MR images. Especially in the distinction between herniation with calcification and herniation without calcification, the HerniationDet model we use can capture more detail about the features that are difficult to observe with the naked eyes in MR images. Hence, it is more advantageous to use the computer vision to detect calcified lumbar disc herniation than the orthopedists. Finally, Fig. 12 shows the CT and MR results of the same patient.

5 DISCUSSION AND CONCLUSION

Deep learning has been widely used in the natural image domains, but seldom used for detecting spinal diseases. In this paper, the proposed approach well detected and classified the spinal disease in CT or MR images which is vital for making the preoperative planning.

Calcified lumbar disc herniation is quite a common problem in adult degenerative diseases. We applied the deep learning technique to automatically detect the location of the spine and classify the spinal diseases. In order to have a better performance, we pre-processed the MR and CT image dataset by normalization and data augmentation. With VGG16 backbone network and multi-modal fusion strategy, the performance of the detect network is competitive comparing with other existing methods. And the performance of the network is close to the experts who have the significant clinical experience on spine.

However, there are some limitations with our research. Firstly, because it is difficult to obtain and label medical images of the lumbar disc, collecting numerous labeled lumbar disc images is not easy. Currently, the number of images among different categories is extremely unbalanced. In the future, we plan to construct a more complete dataset for further study. In addition, the research of algorithms in the field of object detection has changed rapidly, more and more improvements have been proposed. It is worth to combine the new structure with HerniationDet to improve the performance of the network. In addition, further depth research in multi-modal machine learning is needed, which we will do in the following research.

Table 3: The detection results of HerniationDet on the test set.

	Normal	Herniation	Calcification
Number of detected objects	21	19	14
Number of true objects	20	16	15
TP	19	14	10
FP	2	5	4
FN	1	2	5
Precision(%)	90.47	73.68	71.43
Recall(%)	95.00	87.50	66.67
F1(%)	92.67	80.00	68.97

Table 4: The diagnosis of orthopedic experts on test set.

	Normal	Herniation	Calcification
Number of detected objects	19	18	14
Number of true objects	20	16	15
TP	18	13	9
FP	1	5	5
FN	2	3	6
Precision(%)	94.74	72.22	64.29
Recall(%)	90.00	81.25	60.00
F1(%)	92.31	76.47	55.17

## ACKNOWLEDGMENTS

This work was supported by Zhejiang Provincial Natural Science Foundation of China (Nos. LY21F020015, LY20F020015), National Natural Science Foundation of China (No. 61972121). The authors would like to thank the reviewers for their comments and suggestions in advance.

## REFERENCES

- [1] S. Schmidt, J. Kappes, M. Bergtholdt, V. Pekar, S. Dries, D. Bystrov, C. Schnorr, Spine detection and labeling using a parts-based graphical model, in: *Biennial International Conference on Information Processing in Medical Imaging*, Springer, 2007, pp. 122–133.
- [2] A. B. Oktay, Y. S. Akgul, Localization of the lumbar discs using machine learning and exact probabilistic inference, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2011*, 2011, pp. 158–165.
- [3] B. Glocker, D. Zikic, E. Konukoglu, D. R. Haynor, A. Criminisi, Vertebrae localization in pathological spine CT via dense classification from sparse annotations, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, 2013, pp. 262–270.
- [4] Bounds, Lloyd, Mathew, Waddell, A multilayer perceptron network for the diagnosis of low back pain, in: *IEEE 1988 International Conference on Neural Networks*, 1988, pp. 481–489.
- [5] S. Koompaiojin, K. A. Hua, C. Bhadrakom, Automatic classification system for lumbar spine X-ray images, in: *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, IEEE, 2006, pp. 213–218.
- [6] S. Hao, J. Jiang, Y. Guo, H. Li, Active learning based intervertebral disk classification combining shape and texture similarities, *Neurocomputing* 101 (2013) 252–257.
- [7] S. Ruiz-Espana, E. Arana, D. Moratal, Semiautomatic computer-aided classification of degenerative lumbar spine disease in magnetic resonance imaging, *Computers in biology and medicine* 62 (2015) 196–205.
- [8] I. Castro-Mateos, J. M. Pozo, A. Lazary, A. F. Frangi, 2D segmentation of intervertebral discs and its degree of degeneration from T2-weighted magnetic resonance images, in: *Medical Imaging 2014: Computer-Aided Diagnosis*, Vol. 9035, International Society for Optics and Photonics, SPIE, 2014, pp. 310–320.
- [9] C. W. Pfirrmann, A. Metzendorf, M. Zanetti, J. Hodler, N. Boos, Magnetic resonance classification of lumbar intervertebral disc degeneration, *Spine* 26 (17) (2001) 1873–1878.
- [10] L. Ramirez, N. G. Durdle, V. J. Raso, D. L. Hill, A support vector machines classifier to assess the severity of idiopathic scoliosis from surface topography, *Ieee transactions on information technology in biomedicine* 10 (1) (2006) 84–91.
- [11] L. Seoud, M. M. Adankon, H. Labelle, J. Dansereau, F. Cheriet, Prediction of scoliosis curve type based on the analysis of trunk surface topography, in: *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, IEEE, 2010, pp. 408–411.
- [12] H. Yu, F. He, Y. Pan, A novel segmentation model for medical images with intensity inhomogeneity based on adaptive perturbation, *Multimedia Tools and Applications* 78 (9) (2019) 11779–11798.
- [13] C. Chen, D. Belavy, W. Yu, C. Chu, G. Armbrrecht, M. Bansmann, D. Felsenberg, G. Zheng, Localization and segmentation of 3D intervertebral discs in MR images by data driven estimation, *IEEE transactions on medical imaging* 34 (8) (2015) 1719–1729.
- [14] A. Suzani, A. Seitel, Y. Liu, S. Fels, R. N. Rohling, P. Abolmaesumi, Fast automatic vertebrae detection and localization in pathological CT scans - a deep learning approach, in: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015, pp. 678–686.
- [15] D. Yang, T. Xiong, D. Xu, S. K. Zhou, Z. Xu, M. Chen, J. Park, S. Grbic, T. D. Tran, S. P. Chin, et al., Deep image-to-image recurrent network with shape basis learning for automatic vertebra labeling in large-scale 3D CT volumes, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2017, pp. 498–506.
- [16] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25*, Curran Associates, Inc., 2012, pp. 1097–1105.
- [17] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: *European conference on computer vision*, 2014, pp. 818–833.
- [18] K. Simonyan, A. Zisserman, Very deep convolutional networks for largescale image recognition, *arXiv preprint arXiv:1409.1556*.
- [19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabino-vich, Going deeper with convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [20] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [21] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [22] H. Chen, Q. Dou, X. Wang, J. Qin, J. C. Y. Cheng, P.-A. Heng, 3D fully convolutional networks for intervertebral disc localization and segmentation, in: *Medical Imaging and Augmented Reality*, 2016, pp. 375–382.
- [23] N. Lessmann, B. van Ginneken, I. Isgum, Iterative convolutional neural networks for automatic vertebra identification and segmentation in CT images, in: *Medical Imaging 2018: Image Processing*, Vol. 10574, 2018, p. 1057408.
- [24] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [25] R. Girshick, Fast R-CNN, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [26] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, in: *Advances in neural information processing systems*, 2015, pp. 91–99.
- [27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, SSD: Single shot multi-box detector, in: *European conference on computer vision*, Springer, 2016, pp. 21–37.
- [28] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [29] Tzutalin, *LabelImg*, Free Software: MIT License (2015). URL <https://github.com/tzutalin/labelImg>
- [30] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [31] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE transactions on pattern analysis and machine intelligence* 37 (9) (2015) 1904–1916.
- [32] J. Du, W. Li, K. Lu, B. Xiao, An overview of multi-modal medical image fusion, *Neurocomputing* 215 (2016) 3–20.

# MCSC-UTNet: Honeycomb lung segmentation algorithm based on Separable Vision Transformer and context feature fusion

Wei Jianjian,  
College of Software, Taiyuan  
University of Technology, Jinzhong,  
China  
18406586202@163.com

Li Gang,  
College of Software, Taiyuan  
University of Technology, Jinzhong,  
China  
tx2090@126.com

He Kan,  
College of Mathematics, Taiyuan  
University of Technology, Jinzhong,  
China  
hekanquantum@163.com

Li Pengbo,  
College of Software, Taiyuan  
University of Technology, Jinzhong,  
China  
553121336@qq.com

Zhang Ling,  
College of Software, Taiyuan  
University of Technology, Jinzhong,  
China  
zl2090@126.com

Wang Ronghua,  
Shanxi Bethune Hospital, Taiyuan,  
China  
wangronghuayaya@163.com

## ABSTRACT

**Abstract:** Due to the problems of more noise and lower contrast in X-ray tomography images of the honeycomb lung, and the poor generalization of current medical segmentation algorithms, the segmentation results are unsatisfactory. We propose an automatic segmentation algorithm MCSC-UTNet based on SepViT with contextual feature fusion for honeycomb lung lesions to address these problems. Firstly, a Multi-scale Channel Shuffle Convolution (MCSC) module is constructed to enhance the interaction between different image channels and extract the local lesion feature at different scales. Then, a Separable Vision Transformer (SepViT) module is introduced at the bottleneck layer of the network to enhance the representation of the global information of the lesion. Finally, we add a context-aware fusion module to relearn the encoder feature and strengthen the contextual relevance of the encoder and decoder. In comparison experiments with eight prevalent segmentation models on the honeycomb lung dataset, the segmentation metrics of this method, Jaccard coefficient, mIoU, and DSC are 90.85%, 95.32%, and 95.07%, with Jaccard coefficient improving by 3.56% compared with that before. Compared with medical segmentation models such as TransUNet, Sharp U-Net, and SETR, this paper's method has improved results and segmentation performance.

## KEYWORDS

honeycomb lung segmentation, medical image processing, Transformer, feature fusion, multi-scale channel shuffle convolution

### ACM Reference Format:

Wei Jianjian., Li Gang., He Kan., Li Pengbo., Zhang Ling., and Wang Ronghua., 2023. MCSC-UTNet: Honeycomb lung segmentation algorithm based on Separable Vision Transformer and context feature fusion. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590093>

(CACML 2023), March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 7 pages. <https://doi.org/10.1145/3590003.3590093>

## 1 INTRODUCTION

Honeycomb lung is a fatal, insidious, and devastating diffuse lung disease that presents with persistent alveolar epithelial damage leading to pulmonary fibrosis and irreversible damage to the lungs<sup>[1]</sup>. Today, most clinical diagnoses of honeycomb lung disease rely on specialist radiologists to analyze patients' CT images of the lung for lesions<sup>[2]</sup>. In diagnosing the condition, professional physicians perform manual visual recognition of CT images based on the knowledge and experience they already possess. Still, the continuous mental effort and long work hours are prone to visual fatigue and inaccuracy. It will result in high subjectivity in the diagnostic results, accompanied by misdiagnosis and missed diagnoses. It also increases the difficulty of later treatment of patients. Therefore, it is essential to use image segmentation methods to automatically segment lung lesions to assist physicians in making accurate diagnoses of the extent of a patient's condition, thereby providing appropriate solutions to guide clinical decision-making and prognostic treatment<sup>[3-4]</sup>. For clinical purposes, it has significant research value.

In recent years, CNN-based methods have been widely used in medical image processing due to the powerful feature representation capability of deep learning and the ability to model complex tasks. In particular, U-shaped convolutional neural networks with an encoder-decoder structure have achieved remarkable performance in medical image segmentation. The encoder-decoder network model UNet<sup>[5]</sup> is first proposed to obtain better segmentation results on three medical datasets by fusing high-level feature from the upsampling stage with the low-level feature from the downsampling stage using a skip connection. To address the problem of easy loss of spatial feature in convolutional computation, CE-Net<sup>[6]</sup> generates more semantic feature maps by fusing the contextual connectors of dense convolutional blocks. The network could reduce information loss and obtain fine edges of the target. Sharp U-Net<sup>[7]</sup> applies sharpened convolution kernels to generate intermediate feature maps to solve the over-segmentation problem caused by semantic gaps. To address the drawback of missing information

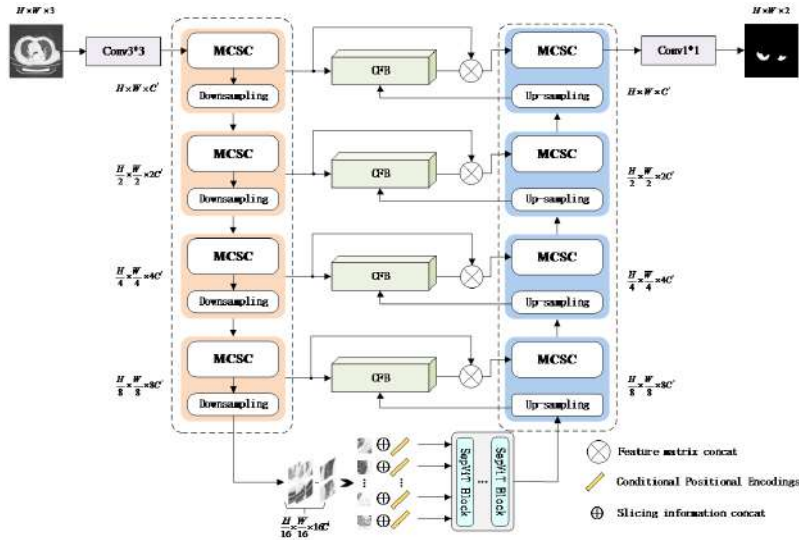


Figure 1: network structure diagram

at multiple scales in the network, HDA-ResUNet<sup>[8]</sup> replaces the bottom convolutional layer of UNet with an expanded convolution layer containing a channel attention mechanism to fuse information from different size perceptual fields. The above CNN-based methods have demonstrated the applicability of UNet networks in medical segmentation.

Although the CNN method based on the encoder-decoder structure has excellent advantages in extracting local feature of images, the traditional convolution operation in whole convolutional networks leads to duplicate feature redundancy and affects the segmentation effect. The limitations of convolutional operations pose challenges for studying global information in an image, especially for pixel-level tasks such as semantic segmentation, which is crucial.

Recently, the Transformer family has broken the absolute status of CNNs in computer vision tasks, and these works<sup>[9–11]</sup> have demonstrated the excellent modeling capabilities of the Transformer for global information in images. Vision in Transformer (ViT)<sup>[12]</sup> is one of the many variants of Transformer that excels in medical image segmentation tasks. TransUNet<sup>[13]</sup> uses the ViT module as the encoder base module of the segmentation network, using the Transformer’s excellent global information modeling capability to achieve precise localization of the lesion site. To segment target edges more accurately, Medical Transformer<sup>[14]</sup> uses the gated axial Transformer module to obtain more accurate location information. Therefore, the Transformer architecture-based medical segmentation network can establish global relationships in images through the computation of attention mechanisms<sup>[15]</sup>, fully extracting global feature and completing the segmentation task of medical targets with high quality.

The main contributions of this paper are as follows.

(1) To construct a lightweight Multi-scale Channel Shuffle Convolution (MCSC) to obtain multi-scale feature information in images at a small cost, increase the correlation between feature map channels.

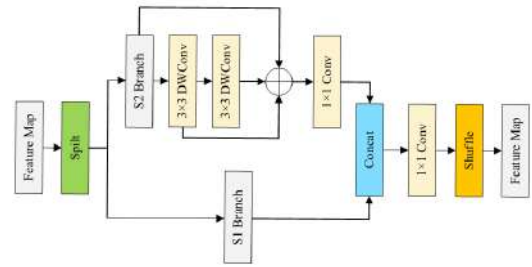


Figure 2: channel shuffle convolution block

(2) We introduce SepViT as a feature connector between the encoder and the decoder to enhance the feature representation of global information and improve the segmentation accuracy of the lesion region.

(3) To reduce the semantic gap due to a mismatch in the receptive domain, we use the Context-aware fusion Block (CFB) to reconstruct the feature distribution in the encoder, enhance the semantic correlation between the encoder and decoder, and achieve contextual feature fusion for semantic matching.

(4) MCSC-UTNet model achieves better segmentation performance on the honeycomb lung dataset.

## 2 MATERIALS AND METHODS

### 2.1 Proposed MCSC-UTNet Architecture

The overall structure of the network model MCSC-UTNet in this paper is shown in Figure 1. The model proposed in this paper consists of four parts: encoder network, bottleneck layer, decoder network and skip connection. The encoder network contains a 3x3 convolution layer and four multi-scale channel shuffle convolution blocks. At the bottleneck layer of the network, the SepViT module is adopted to improve the interaction of global information in the

feature. In the skip connection, the four feature maps obtained by the encoder convolution block corresponded to the feature maps sampled on the decoding network part. They are fed separately into the context-aware fusion module of the corresponding stage for feature enhancement operations. The improved feature fusion scheme for low-level and high-level feature can bridge the semantic gap between feature. The feature fusion operation could avoid losing the original feature due to direct summation operations.

## 2.2 Multi-scale channel shuffle convolution blocks

Most existing segmentation networks usually use high-level channel convolution to extract deep abstract feature of lesions in medical images. However, simple stacking of fixed-size convolutional layers increases the model's computational effort and extracts single, redundant feature, thus reducing segmentation accuracy. Therefore, this paper improves on the Shuffle Block in the ShuffleNet-V2 network<sup>[16]</sup> and proposes a novel multi-scale channel shuffle convolution block. The module structure is shown in Figure 2. After separating the feature map, the module is divided into two branches S1 and S2. Branch S2 has two  $3 \times 3$  depth separable convolution and  $1 \times 1$  convolution. It could obtain the multi-scale feature in the image using a residual join operation. Afterward, the convolved feature map is fused with the feature map of branch S1. At this point, a channel shuffle operation is applied to it to exchange the information between the various channels of the feature map. The convolution module increases the perceptual field and realizes the interaction between channels by separating and exchanging the feature map channels. Using  $1 \times 1$  convolution can effectively reduce the convolution calculation and similarity of focal feature and improve the segmentation effect of the honeycomb lung.

## 2.3 SepViT-based contextual connectors

To build a feature extractor that focuses more on global communication and have larger perceptual field, the SepViT module<sup>[17]</sup> is introduced into the bottleneck layer of the UNet network. The communication between local and global information in different subspaces of the network is enriched by using depth-separable self-attention to increase the global feature weight assignment. Thereby the global perceptual field of the network becomes more extensive. The module's structure is shown in Figure 3. The SepViT consists of the Depthwise Separable Self-Attention (DSSA) layer, MLP and LN Layer. The Depthwise Separable Self-Attention layer consists of Depthwise Self-Attention and PointWise Self-Attention. Depthwise Self-Attention can receive slicing and conditional location encoding information. PointWise Self-Attention is used to fuse information between different slices, receiving the feature and dependent location encoding information from the Depthwise Self-Attention extraction. The MLP Layer learns the non-linear relationships between feature. To prevent gradient disappearance and explosion problems and speed up the model's convergence, residual connectivity and layer normalization operations are used in the SepViT block to train deeper networks.

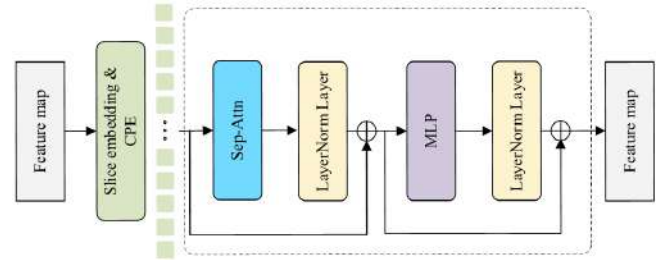


Figure 3: module structure diagram

## 2.4 Context-aware fusion Block

In traditional UNet networks, only the encoder and decoder feature are directly connected to fuse low-level semantic information with the high-level abstract feature. They ignore the contextual relationship between feature, and the representation of essential feature is weakened. Inspired by Meng<sup>[18]</sup>, a context-aware fusion block is introduced to reconstruct the jump connection structure.

The context-aware fusion module is shown in Figure 4, which receives the low-level feature map  $F_h$  from the encoder and the high-level feature map  $F_l$  from the decoder. Firstly, the global average pooling operation (GAP) generates feature maps with global spatial information in the first step. Then, the multilayer perceptron layer models the context information in the low-level and high-level feature maps to generate vectors  $h$  and  $l$  (where  $h$  and  $l$  denote the weight vectors generated by the high-level and low-level feature maps, respectively). In the second step, the weight vectors are multiplied with the two feature maps to generate redistributed feature maps  $F_l'$  and  $F_h'$ . In the third step, two  $3 \times 3$  convolution are used to achieve weighted feature fusion, using residual connections to accept information from high-level feature and construct a deeper network.

# 3 EXPERIMENT

## 3.1 Datasets introduction

In the present study, we collect CT data from a number of patients with honeycomb lung from different tomographic scanners provided by Baiqien Hospital in Shanxi Province, with 2350 CT images from 163 patients with interstitial lung disease. All patient data is obtained from the same center to avoid discrepancies in the data obtained. All CT image data are anonymised. Moreover, each honeycomb lung image is annotated by an experienced thoracic radiologist. We used Gaussian noise with contrast enhancement to amplify the dataset. The total number is 7050. The CT images of the patient is shown in Figure 5 (The first column is the original image, the second column is the image containing Gaussian noise, the third column is the contrast-enhanced image, and the fourth column is the GT).

## 3.2 Experiment environment

The experiments are conducted under the deep learning framework PyTorch, with a 64-bit Ubuntu 18.04.5 operating system, an NVIDIA TitanV Volta GPU server with 12GB of video memory, and python 3.7 and PyTorch 1.6.0. The optimal segmentation network is trained

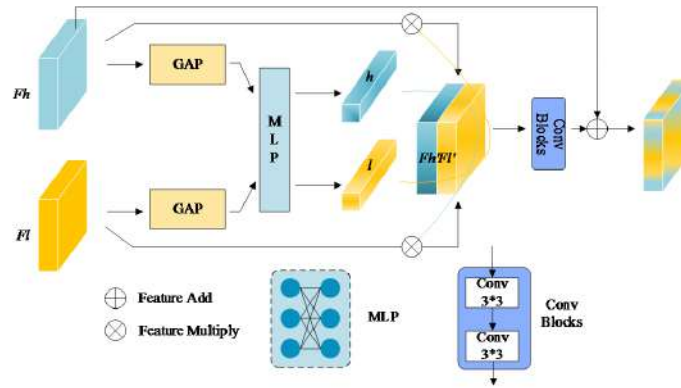


Figure 4: fusion module diagram

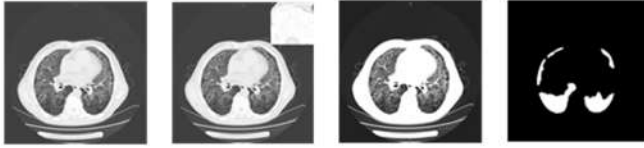


Figure 5: lung CT images

using an SGD optimizer. We set the momentum coefficient momentum to 0.7, the initial learning rate  $lr$  to 0.05, the experiment to decay by 50% every 50 Epochs, the batch size to 4, and the number of training iterations Epoch to 200. To comprehensively and objectively evaluate the segmentation performance of the proposed method, we use the Jaccard coefficient, mIoU, and DSC as evaluation metrics to analyze the prediction results of the network from different perspectives.

### 3.3 Ablation experimental analysis

The ablation experiments are conducted in this paper using the honeycomb lung dataset, and the results are shown in Table 1. The table shows that the segmentation performance of the U-network with a SepViT-based contextual connector is significantly improved compared with the original UNet. The Jaccard coefficient, mIoU, and DSC are 88.32%, 94.02%, and 93.68%. The above data proves that SepViT can capture the long-distance relationship. Thereby allowing the link between global and local feature information to be facilitated.

### 3.4 Analysis of experimental results for different values of the number of SepViT blocks

From the first loss curve subgraph in Figure 6, we can see that when the number of SepViT blocks is 3 and 4, the loss curve oscillates more points and the loss value is large. When the number of SepViT blocks is 5, the loss function reaches the minimum loss value in the regional smoothing stage; in the second loss curve subgraph in Figure 6, the loss curves for  $N=7$  and 8 still do not reach convergence when epoch > 100. Then, the third loss curve subgraph in Figure 6 compares the two setup parameters with better convergence of the loss values, from which it can be seen that the loss at  $N=5$  has reached convergence at iteration 125 and the curve tends to be smooth; the loss curve for  $N=6$  still has a small oscillation point at epoch=175.

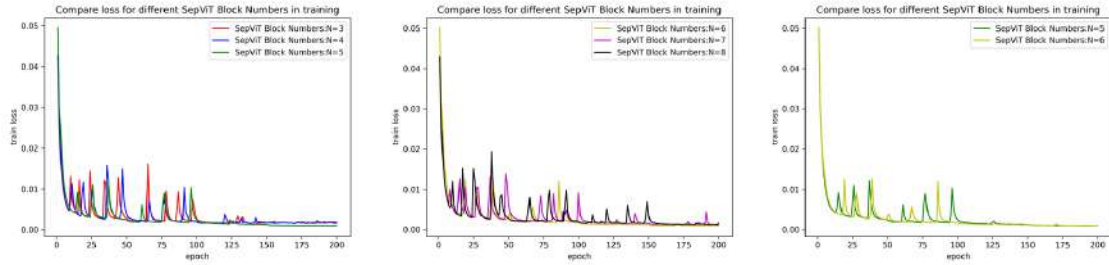
In order to verify the effect of the number of SepViT blocks on the segmentation effect, a set of comparison experiments are set up. The test results are shown in Table 2. When the number of coding blocks was set to 5 and 6, the more vital local and global information extraction ability of the SepViT module was activated. The experimental model with the number of coding blocks  $N=5$  was within 0.01% error in all evaluation metrics, such as the Jaccard coefficient, compared to the model with  $N=6$ . Therefore, to make the segmentation of honeycomb lung the best, the number of SepViT Blocks is selected to be 5 in the experiment to evaluate the segmentation accuracy and the computational cost of the model in all six groups of experiments.

Table 1: experiments

Method	Jaccard coefficient(%)	mIoU(%)	DSC(%)
UNet	87.29	93.50	92.85
UNet+SepViT	88.32	94.02	93.68
UNet+SepViT+CFB	89.01	94.38	94.09
UNet+SepViT+CFB+CSC	90.85	95.32	95.07

**Table 2: results of SepViT Blocks on segmentation mode**

SepViT Block Numbers(N)	Jaccard coefficient(%)	mIoU(%)	DSC(%)
3	77.16	88.21	84.95
4	76.33	87.73	84.02
5	90.85	95.32	95.07
6	90.86	95.31	95.02
7	76.36	87.79	83.99
8	78.89	89.12	86.32

**Figure 6: lung CT images of different patients****Table 3: experiments**

Method	Jaccard coefficient(%)	mIoU(%)	DSC(%)	Params/M
UNet <sup>[5]</sup>	87.29	93.50	92.85	31.04
CE-Net <sup>[6]</sup>	89.82	94.79	94.54	13.39
Sharp U-Net <sup>[7]</sup>	90.24	95.01	94.77	31.06
HDA-ResUNet <sup>[8]</sup>	90.63	95.20	94.99	33.73
TransUNet <sup>[13]</sup>	87.87	93.79	93.38	45.68
DeepLabv3+ <sup>[19]</sup>	88.11	93.91	93.50	59.33
R2U-Net <sup>[20]</sup>	62.75	80.97	74.53	29.89
SETR <sup>[21]</sup>	85.50	92.57	91.98	43.25
Proposed Method	90.85	95.32	95.07	34.15

### 3.5 Analysis of comparative experiments

The experiments are conducted to compare several segmentation models from four different perspectives to evaluate the effectiveness of this paper's method MCSC-UTNet. As can be seen from Table 3, compared with DeepLabv3+ and CE-Net networks, the Jaccard coefficient, mIoU, and DSC of the proposed MCSC-UTNet are increased by about 1%. For UNet variant networks such as Sharp U-Net and HDA-ResUNet, the improvement in the Jaccard coefficient is still greater than 1%, although the improvement in both mIoU and DSC is not significant. The evaluation metrics of TransUNet and SETR networks are lower than the method in this paper due to the self-attentive mechanism in Transformer, which pays less attention to local information. In summary, the proposed MCSC-UTNet network's segmentation performance has improved compared with other networks.

In the comparison experiments, the CT images of five patients are used for testing in this paper, and the segmentation results of

other networks are shown in Figure 7. As seen in Figure 7, the UNet and Sharp U-Net networks do not show discontinuities in the lesion results. This results in a loss of edge information in the lesion region, resulting in multiple jagged segmentation results. CENet, TransUNet, and SETR networks mis-segmented the background in the middle of the lung lobe as the target region. For CENet and TransUNet, the top of the lesion region in the second original image appears to be over-segmented. Compared to the other networks, MCSC-UTNet improves the lesion shape and edges. The segmented lesion shape is identical to the annotated image, allowing satisfactory segmentation results to be obtained.

The comparison experiments also include the size of each model covariate. The smaller the number of model parameters, the higher the DSC value, which proves that the generalization and applicability of the model is better. Compared with the lightweight models CENet and UNet, the number of model parameters in this paper has slightly increased, but the evaluation index DSC has the highest

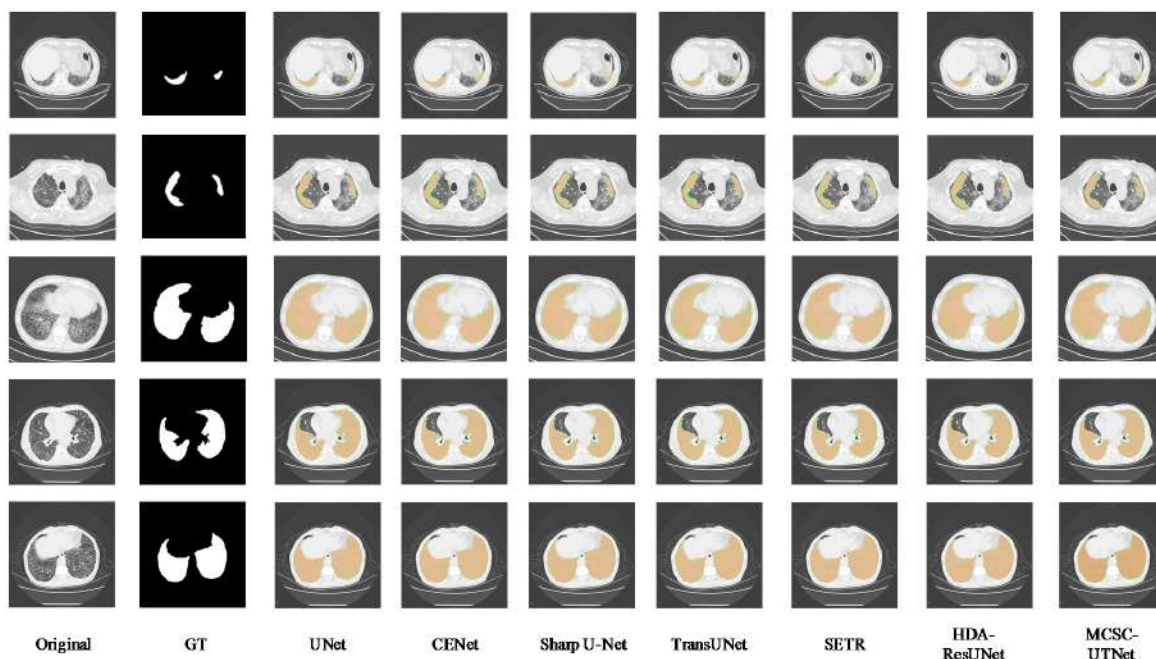


Figure 7: renderings of multiple networks

value; compared with the latest SharpUNet and HAD-ResUNet models, the difference in the number of parameters is only 0.3M, and this method still has the best segmentation effect. Compared with the Transformer architecture SETR, TransUNet and other models, the difference in the size of model parameters is nearly 30%. The difference in the number of parameters is nearly 30%, and the value of the segmentation index DSC is the highest, so this method can achieve the best performance.

## 4 CONCLUSION

This paper proposes a U-shaped image segmentation network combining multi-scale channel shuffle convolution, SepViT and context-aware fusion modules. The model solves the problem of low segmentation accuracy due to irregular shape of lesion sites and unbalanced data categories in honeycomb lung CT images. Furthermore, it can improve the problem of missing edges of lesion sites in the segmentation process and increase the segmentation accuracy. Ablation experiments are conducted by performing on a honeycomb lung dataset. Compared with other networks, MCSC-UTNet achieves superior results in all three evaluation metrics. Thus, the experiments demonstrate the advanced nature of the method in this paper.

However, the network in this paper still suffers from high computational cost and loss of some boundary information. In future work, we will adopt methods such as model compression and parameter sharing to simplify the model size. With the focus on lesion boundary segmentation, the feature extraction capability and generalization of the algorithm will be improved to further promote the development of applications in the field of automatic medical image segmentation.

## ACKNOWLEDGMENTS

This research was supported by the Central Project Leading Local Science and Technology Development Fund (Grant No. YDZJSX2021C004, YDZJSX2022A016) and the Natural Science Foundation of Shanxi Province (Grant No. 20210302124554). We also would like to thank Shanxi Bethune Hospital for contribution of honeycomb lung medical images and annotated images.

## REFERENCES

- [1] YOO S A, PARK H E, KIM M, *et al.* A case report of pirfenidone-induced lichenoid drug eruption in a patient with idiopathic pulmonary fibrosis. *Ann Dermatol.* 2022; 34(2): 136-8.
- [2] MAHER T M, BENDSTRUP E, DRON L, *et al.* Global incidence and prevalence of idiopathic pulmonary fibrosis. *Respir Res.* 2021; 22(1): 197.
- [3] TOMIOKA H, AMIMOTO H, FUJII H, *et al.* Asymmetrical interstitial lung disease suggested to be due to hypoplasia of the unilateral pulmonary artery: A Case Report with a 20-year Follow-up. *Internal Medicine.* 2021; 60(8): 1265-70.
- [4] ALJABRI M, ALGHAMDI M. A review on the use of deep learning for medical images segmentation. *Neurocomputing.* 2022; 506: 311-35.
- [5] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//*Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention.* Munich, GERMANY: Springer, 2015: 234-241.
- [6] GU Z, CHENG J, FU H, *et al.* CE-Net: Context Encoder Network for 2D Medical Image Segmentation. *IEEE TRANSACTIONS ON MEDICAL IMAGING.* 2019; 38(10): 2281-92.
- [7] ZUNAIR H, BEN HAMZA A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *COMPUTERS IN BIOLOGY AND MEDICINE.* 2021; 136.
- [8] WANG Z, ZOU Y, LIU P X. Hybrid dilation and attention residual U-Net for medical image segmentation. *COMPUTERS IN BIOLOGY AND MEDICINE.* 2021; 134.
- [9] Xu X, Feng Z, Cao C, *et al.* An improved swin transformer-based model for remote sensing object detection and instance segmentation. *Remote Sensing.* 2021; 13(23): 4779.
- [10] Jin Y, Han D, Ko H. Trseg: Transformer for semantic segmentation. *Pattern Recognition Letters.* 2021; 148: 29-35.

- [11] Wang H, Xie S, Lin L, *et al.* Mixed transformer u-net for medical image segmentation[C]//ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022: 2390-2394.
  - [12] Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. [published online June 6, 2021].
  - [13] CHEN J, LU Y, YU Q, *et al.* TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. [published online February 8, 2021]. <https://arxiv.org/abs/2102.04306>.
  - [14] VALANARASU J M J, OZA P, HACIHALILOGLU I, *et al.* Medical Transformer: Gated Axial-Attention for Medical Image Segmentation [C]//International Conference on Medical Image Computing and Computer Assisted Intervention. Strasbourg, France: Springer, 2021: 36-46.
  - [15] Guo M H, Xu T X, Liu J J, *et al.* Attention mechanisms in computer vision: A survey. Computational Visual Media. 2022; 1-38.
  - [16] MA N, ZHANG X, ZHENG H-T, *et al.* ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture De-sign[C] //Proceedings of 15th European Conference on Computer Vision. Munich, GERMANY: COMPUTER VISION, 2018: 122-138.
  - [17] Li W, Wang X, Xia X, Wu J, Xiao X, Zheng M, Wen S. SepViT: Separable Vision Transformer. [published online May 7, 2022]. <https://arxiv.org/abs/2203.15380>
  - [18] LOU M, MENG J, QI Y, *et al.* MCRNet: Multi-level con-text refinement network for semantic segmentation in breast ultrasound imaging. Neurocomputing. 2022; 470: 154-69.
  - [19] CHEN L C E, ZHU Y K, PAPANDREOU G, *et al.* Encoder-Decoder with Atrous Separable Convolution for Seman-tic Image Segmentation [C]//Proceedings of 15th Euro-pean Conference on Computer Vision. Munich, GERMANY: Springer, 2018: 833-851.
  - [20] ALOM M Z, YAKOPCIC C, HASAN M, *et al.* Recurrent residual U-Net for medical image segmentation. Journal of medical imaging. 2019; 6(1): 014006.
  - [21] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, *et al.* Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2021: 6877-6886.
- Li Gang,, He Kan,, Li Pengbo,, Zhang Ling,, Wang Ronghua,,

# Research on Colorization of Qinghai Farmer Painting Image Based on Generative Adversarial Networks

Chunyan Peng, C.P, and Peng\*  
Department of Computer Qinghai  
Normal University Xining, China;  
State Key Laboratory of Tibetan  
Intelligent Information Processing  
and Application, Qinghai Normal  
University, Xining, China  
pcy@qhnu.edu.cn

Xueya Zhao, X.Z, and Zhao  
Department of Computer Qinghai  
Normal University Xining, China;  
State Key Laboratory of Tibetan  
Intelligent Information Processing  
and Application, Qinghai Normal  
University, Xining, China  
615348915@qq.com

Guangyou Xia, G.X, and Xia  
Department of Computer Qinghai  
Normal University Xining, China;  
State Key Laboratory of Tibetan  
Intelligent Information Processing  
and Application, Qinghai Normal  
University, Xining, China  
1978644276@qq.com

## ABSTRACT

At present, deep learning method is widely used in the field of gray image colorization. Qinghai farmer painting has distinct national characteristics. The farmer painting has bright colors, high saturation, chaotic color distribution and low color contrast, so it is difficult to restore the image color with high fidelity by using the general deep learning image colorization method. The Pix2Pix generation adversarial network of grayscale image colorization method uses the Leaky ReLU function as the activation function. The proposal algorithm replaces the maximum pooling layer with the convolution layer to retain more image feature information and further to improve the color simulation effect. Meanwhile, in view of the lack of relevant Qinghai farmer painting data set, the data set of Qinghai farmer paintings is constructed to meet the needs of network training. The experimental results show that the improved method further improves the color effect and can generate high quality color images of Qinghai farmer paintings with more real color distribution.

## CCS CONCEPTS

• **Computing methodologies** → Machine learning; Machine learning algorithms; Artificial intelligence; Computer vision; Computer vision problems; Image segmentation.

## KEYWORDS

Qinghai farmer's painting, gray-scale image, colorization, generative adversarial networks

### ACM Reference Format:

Chunyan Peng, C.P, and Peng, Xueya Zhao, X.Z, and Zhao, and Guangyou Xia, G.X, and Xia. 2023. Research on Colorization of Qinghai Farmer Painting Image Based on Generative Adversarial Networks. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*,

\*Place the footnote text for the author (if applicable) here.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590094>

March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 9 pages.  
<https://doi.org/10.1145/3590003.3590094>

## 1 INTRODUCTION

Qinghai farmer's painting is a unique kind of painting in Qinghai traditional folk art. However, at present, the number of inheritors of the older generations of Qinghai farmer's painting is decreasing, and most of the younger generations choose to go out to work for a living. The creative skills of Qinghai farmer's painting are facing the risk of no one inheriting them. Moreover, most of the relevant images recorded in some ancient documents and publications are gray-scale images, which are not conducive to the research and digital protection of Qinghai farmer's painting. Therefore, it is of great significance to use the gray-scale image colorization technology to convert the gray-scale images of Qinghai farmer's painting into high-quality color images.

Gray-scale image colorization technology has been one of the research hotspots in the field of digital image processing. Gray-scale image colorization technology has high application value and academic research value in the fields of medicine, film and television, and color restoration art images. The current gray-scale image colorization methods are mainly divided into traditional colorization methods and colorization methods based on depth learning [1–3].

The traditional gray-scale image colorization methods mainly include user interaction based on colorization method and reference image based color transfer method.

(1) Colorization method based on user interaction: this method mainly marks color points on the gray-scale image by the user, and then spreads the color marked by the user to the whole image according to the similarity of pixel information of the gray-scale image [4]. Levin et al. [5] believe that pixels with similar brightness information in the image should have similar colors, and realize the color diffusion through the global optimization algorithm; Huang et al. [6] proposed an adaptive edge detection algorithm, which extracts the edge of the object in the gray-scale image by processing the edge information of the gray-scale image, and then carries out color diffusion in the segmented image area, so as to alleviate the problem of color leakage in the coloring process of the gray-scale image; Yatziv et al. [7] believe that the smaller the geodesic distance between the two pixels indicates that the chromaticity information of the two pixels is more similar. Through the distance function and the geodesic distance, determine the weight of the

color value of the shaded area closest to the uncolored pixel, and finally calculate the color value of the uncolored pixel; Qu et al. [8] proposed a colorization method for black-and-white comics, which uses Gabor wavelet filter to calculate the texture similarity of image area, reduces the number of color marks added by the user, and uses the method of continuous level set to propagate the color of user marks. At the same time, by calculating the change intensity of image gradient, the problem of color leakage in the process of color propagation is alleviated; Luan et al. [9] grouped the pixels with similar brightness and texture in the image according to the intensity continuity and texture similarity of the image, divided the image into several coherent regions, marked the selected pixels in each coherent region with color, and then colored the remaining pixels in the coherent region through color mapping, so as to realize the colorization of gray-scale image; Heu et al. [10] proposed a color transfer method based on priority. By calculating the priority of non-source pixels adjacent to the source pixels, and using the priority recognition algorithm, the color marked by the user on a group of source pixels is transferred to its adjacent non-source pixels. The non-source pixels become source pixels after coloring, and then the gray-scale image is colored by repeating this operation.

The colorization method based on user interaction is prone to the problem of color leakage. It needs to add a large number of color markers to color the gray-scale images with complex textures and scenes. The user interaction process is cumbersome, the operation is difficult and the efficiency is not high.

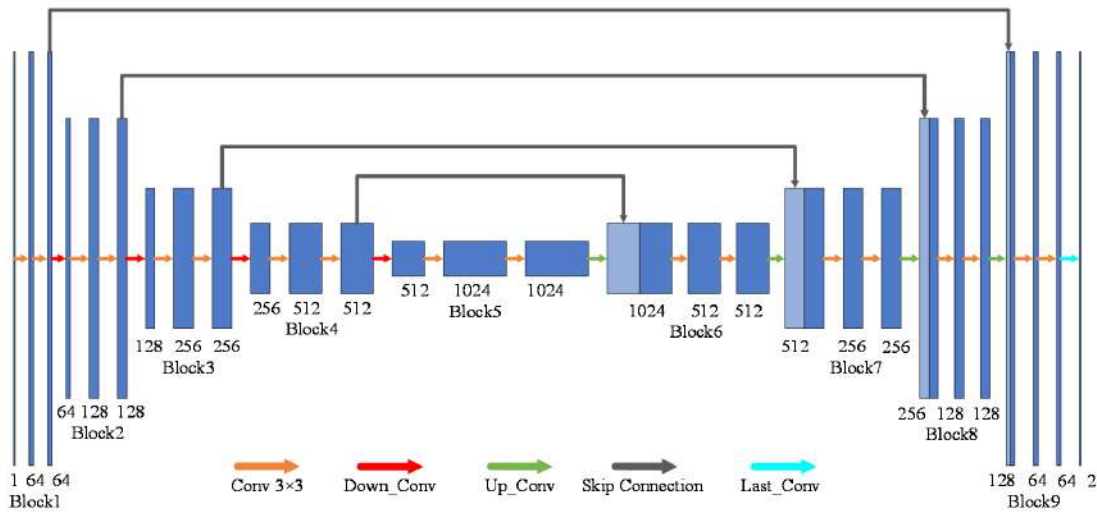
(2) Colorization method based on reference image: Welsh et al. [11] proposed a technology to transfer the color of the reference image to the target image to complete the colorization of the target image, reducing the workload of user interaction in the coloring process [12]; Chapiat et al. [13] used the image segmentation algorithm to achieve the global consistency of the image, so as to reduce the impact of the spatial coherence of the pixels on the color of the pixels. By estimating the probability distribution of all possible colors of each pixel of the target image, the pixel color with the highest matching degree in the reference image was transferred to the target image; Li et al. [13, 14] proposed a cross-scale texture matching method, locally considering the appropriate matching scale to minimize the matching error and spatial change of scale, and detecting and correcting unreasonable matching results through image position information, so as to improve the robustness and quality of coloring results; Cao et al. [15] proposed a locally adaptive gray-scale image coloring method, which uses support vector machine [16] and improved simple linear iterative clustering algorithm [17] to classify the reference image and gray-scale image respectively, and uses the classification results and the similarity of texture information between images to find pixels matching the reference image in the target image, Then the reference pixel color is transferred to the corresponding gray-scale image pixels with high confidence. At the same time, the colorization effect is further improved by combining with manual interaction.

The gray-scale image colorization effect of this method largely depends on the selection of color reference image. Although the number of manual interactions is reduced to a certain extent, the operation is still difficult.

Because the traditional gray-scale image colorization methods need tedious manual interaction, the operation is difficult, the overall efficiency is low, and the problem of color leakage is easy to occur, so the traditional colorization methods have been replaced by the colorization methods based on deep learning.

Among the current deep learning methods, Generative Adversarial Networks (GAN) [18] has made great achievements in the field of image generation, and can generate higher quality images [19]. In view of the excellent performance of the generative adversarial networks, many scholars have applied them to the colorization of gray-scale images. Nazeri et al. [20] proposed a gray-scale image colorization method based on deep convolution generative adversarial networks (DCGAN), and compared the colorization effects of the network on CIFAR-10 image dataset and Places365 image dataset; Liu et al. [21] used the Cycle-GAN [22] to color the gray-scale image of the portrait, adopted the joint consistent loss training network model, and introduced the method of multi feature fusion in the discriminative network, which improved the colorization accuracy of the gray-scale image of the portrait under the complex background; Liang et al. [23] proposed an improved DualGAN [24] near-infrared image colorization method, which uses the integral of the generative network loss to train the discriminative network to reduce the probability of the wrong image generated by the generative network and improve the generalization ability of the discriminative network; Zhao et al. [25] proposed a video colorization technology (VCGAN) combined with hybrid generative adversarial networks, using two pre-trained ResNet-50-In networks [26] as global feature extractor and placeholder feature extractor respectively to improve video colorization quality and maintain video time-space consistency; Zhang et al. [27] used the deep aggregation structure network to realize the colorization of gray-scale images, adopted the generative adversarial networks structure, added long connections to the traditional network structure, improved the stability of the network structure and the utilization of image feature information, and improved the colorization performance to a certain extent; Wu et al. [28] proposed a gray-scale image colorization method based on foreground semantics, which uses the foreground network to extract the low-level and high-level feature information of the foreground part of the image, and fuses the extracted foreground semantic information with the panoramic feature information of the image, so as to achieve a more natural colorization effect; Wan et al. [29] used U-Net as the generative network, introduced the adaptive feature fusion module and attention mechanism while deepening the network depth, so as to alleviate the color overflow problem in the colorization process and improve the colorization performance of the generative network.

Although the colorization methods based on deep learning can realize the automatic colorization of gray-scale images without manual intervention and tedious manual interaction, the current colorization methods based on deep learning are used to color the gray-scale images of Qinghai farmer's painting. The generated images have the problems of chaotic color distribution and low color contrast. Therefore, this paper proposes a gray-scale image colorization method for Qinghai farmer's painting based on Pix2Pix generative adversarial network [30], improves the Pix2Pix generative adversarial network, uses Leaky ReLU as the activation function, and uses the convolution layers to replace the maximum



**Figure 1: According to the characteristics of Qinghai peasant paintings, this paper improves the generation network of Qinghai peasant paintings to further enhance the color effect**

pool layers of the original generative network, so as to further improve the utilization of image feature information by the network, and then restore more color details of the image, the problem of low color contrast of the generated color image is alleviated.

## 2 THE COLORIZATION METHOD OF GRAY-SCALE IMAGE OF QINGHAI FARMERS' PAINTING BASED ON PIX2PIX GENERATIVE ADVERSARIAL NETWORK

### 2.1 Generative Network Structure

The traditional generative adversarial networks model has the defect of low utilization of image feature information. Compared with the traditional generative adversarial networks, the Pix2Pix generative adversarial network has a higher utilization of image feature information. Its generative network adopts the U-Net network structure, and the U-Net network structure adds a jump connection mechanism. It can fuse the feature map output from the convolution layers in the down-sampling process with the corresponding feature map output from the deconvolution layers in the image reconstruction process, so as to improve the reuse rate of image feature information in the network. Compared with the traditional generative adversarial networks, Pix2Pix generative adversarial network uses U-Net as the generative network can not only improve the utilization of image feature information, but also restore more color and detail information of the image.

The Pix2Pix generative adversarial network uses ReLU activation function and maximum pooling. The ReLU activation function can accelerate the convergence speed of the network, but if its input value is less than 0, the neurons may not be able to update the parameters, thus affecting the stability of the network in the training process. Therefore, this paper uses Leaky ReLU function as the activation function on the basis of the original Pix2Pix generative

adversarial network. Its expression is:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{if } x \leq 0 \end{cases} \quad (1)$$

The advantage is that a slope  $\alpha$  can be added to the negative value of the input, so as to avoid the problem that the neuron cannot update the parameters caused by the function gradient being 0 when the input is negative. The value of  $\alpha$  set in this paper is 0.2, and the Tanh activation function is used in the last convolution layer. At the same time, due to the bright colors and various colors of Qinghai farmer's painting, and the image structure and color distribution are also relatively complex. If the maximum pooling operation is used, part of the image feature information will be lost, thus limiting the colorization effect of the network. Therefore,

this paper removes the maximum pooling operation in the generative network. The maximum pooling operation in the network is replaced by a convolution operation with a convolution kernel of  $4 \times 4$  and a stride of 2, which retains more image feature information during the down-sampling process, thereby further improving the colorization effect.

Figure 1 is the structure diagram of the generative network of this paper. The generated network parameters are shown in Table 1

The process of image feature extraction includes 5 convolution blocks: Block1~Block5. After the gray-scale image is input into the generative network, each convolution block will have two convolution layers to process it. Each convolution block includes two convolution layers, including convolution operation, batch normalization and activation processing operations. The feature map output from convolution blocks other than Block5 will be input to Down\_Conv convolution, the Down\_Conv convolution layer includes convolution operation and batch normalization and activation processing operations. After Down\_Conv convolution layer processing, the size of the feature map is halved and the dimension remains unchanged, and then it is used as the input of the next convolution block.

**Table 1: Generative Network Parameters of This Paper**

Structure name	Convolution kernel size	Stride	Output
Block1	3×3	1	64
Down_Conv1	4×4	2	64
Block2	3×3	1	128
Down_Conv2	4×4	2	128
Block3	3×3	1	256
Down_Conv3	4×4	2	256
Block4	3×3	1	512
Down_Conv4	4×4	2	512
Block5	3×3	1	1024
Up_Conv1	4×4	2	512
Block6	3×3	1	512
Up_Conv2	4×4	2	256
Block7	3×3	1	256
Up_Conv3	4×4	2	128
Block8	3×3	1	128
Up_Conv4	4×4	2	64
Block9	3×3	1	64
Last_Conv	1×1	1	2

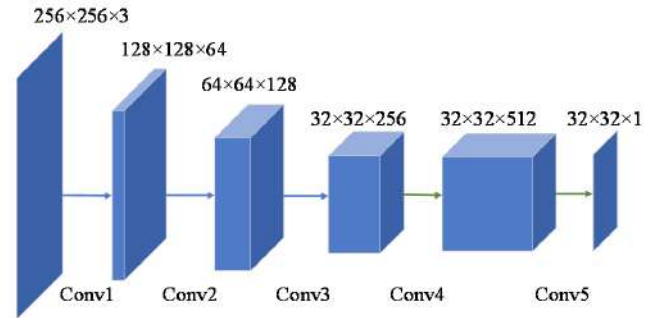
**Table 2: Discriminative Network Parameters in This Paper**

Structure name	Convolution kernel size	Stride	Output
Conv1	4×4	2	64
Conv2	4×4	2	128
Conv3	4×4	2	256
Conv4	4×4	1	512
Conv5	4×4	1	1

The process of image reconstruction includes four convolution blocks: Block6~Block9. The structure of each convolution block is the same as that of the convolution block in the process of image feature extraction. After the extraction of the image feature information of the other convolution blocks except Block9 is completed, the output feature map will be processed by the transposed convolution layer, and the size of the feature map after processing will be doubled. Then, it is fused with the feature map of the corresponding convolution layer output in the down-sampling process as the input of the next convolution block, and the output of Block9 will be output by Last\_Conv convolution treatment, Last\_Conv convolution layer includes convolution and activation processing operations, and then outputs the color channels of the image to complete the prediction of the values of color channels of the gray-scale image. Finally, the gray-scale image and the color channels can be fused to generate a color image.

## 2.2 Discriminative Network Structure

The discriminative network used in this paper is the PatchGAN [31]. The traditional discriminative network for generative adversarial networks usually adds a full connection layer at the end of the network to map the input generated image into a real number, which represents the probability that the input image is a real image. PatchGAN is obviously different from the traditional discriminator.

**Figure 2: The above is the schematic diagram of the discriminative network structure in this paper**

PatchGAN is a full convolution network, which maps the input generated image into  $N \times N$  matrix  $M$ , where  $M_{ij}$  represents the probability that the distribution of the corresponding image area in the generated image is the distribution of the real image. Finally, the mean value of matrix  $M$  represents the probability that the generated image is the real image.

The discriminative network structure is shown in Figure 2  
The discriminative network parameters are shown in Table 2

As the input of the discriminative network, the real image and the generated image will pass through five convolution layers. The first three convolution layers contain convolution and activation processing operations. Only the Conv1 convolution layer does not contain batch normalization. The size of each convolution layer image is halved and the dimension is increased. The activation function used is LeakyReLU function; the convolution step of Conv4 and Conv5 convolution layers is 1, and only Conv4 convolution layer contains batch normalization operation and activation processing operation; Finally, the fifth convolution layer outputs a  $32 \times 32$  matrix.

The discriminative network of traditional generative adversarial networks is to judge whether the overall distribution of the generated image conforms to the distribution of the real image. PatchGAN outputs a matrix. Each value in the matrix represents the probability that the distribution of the corresponding image area in the generated image conforms to the distribution of the real image. Finally, the probability that the generated image is a real image is expressed according to the mean value of the output matrix.

### 2.3 Loss Function

In this paper, the loss function of the network model is composed of conditional generative adversarial loss function and L1 loss function. The L1 loss function is defined by the L1 distance between the pixel values of the generated image and the real image, which enables the generative network to generate a color image closer to the real image. The introduction of L1 loss function can accelerate the convergence of the network model and improve the training efficiency of the network model. The L1 loss function can better restore the low resolution part of the image and ensure that the network model can generate a clear color image. The loss function of the network model in this paper can be expressed by Eq(2-4):

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L1}(G) \quad (2)$$

$$L_{cGAN}(G, D) = E_{x,y} [\log D(x, y)] + E_x [\log (1 - D(x, G(x)))] \quad (3)$$

$$L_{L1}(G) = E_{x,y} [\|y - G(x)\|_1] \quad (4)$$

Where  $x$  and  $y$  are the input gray-scale image and real image respectively,  $E$  represents the expected value, is the conditional generative adversarial loss function, and is the L1 loss function,  $\lambda$  is the corresponding weight.

## 3 EXPERIMENTAL COMPARISON AND ANALYSIS

### 3.1 Qinghai Farmer's Painting Image Dataset

This paper studies the method of deep learning to realize the colorization of gray-scale images of Qinghai farmer's painting. It needs to use the image dataset related to Qinghai farmer's painting to train the neural network model, but there are no datasets related to Qinghai farmer's painting at present. Therefore, this paper collects the image data related to Qinghai farmer's painting through field investigation, online collection and consulting relevant literature, more than 1800 images of Qinghai farmer's painting have been collected, and a preliminary image dataset of Qinghai farmer's



Figure 3: The above are some relevant images in the preliminary Qinghai farmer's painting image dataset we have constructed.



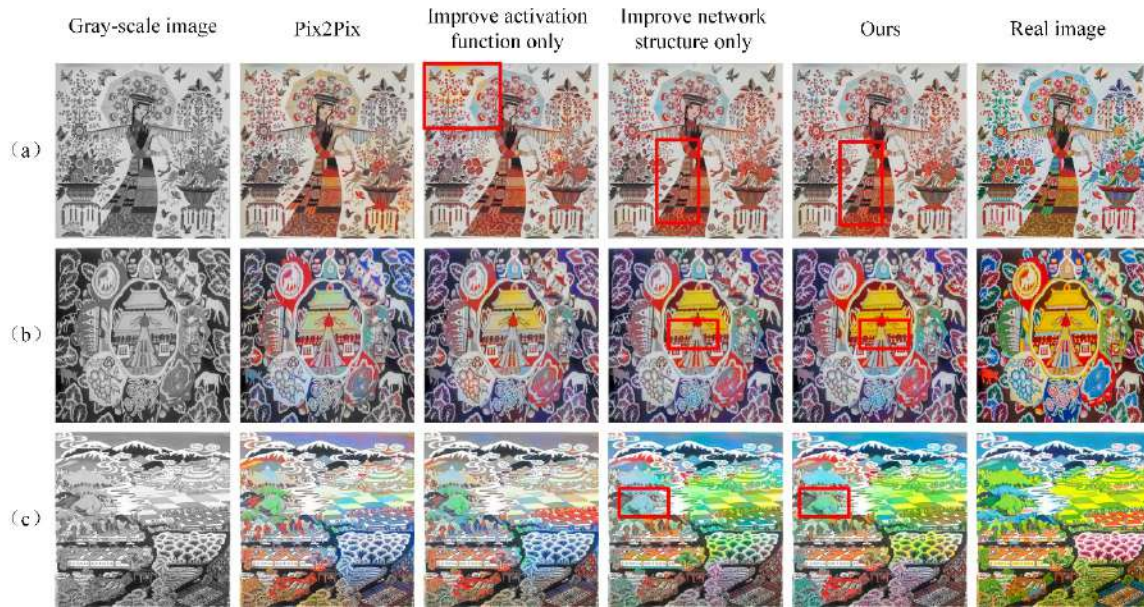
Figure 4: Through the standard processing and screening of images, we construct a Qinghai farmer's painting image dataset suitable for network model training.

painting has been constructed. Some of the image data are shown in Figure 3

At present, the format and size of the collected images of farmer's paintings are different, and the number is small, which cannot meet the training needs of the neural network model. Therefore, it is necessary to process the images through the computer, convert the images into a unified format and size, and select the images suitable for the training of the network model by manual screening. At the same time, the characteristic pattern elements are extracted by manually cutting the collected Qinghai farmer's painting images, and the Qinghai farmer's painting image dataset is expanded to build a Qinghai farmer's painting image dataset that meets the training needs of network model. At present, there are 9574 images related to Qinghai farmer's painting in the constructed farmer's painting image dataset.

To train the generative adversarial networks model constructed in this chapter using the Qinghai farmer's painting image dataset, it is necessary to carry out standard processing on the images in the dataset, and change the image size in the dataset to  $256 \times 256$  size. Partial images of dataset used for network model training are shown in Figure 4

A standard dataset plays an important role in the application of deep learning methods in various research fields. The construction of Qinghai farmer's painting image dataset can provide strong data



**Figure 5:** Through the corresponding ablation experiment, we can see the influence of each improved part of our method on the experimental results, and then show the improvement of our method in the colorization effect of the gray-scale images of Qinghai farmer's painting.

support for the application of deep learning methods in the research of Qinghai farmer's painting.

### 3.2 Dataset Dataset and Experimental Configuration

Our method uses Lab color space, and the image dataset used is the Qinghai farmer's painting image dataset constructed in this paper. 9274 images in the image dataset are used to train the network model, and 300 images are used for testing.

**Network parameter configuration:** This paper uses the improved Pix2Pix generative adversarial network model to realize the colorization of the gray-scale images of Qinghai farmer's painting, and uses Adam optimizer to update the network model parameters. The learning rate of the generative network and the discriminative network is 0.0001, batch\_size is set to 25, the network model is trained for 100 rounds.

**Hardware environment:** the CPU used in the experiment is Intel (R) core (TM) i5-9400, the GPU is NVIDIA GeForce RTX 3060, and the memory capacity is 8GB.

**Software environment:** the operating system used in the experiment is Windows10 64 bit, and the programming environment is Python3.6.

### 3.3 Figures Image Evaluation Criteria

In order to analyze the colorization effect of the gray-scale image colorization method of Qinghai farmer's painting based on the improved Pix2Pix generative adversarial network proposed in this paper, this paper selects the traditional generation counter-measure network model, the gray-scale image coloring method

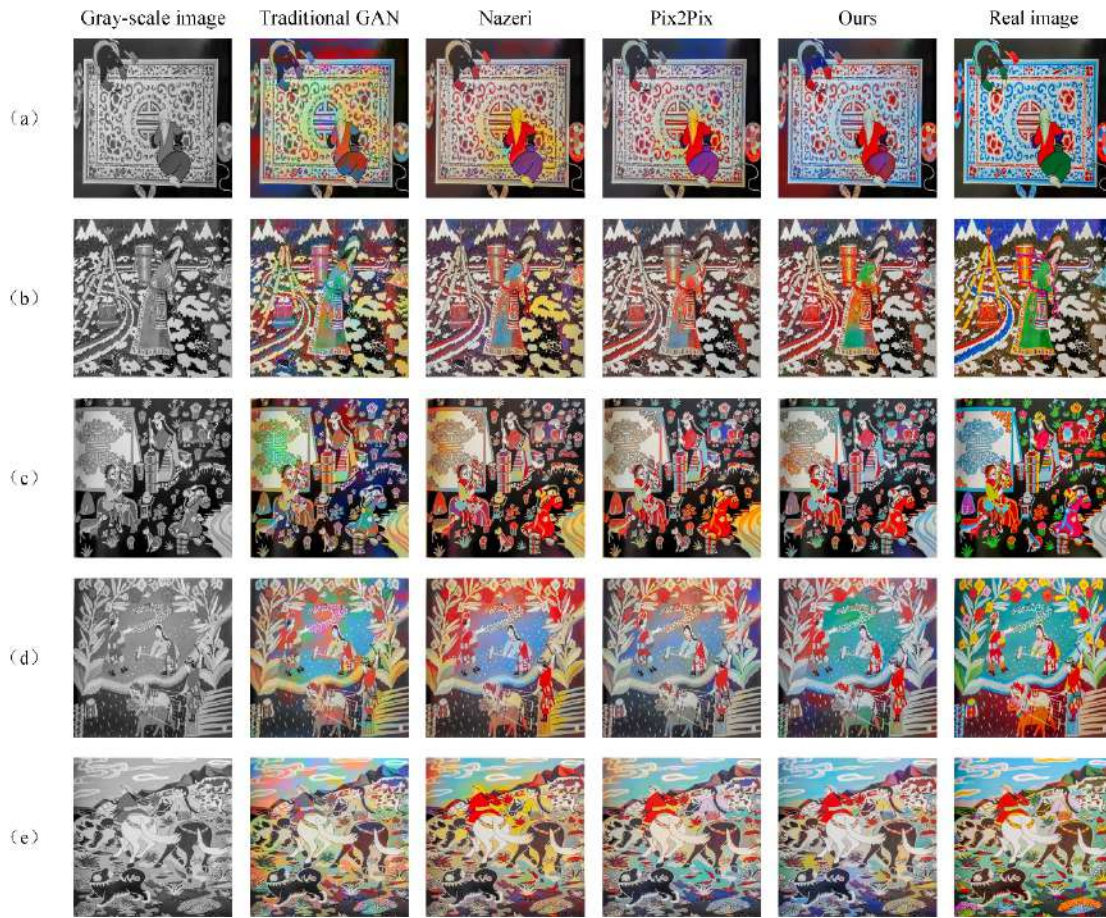
proposed by Nazeri et al. [20] and the Pix2Pix generative adversarial network model to compare with our method, and selects the root mean square error (RMSE) Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) were used as the evaluation criteria of experimental results. By comparing the experimental results of each algorithm, the advantages and disadvantages of each algorithm in the gray-scale image colorization effect of Qinghai farmer's painting are analyzed.

#### Comparison and Analysis of Experimental Results

In order to analyze the influence of each improved part of our method on the experimental results, the corresponding colorization ablation experiments were carried out. The experimental results are shown in Figure 5.

It can be seen from the experimental results that, compared with Pix2Pix generative adversarial network, the color contrast of the image is improved after using Leaky ReLU activation function, but the color images of Qinghai farmer's painting generated by it are still quite different from the real images. After improving the generative network structure, more accurate colorization is achieved, and the overall colorization effect is significantly improved, but there is a deviation for the colorization of some detail areas of the image. Compared with the above experimental results, our method is more accurate for the colorization of image details, and the color information of the color image of Qinghai farmer's painting is more restored, which further improves the colorization effect.

At the same time, we compare our method with the traditional generative adversarial networks model, the colorization method proposed by Nazeri and the original Pix2Pix generative adversarial network model, and further analyzes the colorization effect of our method on the gray-scale images of Qinghai farmer's painting.



**Figure 6: Compared with other methods, our method further improves the colorization effect, and the color images of Qinghai farmer's painting generated by our method are closer to the real images.**

The comparison of experimental results of each method is shown in Figure 6. Through the experimental results, it can be seen that the traditional generative adversarial networks model does not restore the image color information well, and the generated color images are prone to produce more color spots and color blocks. The colorization effect of the gray-scale images of Qinghai farmer's painting is not ideal, there are many chaotic colors in the image background, and the generated color images have the problem of different degrees of coloring confusion, there is a big difference between the color images of Qinghai farmer's painting generated by the traditional generative adversarial networks model and the real images. Compared with the traditional generative adversarial networks, the method of Nazeri alleviates the problem of disordered color distribution in the process of traditional generative adversarial networks coloring gray-scale images of Qinghai farmer's painting, but it tends to generate warm tone color images, and the overall color of the images is relatively single and boring. Compared with the traditional generative adversarial networks, the color images generated by Pix2Pix generative adversarial network model has a more uniform color distribution; compared with the method of Nazeri, the Pix2Pix generative adversarial network model improves

the color richness of the generated color images, and the color information restored from the color images of Qinghai farmer's painting generated by Pix2Pix is more accurate. Through the comparison between the experimental results and the real images, it can be seen that the color images generated by the Pix2Pix generative adversarial network model are dark in brightness, compared with the real images, the color contrast is low and the visual effect is dark.

The colorization effect of the gray-scale image of Qinghai farmer's painting realized by our method is more ideal. Compared with the traditional generative adversarial networks, the color distribution of the color images of Qinghai farmer's painting generated by our method is more consistent with the color distribution of the real images. Compared with the method of Nazeri, the color images of Qinghai farmer's painting generated by our method are more colorful. Compared with the original Pix2Pix generative adversarial network model, the color images of Qinghai farmer's painting generated by our method are closer to the real images in color contrast, and the color information has a higher degree of restoration.

**Table 3: Comparison of RMSE Values of Each Method**

Order Number	Traditional GAN	Nazeri	Pix2Pix	Ours
(a)	51.3291	48.1345	45.4450	42.1877
(b)	44.9812	41.5124	40.4854	38.6916
(c)	52.9889	44.2862	44.5475	41.2409
(d)	51.8114	52.4287	50.8441	45.8850
(e)	42.9969	39.9505	34.1487	34.0989

**Table 4: Comparison of PSNR Values of Each Method**

Order Number	Traditional GAN	Nazeri	Pix2Pix	Ours
(a)	13.9235	14.4817	14.9811	15.6271
(b)	15.0702	15.7672	15.9848	16.3785
(c)	13.6471	15.2054	15.1543	15.8242
(d)	13.8423	13.7394	14.0060	14.8974
(e)	15.4621	16.1003	17.4633	17.4760

**Table 5: Comparison of SSIM Values of Each Method**

Order Number	Traditional GAN	Nazeri	Pix2Pix	Ours
(a)	0.9753	0.9769	0.9786	0.9778
(b)	0.9828	0.9844	0.9854	0.9855
(c)	0.9587	0.9677	0.9704	0.9702
(d)	0.9644	0.9668	0.9681	0.9710
(e)	0.9804	0.9817	0.9831	0.9845

**Table 6: Comparison of PSNR, SSIM and RMSE of Each Method**

Order Number	Traditional GAN	Nazeri	Pix2Pix	Ours
RMSE	45.6886	39.2319	38.5845	37.8719
PSNR	15.0601	16.5071	16.6435	16.7662
SSIM	0.9788	0.9816	0.9825	0.9831

Table 3, Table 4 and Table 5 respectively show the RMSE value, PSNR value and SSIM value of the experimental results obtained by each method.

At the same time, in order to evaluate the traditional generative adversarial networks model, the gray-scale image colorization method proposed by Nazeri. Pix2Pix generative adversarial network model and our method in general, this paper calculates the mean value of RMSE, PSNR and SSIM of all experimental results of each method. The data are shown in Table 6.

According to the comparison of the experimental results and the experimental data, it can be seen that the image generated by the traditional generative adversarial networks model has the problem of chaotic color distribution to varying degrees, which is easy to destroy the original structure information of the images, resulting in a large RMSE of the experimental results, and its PSNR and SSIM generally lag behind the methods of Nazeri et al. [20], Pix2Pix generative adversarial network model and our method. Although the method of Nazeri et al. [20] alleviates the problems of the traditional generative adversarial networks model, the overall color of

the color images of Qinghai farmer’s painting is relatively single and boring; Pix2pix generative adversarial network model alleviates the problems of traditional generative adversarial networks model and the method of Nazeri et al. [20], but the color images of Qinghai farmer’s painting generated by Pix2Pix have the problem of low color contrast, and there is still a big difference between the color images of Qinghai farmer’s painting generated by it and the real images. Compared with the above three methods, our method alleviates the problems of the above methods. The color distribution and color contrast of the color images of Qinghai farmer’s painting generated by our method are closer to the real images. The color information and structure information of the experimental results have a higher degree of restoration, and the colorization effect is more ideal. From the overall experimental results, our method outperforms the other three methods in terms of RMSE , PSNR and SSIM.

## 4 CONCLUSIONS

Digital image processing technology based on deep learning has opened up a new way for the protection of intangible cultural heritage. This paper takes the gray-scale images of Qinghai farmer's painting as the research object. In view of the problems that the current colorization methods based on deep learning tend to generate color images with relatively single color, the image color distribution is chaotic, and the color contrast of the generated images is low in the process of coloring the gray-scale images of Qinghai farmer's painting, this paper improves the Pix2Pix generative adversarial network, and selects the Leaky ReLU function as the activation function of the generative network. The convolution layers are used to replace the maximum pooling layers, so as to retain more image feature information and further improve the colorization effect. RMSE, PSNR and SSIM are selected to evaluate the colorization results. The experiments show that the quality and colorization effect of the color images of Qinghai farmer's painting generated by the improved Pix2Pix generative adversarial network are improved.

## ACKNOWLEDGMENTS

Focus on research and development and achievement transformation project in qinghai province (Grant no: 2022-GX-155); National Key Research and Development Program (Grant no: 2020YFC1523305) and National Natural Science Foundation Project (Grant no: 62262056)

## REFERENCES

- [1] Su J W, Chu H K, Huang J B. 2020. Instance-aware image colorization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. Available at: <https://doi.org/10.1109/cvpr42600.2020.00799>.
- [2] Žeger I, Grgić S, Vuković J, *et al.* 2021. Grayscale image colorization methods: Overview and evaluation. *IEEE Access*, 9, pp. 113326–113346. Available at: <https://doi.org/10.1109/access.2021.3104515>.
- [3] Lee J, Kim E, Lee Y, *et al.* 2020. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [Preprint]. Available at: <https://doi.org/10.1109/cvpr42600.2020.00584>.
- [4] Casaca W, Colnago M, Nonato L G. 2015. Interactive image colorization using Laplacian coordinates. *Computer Analysis of Images and Patterns*, pp. 675–686. Available at: [https://doi.org/10.1007/978-3-319-23117-4\\_58](https://doi.org/10.1007/978-3-319-23117-4_58).
- [5] LEVIN A, LISCHINSKI D, WEISS Y. 2004. Colorization using optimization," *ACM SIGGRAPH 2004 Papers* [Preprint]. Available at: <https://doi.org/10.1145/1186562.1015780>.
- [6] Huang Y C, Tung Y S, Chen J C, *et al.* 2005. An adaptive edge detection based colorization algorithm and its applications. *Proceedings of the 13th annual ACM international conference on Multimedia* [Preprint]. Available at: <https://doi.org/10.1145/1101149.1101223>.
- [7] Yatziv L, Sapiro G. 2006. Fast image and video colorization using chrominance blending. *IEEE Transactions on Image Processing*, 15(5), pp. 1120–1129. Available at: <https://doi.org/10.1109/tip.2005.864231>.
- [8] Qu Y, Wong T T, Heng P A. 2006. Manga colorization. *ACM SIGGRAPH 2006 Papers on - SIGGRAPH '06* [Preprint]. Available at: <https://doi.org/10.1145/1179352.1142017>.
- [9] Luan Q, Wen F, Cohen-Or D, *et al.* 2007. Natural image colorization. *Proceedings of the 18th Eurographics conference on Rendering Techniques*. 2007: 309–320.
- [10] HEU J H, HYUN D Y, KIM C S, *et al.* 2009. Image and video colorization based on prioritized source propagation. 2009 16th IEEE International Conference on Image Processing (ICIP) [Preprint]. Available at: <https://doi.org/10.1109/icip.2009.5414371>.
- [11] Welsh T, Ashikhmin M, Mueller K. 2002. Transferring color to greyscale images. *Proceedings of the 29th annual conference on Computer graphics and interactive techniques* [Preprint]. Available at: <https://doi.org/10.1145/566570.566576>.
- [12] Li B, Zhao F, Su Z, *et al.* 2017. Example-based image colorization using locality consistent sparse representation. *IEEE Transactions on Image Processing*, 26(11), pp. 5188–5202. Available at: <https://doi.org/10.1109/tip.2017.2732239>.
- [13] Charpiat G, Hofmann M, Schölkopf B. 2008. Automatic image colorization via multimodal predictions. *Lecture Notes in Computer Science*, pp. 126–139. Available at: [https://doi.org/10.1007/978-3-540-88690-7\\_10](https://doi.org/10.1007/978-3-540-88690-7_10).
- [14] Li B, Lai Y K, John M, *et al.* 2019. Automatic example-based image colorization using location-aware cross-scale matching. *IEEE Transactions on Image Processing*, 28(9), pp. 4606–4619. Available at: <https://doi.org/10.1109/tip.2019.2912291>.
- [15] Cao Liqin, Shang Yongxing, Liu Tingting, Li Zhijiang, Ma Ailong. 2019. Locally Adaptive Grayscale Image Colorization. *Chinese Journal of Image Graphics*. 24(08): 1249–1257.
- [16] Cortes C, Vapnik V. 1995. Support-vector networks. *Machine learning*. 20(3): 273–297.
- [17] Achanta R, Shaji A, Smith K, *et al.* 2012. Slc superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 34(11), pp. 2274–2282. Available at: <https://doi.org/10.1109/tpami.2012.120>.
- [18] Goodfellow I J, Pouget-Abadie J, Mirza M, *et al.* 2014. *Generative Adversarial Networks*. Cambridge University Press Ebooks. 153–173. <https://doi.org/10.1017/9781108891530.013>
- [19] Han L, Min M R, Stathopoulos A, *et al.* 2021. Dual Projection Generative Adversarial Networks for Conditional Image Generation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). <https://doi.org/10.1109/iccv48922.2021.01417>
- [20] Nazeri K, Ng E, Ebrahimi M. 2018. Image Colorization Using Generative Adversarial Networks. *Lecture Notes in Computer Science*. 85–94. [https://doi.org/10.1007/978-3-319-94544-6\\_9](https://doi.org/10.1007/978-3-319-94544-6_9)
- [21] Liu Changtong, Cao Lin, Du Kangning. 2020. Portrait Coloring Based on Joint Consistent Cyclic Generative Adversarial Networks. *Computer Engineering and Applications*. 56(16): 183–190.
- [22] Zhu J Y, Park T, Isola P, *et al.* 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *International Conference on Computer Vision*. <https://doi.org/10.1109/iccv.2017.244>
- [23] Liang W, Ding D, Wei G. 2021. An improved DualGAN for near-infrared image colorization. *Infrared Physics & Technology*. 116, 103764. <https://doi.org/10.1016/j.infrared.2021.103764>
- [24] Yi Z, Zhang H, Tan P, *et al.* 2017. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *ArXiv (Cornell University)*. <https://doi.org/10.1109/iccv.2017.310>
- [25] Zhao Y, Po L M, Yu W Y, *et al.* 2022. VCGAN: Video Colorization with Hybrid Generative Adversarial Network. *IEEE Transactions on Multimedia*, 1. <https://doi.org/10.1109/tmm.2022.3154600>
- [26] He K, Zhang X, Ren S, *et al.* 2016. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/cvpr.2016.90>
- [27] Zhang Yi, Wei Wenwen, Gong Zhiyuan. 2021. Gray-scale image colorization method based on deep aggregate structure network. *Computer Application Research*. 38(03): 923–927.
- [28] Wu Lidan, Xue Yuyang, Tong Tong, Du Min, Gao Qinqian. 2021. Image Coloring Algorithm Based on Foreground Semantic Information. *Computer Applications*. 41(07): 2048–2053.
- [29] Wan Yuanyuan, Wang Yuqing, Zhang Xiaoning, Li Yuqun, Chen Xiaolin. 2021. Adversarial grayscale image colorization combined with global semantic optimization. *Yeijing Yu Xianshi*. <https://doi.org/10.37188/cjcd.2021-0012>
- [30] Isola P, Zhu J Y, Zhou T, *et al.* 2017. Image-to-Image Translation with Conditional Adversarial Networks. *Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/cvpr.2017.632>
- [31] Li C, Wand M. 2016. Precomputed Real-Time Texture Synthesis with Markovian Generative Adversarial Networks. *Lecture Notes in Computer Science*. 702–716. [https://doi.org/10.1007/978-3-319-46487-9\\_43](https://doi.org/10.1007/978-3-319-46487-9_43)

# A Histo-Puzzle Network for Weakly Supervised Semantic Segmentation of Histological Tissue Type

Tengyun, Ma

Chongqing Institute of Green and Intelligent Technology,  
CAS  
matengyun@cigit.ac.cn

Lin, Chen\*

Chongqing Institute of Green and Intelligent Technology,  
CAS  
chenlin@cigit.ac.cn

Guotian, He\*

Chongqing Institute of Green and Intelligent Technology,  
CAS  
heguotian@cigit.ac.cn

Yuanchang, Lin

Chongqing Institute of Green and Intelligent Technology,  
CAS  
lyc@cigit.ac.cn

## ABSTRACT

Digital pathological images with a large range of Histological Tissue Types (HTTs) contain more sophisticated contours than natural images. In recent years, deep learning algorithms have been widely applied to assist HTT analysis in a weakly-supervised manner by exploiting the class activation maps (CAM). However, the previous methods tend to confusedly activate the most discriminative regions of feature maps, resulting in incomplete segmented contour. This paper proposes a Histo-Puzzle network to improve the HTTs classification and segmentation based on patch-level self-supervised learning. Specifically, our model separates the HTT images into tiled patches by a puzzle module. Then we train a classifier on the supervision of reconstructed CAMs and image-level labels simultaneously. Experiments are conducted on the digital pathology database with 51 hierarchical HTTs. The experimental results show that our proposed method outperforms previous state-of-the-art methods on segmentation tasks of morphological and functional types.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Image segmentation;

## KEYWORDS

Histological tissue type (HTT) analysis, Class activation maps, Self-supervised learning

## ACM Reference Format:

Tengyun, Ma, Guotian, He, Lin, Chen, and Yuanchang, Lin. 2023. A Histo-Puzzle Network for Weakly Supervised Semantic Segmentation of Histological Tissue Type. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590095>

\*Corresponding author.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).  
CACML 2023, March 17–19, 2023, Shanghai, China  
© 2023 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-9944-9/23/03.  
<https://doi.org/10.1145/3590003.3590095>

## 1 INTRODUCTION

In the field of computer-aided diagnosis (CADs), computational pathology is an essential component that mainly aims to assist human pathologists in improving the accuracy and efficiency of diagnosis based on digital pathology images. For digital histopathological analysis, the main task is to analyze different Histological Tissue Types (HTTs). In recent years, many convolutional neural networks (CNN) architectures [1-3] have been widely used to analyze HTTs automatically with multi-label classifiers. However, the annotations become relatively short for a supervised learning model as the analysis task evolves into more complex forms, and image patterns are diverse about specific organs. Accordingly, weakly supervised semantic segmentation (WSSS) has been widely introduced to use image-level annotations to infer pixel-level labels by training a classification task based on Class Activation Map (CAM) [4, 5], alleviating pathologists' annotation burden.

However, the prevailing CAM-based approaches generally activate the most decisive regions and ignore other confusing locations with high classification confidence, which results in fractured regions that are too small to generalize integrated results. Furthermore, many WSSS methods based on CAMs focus on generating the corresponding regions more accurately in semantic segmentation tasks [6-8] by using complex attention mechanisms or graph theory to yield good results. Recently, some studies [9] found that the CAMs of isolated patches in the tiled images are more generalizable than CAMs with original images. Furthermore, the tiled CAMs can be trained directly during classification learning.

Inspired by the tiled CAMs in [9], we propose a Histo-Puzzle method to tile the histological images to generate more integrated CAMs. Firstly, we apply a puzzle module to split the original histological images and merge the generated CAMs at the network's backbone. Then we use the reconstructed regularization loss to constrain the CAMs from the tiled patches similar to those from the original images. The proposed method can self-supervise the classification model to narrow the gap between the original and merged CAMs and does not require any additional modifications to the existing CNN architectures. Our experiments show that the proposed approach achieves superior results on both classification and segmentation tasks.

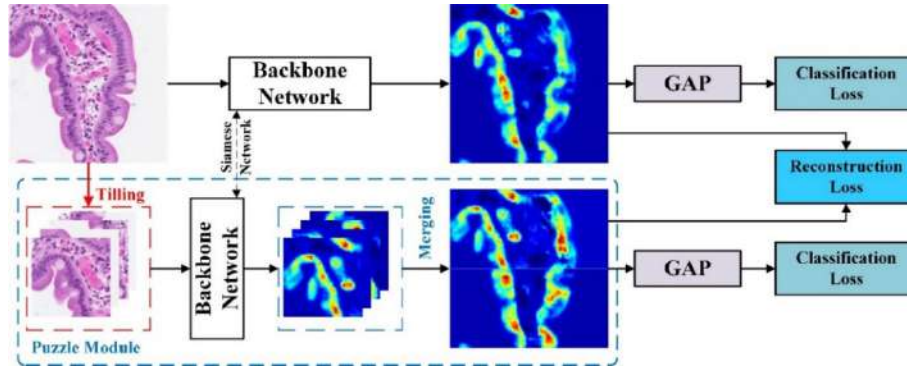


Figure 1: The architecture of our proposed Histo-Puzzle method, which integrates the standard CAMs and Puzzle modules. These two networks share the same weights.

## 2 RELATED WORK

### 2.1 Weakly Supervised Semantic Segmentation

WSSS aims to reduce the model’s reliance on pixel-level annotation so that these methods can use some low-level labels, for instance, image-level classification labels [6, 10] and bounding boxes [11]. Some approaches achieved remarkable performance after the CAMs [4, 5] were proposed. Most of the methods train a classifier to generate the CAMs as the segmentation regions and apply some additional modules to refine them as final results. Ahn et al. [6] apply AffinityNet to retrain the semantic affinity between a pair of pixels and spread the predicted area based on CAMs. For histopathological semantic segmentation, some methods [12, 13] also utilize the CAMs as segmentation predict results, while other methods [14] utilize CAMs as pseudo masks and retrain a Fully Convolutional Networks (FCN) model to obtain fine-grained segmentation contours.

### 2.2 Attention Mechanisms in WSSS

Some studies of WSSS apply attention mechanisms to fine-tune the CAMs. Simonyan et al. [15] conduct a visualization model of image classification, which explicitly visualizes the semantic regions of each class. The methods based on CAM [4, 5] also focus on the models’ attention regions. SEAM [7] proposed a self-supervised attention mechanism to train classifiers for generating CAMs, then refined-trained them on AffinityNet [6], and finally trained the segmentation model based on DeepLab [16], and got remarkable results on VOC 2012 dataset. In General, attention mechanisms approaches are usually too complex to implement even though they can achieve considerable performance. In this work, we propose a straightforward attention mechanism without any additional complex components by training the model based on the supervision of split and merged images.

## 3 PROPOSED METHOD

The problem with traditional CAMs generation is that only the most discriminative regions are activated, while other significant regions of equal importance are often ignored to prevent the over-fitting of the classifier. The most intuitive idea is to directly reduce the image size to be classified so that the model can activate more

representative regions on each small image. Inspired by that, this study proposed a Histo-Puzzle method to improve the WSSS. More specially, we split the original input images into separate patches and retrain the backbone networks by a puzzle module. Then we merge the CAMs of separate patches into a reconstructed CAM, which will obtain more uniform activation regions than those of the model using the original images. Specifically, we propose the self-supervised reconstruction loss to reduce the gap between the original and the reconstructed CAMs. As shown in Figure 1, our method contains the classification loss, which supervises the classification training, and reconstruction loss, which helps to obtain more generalized CAMs.

### 3.1 Puzzle Module

First, we employ the puzzle module to tile and split the original histological images, and merge the CAMs at the end of the network. The puzzle module tile the input image  $I$  with size  $W \times H$  into split non-overlapping patches  $\{I^{1,1}, I^{1,2}, I^{2,1}, I^{2,2}\}$  with size  $W/2 \times H/2$ . We input original images  $I$  to get CAM  $A^o$ . And then, we input these patches, the CAMs generated from the last layer of the network as  $\{A^{1,1}, A^{1,2}, A^{2,1}, A^{2,2}\}$  for each patch. The puzzle module merges the split CAMs  $A^{i,j}$  into a reconstruction CAM as  $A^{re}$ , which has the same size as  $A^o$ .

### 3.2 Generating CAMs

We use the simple WSSS method proposed in [4] to generate CAMs. We denote a network for feature extracting as  $F(\cdot)$ . For each input image  $I$ , the ideal network predicts the classification results  $y_p$  as  $y_p = G(F(I))$ , where  $G(\cdot)$  is the global average pooling (GAP) layer. So the activation map of each class is their feature map, represented as  $A = f = F(I)$ . During training, we calculate the loss between the feature map of the original image and merged CAMs, i.e., we set the generated CAMs as segmentation results. To get the CAMs of all classes  $\hat{A}$ , we further weigh the feature map  $A_c^{re}$  (where  $c \in \mathbb{R}^C$ , and  $C$  denotes the total number of HTT classes) with the confidence scores, which predict by the trained classifier:

$$\hat{A} = y_p A^{re} \quad (1)$$

where  $y_p$  is the predicted vector of this classifier, and  $A^{re}$  denotes the reconstructed CAMs. Moreover, the generated CAMs  $A^{re}$  are normalized by the maximum and minimum value of each  $c$  channel.

### 3.3 Loss Design

We only use the image-level classification label  $y_l$  to supervise the training of the classifier. The multi-label soft margin loss function is employed to calculate the loss between true label  $y_l$  and predicted label  $y_p$ , which is defined as:

$$\hat{y}_p = \begin{cases} y_p, & \text{if } y_p = 1 \\ 1 - y_p, & \text{otherwise} \end{cases} \quad (2)$$

$$\ell_{cls}(y_p, y_l) = -\log(\hat{y}_p) \quad (3)$$

The GAP layer is applied to generate the CAMs of the original input  $A^o$  and merged CAMs  $A^{re}$  as  $y_p^o = G(A^o)$  and  $y_p^{re} = G(A^{re})$ . So the classification loss respectively calculates the original and reconstructed CAM as:

$$\mathcal{L}_{cls} = \ell_{cls}(y_p^o, y_l) \quad (4)$$

$$\mathcal{L}_{re-cls} = \ell_{re-cls}(y_p^{re}, y_l) \quad (5)$$

The model can learn more about histological features and improve classification performance from detached patches with two classification loss functions. However, the CAMs from the two images also need an implicit constraint for obtaining better segmentation regions. Thus, we propose a reconstruction regularization to reduce the gap between them:

$$\mathcal{L}_{re} = \|A^o - A^{re}\|_1 \quad (6)$$

This loss provides additional self-supervision information and forces the CAMs to activate more generalized regions. Therefore the CAMs will achieve better results for HTTs segmentation tasks. Finally, the total losses of our method as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{re-cls} + \alpha \mathcal{L}_{re} \quad (7)$$

where  $\alpha$  denotes the balance factor of  $\mathcal{L}_{re}$ . Because  $\mathcal{L}_{re}$  is the L1 norm and has unbalanced order of magnitude. The classification losses of the original input  $\mathcal{L}_{cls}$  and remerged CAM  $\mathcal{L}_{re-cls}$  are used to roughly activate the HTTs regions. And reconstruction loss  $\mathcal{L}_{re}$  is used to narrow the gap between pixel-level and image-level supervision.

### 3.4 Segmentation post-processing

We conduct the model on a digital pathology database, i.e., the Atlas of Digital Pathology (ADP) database [17]. The ADP dataset includes 51 histological tissue types from different healthy organs for the classification task and 31 for segmentation. The segmentation task should also identify two additional categories: 'Background' for non-tissue regions and 'Other' for non-functional tissue regions. The WSSS cannot activate the 'background' and 'Other' types. Therefore, we follow the segmentation post-processing in [12] after generating the CAMs of 31 classes. Finally, we employ the fully-connected Conditional Random Field (CRF) to produce more continuous contours and get segmentation results.

## 4 EXPERIMENT

### 4.1 Experiment Environment

The ADP dataset [17] consists of 17618 images for classification and WSSS segmentation, 14100 images for training, 1759 for validation, and 1759 for testing. And 50 hand-segmented images are left for segmentation results evaluation. It is noteworthy that the segmentation task of ADP is divided into morphological and functional types. The input histopathologic images are resized into  $512 \times 512$  without special transform. We fine-tuned the maximum value of balance factor  $\alpha$  on different experiments, and the factor linearly increased to  $\alpha$  by half epochs. All the tested models are implemented with PyTorch on two GeForce GTX 1080Ti GPUs.

Texture analysis is commonly done in other image classification and segmentation studies fields, like surface defects recognition [18, 19]. To further understand the peculiarity of pathological tissue segmentation, we briefly illustrate some morphological features of ADP tissues by their segmentation ground truth in Figure 2. It can be seen that some tissues have very small and punctate or spindle-shaped textural areas (like E.M.S and T), while the other tissues have larger areas (like E.M.O and G.O).

### 4.2 The Results of HTTs Classification

The performance of the classification model is vital for WSSS. We evaluate the performance of different classifiers by ablation studies to determine the optimal model for WSSS. The true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), accuracy (ACC), F1-score (F1), and mean area under the curve (mAUC) are used as evaluation metrics. We first evaluate the effect of loss composition. As shown in Table 1, the baseline model based on the original  $\mathcal{L}_{cls}$  achieves 0.9553 mAUC values. The single remerged CAM loss  $\mathcal{L}_{re-cls}$  or reconstruction loss  $\mathcal{L}_{re}$  plays a similar role to  $\mathcal{L}_{cls}$ . However, when we simultaneously train the model with the self-supervised learning of  $\mathcal{L}_{re-cls}$  and  $\mathcal{L}_{re}$ , the classification performance can be boosted and achieves 0.9666 in terms of mAUC. Our integrated model outperforms every single baseline in most evaluation metrics.

Then we further represent the performance of models with different backbones in Table 2. We select the ResNet101 and ResNet50 as the CAMs generators in segmentation because they outperform the other models.

### 4.3 The Results of HTTs Segmentation

We further evaluate the segmentation results of HTTs in terms of mean Intersection-Over-Union (mIoU) and inverse log frequency-weighted Intersection-Over-Union (fIoU). First, we compare the HTTs segmentation generated from our proposed methods with other state-of-the-art (SOTA) CAM methods in Table 3. The segmentation results based on our Histo-Puzzle are closest to the ground truth. Moreover, we report the influence of the post-processing method, and the results are shown in Table 4. The post-processing stages considerably influence the final segmentation results, but the most impactful step is still CAMs generation.

We reimplement the other existing SOTA methods and test them on the same ADP dataset for comparison. As shown in Table 5, compared to the other SOTA methods, our Histo-Puzzle method

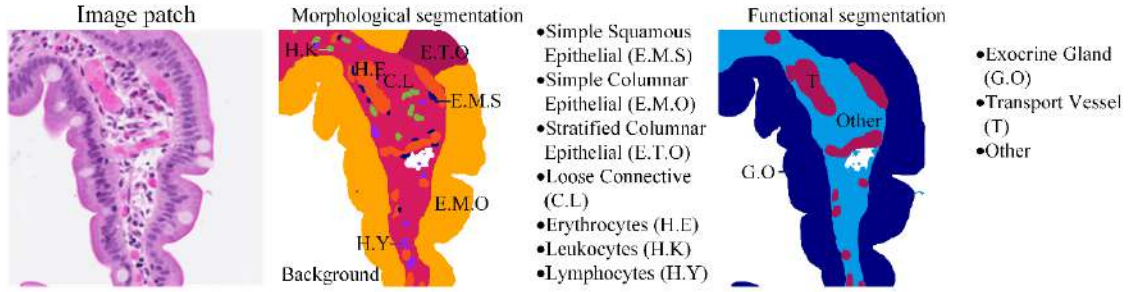


Figure 2: The morphological characteristics of different pathological tissues.

Table 1: Ablation study for loss composition with ResNet101 as the backbone.

Loss	TPR	FPR	TNR	FNR	ACC	F1	mAUC
$\mathcal{L}_{cls}$	0.8616	0.0260	0.9740	0.1384	0.9564	0.8611	0.9553
$\mathcal{L}_{cls} + \mathcal{L}_{re-cls}$	<b>0.8780</b>	0.0242	0.9758	<b>0.1220</b>	0.9604	0.8745	0.9639
$\mathcal{L}_{cls} + \mathcal{L}_{re}$	0.8747	0.0234	0.9766	0.1253	0.9606	0.8744	0.9636
$\mathcal{L}_{cls} + \mathcal{L}_{re-cls} + \mathcal{L}_{re}$	0.8763	<b>0.0206</b>	<b>0.9794</b>	0.1237	<b>0.9632</b>	<b>0.8821</b>	<b>0.9666</b>

Table 2: Experiments of various backbones for segmentation model selection.

Backbone	TPR	FPR	TNR	FNR	ACC	F1	mAUC
ResNet-50	0.8740	0.0224	0.9776	0.1260	0.9614	0.8765	0.9663
ResNet-101	0.8763	<b>0.0206</b>	<b>0.9794</b>	0.1237	<b>0.9632</b>	<b>0.8821</b>	<b>0.9666</b>
ResNest-50	<b>0.8845</b>	0.0229	0.9771	<b>0.1155</b>	0.9626	0.8812	0.9665
ResNest-101	0.8745	0.0226	0.9774	0.1255	0.9613	0.8764	0.9664

Table 3: Comparison of HTTs segmentation results generated by Histo-Puzzle and other existing methods based on CAM.

Methods	Backbone	morph		func	
		fIoU	mIoU	fIoU	mIoU
CAM [4]	ResNet-101	0.2202	0.2356	0.2058	0.1976
Grad-CAM [5]	ResNet-101	0.2201	0.2355	0.2100	0.2018
Grad-CAM++ [20]	ResNet-101	0.2178	0.2324	<b>0.2205</b>	<b>0.2134</b>
Histo-Puzzle	ResNest-50	<b>0.2525</b>	<b>0.2640</b>	0.2134	0.2047
Histo-Puzzle	ResNet-101	0.2441	0.2600	0.2005	0.1925

improves the mIoU values on both morphological and functional tissue types, i.e., our model achieves the best mIoU values of 0.2947 for morphological types and produces the best mIoU values of

0.5624 for functional types, by using the ResNest-50 and ResNet101 as the backbone, respectively.

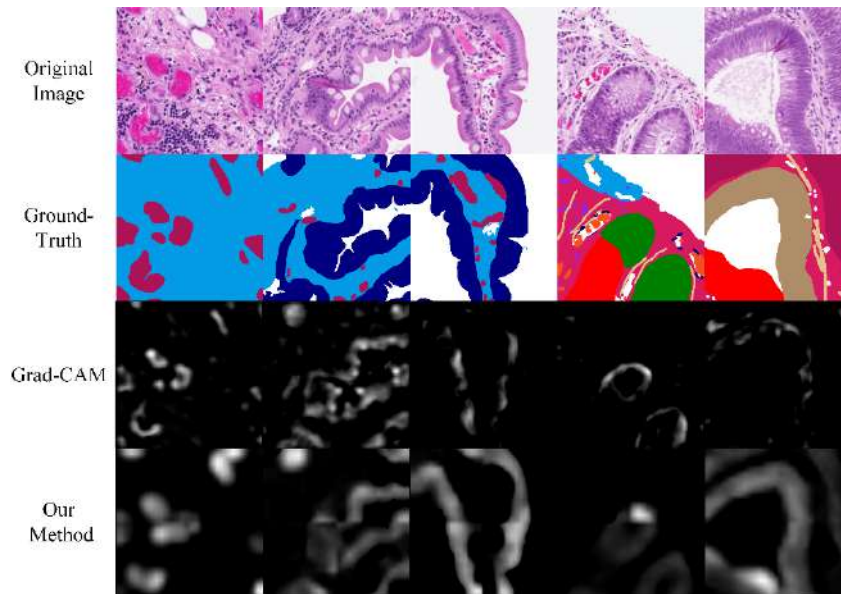
The pixel-level segmentation results of HTTs are demonstrated in Figure 3. The difference in the activation regions proves that direct

Table 4: Stage ablative experiments for post-processing with ResNest-50 and ResNet101 as the backbone.

Stage	morph(ResNest-50)		func(ResNet-101)	
	fIoU	mIoU	fIoU	mIoU
Generated CAM	0.2525	0.2640	0.2005	0.1925
Adjustments	0.2547	0.2688	0.5330	0.5407
Post-processing	<b>0.2787</b>	<b>0.2947</b>	<b>0.5540</b>	<b>0.5624</b>

**Table 5: Comparison of Histo-Puzzle and other existing state-of-the-art methods based on WSSS.**

Methods	Backbone	morph	func
		mIoU	mIoU
SEC [21]	VGG-16	0.1628	0.3225
DSRG [22]	VGG-16	0.1375	0.4732
HistoSegNet[12]	HistoSegNet	0.2206	0.5505
SEAM [7]	Wide-ResNet-38	0.2539	0.5051
MPS-PDA [14]	Wide-ResNet-38	0.0939	0.3058
Grad-CAM++ [20]	ResNet-101	0.2074	0.5452
Histo-Puzzle	ResNest-50	<b>0.2947</b>	0.5173
Histo-Puzzle	ResNet-101	0.2223	<b>0.5624</b>

**Figure 3: The segmentation results of HTTs are obtained by our proposed Histo-Puzzle method and Grad-CAM. Each category in the images corresponds to one of the most representative categories in the Ground-truth.**

supervision of CAM by the Histo-Puzzle network can generate a more integrated mask, while the Grad-CAM only activates more scattered pixels by using the most discriminative information.

## 5 CONCLUSIONS

In this paper, we proposed a novel Histo-Puzzle method for WSSS of HTTs. To generate more integrated CAMs, we apply a puzzle module to tile the split images and train the HTTs classification model with a reconstruction loss. The segmentation contours generated from the Histo-Puzzle network can be optimized to be more generalized and closer to ground truth. Compared with the other existing WSSS methods on the ADP benchmark dataset, our Histo-Puzzle network obtains state-of-the-art performances on both morphological and functional types, proving our method’s effectiveness.

## ACKNOWLEDGMENTS

This research is supported by: the National Nature Science Foundation of China under grant No. 61902370; Cooperation Projects between Chongqing Universities in Chongqing and Institutions Affiliated with the Chinese Academy of Sciences (HZ2021011); Chongqing Technology Innovation and Application Development Special (cstc2021jscx-cylhX0009); Chongqing Technology Innovation and Application Development Special Major Theme Special (cstc2019jscx-zdztzxX0014).

## REFERENCES

- [1] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from <http://arxiv.org/abs/1409.1556>
- [2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. IEEE, 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [3] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola.

2022. ResNeSt: Split-Attention Networks. In IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, 2022. IEEE, 2735–2745. <https://doi.org/10.1109/CVPRW56347.2022.00309>
- [4] Bolei Zhou, Aditya Khosla, Gata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016. IEEE, 2921–2929. <https://doi.org/10.1109/CVPR.2016.319>
- [5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017. IEEE Computer Society, 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- [6] Jiwoon Ahn and Suha Kwak. 2018. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. Computer Vision Foundation / IEEE Computer Society, 4981–4990. <https://doi.org/10.1109/CVPR.2018.00523>
- [7] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. 2020. Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, 2020. IEEE, 12272–12281. <https://doi.org/10.1109/CVPR42600.2020.01229>
- [8] Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, and Xiao Han. 2022. Node-aligned Graph Convolutional Network for Whole-slide Image Representation and Classification. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, 2022. IEEE, 18791–18801. <https://doi.org/10.1109/CVPR52688.2022.01825>
- [9] Sanghyun Jo and Injae Yu. 2021. Puzzle-CAM: Improved Localization Via Matching Partial And Full Features. In 2021 IEEE International Conference on Image Processing, Anchorage, AK, 2021. IEEE, 639–643. <https://doi.org/10.1109/ICIP42928.2021.9506058>
- [10] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. 2019. FickleNet: Weakly and Semi-Supervised Semantic Image Segmentation Using Stochastic Inference. In IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, 2019. IEEE, 5267–5276. <https://doi.org/10.1109/CVPR.2019.00541>
- [11] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. 2017. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017. IEEE, 1665–1674. <https://doi.org/10.1109/CVPR.2017.181>
- [12] Lyndon Chan, Mahdi S. Hosseini, Corwyn Rowsell, Konstantinos N. Plataniotis, and Savvas Damaskinos. 2019. HistoSegNet: Semantic Segmentation of Histological Tissue Type in Whole Slide Images. In 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019. IEEE, 10661–10670. <https://doi.org/10.1109/ICCV.2019.01076>
- [13] Kangning Liu, Yiqiu Shen, Nan Wu, Jakub Piotr Chledowski, Carlos Fernandez-Granda, and Krzysztof J. Geras. 2021. Weakly-supervised High-resolution Segmentation of Mammography Images for Breast Cancer Diagnosis. In Medical Imaging with Deep Learning, 2021, Lubeck, Germany. PMLR, 451–472. <https://proceedings.mlr.press/v143/liu21b.html>
- [14] Chu Han, Jiatai Lin, Jinhai Mai, Yi Wang, Qingling Zhang, Bingchao Zhao, Xin Chen, Xipeng Pan, Zhenwei Shi, Zeyan Xu, Su Yao, Lixu Yan, Huan Lin, Xiaomei Huang, Changhong Liang, Guoqiang Han, and Zaiyi Liu. 2022. Multi-layer pseudo-supervision for histopathology tissue semantic segmentation using patch-level classification labels. Medical Image Analysis. 80 (2022), 102487. <https://doi.org/10.1016/j.media.2022.102487>
- [15] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034. Retrieved from <http://arxiv.org/abs/1312.6034>
- [16] Liangchih Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE transactions on pattern analysis and machine intelligence 40, 4 (2018), 834–848. <https://doi.org/10.1109/TPAMI.2017.2699184>
- [17] Mahdi S. Hosseini, Lyndon Chan, Gabriel Tse, Michael Tang, Jun Deng, Sajad Norouzi, Corwyn Rowsell, Konstantinos N. Plataniotis, and Savvas Damaskinos. 2019. Atlas of Digital Pathology: A Generalized Hierarchical Histological Tissue Type-Annotated Database for Deep Learning. In IEEE Conference on Computer Vision and Pattern Recognition, Long Beach. IEEE, 11747–11756. <https://doi.org/10.1109/CVPR.2019.01202>
- [18] Ihor Konovaleenko, Pavlo Maruschak, Vitaly Brevus, and Olegas Prentkovskis. 2021. Recognition of Scratches and Abrasions on Metal Surfaces Using a Classifier Based on a Convolutional Neural Network. Metals. 11, 4. <https://www.mdpi.com/2075-4701/11/4/549>
- [19] Ihor Konovaleenko, Pavlo Maruschak, Janette Brezinová, Olegas Prentkovskis, and Jakub Brezina. 2022. Research of U-Net-Based CNN Architectures for Metal Surface Defect Detection. Machines. 10, 5. <https://www.mdpi.com/2075-1702/10/5/327>
- [20] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. 2018. Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In 2018 IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, 2018. IEEE Computer Society, 839–847. <https://doi.org/10.1109/WACV.2018.00097>
- [21] Alexander Kolesnikov and Christoph H. Lampert. 2016. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, 2016, Proceedings, Part IV. Springer, 695–711. [https://doi.org/10.1007/978-3-319-46493-0\\_42](https://doi.org/10.1007/978-3-319-46493-0_42)
- [22] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. 2018. Weakly-Supervised Semantic Segmentation Network With Deep Seeded Region Growing. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018. IEEE, 7014–7023. <https://doi.org/10.1109/CVPR.2018.00733>

# Pose Estimation of Space Targets Based on Geometry Structure Features

Xiwen Liu  
liuxiwen20@mails.ucas.ac.cn  
University of Chinese Academy of Sciences;  
Institute of Software, Chinese Academy of Sciences  
Haidian Qu, Beijing Shi, China

Shuling Hao  
Institute of Software, Chinese Academy of Sciences  
Haidian Qu, Beijing Shi, China  
shuling@iscas.ac.cn

Kefeng Xu  
Institute of Software, Chinese Academy of Sciences  
Haidian Qu, Beijing Shi, China  
kefeng@iscas.ac.cn

## ABSTRACT

The pose estimation of space targets is of great significance for space target state assessment, anomaly detection, fault diagnosis, etc. With the development of adaptive optics technology, the imaging quality of ground-based optical systems has been greatly improved, and we can use the observed images to estimate the pose of space targets. However, the imaging process of the ground-based optical system is still affected by various noises and disturbances, which makes the images degrade. Aiming at the space target pose estimation with these degraded images, we propose a new pose estimation pipeline based on robust geometry structure features. By associating the corresponding geometry structure feature between consecutive frames, we can get the target pose by optimization method. This paper will explain the definition and extraction of the proposed geometry structure feature. We propose a geometry structure feature prediction method base on set prediction in a multi-task way with target components classification and segmentation. Experiments show that our structure feature prediction network achieves competitive results on the simulated photo-realistic SpaceShuttle dataset which is rendered according to the physics imaging process.

## CCS CONCEPTS

• **Computing methodologies** → **Shape inference**; *Image segmentation*; *Reconstruction*.

## KEYWORDS

space targets, pose estimation, geometry structure feature, multi-task

### ACM Reference Format:

Xiwen Liu, Shuling Hao, and Kefeng Xu. 2023. Pose Estimation of Space Targets Based on Geometry Structure Features. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590096>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590096>

## 1 INTRODUCTION

In recent years, as more and more satellites have been sent into space, a large number of non-cooperative satellites and space debris have accumulated in near-Earth orbit, resulting in a gradually deteriorating space environment. Some countries have set up real-time surveillance systems to prevent satellite collisions, such as the United States Space Surveillance Network (SSN) and the Russian Space Surveillance Network (SSS). Surveillance systems can be divided into two types: space-based surveillance systems and ground-based surveillance systems. Estimating the pose is an important part of space target monitoring, and it is of great significance to the real-time target state assessment, anomaly detection, fault diagnosis, etc.

With the improvement of the performance of optoelectronic telescopes and the development of adaptive optics technology, ground-based surveillance systems can acquire high-resolution images of space targets, making it possible to estimate space target poses from observed images. Due to its prominence in computer vision problems, the academic community also favors deep learning models to solve pose estimation problems. Deep learning models rely on large annotated datasets. Although there are some large-scale datasets such as the LINEMOD dataset [3] and the YCB video dataset [18], ground-based optics datasets for spatial targets are lacking. The main reason is that ground-based optical systems are limited by geographical location and meteorological conditions, which makes it difficult to obtain real images with rich poses. It is also difficult to obtain accurate annotations by manual annotation. To address this problem, several methods [2, 8, 12, 13, 17] have been proposed to generate high-quality images with accurate annotations that can be used for training. There are domain differences between the simulated data obtained by the above methods and the real data, which will inevitably affect the generalization ability of the model. Therefore, quite a few datasets with degraded processing operations on the simulated data are proposed. [5] proposed the SPEED dataset, which uses Gaussian blur and zero-mean Gaussian white noise to process images. [9] proposed a SPEED+ dataset acquired with TRON equipment using light boxes and sun lamps to simulate a real illumination environment to reduce the domain gap with real data. [16] considered the effects of atmospheric turbulence and noise during ground-based optical imaging, and proposed a sequential dataset for pose estimation. However, most of the current studies of space target simulation datasets do not consider the physical process of ground-based optical imaging finely enough, resulting in the simulation data still having large differences from

the real data. In Section 3, we propose a simulation data rendering and processing method based on Blenderproc. The simulation data is closer to the real data in terms of visual characteristics.

The main challenge of space target pose estimation is that ground-based optical telescopes are affected by various factors during imaging, including atmospheric turbulence, atmospheric scattering as well as various noises, resulting in image blurring and degradation. As a branch of computer vision, the spatial targets pose estimation methods based on deep learning can be divided into direct approaches and two-stage approaches. Direct approaches treat the pose estimation problem as a regression or classification task, predicting pose-related parameters directly from the input images. Sharma et al. [14] proposed the application of convolutional neural networks to spacecraft pose estimation based on hard viewpoint classification, quantifying the pose labels into 3000 category labels. Proenca et al. [10] proposed a deep learning framework for pose estimation based on ResNet networks to directly regress the output position. Sharma et al. [15] proposed a Spacecraft Pose Network (SPN) to provide more accurate pose estimation with a hybrid classification-regression approach. Direct approaches often rely on time-consuming pose refinement operations to improve performance. Some recent approaches use a two-stage approach to estimate pose, first using CNN to regress 2-dimensional key points and then using the Perspective-n-Point (PnP) algorithm to compute pose parameters. In other words, the detected key points can serve as intermediate features for pose estimation. Huo et al. [4] designed a lightweight YOLOv3 network for predicting keypoint locations, followed by regression to generate a heat map, finally, the pose was obtained and optimized using the PnP and EKF methods. Song et al. [16] proposed a YOLO-6D-based keypoint prediction network with weak supervision of the depth image and finally solved the pose using the EPnP algorithm. Although two-stage approaches show better results compared to the direct approaches, the key points are susceptible to noise interference and less robustness on low-quality ground-based optical images. In this paper, unlike existing two-stage methods based on keypoints, we focus on extracting geometry structure features from monocular images for pose estimation, and supervising the category and segmentation map of target components.

The contributions of this paper are as follows:

- We introduce a pose estimation method using geometry structure feature as an intermediate representation. Compared to keypoint-based methods, our approach can be applied to different space targets without the need for a known target model. The structure feature is advantageous for solving the non-cooperative target pose estimation of degraded images.
- We propose a new framework for detecting geometry structure features of spatial targets. The structure feature prediction problem is described as a multi-task problem that predicts not only the geometry structure features of the target but also the classification and segmentation of the target components.
- Based on the real physical imaging process, we propose a photo-realistic simulation dataset SpaceShuttle, so that the model we trained on the simulation dataset can be easily transferred to real images.

The rest of the paper is organized as follows. The network architecture for the prediction of geometry structure features is presented in Section 2. In Section 3 we simulate the photo-realistic dataset following the ground-based optical imaging process and do detailed comparative experiments for structure feature prediction. Finally, Section 4 concludes the paper.

## 2 APPROACH

In this section, we present the method for detecting geometry structure features of space targets. Figure 1 shows the architecture of the proposed structure feature detection network. We first present how to formalize the geometry structure feature detection problem, and then introduce the details of the network. After that, we describe our training objectives and inference process.

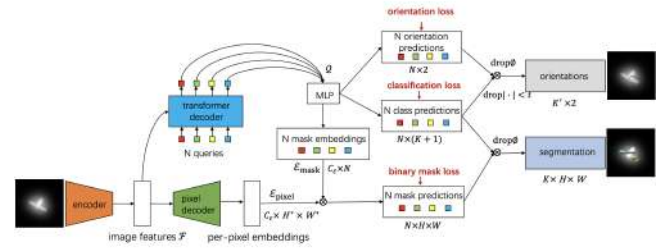


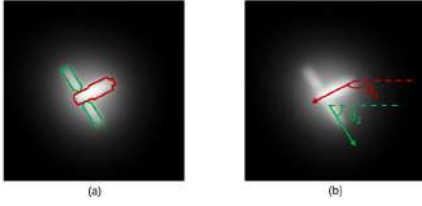
Figure 1: The network architecture of our approach.

### 2.1 Problem Definition

For the space target, as shown in Figure 2(a), the composition component mainly includes the backbone and the solar cell wing. In this paper, the geometry structure feature is the orientation of each component that makes up the spatial target. In other words, it is the angle  $\theta$  between the long axis of each component and the horizontal axis of the pixel coordinate system, where  $\theta \in [0^\circ, 180^\circ)$ , as shown in Figure 2(b). Although the straightforward way to represent the orientation of each component is using an angle, we do not adopt this strategy because the angle is discontinuous. For example, a structure with angle  $0^\circ$  and a structure with  $180^\circ$  both indicate horizontal directions, but their angles are quite different. This makes the network hard to learn to detect the orientation. Instead of using an angle to represent geometry structure feature, we use the sine and cosine values of the twice angle ( $\cos 2\theta, \sin 2\theta$ ) to represent orientation.

### 2.2 Network Architecture

Essentially, our proposed model is based on MaskFormer [1], and by adding a geometry structure feature orientation prediction branch, the orientation of geometry structural feature is predicted as an intermediate representation of the pose. To achieve this, a structure feature prediction network is proposed based on the encoder-decoder architecture, as shown in Figure 1. The encoder, the category prediction branch, and the mask prediction branch are the same as MaskFormer's network architecture. The input of the model is a gray-scale image  $I \in \mathbb{R}^{3 \times 512 \times 512}$ , the output of branch 1 is the prediction result of the geometry structure feature



**Figure 2: Definition of geometry structure features.** a) Basic component of the space target. The red line enclosed indicates the backbone, and the green line encloses the solar cell wing. b) Definition of the geometry structure orientation.  $\theta_1$  indicates the angle between the long axis of the backbone and the horizontal axis, and  $\theta_2$  indicates the angle between the long axis of one of the solar cell wings and the horizontal axis.

orientation, the output of branch 2 is the category prediction result, and the output of branch 3 is the mask prediction result of the components. The number of structure features is equal to the number of basic components in the image, and the size of the mask is equal to the size of the input image.

**Encoder.** For the public encoder, we continue to use ResNet50 to generate low-resolution image feature maps  $\mathcal{F} \in \mathbb{R}^{2048 \times 16 \times 16}$ .

**Geometry structure feature orientation prediction branch.** The purpose of this branch is to use the Transformer decoder to decode the feature map  $\mathcal{F}$  to generate the orientation of the structure in the image. The decoder is stacked with 6 decoding layers, and the self-attention layer of the decoding layer has 8 heads. The decoding leads to the segment embedding  $Q \in \mathbb{R}^{N \times 256}$ ,  $N$  indicates the number of queries, here the default  $N$  is 20.  $Q$  is then fed into a 2-layer fully connected network with a ReLU activation function. Then the norm function is used so that each structure feature orientation of the output is a unit vector satisfying the condition  $\sin^2 2\theta + \cos^2 2\theta = 1$ . Finally, the branch yields a orientation prediction result  $\{o_i \in \hat{\mathbb{I}}\}_{i=1}^N$ . Here,  $\hat{\mathbb{I}}$  denote the set of 2-dimensional unit vectors.

**Category prediction branch.** The category prediction branch extends the structure of MaskFormer. This branch applies a linear classifier on the segment embedding  $Q$ , followed by a softmax activation function to generate the category probability prediction  $\{p_i \in \Delta^{K+1}\}_{i=1}^N$ . Here  $K$  is the number of categories of the component. The classifier additionally predicts a no-object category ( $\emptyset$ ) to satisfy the case of not corresponding to any component.

**Mask prediction branch.** For mask prediction, a Multi-Layer Perceptron (MLP) with 2 hidden layers converts each segment embedding  $Q$  into a mask embedding  $\mathcal{E}_{\text{mask}}$ . finally, we obtain each binary mask prediction  $m_i \in [0, 1]^{512 \times 512}$  by the dot product between the mask embedding and each pixel embedding  $\mathcal{E}_{\text{pixel}}$  computed by the pixel decoder module. The dot product is followed by a sigmoid activation, i.e.,  $m_i[h, w] = \text{sigmoid}(\mathcal{E}_{\text{mask}}[:, i]^T \cdot \mathcal{E}_{\text{pixel}}[:, h, w])$ , here  $i \in \{1, \dots, N\}$ .

### 2.3 Multi-task Training Objective

We divide the structural feature detection task into 3 subtasks: component classification, component segmentation and geometry

structure feature regression, and define the output  $z$  as a set of  $N$  probability-mask-orientation pairs, i.e.,  $z = \{(p_i, m_i, o_i)\}_{i=1}^N$ . To train the model, the matching  $\sigma$  between a set of outputs  $z$  and a set of ground truth  $z^{\text{gt}} = \{(c_i^{\text{gt}}, m_i^{\text{gt}}, o_i^{\text{gt}}) \mid c_i^{\text{gt}} \in \{1, \dots, K\}, m_i^{\text{gt}} \in \{0, 1\}^{512 \times 512}, o_i^{\text{gt}} \in \hat{\mathbb{I}}\}_{i=1}^{N^{\text{gt}}}$  needs to be predicted by the Hungarian algorithm. Here  $N \geq N^{\text{gt}}$ , a set of ground truth tags are filled with no-object tags to allow one-to-one matching.

Given a match  $\sigma$ , the loss  $\mathcal{L}$  consists of a cross-entropy classification loss, a binary mask loss  $\mathcal{L}_{\text{mask}}$ , and a geometry structure orientation regression loss  $\mathcal{L}_{\text{ori}}$  for each prediction segment.

$$\mathcal{L}(z, z^{\text{gt}}) = -\log p_{\sigma}(c^{\text{gt}}) + \lambda_1 t \mathcal{L}_{\text{mask}}(m_{\sigma}, m^{\text{gt}}) + \lambda_2 t \mathcal{L}_{\text{ori}}(o_{\sigma}, o^{\text{gt}}) \quad (1)$$

Here  $\lambda_1$  and  $\lambda_2$  are the balance parameters controlling the trade-off between the three terms, setting  $\lambda_1 = 1.0$  and  $\lambda_2 = 10.0$ . And  $t$  is an indicator for category labeling. No-object is labeled as 0 ( $t = 0$ ) and the others are labeled as 1 ( $t = 1$ ). We use the focal loss [6] and the dice loss [7] as the mask loss  $\mathcal{L}_{\text{mask}}(m, m^{\text{gt}}) = \lambda_{\text{focal}} \mathcal{L}_{\text{focal}}(m, m^{\text{gt}}) + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}}(m, m^{\text{gt}})$  and set the hyperparameters to  $\lambda_{\text{focal}} = 20.0$  and  $\lambda_{\text{dice}} = 1.0$ .

The geometry structure orientation prediction loss  $\mathcal{L}_{\text{ori}}$  is defined as the mean squared loss of the output orientation and ground truth orientation.

$$\mathcal{L}_{\text{ori}}(o_{\sigma(i)}, o_i^{\text{gt}}) = \frac{1}{N^{\text{gt}}} \sum_{i=1}^{N^{\text{gt}}} (o_{\sigma(i)} - o_i^{\text{gt}})^2 \quad (2)$$

### 2.4 Mask Classification and Orientation Regression Inference

**Mask classification inference.** Semantic segmentation inference performs matrix multiplication of category prediction output and mask prediction output, i.e.,  $\arg \max_{c \in \{1, \dots, K\}} \sum_{i=1}^N p_i(c) \cdot m_i[h, w]$ . This strategy returns the category probability for each pixel, but directly maximizing the per-pixel category probability results in indistinguishable from component to the background. Therefore, our mask classification inference sets 0.1 as the threshold value, and if the maximum pixel category probability is less than 0.1, the pixel is marked as background label  $K + 1$ .

$$C[h, w] = \begin{cases} \arg \max_{c \in \{1, \dots, K\}} \sum_{i=1}^N p_i(c) \cdot m_i[h, w] & \text{if } \max_{c \in \{1, \dots, K\}} \sum_{i=1}^N p_i(c) \cdot m_i[h, w] \geq 0.1 \\ K + 1 & \text{otherwise} \end{cases} \quad (3)$$

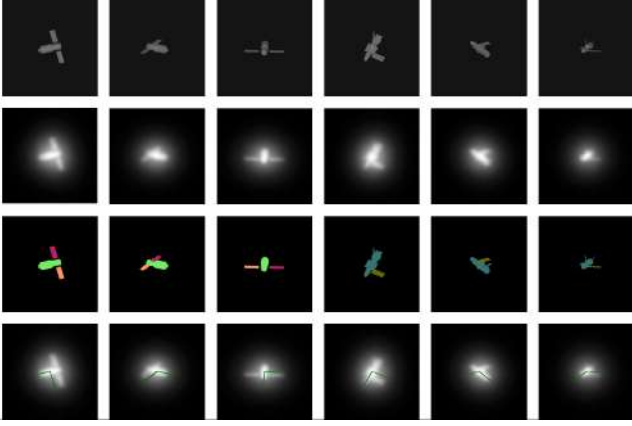
**Orientation regression inference.** Orientation regression inference performs a matrix multiplication of the category prediction output and the orientation prediction output. We set 0.5 as the threshold, and if the orientation modulus length of each category is less than 0.5, it is considered that there is no structure of this category. The final output is the orientation unit vector.

$$\text{Ori}(c) = \begin{cases} \frac{\sum_{i=1}^N p_i(c) \cdot o_i}{\left\| \sum_{i=1}^N p_i(c) \cdot o_i \right\|} & \text{if } \left\| \sum_{i=1}^N p_i(c) \cdot o_i \right\| \geq 0.5 \\ \text{None} & \text{otherwise} \end{cases} \quad (4)$$

## 3 EXPERIMENTS AND DISCUSSION

**Dataset.** To evaluate the performance of our model on different space targets, taking Tianzhou-2 and a space shuttle as examples,

our data are images of two kinds of space targets. The sequence images with annotations are obtained according to the rendering method based on Blenderproc proposed by Song et al. [16]. To narrow the gap with the real data, in addition to simulating the influence of atmospheric turbulence and noise, we also simulated the atmospheric scattering effect using a Gaussian statistical model during the image degradation process. Then we get the simulated photo-realistic SpaceShuttle dataset. The rendered images, simulation images, segmentation image annotations and geometry structure feature annotations are shown in Figure 3. We divided 12,000 simulation images into 10,800 and 1,200 images for training and testing, respectively. The size of the image is  $512 \times 512$ .



**Figure 3: Simulated SpaceShuttle dataset. The first row is the original image rendered by BlenderProc, the second row is the degraded images, the third row is the component segmentation annotation images, and the last row is the geometry structure feature annotation images.**

**Evaluation metrics.** We use five metrics to evaluate our model, namely, Orientation Error (OE) metric, Mean Intersection Over Union (mIoU), Frequency Weighted Intersection Over Union (floU), Pixel Accuracy (pAcc), and Mean Accuracy (mAcc). The orientation error metric is used to count the angle error between the predicted orientation and the ground truth orientation of the geometry structure feature. If the angle error is less than the threshold, the prediction is considered correct. We set  $1^\circ$ ,  $2^\circ$ ,  $5^\circ$ , and  $10^\circ$  as the thresholds, respectively. That is, this metric is used to measure the percentage of correct estimates of the predicted orientation of the structure features. The angular error is shown in Equation 5.

$$\text{angle error} = \frac{1}{2} \frac{180}{\pi} \arccos(o^T \cdot o^{gt}) \quad (5)$$

mIoU, floU, pAcc, and mAcc are evaluation metrics commonly used in semantic segmentation to measure the results of target component segmentation.

**Implementation details.** During the experiments, standard random cropping and random color dithering between 0.5 and 2.0 are used as data augmentation. For the network, we use the result of pre-training on ImageNet-1K [11] to initialize the network parameters of the encoder, and other network parameters are initialized in a default manner. The initial learning rate is  $10^{-4}$ , the batch size

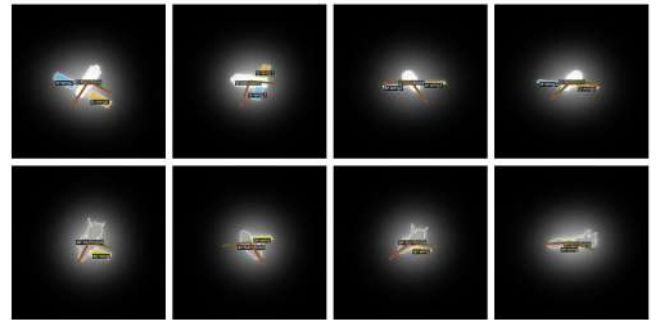
**Table 1: Comparison of our approach with MaskFormer On SpaceShuttle dataset**

Metrics	MaskFormer	Ours
OE metric- $1^\circ$	/	96.4667
OE metric- $2^\circ$	/	98.0000
OE metric- $5^\circ$	/	98.8333
OE metric- $10^\circ$	/	99.2667
mIoU	97.5563	97.6881
floU	98.3049	98.4140
pAcc	98.6349	98.7325
mAcc	99.1423	99.1979

is 4, and the model is trained for 200,000 iterations. We use the AdamW optimizer with a momentum of 0.9 and a weight decay of  $10^{-4}$  at 50,000 iterations. All experiments are performed on a Titan X (pascal) GPU with CUDA 11.3.

**Result analysis.** We set up comparative experiments to verify the performance of our network. In Table 1, we use MaskFormer as a baseline to analyze the results of component segmentation. Our method outperforms MaskFormer by 0.1% for Mean Intersection Over Union and Frequency Weighted Intersection Over Union. For Pixel Accuracy and Mean Accuracy, our method also achieves better performance, which is mainly attributed to our improvements in mask classification inference. The improvement of these performance means that our model can also show the good ability of component classification and component segmentation on degraded images. We show the segmentation prediction results in Figure 4.

In terms of the Orientation Error metric, our model predicts the structural feature orientation with 96.46% accuracy within  $1^\circ$  error, 98% accuracy within  $2^\circ$  error, and 99.26% accuracy within  $10^\circ$  error. This shows that the structural feature orientation prediction model can show well robustness on degraded images, and our proposed structural features can also be used as robust intermediate features for pose estimation of spatial targets. In Figure 4, we show the geometry structure orientation prediction results.



**Figure 4: Component segmentation and orientation prediction results. The green lines represent the ground truth structure orientations, and the red lines are the predicted structure orientations.**

## 4 CONCLUSION

In this paper, we work on the problem of space target pose estimation of ground-based optical images. First, we propose an image degradation method to simulate atmospheric scattering and create a photo-realistic SpaceShuttle simulation dataset. Second, we propose a new geometry structure feature for pose estimation. Then a structure feature prediction network is proposed to predict the geometry structure features, and a multi-task approach is used to learn the relevant information, which provides more constraints to the network. Also, the simulation data we use makes the network easy to migrate to real data. Experimental results show that the method achieves good performance on low-quality images.

Space target pose estimation based on monocular images is a complex problem. The prediction of structural features in certain extreme cases is also difficult due to the complex environment in space and multiple disturbances in the imaging process of ground-based optical systems. In the future, we can try to extract the geometry structure feature between frames by multi-frame image feature prediction methods to overcome the interference caused by extreme environments.

## REFERENCES

- [1] Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems* 34 (2021), 17864–17875.
- [2] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. 2019. BlenderProc. *arXiv preprint arXiv:1911.01911* (2019).
- [3] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. 2013. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision—ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5–9, 2012, Revised Selected Papers, Part I* 11. Springer, 548–562.
- [4] Yurong Huo, Zhi Li, and Feng Zhang. 2020. Fast and Accurate Spacecraft Pose Estimation From Single Shot Space Imagery Using Box Reliability and Keypoints Existence Judgments. *IEEE Access* 8 (2020), 216283–216297. <https://doi.org/10.1109/ACCESS.2020.3041415>
- [5] Mate Kisantal, Sumant Sharma, Tae Ha Park, Dario Izzo, Marcus Märtens, and Simone D’Amico. 2020. Satellite pose estimation challenge: Dataset, competition design, and results. *IEEE Trans. Aerospace Electron. Systems* 56, 5 (2020), 4083–4098.
- [6] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.
- [7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*. Ieee, 565–571.
- [8] Nathan Morrical, Jonathan Tremblay, Yunzhi Lin, Stephen Tyree, Stan Birchfield, Valerio Pascucci, and Ingo Wald. 2021. NVISII: A Scriptable Tool for Photorealistic Image Generation. *arXiv:2105.13962 [cs.CV]*
- [9] Tae Ha Park, Marcus Märtens, Gurvan Lecuyer, Dario Izzo, and Simone D’Amico. 2022. SPEED+: Next-generation dataset for spacecraft pose estimation across domain gap. In *2022 IEEE Aerospace Conference (AERO)*. IEEE, 1–15.
- [10] Pedro F Proença and Yang Gao. 2020. Deep learning for spacecraft pose estimation from photorealistic rendering. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 6007–6013.
- [11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115 (2015), 211–252.
- [12] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9339–9347.
- [13] Max Schwarz and Sven Behnke. 2020. Stilleben: Realistic scene synthesis for deep learning in robotics. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 10502–10508.
- [14] Sumant Sharma, Connor Beierle, and Simone D’Amico. 2018. Pose estimation for non-cooperative spacecraft rendezvous using convolutional neural networks. In *2018 IEEE Aerospace Conference*. IEEE, 1–12.
- [15] Sumant Sharma and Simone D’Amico. 2019. Pose estimation for non-cooperative rendezvous using neural networks. *arXiv preprint arXiv:1906.09868* (2019).
- [16] Jingrui Song, Shuling Hao, and Kefeng Xu. 2021. Uncooperative Satellite 6D Pose Estimation with Relative Depth Information. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4–6, 2021, Proceedings, Part II*. Springer, 166–177.
- [17] Thang To, Jonathan Tremblay, Duncan McKay, Yukie Yamaguchi, Kirby Leung, Adrian Balan, Jia Cheng, William Hodge, and Stan Birchfield. 2018. NDDS: NVIDIA Deep Learning Dataset Synthesizer. [https://github.com/NVIDIA/Dataset\\_Synthesizer](https://github.com/NVIDIA/Dataset_Synthesizer).
- [18] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2017. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199* (2017).

# Detecting Respiratory Events with End-to-End ConvNet

Yanping Shuai

Shenzhen International Graduate School of Tsinghua  
University  
Shenzhen Shi, Guangdong Shen, China

Xingjun Wang\*

Tsinghua University  
Shenzhen Shi, Guangdong Shen, China

Zhangbo Li

Dongguan Jianda Information Technology Co., LTD  
Shenzhen Shi, Guangdong Shen, China

Hanrong Cheng\*

Institute of Respiratory Diseases, Shenzhen People's  
Hospital, The Second Clinical Medical College of Jinan  
University, The First Affiliated Hospital of Southern  
University of Science and Technology  
Shenzhen Shi, Guangdong Shen, China

## ABSTRACT

Detecting respiratory events in sleep requires much attention and is labor consuming conventionally. With the development of technology, some kinds of software that can automatically detect the respiratory events was designed to help simplify and improve this process. However, in order to ensure its accuracy of the detection, it is necessary to provide appropriate key parameters before using it. After that the interval adjustment also needs to be done manually, which still takes a lot of time and means high demands on the technicians. In this paper, an end-to-end ConvNet was used to detect the respiratory events which does not need to provide any extra parameters. Its performance was further compared with widely used events detection software, Philips Sleepware G3 with Smonolyzer. The results show that ConvNet has higher accuracy than G3 with Smonolyzer in event detection. Such a ConvNet-based analysis system is sufficiently accurate for event detection according to the AASM classification criteria.

## CCS CONCEPTS

• Computing methodologies → Object detection.

## KEYWORDS

Respiratory event, End-to-End ConvNet, Smonolyzer

### ACM Reference Format:

Yanping Shuai, Zhangbo Li, Xingjun Wang, and Hanrong Cheng. 2023. Detecting Respiratory Events with End-to-End ConvNet. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590098>

\*Corresponding authors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590098>

## 1 INTRODUCTION

With the increasing pace of life, the occurrence of sleep breathing disorders is becoming more and more common[21]. It affects all aspects of life, causes various diseases, and can even be life-threatening in serious cases[4, 12, 20, 24, 25, 27]. Polysomnography (PSG) is the international standard for sleep breathing disorder detection[8]. It is a variety of physiological signals collected during sleep monitoring, which usually include EEG, EOG, EMG, ECG, and Airflow[11]. Sleep respiratory event detection from PSG signals refers to the use of physiological signal analysis techniques to analyze respiratory signals during sleep monitoring to detect respiratory events that occur during sleep[22]. This method can help physicians assess a patient's breathing and detect possible disorders such as sleep breathing disorders in a timely manner. The manual detection of respiratory events is a time-consuming and labor-intensive task[11]. With the development of technology, some software for automatic detection of respiratory events has been developed[1]. But the use of these software requires appropriate parameters to be set in advance and manual adjustment of the event interval afterwards, which is also time-consuming and requires a certain level of technician skill, a more efficient and accurate respiratory event detection algorithm is imperative[13, 17].

There are several challenges in the develop of respiratory event detection algorithm. The first is noise interference, respiratory event detection algorithm needs to process the signals from the sensors, which may have various noise interferences (e.g., body motion, ventilator noise, etc.)[15]. The algorithm needs to have a strong anti-interference capability in order to effectively extract the features of respiratory events. Secondly, due to the diversity of respiratory events, there are various types of respiratory events[6] (e.g., apnea, hypoventilation, etc.) and they may be different for each individual. The algorithm needs to be highly adaptable in order to accurately detect different types of respiratory events. Finally, the data volume is huge, with PSG signals often lasting for hours or even days[11][3]. Algorithms need to be computationally efficient in order to process and analyze this data in an acceptable amount of time.

In this paper, an end-to-end convolutional network detection algorithm is proposed. The end-to-end respiratory event detection algorithm can directly learn the process of extracting useful features from the original signal, which eliminates the manual feature

extraction step and makes the algorithm simpler and easier to use. Secondly, end-to-end respiratory event detection algorithms can learn features that are universal and therefore have a strong generalization capability. This means that the algorithm can perform well on various types of data. Lastly, the convolutional network can process large amounts of data quickly using GPU accelerated computing and produce results in a short time.

## 2 DATASET

Data for the study was obtained from 680 patients at Shenzhen People's Hospital who underwent full PSG monitoring using Philips Alice 6. However, after a follow-up examination, 65 of these cases were excluded due to loss of important channels associated with respiratory event detection. To compare the detection results of each method, 55 cases of data were randomly selected by symptom distribution (See Fig 1) as the test set. The remaining 625 cases were used as the training data for the model.

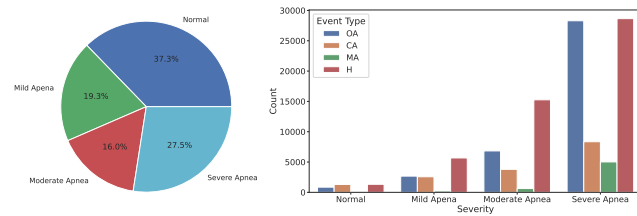


Figure 1: The proportion of each symptom and distribution of respiratory events in dataset

## 3 METHOD

### 3.1 Manual Detection and Smonolyzer Detection

As a benchmark for comparison, full manual inspection was performed in the test set. Three highly trained PSG technicians with Registered Polysomnographic Technologist Certificate scored event detection in accordance with AASM 2012 guidelines[3]. As a comparison, Sleepware G3 with integrated Somnolyzer[2][26] detected the test set as another label. It's the first time that Somnolyzer was applied to the Asian population dataset for such respiratory events detection analysis in research.

### 3.2 Data Preprocessing

Referring to the channels used in the identification of respiratory events in the AASM manual, the airflow signal, the thoracic-abdominal belt signal, the blood oxygen signal and the forehead EEG signal were selected[3] from the PSG signals as shown in Fig 2 (due to the problem of partial signal shedding, only the EEG fc3 channel was selected). In order to unify the input to ConvNet, all signals were resampled to 10 Hz. Then the selected signals were cut into slices that can be fed into the network. To prevent the loss of events at the edge of the slice, overlapping slice[14] was used. That is, incomplete respiratory events at the edge of each slice are not counted, and the second half of the previous slice is used as the latter instead. Referring to the histogram of respiratory event

durations below displayed in Fig 2, all respiratory events can be covered using overlapping slices with a duration of 256 seconds. In order to facilitate network identification, 0 – 1 normalization scaling is performed on the signal before put into the network.

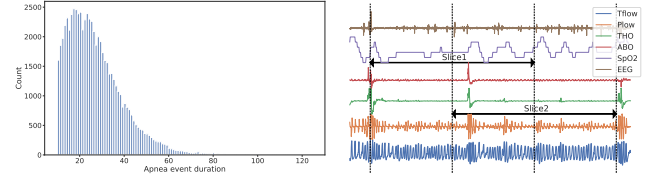


Figure 2: The distribution of the apnea events duration and Two slices of data with selected channel signals

### 3.3 ConvNet Model

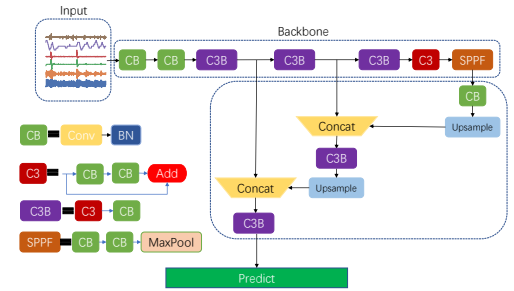


Figure 3: Structure of the ConvNet.

As displayed in Fig 3, the detection network, ConvNet is divided into three parts. After preprocessing, the first step is to extract features of the input with Backbone network. It mainly consists of three residual-based convolutional networks[10]. The network is followed by two convolutional layers immediately after the input layer, then three remaining blocks consisting of two convolutional layers for extracting low-level, mid-level, and high-level features of the signal, respectively. The Neck layer fuses the extracted features by means of upsampling and Concat[16]. In the end,  $40 \times 3$  vector was used to output the result, representing the coordinate offset, event duration and predict confidence.

### 3.4 Loss Function

As show in Fig 4, The signal slices are divided into 40 small regions during training, and each small region has a detection segment corresponding to it for detecting targets. When a target spans multiple grids, it is detected multiple times in all relevant regions. In the training process, the one-dimensional signal use DIoU[28] as the basic loss function to obtain satisfactory results. The overall loss function is

$$Loss = L_{conf} + L_{loc} \quad (1)$$

$$L_{(conf)} = 1 - DIoU(b, event) \quad (2)$$

$$L_{loc} = BCE(obj, DIoU) \quad (3)$$

$$DIoU = \rho^2(A, B)/c^2 \quad (4)$$

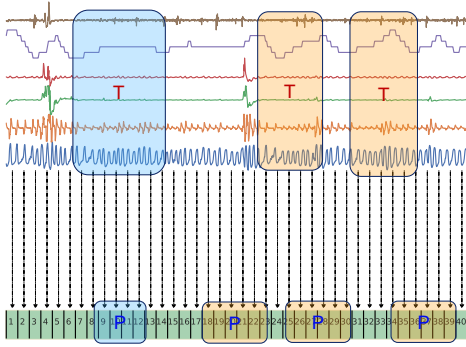


Figure 4: predict vector with target events.

Where,  $\rho(A, B)$  is the distance between the center points of the two events. Through this loss function, the fitted event interval offset and event confidence can be obtained by training with the existing labels.

### 3.5 PostProcessing

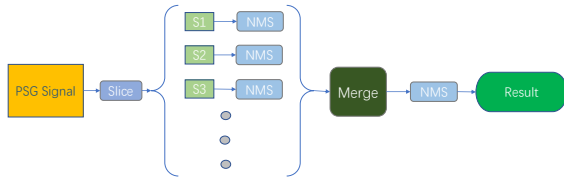


Figure 5: Two stages of NMS.

Through detection, 256s length of the input signals were map the into a 40-bit vector. After zooming, specific coordinates can be obtained. Non-maximum suppression (NMS)[19] is a technique used in image processing to detect and remove areas of strong response in an image. Since the scheme of repeated slicing is adopted when passing through slices before, when merging all slices, the problem of overlap slicing also needs to be considered. Here, NMS was directly applied to all the slices (See Fig 5). Event with the highest confidence was picked out the as the final result. All these picked events were taken as detected respiratory events of a complete case.

## 4 RESULTS

### 4.1 Accuracy

In the general neural network detection scheme, the precision rate (P), recall rate (R), and F1[5] are generally used for evaluation.

Table 1: The accuracy of ConvNet and Smonolyzer when IOU>0.5

	P	R	F1
G3 with Smonolyzer	0.807	0.866	0.835
ConvNet	0.858	0.882	0.870

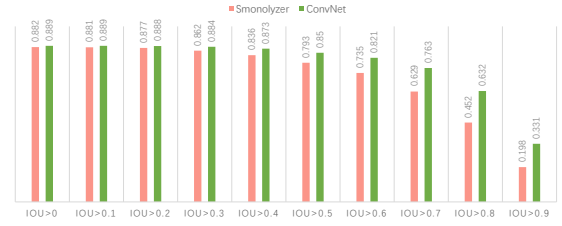


Figure 6: The F1 score in different IOU.

Fig 6 demonstrates the comparison of F1 metrics at each IOU[28]. And it is clear that in all segments, the proposed ConvNet outperforms the metrics compared to Smonolyzer, and the gap increases as the IOU increases, which indicates that ConvNet can better match the manual labeling results and can reduce the workload of subsequent manual adjustments, which shows the positive significance of the scheme. In particular, as shown in Table 1 selected at IOU > 0.5, the proposed scheme outperforms the results of Smonolyzer in all metrics.

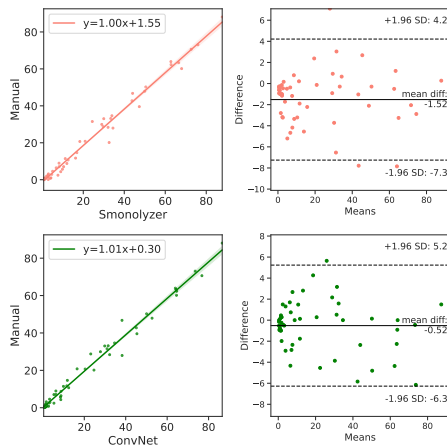
### 4.2 Consistency

AHI stands for apnea-hypopnea index[3]. It is a measure of the severity of sleep apnea and is calculated by dividing the total number of apneas and hypopneas that occur during a sleep study by the total number of hours of sleep. To verify the consistency of AHI between the ConvNet and Smonolyzer with manual annotation, the Kappa consistency coefficient[23] and Bland-Altman[9] plots were utilized.

Table 2: The comparison of Kappa values

	Kappa	Standard error
G3 with Smonolyzer	0.854	0.021
ConvNet	0.858	0.016

The Kappa coefficient was used to measure the degree of agreement and to compare the agreement of the two diagnostic methods for uniform case results. Where Kappa values less than 0.2 indicate poor agreement, between 0.2 and 0.4 indicate moderate agreement, between 0.6 and 0.8 indicate strong agreement, and between 0.8 and 1.0 indicate strong agreement. As can be seen from Table 2, both the result of ConvNet and Smonolyzer have a strong agreement with manual annotation on AHI, with the ConvNet having a slightly higher agreement, which is in consistent with the accuracy of respiratory event detection. The Bland-Altman plot visualizes the consistency of the results analyzed by the two event detection methods, where the horizontal coordinate is the mean of the measurements by the two methods and the vertical coordinate is the difference between the two methods. The middle line indicates the mean of the difference, and the upper and lower lines are the upper and lower limits of the 95% agreement limits (i.e., 1.96 standard deviation upper and lower bound values). If the scatter points basically all fall in the 95% agreement interval, the agreement is good. As can be seen from Fig 7, the detection values of ConvNet and



**Figure 7: Bland-Altman plots of AHI-Manual and AHI-detected in the test dataset. The Red one result is detected by Smonolyzer and the green one is detected by ConvNet.**

Smonolyzer are very close to those of the manual annotation, and both have a considerable effect in assisting treatment.

## 5 CONCLUSION AND FUTURE WORK

An end-to-end fully convolutional network that allows the detection of PSG multichannel signals is presented in this work. With proper signal preprocessing and suitable slicing method, training using clinical labels, and finally combined with dual NMS, it enables fast detection of accurate localization of respiratory events in complete PSG signals. Comparison with the test set, while maintaining the high consistency of AHI, the network provides an advantage over the now commonly used clinical Smonolyzer in terms of accuracy metrics on all IOU segments, and better results as the IOU increases. It indicates that the use of this algorithm results in more accurate test results and closer to manual event intervals, which helps reduce the burden on clinical technologists and contributes to the achievement of fully automated detection.

The detection network demonstrated in this work can be used in future wearable device[18] detection schemes to achieve pre-screening[7] to reduce the medical burden. It also has great potential for the detection of generic timing signals, such as speech detection, engine fault detection, etc. The evaluation of these application areas will be left for future work.

## ACKNOWLEDGMENTS

This research was supported by Shenzhen Municipal Natural Science Foundation (WDZC20200818121348001).

## REFERENCES

- [1] Peter Anderer, Arnaud Moreau, Michael Woertz, et al. 2010. Computer-Assisted Sleep Classification according to the Standard of the American Academy of Sleep Medicine: Validation Study of the AASM Version of the Somnolyzer 24 × 7. *Neuropsychobiology* 62, 4 (2010), 250–264. <https://doi.org/10.1159/000320864> Publisher: Karger Publishers.
- [2] Peter Anderer, Marco Ross, Andreas Cerny, and Edmund Shaw. 2022. Automated Scoring of Sleep and Associated Events. *Advances in the Diagnosis and Treatment of Sleep Apnea* (2022), 107–130. [https://doi.org/10.1007/978-3-031-06413-5\\_7](https://doi.org/10.1007/978-3-031-06413-5_7) Publisher: Springer, Cham.
- [3] Richard B. Berry, Rita Brooks, Charlene E. Gamaldo, et al. 2012. The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications*, Darien, Illinois, American Academy of Sleep Medicine 176 (2012), 2012.
- [4] Rohit Budhiraja, Pooja Budhiraja, and Stuart F. Quan. 2010. Sleep-disordered breathing and cardiovascular disorders. *Respiratory Care* 55, 10 (2010), 1322–1332. ISBN: 0020-1324 Publisher: Respiratory Care.
- [5] Nancy Chinchor. 1992. MUC-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding (MUC4 '92)*. Association for Computational Linguistics, USA, 22–29. <https://doi.org/10.3115/1072064.1072067>
- [6] Heidi Danker-Hopfe, Peter Anderer, Josef Zeitlhofer, et al. 2009. Interrater reliability for sleep scoring according to the Rechtschaffen & Kales and the new AASM standard. *Journal of sleep research* 18, 1 (2009), 74–84. ISBN: 0962-1105 Publisher: Wiley Online Library.
- [7] Marco Fernandez, Kathy Burns, Beverly Calhoun, et al. 2007. Evaluation of a new pulse oximeter sensor. *American Journal of Critical Care* 16, 2 (2007), 146–152. ISBN: 1062-3264 Publisher: AACN.
- [8] American Academy of Sleep Medicine Task Force. 1999. Sleep-related breathing disorders in adults: recommendations for syndrome definition and measurement techniques in clinical research. The Report of an American Academy of Sleep Medicine Task Force. *Sleep* 22, 5 (1999), 667.
- [9] Davide Giavarina. 2015. Understanding bland altman analysis. *Biochemia medica* 25, 2 (2015), 141–151. ISBN: 1330-0962 Publisher: Medicinska naklada.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [11] Conrad Iber. 2007. The AASM manual for the scoring of sleep and associated events: Rules. *Terminology and Technical Specification* (2007). Publisher: American academy of sleep medicine.
- [12] Alex Iranzo and Joan Santamaria. 2015. Sleep in neurodegenerative diseases. *Sleep medicine* (2015), 271–283. Publisher: Springer.
- [13] Amy S Jordan, David G McSharry, and Atul Malhotra. 2014. Adult obstructive sleep apnoea. *The Lancet* 383, 9918 (Feb. 2014), 736–747. [https://doi.org/10.1016/S0140-6736\(13\)60734-5](https://doi.org/10.1016/S0140-6736(13)60734-5)
- [14] Ho-Chan Kim, Jae-Won Choi, Eric MacDonald, and Ryan Wicker. 2010. Slice overlap-detection algorithm for process planning in multiple-material stereolithography. *The International Journal of Advanced Manufacturing Technology* 46, 9 (Feb. 2010), 1161–1170. <https://doi.org/10.1007/s00170-009-2181-x>
- [15] Clete A. Kushida, Michael R. Littner, Timothy Morgenthaler, et al. 2005. Practice parameters for the indications for polysomnography and related procedures: an update for 2005. *Sleep* 28, 4 (April 2005), 499–521. <https://doi.org/10.1093/sleep/28.4.499>
- [16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, et al. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.
- [17] Damien Léger and Virginie Bayon. 2010. Societal costs of insomnia. *Sleep medicine reviews* 14, 6 (2010), 379–389. ISBN: 1087-0792 Publisher: Elsevier.
- [18] Mokhinabonu Mardonova and Yosoon Choi. 2018. Review of wearable device technology and its applications to the mining industry. *Energies* 11, 3 (2018), 547. ISBN: 1996-1073 Publisher: MDPI.
- [19] Alexander Neubeck and Luc Van Gool. 2006. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3. IEEE, 850–855.
- [20] Jan B. Pietzsch, Abigail Garner, Lauren E. Cipriano, and John H. Linehan. 2011. An integrated health-economic analysis of diagnostic and therapeutic strategies in the treatment of moderate-to-severe obstructive sleep apnea. *Sleep* 34, 6 (2011), 695–709. ISBN: 0161-8105 Publisher: Oxford University Press.
- [21] Naresh M. Punjabi. 2008. The Epidemiology of Adult Obstructive Sleep Apnea. *Proceedings of the American Thoracic Society* 5, 2 (Feb. 2008), 136–143. <https://doi.org/10.1513/pats.200709-155MG> Publisher: American Thoracic Society - PATS.
- [22] S. Quan, B. Howard, C. Iber, et al. 1997. The Sleep Heart Health Study: design, rationale, and methods. *Sleep* (1997). <https://doi.org/10.1093/SLEEP/20.12.1077>
- [23] Alan S. Rigby. 2000. Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. *Disability and rehabilitation* 22, 8 (2000), 339–344. ISBN: 0963-8288 Publisher: Taylor & Francis.
- [24] Tracy L. Skaer and David A. Sclar. 2010. Economic implications of sleep disorders. *Pharmacoeconomics* 28, 11 (2010), 1015–1023. ISBN: 1179-2027 Publisher: Springer.
- [25] Michael H. Smolensky, Lee Di Milia, Maurice M. Ohayon, and Pierre Philip. 2011. Sleep disorders, medical conditions, and road accident risk. *Accident Analysis & Prevention* 43, 2 (2011), 533–548. ISBN: 0001-4575 Publisher: Elsevier.
- [26] Laurie Thiesse, Luc Staner, Gil Fuchs, et al. 2022. Performance of Somno-Art Software compared to polysomnography interscorer variability: A multi-center study. *Sleep Medicine* 96 (Aug. 2022), 14–19. <https://doi.org/10.1016/j.sleep.2022.04.013>
- [27] Stephen Tregear, James Reston, Karen Schoelles, and Barbara Phillips. 2009. Obstructive sleep apnea and risk of motor vehicle crash: systematic review and meta-analysis. *Journal of clinical sleep medicine* 5, 6 (2009), 573–581. ISBN:

1550-9389 Publisher: American Academy of Sleep Medicine.

- [28] Zhaohui Zheng, Ping Wang, Wei Liu, et al. 2020. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference*

*on artificial intelligence*, Vol. 34. 12993–13000. Issue: 07.

# Human Activity Recognition based on Transformer in Smart Home

Xinmei Huang

huangxm20@tsinghua.mails.edu.cn  
Shenzhen International Graduate School, Tsinghua  
University  
Shenzhen, Guangdong, China

Sheng Zhang\*

zhang\_sh@mail.tsinghua.edu.cn  
Shenzhen International Graduate School, Tsinghua  
University  
Shenzhen, Guangdong, China

## ABSTRACT

With the advancement of artificial intelligence, smart home has attracted much attention from scholars. Human Activity Recognition (HAR) is a crucial foundation for various applications in smart home. In this paper, to improve the accuracy of HAR and promote the development of applications and services in smart home, we propose a Transformer-based approach that integrates multiple sensor sequence inputs for HAR. We integrate sequence features, collect contextual information, and employ Transformer to recognize various activities for the CASAS Aruba dataset that uses environmental sensors. The validation results on real-world dataset demonstrate its effectiveness compared to traditional machine learning and deep learning methods.

## CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing*; • **Computing methodologies**;

## KEYWORDS

Human activity recognition, Transformer, Smart home

### ACM Reference Format:

Xinmei Huang and Sheng Zhang. 2023. Human Activity Recognition based on Transformer in Smart Home. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590100>

## 1 INTRODUCTION

In recent years, Internet of Things (IoT) has rapidly developed. With the advancement of the Internet, wireless communication and other technologies, smart home has become an essential intelligent environment in IoT, which has greatly contributed to creating intelligent services and greatly facilitated people's daily lives. Human Activity Recognition (HAR) in smart home can utilize user behaviors and inferring user needs, which provide solutions in applications such as home intelligence services and home healthcare for the elderly.

\*Corresponding author.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CACML 2023, March 17–19, 2023, Shanghai, China  
© 2023 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-9944-9/23/03.  
<https://doi.org/10.1145/3590003.3590100>

HAR focuses on identifying and classifying human actions and behaviors by analysing collected data. HAR has a wide range of applications, including behavior analysis[1], healthcare[2], sports[3] and smart home[4]. There are two main types of HAR: vision-based and sensor-based[5]. Vision-based HAR can utilize high-quality visual information but require significant computational resources and raise important user privacy concerns. On the other hand, sensor-based HAR uses data from sensors such as accelerometers, gyroscopes, and magnetometers to recognize human activities. Sensor-based HAR includes wearable, object, environmental, and hybrid sensors[6], as shown in Table 1. Sensor-based HAR is more cost-effective and efficient that can work in real-time. Environmental sensor-based HAR is widely applied in smart home, which is unobtrusively embedded into users' daily lives to minimize negative impact. Building on this foundation, offering users more active and intelligent services becomes possible.

HAR in smart home is usually treated as a classification problem. Currently, many papers have utilized various Machine Learning (ML) and Deep Learning (DL) methods in HAR based on smart home environment sensors, such as KNN[7], SVM[7], LogitBoost[8], CNN[9], LSTM[10, 11]. Transformer has performed well in Natural Language Processing (NLP) and Computer Vision (CV) since it was proposed in 2017. However, using Transformer in HAR based on smart home environmental sensors is rarely explored.

In this paper, to optimize sensor-based HAR in smart homes, we propose a method to apply Transformer to user HAR in smart home, and specifically optimize the work of deep feature extraction. To illustrate its superiority over traditional ML and DL methods, we verify the effectiveness of our approach on the CASAS Aruba dataset, which collects the state of environmental sensors in a smart home and generates the sequences of sensor events.

## 2 METHODS

In this section, we provide details on the principle of Transformer and the specific steps of HAR. In addition, our approach was introduced.

### 2.1 Transformer

Transformer is a powerful deep learning model proposed by Google in 2017 which innovatively adopts the self-attention mechanism[12]. The Transformer has achieved significant performance improvements in many NLP[13] and CV[14] tasks and has become one of the essential foundational models.

Transformer consists of two main parts: the encoder and the decoder, as shown in Fig. 1. The encoder is composed of multiple layers, each consisting of two sub-layers: a self-attention layer and

**Table 1: Different Types of Sensor-based HAR**

Type	Description	Examples
Wearable sensors	Worn by users or integrated into portable devices	Smartphones, smartwatches, smart bands, etc.
Object sensors	Attached to objects to describe related activities	Accelerometer on cup, RFID, etc.
Environmental sensors	Planted in surroundings to reflect human activities	Thermometer, hygrometer, door Sensor, etc.
Hybrid sensors	Combination of different types of sensors	Combination of types planted in smart environments

a position-wise feed-forward layer. The decoder is also composed of multiple layers, each consisting of three sub-layers: a masked self-attention layer, an encoder-decoder attention layer, and a position-wise feed-forward layer.

The formulas for the self-attention and position-wise feed-forward layers are as follows: In the self-attention layer, compute the attention weights as a softmax function of the dot product of query ( $Q$ ), key ( $K$ ), and value ( $V$ ) matrices:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (1)$$

where  $Q$ ,  $K$ , and  $V$  are matrices representing the queries, keys, and values in the self-attention layer.  $d_k$  represents the dimension of the key vectors.

Using multiple attention heads in the Transformer model enables it to simultaneously attend to information from various representation subspaces at different positions. Compute multiple attention scores in parallel by linearly projecting the queries, keys, and values matrices into several heads:

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \quad (2)$$

Concatenate the outputs from each head into a single matrix, and then project the concatenated matrix using a linear layer:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O, \quad (3)$$

where the projections are parameter matrices  $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$ ,  $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$  and  $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$ .

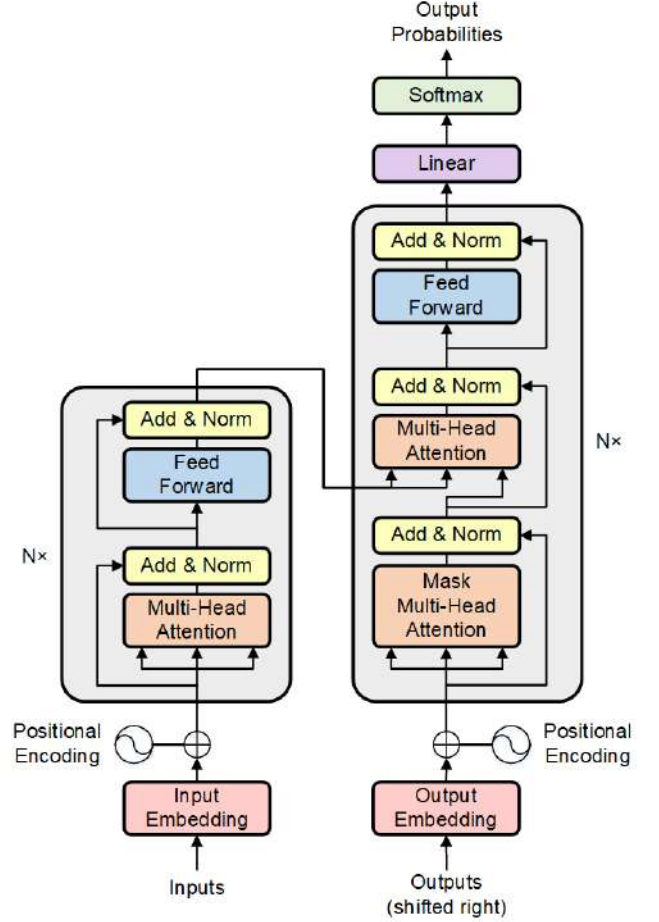
In the position-wise feed-forward layer, apply a two-layer neural network consisting of a ReLU activation function:

$$\text{FFN}(x) = \max(0, xW_1 + b_1) W_2 + b_2, \quad (4)$$

where  $W_1$ ,  $W_2$ ,  $b_1$ , and  $b_2$  are learnable parameters in the position-wise feed-forward layer.

## 2.2 Our Approach

The input of the recognition model is the sequence of events collected by the smart home sensors, and the output is the sequence of human activities. Therefore, Transformer with excellent capabilities in Sequence-to-Sequence (Seq2Seq) implementation is very suitable for sensor-based HAR in smart home. In addition, Transformer allows parallel processing of input sequences and efficient processing of long sequences due to its self-attention mechanism. Compared with the traditional CNN and RNN, it has higher efficiency and scalability in the task of HAR, and is more suitable for the rapid development trend of the number of smart home sensors.

**Figure 1: Model Architecture of Transformer**

Based on the outstanding performance of Transformer in Seq2Seq implementation, our method utilizes Transformer as a recognition model for HAR in smart home. We also designed and implemented processes according to the architecture of sensor-based HAR.

Sensor-based HAR can be regarded as a classification problem. It involves several steps: (a) data collection, (b) data preprocessing, (c) feature extraction and model training, and (d) activity recognition, as shown in Fig. 2.

Firstly, various environmental sensors collect data and generate the sequences of sensor events containing a series of information, such as time and sensor states.

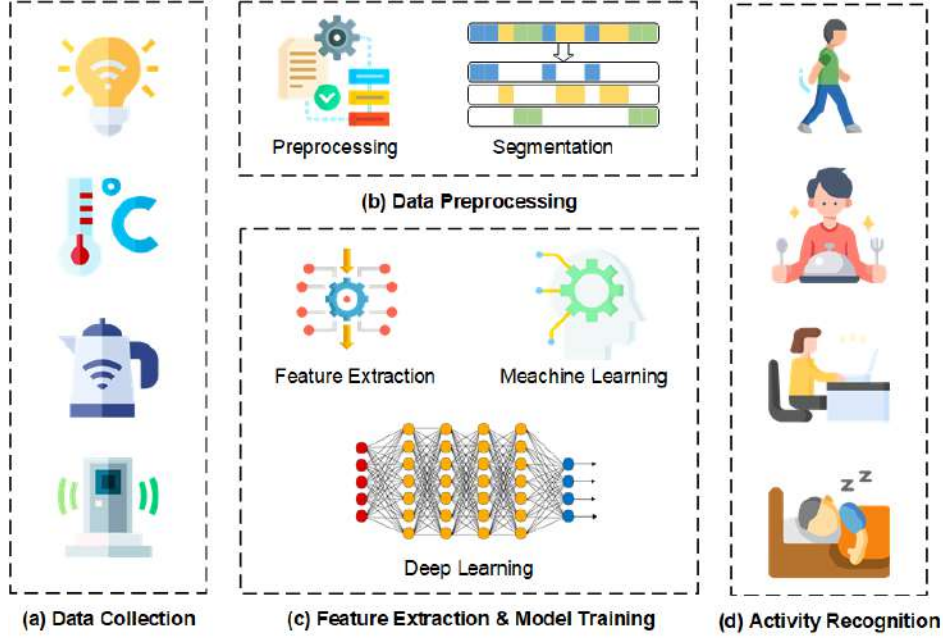


Figure 2: Architecture of Sensor-based HAR

Secondly, we preprocess and segment the raw data. We remove abnormal data from the raw data and standardize the data format. Unlike smartphones and wearable sensors, most smart home sensors do not collect data at a fixed frequency but generate sensor event streams based on user activity triggers. Therefore, using windows with an equal time interval or sensor event is not suitable[15]. Instead, we use the activity-driven sliding windows, which detect the window where user activity occurs and segments the sequence accordingly. The segmentation of activities is shown in Fig. 3. This method is more flexible and accurate.

Next, extract deep features from the input sequences based on the characteristics of the sensor data. For state-type sensors, we extract features such as state, frequency, and time. For numerical-type sensors, we use statistical features such as maximum, minimum, mean, variance, and standard deviation. For improving the performance of the activity recognition model, deep feature extraction can effectively represent the sequence features of the original data, powerfully integrate multi-sensor input features, and fully use contextual information such as time and temperature in the sensor data. After the above processing, the raw sensor data has become a sequence with feature information, which can be used as the input to the recognition model.

Finally, we use Transformer to recognize the human activities. The next section will provide a detailed introduction to the Transformer.

### 3 EXPERIMENT

We evaluate our approach on a CASAS smart home dataset called Aruba. In this section, we describe the Aruba dataset used, the parameter settings, and evaluation metrics, and analyze the experimental results.

#### 3.1 Dataset

We use the Aruba dataset, which is part of the CASAS smart home project[16] implemented by Washington State University (WSU). The Aruba dataset collects sensor data from a smart home environment where a female volunteer resides. It captures sensor events generated by motion sensors, door closure sensors, and temperature sensors. The details of the dataset are shown in Table 2. Since the activity "Resperate" appears only 6 times, we excluded this activity and focused on HAR experiment involving the remaining 10 activities. Fig. 4 presents the employment of sensors within the space, and the collected data are displayed in Table 3.

#### 3.2 Experimental Settings and Evaluation Metrics

We carefully selected and optimized the parameters of the Transformer model, which are provided in detail in Table 4. We then applied our model to perform HAR on the Aruba dataset. The results verify the effectiveness of our approach.

We use precision, recall, F1-score to evaluate the performance of Transformer model. The metrics are defined as follows:

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

where  $TP$ ,  $FP$  and  $FN$  stands for true positives, false positives and false negatives.

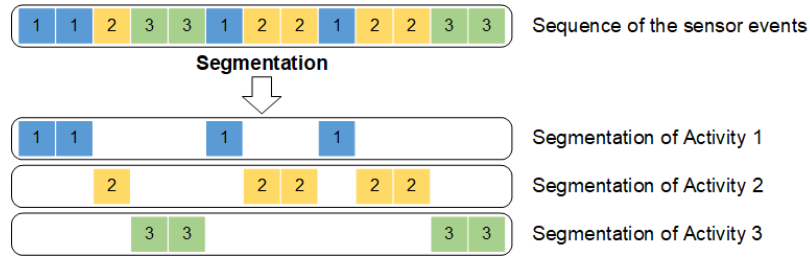


Figure 3: Segmentation of Activities

Table 2: the Details of Aruba Dataset

Type of Sensors	the Number of Sensors	Activities	Activity Occurrences	Duration Days
M, D, T	40	11	6477	219

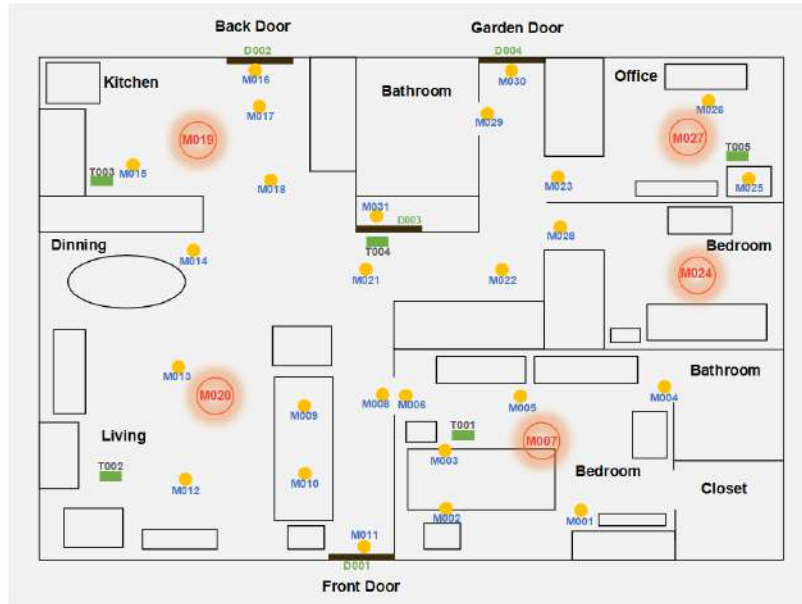


Figure 4: the Employment of Sensors within the Space

Table 3: Collected Data in Aruba Dataset

Date	Time	Sensor ID	State	Event Label
2010-11-04	05:18:27.278517	T001	18.5	.....
2010-11-04	05:40:43.642664	M003	OFF	Sleeping end
2010-11-04	05:40:44.223548	M003	ON	
2010-11-04	05:40:45.939846	M005	ON	
2010-11-04	05:40:46.310862	M003	OFF	
2010-11-04	05:40:51.303739	M004	ON	Bed_to_Toilet begin
.....				
2010-11-04	05:43:30.279021	M004	OFF	Bed_to_Toilet end
.....				
2010-11-04	05:43:45.7324	M003	ON	Sleeping begin
.....				

**Table 4: Parameters Settings of the Transformer**

Parameter	Setting
d_model	40
number of Attention heads	4
batch_size	128
sequence length	200

**Table 5: Performance Comparison in Single-activity Recognition Tasks**

Approach	Metrics	MP	Rlx	Eat	Wk	Slp	WD	BT	EH	LH	Hk
KNN	Precision	0.951	0.941	0.917	0.875	1.000	0.000	0.857	0.875	0.830	0.000
	Recall	0.969	0.990	0.846	0.778	1.000	0.000	0.750	0.795	0.886	0.000
	F1-score	0.960	0.965	0.880	0.824	1.000	0.000	0.800	0.833	0.857	0.000
GBDT	Precision	0.952	1.000	1.000	1.000	1.000	0.000	1.000	0.878	0.830	0.667
	Recall	0.981	0.990	1.000	1.000	1.000	0.000	1.000	0.818	0.886	1.000
	F1-score	0.966	0.995	1.000	1.000	1.000	0.000	1.000	0.847	0.857	0.800
CNN	Precision	0.946	0.997	0.963	1.000	1.000	0.000	1.000	0.900	<b>0.833</b>	0.667
	Recall	0.988	0.997	1.000	1.000	1.000	0.000	1.000	0.818	<b>0.909</b>	0.500
	F1-score	0.967	0.997	0.981	1.000	1.000	0.000	1.000	0.857	<b>0.870</b>	0.571
RNN	Precision	0.953	1.000	1.000	1.000	1.000	0.000	1.000	0.872	0.796	0.750
	Recall	1.000	1.000	0.962	1.000	1.000	0.000	1.000	0.773	0.886	0.750
	F1-score	0.976	1.000	0.980	1.000	1.000	0.000	1.000	0.819	0.839	0.750
<b>Our Approach</b>	Precision	<b>0.958</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.000	<b>1.000</b>	<b>0.878</b>	0.830	<b>1.000</b>
	Recall	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.000	<b>1.000</b>	<b>0.818</b>	0.886	<b>1.000</b>
	F1-score	<b>0.979</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	<b>1.000</b>	0.000	<b>1.000</b>	<b>0.847</b>	0.857	<b>1.000</b>

### 3.3 Results and Analysis

Our method shows excellent performance in single-activity recognition tasks, as shown in the table 5. We use words such as "MP" "Rlx" in table 5 to briefly represent 10 activities: "Meal\_Preparation" "Relax" "Eating" "Work" "Sleeping" "Wash\_Dishes" "Bed\_to\_Toilet" "Enter\_Home" "Leave\_Home" and "Housekeeping". Out of 10 activities, the precision, recall, and F1-score of 6 activities each achieved a perfect score of 1.000. Due to the small sample size of "Wash\_Dishes," it was difficult to achieve accurate classification. Additionally, we compared our approach with KNN, GBDT, CNN, and RNN as recognition models on this dataset, and the results are shown in the table 6. Through our experiments, we demonstrate the effectiveness of our approach compared with traditional ML and DL methods, which achieved high accuracy in recognizing human activities from sensor data.

**Table 6: Performance Comparison of KNN, GBDT, CNN, RNN, Our Approach**

Approach	Precision	Recall	F1-score
KNN	0.915	0.931	0.922
GBDT	0.956	0.96	0.958
CNN	0.953	0.963	0.958
RNN	0.954	0.963	0.958
<b>Our Approach</b>	<b>0.959</b>	<b>0.969</b>	<b>0.964</b>

### 4 CONCLUSION

We propose a HAR method based on Transformer which can be applied in smart home environment. We processed raw data based on the characteristics of environmental sensors, extracted multiple features and contextual information, and innovatively adopted Transformer as the recognition model.

Transformer-based approach shows its power in HAR in smart home. Experimental results on the Aruba dataset demonstrate that it significantly outperforms traditional ML and DL models. It may be due to the unique self-attention mechanism of Transformer that gives it an advantage in processing sensor sequences.

Overall, HAR based on Transformer in smart home is a highly effective method for recognizing user behaviors and activities. These results contribute to the growing body of research exploring the use of advanced techniques for HAR in smart home, and provide a promising avenue for further investigation in this field.

### ACKNOWLEDGMENTS

This work was supported by Shenzhen Science and Technology Program (JCYJ20180508152046428) in China. This work was also supported by Key Laboratory of Advanced Sensor and Integrated System, Shenzhen International Graduate School, Tsinghua University.

### REFERENCES

- [1] Praneeth Vepakomma, Debraj De, Sajal K Das, and Shekhar Bhansali. 2015. A-wristocracy: Deep learning on wrist-worn sensing for recognition of user complex

- activities. In *2015 IEEE 12th International conference on wearable and implantable body sensor networks (BSN)*. IEEE, 1–6.
- [2] Yan Wang, Shuang Cang, and Hongnian Yu. 2019. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications* 137 (2019), 167–190.
  - [3] Thomas Kautz, Benjamin H Groh, Julius Hannink, Ulf Jensen, Holger Strubberg, and Bjoern M Eskofier. 2017. Activity recognition in beach volleyball using a deep convolutional neural network: Leveraging the potential of deep learning in sports. *Data Mining and Knowledge Discovery* 31 (2017) 1678–1705.
  - [4] Asma Benmansour, Abdelhamid Bouchachia, and Mohammed Feham. 2015. Multioccupant activity recognition in pervasive smart home environments. *ACM Computing Surveys (CSUR)* 48, 3 (2015), 1–36.
  - [5] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern recognition letters* 119 (2019), 3–11.
  - [6] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition* 108 (2020), 107561.
  - [7] Nawel Yala, Belkacem Fergani, and Anthony Fleury. 2017. Towards improving feature extraction and classification for activity recognition on streaming data. *Journal of Ambient Intelligence and Humanized Computing* 8, 2 (2017), 177–189.
  - [8] Janns Alvaro Patiño-Saucedo, Paola Patricia Ariza-Colpas, Shariq Butt-Aziz, Marlon Alberto Piñeres-Melo, José Luis López-Ruiz, Roberto Cesar Morales-Ortega, and Emiro De-la Hoz-Franco. 2022. Predictive model for human activity recognition based on machine learning and feature selection techniques. *International Journal of Environmental Research and Public Health* 19, 19 (2022), 12272.
  - [9] Deepika Singh, Erinc Merdivan, Sten Hanke, Johannes Kropf, Matthieu Geist, and Andreas Holzinger. 2017. Convolutional and recurrent neural networks for activity recognition in smart environment. In *Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers* Springer, 194–205.
  - [10] Khaled A Alaghbari, Mohamad Hanif Md Saad, Aini Hussain, and Muhammad Raisul Alam. 2022. Activities recognition, anomaly detection and next activity prediction based on neural networks in smart homes. *IEEE Access* 10 (2022), 28219–28232.
  - [11] Daniele Liciotti, Michele Bernardini, Luca Romeo, and Emanuele Frontoni. A sequential deep learning application for recognising human activities in smart homes. 2020. *Neurocomputing* 396 (2020), 501–513.
  - [12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
  - [13] Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 12462–12470.
  - [14] Yi Liu, Hao Yuan, Zhengyang Wang, and Shuiwang Ji. 2020. Global pixel transformers for virtual staining of microscopy images. *IEEE Transactions on Medical Imaging* 39, 6 (2020), 2256–2266.
  - [15] Jie Wan, Michael J O’grady, and Gregory MP O’Hare. 2015. Dynamic sensor event segmentation for real-time activity recognition in a smart home context. *Personal and Ubiquitous Computing* 19 (2015), 287–301.
  - [16] Diane J Cook. 2012. Learning setting-generalized activity models for smart spaces. *IEEE intelligent systems* 27, 1 (2012), 32.

# Research on Constant Perturbation Strategy for Deep Reinforcement Learning

Jiamin Shen  
Shenyang Aerospace University  
Shenbei Qu, Shenyang Shi, China  
jiaminshen163@163.com

LI XU\*  
Shenyang Aerospace University  
Shenbei Qu, Shenyang Shi, China  
xulibak@163.com

XU WAN  
Shenyang Aircraft Design and  
Research Institute  
Huanggu Qu, Shenyang Shi, China

JIXUAN CHAI  
Shenyang Aerospace University  
Shenbei Qu, Shenyang Shi, China

Chunlong Fan  
Shenyang Aerospace University  
Shenbei Qu, Shenyang Shi, China

## ABSTRACT

The development of attack algorithms for deep reinforcement learning is an important part of its security research. In this paper, we propose a deep reinforcement constant perturbation strategy approach for deep reinforcement learning with long-range time-series dependence from the perspective of the sequence of interaction between an agent and its environment. The algorithm is based on a small amount of historical interaction information, and a constant perturbation is designed to disrupt the long-range temporal association of the deep reinforcement learning algorithm based on sensitive region selection to achieve the attack effect. The experimental results show that the constant perturbation based on time series has a good effect, i.e. inducing agents to make frequent wrong decisions and get minimal reward. At the same time, this algorithm still has an attacking effect on the defensively trained agents, and it effectively reduces the number of computations adversarial perturbations.

## CCS CONCEPTS

• **Computing methodologies** → *Value iteration*; **Intelligent agents**.

## KEYWORDS

Deep reinforcement learning, robustness of models, constant perturbation strategies, time series dependence.

## ACM Reference Format:

Jiamin Shen, LI XU, XU WAN, JIXUAN CHAI, and Chunlong Fan. 2023. Research on Constant Perturbation Strategy for Deep Reinforcement Learning. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590101>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590101>

## 1 INTRODUCTION

Since Mnih proposed Deep Q Network (DQN)[15, 16] based on Deep Neural Network (DNN), Deep Reinforcement Learning (DRL) has become an important aspect of artificial intelligence research. DRL is an end-to-end learning system that combines Reinforcement Learning(RL) techniques with the good feature extraction capabilities of deep learning to achieve end-to-end learning from raw data input to decision result output, and its performance has surpassed the human level in many complex applications, such as games[6, 21], robot control[1], and computer vision[8, 18], etc.

Although DNN models perform well in a variety of machine learning tasks, DNNs are very vulnerable to adversarial sample attacks, meaning that malicious input perturbations can cause DNN models to make false predictions. Goodfellow [9] et al. create adversarial samples to attack DNNs using fast gradient algorithms(FGSM) to cause the models to make false predictions labels. Therefore, there are also a number of security issues associated with intelligence trained using DNNs.

In the DRL system, the agent's decision-making ability is continuously improved by the training algorithms[10, 17, 20, 23, 25], but its defensive ability is not improved accordingly. Therefore, adversarial training [2, 4, 13] and robust learning[5, 24] have been performed on the agent, and finally the agent can effectively fight against some attacks. We believe that the agent after robust learning, already have safety bounds to defend against attacks, and in the context of the DRL with safety bounds, we want to verify the vulnerability of the safety bound algorithm from the time-dependent perspective of the agent.

The agent achieves the final victory by continuously making decisions for each step in each episode of the task, while obtaining information about the state of the environment through observations. We make the agent make the wrong decision and eventually fail the task by adding a constant perturbation to the agent's observations. We believe that adding the same perturbation to different regions of the observation will produce different degrees of perturbation to the agent, so the observation region with a large degree of perturbation is the sensitive region.

We explore whether the agent will continuously make deviations from the correct decision by continuously adding a constant perturbation to the observed sensitive region through the historical information of the agent's interaction with the environment. Adding perturbations to the sensitive region of observations means that only the observations of the agent are modified without changing

the real environment state, and this setting is consistent with many algorithms that perform adversarial attacks on observations [11, 14]. Our main contribution is the constant perturbation strategy, an algorithm that verifies the presence of insecurity in the agent. It arises from a simple attack that :

- 1) It is possible to obtain the sensitive area of the agent's observation from a small number of samples;
- 2) Only the agent's observation of the state is changed, without changing the agent's real environmental;
- 3) A constant perturbation only needs to be computed once, and no additional computation time is needed in the subsequent process, which can be used for real-time setup.

## 2 RELATED WORK

Usually, DRL is divided into value-based DRL and policy-based DRL. Value-based DRL algorithms include deep  $Q$  networks (DQN), double deep  $Q$  network (DDQN), prioritized experience replay  $Q$  networks (Prioritized DQN), and dyadic deep  $Q$  network (Dueling DQN), etc, which mainly approximate the target action value function by DNN, and the  $Q$  value derived from the state and action indicates the cumulative reward obtained by reaching a certain state. Because DQN suffers from excessive bias in  $Q$ -value estimation and training instability, DDQN, Prioritized DQN, and Dueling DQN are all improvements on the DQN method. DDQN indicates network selection actions and target networks for valuation; Prioritized DQN defines the priority level so that the agent prioritizes learning the experience with high level; Dueling DQN divides the  $Q$ -value into two parts to improve the learning efficiency, which are the state mechanism function  $V$  and the relative value function  $A$ .

Since value-based DRL cannot handle the stochastic policy problem, policy-based DRL methods have been investigated, including asynchronous advantage actor critic (A3C) and trust region policy optimization (TRPO). Policy-based DRL learns the probability of taking different actions corresponding to each state directly from the DNN, and thus searches for the optimal policy in the policy space.

DRL-oriented attacks include observation attacks, reward attacks, action attacks, environment attacks, and strategy attacks, which are mainly implemented on Atari game scenarios. In contrast, most of the DRL-oriented attack algorithms are based on observations, which is due to Goodfellow that image inputs in high-dimensional space are easy to attack. In the attack algorithm of Huang, which is similar to the fast gradient symbolic method (FSGM), a perturbation is added in the direction of the largest change in the gradient of the model. Although this method can achieve the attack effect, it does not take into account the high relevance of the reinforcement learning problem in continuous time and only considers the wrong predictions of the agent at each step. In addition, FGSM is a single-step attack method that requires generating a new adversarial sample at each time step, which is not feasible in practical applications.

In the policy induction attacks, Behzadan[3] used replica models and reward function to create adversarial samples to disrupt the training of the target model, but this attack method is still limited

to traditional machine learning in the form of computing the adversarial samples independently at each time step, while training replica models requires significant additional cost. Sun [22] built models to predict future states and corresponding actions. Then, the damage value of the attack strategy is evaluated by the damage awareness metric, and finally, the agent is induced to be in a bad state. Although this attack approach takes into account the continuous nature of DRL, it requires sufficient knowledge of the environment to define the distance metric of the attack state, and it is time consuming and difficult to train the prediction model.

At the same time, as the defensive capabilities of DRL continue to improve, some traditional attack methods can no longer achieve the effect of attacking an agent with security boundaries. Huan Zhang [26] proposed a framework of State-Adversarial Markov Decision Process (SA-MDP), which first computes the upper and lower bounds of perturbations for observations, then brings them into the training of the agent and adds theoretically principled robust policy regulariser, finally trains an agent capable of defending against white-box attack algorithms such as Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [19]. However, we believe that the DRL model with security bounds is still vulnerable in terms of time-dependent defence, and further research is needed.

## 3 METHOD

### 3.1 Problem definition

In RL, an agent learns optimal behaviour by interacting with its environment. The RL problem can usually be modelled as a Markov decision process (MDP). MDP is defined as  $(S, A, P, R, \gamma)$ , where  $S$  is the set of states,  $A$  is the set of actions,  $P : S \times A \times S \rightarrow [0, 1]$  is the state transition probability,  $R$  is the reward function after state transition, and  $\gamma \in (0, 1)$  is the discount factor that makes each reward value decay over time by a discount factor. At the beginning of each time step  $t$ , the agent observes the environment to obtain the current state  $s_t$  and performs an action  $a_t$  according to the current policy  $\pi$ . At the end of  $t$ , the agent receives its reward  $r_t$  and the next observed state  $s_{t+1}$ . The goal of MDP is to find the best sequence of actions to maximise the average reward in the long run.

The DQN algorithm is a typical value-based RL algorithm that mainly approximates the target action value function by the DNN, which represents the cumulative reward obtained by reaching a certain state. The agent chooses an action according to the state-action-value function ( $Q$  function), where the input of the  $Q$  function is a state and an action, and the output is a value that represents the expected value of the cumulative reward that the agent will receive for taking the action in the current state. The principle of decision making of the agent is as follows:

$$\pi(s) = \underset{a \in A}{\operatorname{argmax}} Q(s, a) \quad (1)$$

where  $\pi$  is the agent's policy,  $s$  is the current state,  $Q$  is the state action value function, and  $a$  is the action corresponding to the maximum value of the  $Q$  function.

In the DQN algorithm, the optimal action of the agent is the one corresponding to the maximum  $Q$  value, with respect to Eq. 1. In contrast, the constant perturbation strategy aims to induce the agent to achieve the minimum  $R$ . It adds a constant perturbation to

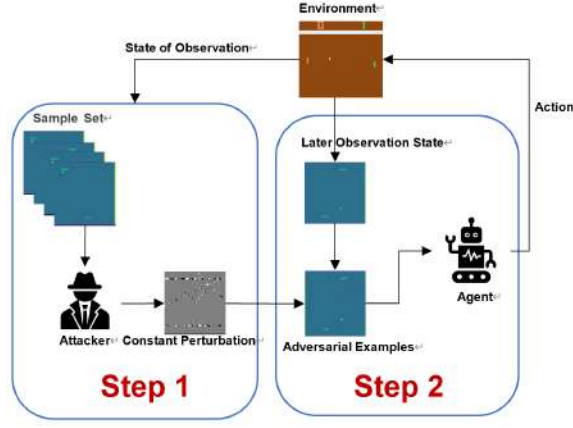


Figure 1: Flowchart of the constant perturbation strategy

all subsequent observations of the agent, which causes the agent to continuously make incorrect action commands. Therefore, in the DRL-oriented constant perturbation strategy attack, the mathematical expression for the adversarial sample ( $\hat{s}$ ) resulting from adding a constant perturbation ( $\delta$ ) to the original observation ( $s$ ) is as follows:

$$\hat{s} = s + \arg \min_{\delta} \|\delta\|_2 \quad \text{while} \quad \pi(\hat{s}) = \arg \min_{a \in A} Q(s, a) \quad (2)$$

where  $\|\cdot\|_2$  denotes the 2-norm of the perturbation, and the smaller it is, the more difficult it is to detect the perturbation with the naked eye. The adversarial sample consists of the original sample and the perturbation, and the agent is induced to choose the action with the smallest  $Q$  value.

### 3.2 Constant perturbation strategy

The main idea of the constant perturbation strategy is to first obtain the observation sensitive regions of the agent's in the whole episode by a small number of consecutive observations, then iteratively calculate the constant perturbation that can induce errors in the agent's decision. The observation sensitive area is the observation area that can produce the maximum perturbation to the agent when the same perturbation is applied to different areas of the agent's observation. The flowchart of the calculation and application of the constant perturbation is shown in Fig. 1 which mainly includes two steps. Step 1: Calculate the constant perturbation using the constant perturbation strategy; Step 2: Add the constant perturbation to the subsequent continuous observation and record the perturbation level of the constant perturbation to the agent.

**3.2.1 Sensitive area algorithm.** Due to the partially observable Markov decision process (POMDP), the tasks in DRL typically have incomplete observational information. For example, the Atari Pong game in the OpenAI Gym library [7] simulates two table tennis players and trains an agent to control a paddle to compete with another paddle controlled by the computer through RL. The agent's observation contains only the position of the paddle and the ball, but not the speed and forward direction of the ball. If the trajectory

of the ball is obtained based on the historical interaction information, then we can estimate the information about the speed as well as the forward direction of the ball, which is crucial to interfere with the agent's decision. Therefore, the sensitive region algorithm obtains the ball motion trajectories by fetching or manipulating the successive collected samples based on the high continuity and similarity of DRL observations, and if perturbations are added to these motion trajectories, the agent will be disturbed to a large extent.

The mathematical expression for finding the  $k$ th sample sensitive region is shown below, where the number of samples in the sample set is  $N$ :

$$c^{k+1} = c^k \mid s^{k+1}, k = 0, \dots, N-1 \quad \text{define} : c^0 = s^0 \quad (3)$$

Where  $c^k$  is the sensitive area calculated from  $k$  samples, the final key area is  $\tilde{s}$ , while  $\tilde{s} = c^N$ . From Eq.3 we can conclude that: the larger  $N$ , the more pixels the sensitive area contains, so the final interference success rate is higher.

**3.2.2 Constant perturbation algorithm.** The constant perturbation algorithm is a constant perturbation method calculated in the observed sensitive region that can continuously interfere with the agent's decision to make an error, which uses the gradient of the model to provide the direction to minimise the  $Q$  value and thus induce the agent to make the worst action command, so the objective function is defined as:

$$L(\hat{s}, \pi(s)) = Q_{\max}(\hat{s}, \pi(s)) - Q_{\min}(\hat{s}, \pi(s)) + c * \min \|\hat{s} - s\|_2 \quad (4)$$

The first part is  $Q_{\max}(\hat{s}, \pi(s)) - Q_{\min}(\hat{s}, \pi(s))$ , which means that the adversarial sample  $\hat{s}$  minimises the value of the optimal action and maximises the  $Q$  value of the worst action of the  $Q$  function, and the second part is  $\min \|\hat{s} - s\|_2$ , which means that the adversarial sample differs from the original sample by the smallest 2-norm of the perturbation, where  $c$  is defined as the hyperparameter. After our experiments we found that setting  $c$  to 0.0001 gave the best effect, so in this experiment we set  $c = 0.0001$ .

**Theory 1.** We assume that the decision surface around the collected original sample is linear. Under this assumption, the optimal perturbation of the input  $s$  is in the direction of maximising the loss function. If we restrict the adversarial perturbation to the region of infinite parametric module length  $L_\infty$ , the mathematical expression for  $\hat{s}$  is as follows:

$$\hat{s} = s + \nabla_{\tilde{s}} L(\tilde{s}, \pi(s)) \quad (5)$$

This theory is the theoretical basis for calculating a corresponding perturbation for each state, but the generalisability of the perturbation cannot be achieved. This means that the adversarial perturbation computed based on the current state does not necessarily have to attack the next state. In contrast, the constant perturbation is a perturbation involving several consecutive related samples, so the constant perturbation algorithm sets a perturbation constraint rate  $\beta$  to achieve the effect that the constant perturbation can play an attack on consecutive different samples. The mathematical expression for the perturbation constraint rate is:

$$\delta_i = \delta_i + \beta \cdot \delta_{i-1} \quad \text{define : } \delta_i = \nabla_{\hat{s}_i} L(\hat{s}_i, \pi(s)) \quad (6)$$

$\beta$  is the constraint rate of perturbation, and  $\delta_i$  is the gradient perturbation of the  $i$ th sample. The mathematical expression of the adversarial example is as follows:

$$\hat{s} = s + \xi \quad \text{define : } \xi = \delta_N \quad (7)$$

where  $\xi$  is the last constant perturbation,  $N$  is the number of samples in the sample set, and  $\delta_N$  is the adversarial perturbation of the last sample in the sample set.

**Theory 2.** In theory 1, the solution of the maximum perturbation with a linear loss function with only one backpropagation does not have a satisfactory attack. Generally, iterative attacks can achieve higher success rates than single-step attacks in a white-box setup. Also, the constant perturbation is not calculated from a single sample, so it is necessary to accumulate direction vectors in the gradient direction of successive original samples to calculate the perturbation iteratively. The mathematical expression is:

$$g_i^t = \frac{\delta_i^t}{\sqrt{\|\delta_i^t\|_2^2}} \quad (8)$$

where  $g_i^t$  denotes the unit direction vector of the  $i$ th sample in the sample set after the  $t$ th iteration, then the cumulative direction vector obtained after each episode of iteration is:

$$v^t = \frac{1}{M} * \sum_{i=1}^N \frac{1}{N} * (g_i^t + \beta \cdot g_{i-1}^t) \quad (9)$$

where  $v^t$  is the direction vector of the  $t$ th iteration of the sample,  $M$  is the total number of iterations,  $N$  is the total number of samples, and  $\beta$  is the constraint rate of the perturbation. Then, after each episode of iterations, the mathematical expression of the perturbation is obtained as follows:

$$v^t = v^t + v^{t-1} \quad (10)$$

**Theory 3.** In theory 2, as the number of iterations increases, some pixel values may overflow the specified range, so we limit the absolute value of the maximum perturbation (infinite parametric modulus length  $L_\infty$ ) in the constant perturbation to no more than  $\epsilon$ . The smaller it is, the harder it is to detect the perturbation with the naked eye. For the part of the gradient larger than the threshold, we use the projected gradient descent method, which projects the adversarial samples into clean samples after each update iteration, and specifies a perturbation step size smaller than  $\epsilon$ ; For the part of the gradient smaller than the threshold, we make the perturbation as large as possible. Therefore, we choose a direct symbolic function that limits the value of the gradient by defining the perturbation step. Then, according to Eq. 6,  $\delta$  can be expressed as:

$$\delta_i = \alpha \cdot \text{sign}(\nabla_{\hat{s}_i} L(\hat{s}_i, \pi(s_i))) \quad (11)$$

For example, the mathematical expression for the  $k$ th iteration of the update for the calculation of the adversarial perturbation is:

$$\hat{s} = \prod_{s+\epsilon} \{s + v^k\}, v^k = 0, \dots, M-1 \quad (12)$$

where  $s$  is the original sample,  $\hat{s}$  is the adversarial sample of  $s$ ,  $\prod_{s+\epsilon}\{\cdot\}$  constrains the adversarial perturbation to be within the  $L_\infty$  sphere, and  $v^k$  is the adversarial perturbation after the  $k$ th iteration.

The process of the DRL-oriented constant perturbation strategy algorithm is clarified by the description of the theoretical algorithm above. A concrete description is given in Alg. 1 below, where  $\tilde{s}$  represents the recalculation of the observed sensitive region in each episode of the algorithm;  $v^e$  is the constant perturbation of the sensitive region in each episode.

---

**Algorithm 1:** constant perturbation strategy algorithm

---

**input :** the  $Q$  function of policy  $\pi$ ,  $\alpha$ : perturbation step size,  $\beta$ : perturbation constraint rate,  $o$ : success rate threshold.

**output:** reward  $\pm$  std: reward and standard deviation of the average of 50 episodes, Frame, SR and  $v$

```

1 Initialization Seed  $\leftarrow$  2022, Episodes  $\leftarrow$  50;
  // Set the environment random seed to 2022 and
  the number of episodes to 50
2 Run the environment;
3 for  $e \leftarrow 1$  to Episodes do
4   Initialization  $\tilde{s} \leftarrow 0, \delta \leftarrow 0, v \leftarrow 0$ ;
5    $M(s) = \{s_0, \dots, s_N\}$  // Collection of sample sets.
6   ;
7    $\tilde{s} = s_0 | s_1 | \dots | s_N$ ;
8   for  $i \leftarrow 1$  to  $M$  do
9     for  $j \leftarrow 1$  to  $N$  do
10       $\delta_j^i = \alpha \cdot \text{sign}(\nabla_{\tilde{s}_j^i} L(\tilde{s}_j^i, \pi(s_j^i)))$ ;
11       $g_j^i = \frac{\delta_j^i}{\sqrt{\|\delta_j^i\|_2^2}}$ ;
12       $v^i = \frac{1}{M} \cdot (g_j^i + \beta \cdot g_{j-1}^i)$ ;
13    end
14     $v^i = v^i + v^{i-1}$ ;
15     $v^i \leftarrow \text{proj}(v^i)$  //  $\text{proj}()$  is to limit the  $v^i$  to
      the sphere.
16    ;
17    if  $h(v^i) > o$  then
18      //  $h()$  is the interference success rate
19      function
20       $v^e = v^i$ ;
21      BREAK;
22    end
23     $v^e = v^i$ ;
24  end
25   $\hat{s} = s + v^e$ ;
26   $a = \arg\min_{a \in A} Q(\hat{s}, \pi(s))$ ; SR =  $h(v^e)$ ;
27 end
```

---

The method tries to find the perturbation that causes  $\pi(\hat{s})$  to be the worst action, causing the agent to make a wrong decision in all observations. Using the above algorithm, the observations received by the agent of the three models in the experiment are

shown in Fig. 2 below. In the figure, the first column of the figure represents the adversarial sample, the second column represents the original sample, and the third column represents the constant perturbation. Each row represents the agent models trained by the different algorithms described in Section 4.1.

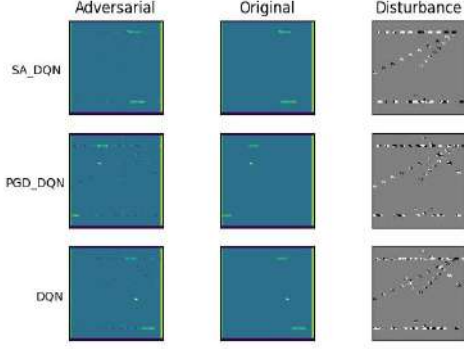


Figure 2: Constant perturbation plots for SA-DQN, PGD-DQN, and DQN models

## 4 EXPERIMENTAL ANALYSIS

### 4.1 Experimental setup

We now test our constant perturbation strategy on the agent of the Pong game in the OpenAI Gym library, using the three agent models from paper [26]:

**DQN model:** Using Double DQN and the prioritised experience replay algorithm and 6 million time steps to train the agent, it was run for 15 hours on a computer configured with a 1080Ti GPU, without robustness training of the agent.

**PGD-DQN model:** Based on the DQN model, the PGD attack algorithm was used for counter training to improve the model’s defence against PGD attacks.

**SA-DQN model:** Using Double DQN and the prioritised experience replay algorithm and training 6 million time steps on the agent while applying robust learning to the agent using the SA-DQN framework, it was run for 50 hours on a computer configured with a 1080Ti GPU.

The network model has the following structure:

The deep Q networks have 3 CNN layers followed by 2 fully connected layers. The first CNN layer has 32 channels, a kernel size of 8, and stride 4. The second CNN layer has 64 channels, a kernel size of 4, and stride 2. The third CNN layer has 64 channels, a kernel size of 3, and stride 1. The fully connected layers have 512 hidden neurons for both value and advantage heads.

Our experiments were performed on a workstation equipped with an NVIDIA GeForce RTX2080Ti GPU, using a Python programming environment and the Pytorch framework, with python version 11.5.

Table 1: The natural  $\text{Reward} \pm \text{std}$  and Frames of the three models

	DQN	PGD-DQN	SA-DQN
Reward $\pm$ std	21.0 $\pm$ 0.0	21.0 $\pm$ 0.0	21.0 $\pm$ 0.0
Frame	1688.80	1637.46	1634.26

### 4.2 Evaluation indicators

In this paper, the experimental benchmark [12] uses the average reward value over 50 episodes, the standard deviation of the reward, the attack success rate (SR), the  $L_2$ -norm and the  $T$  to evaluate the effectiveness of the constant perturbation and thus the robust security of the model:

- The average reward evaluation mechanism is to calculate the average episode reward of the target agent after several episodes of running the target agent. It is mainly used to evaluate the effect of a constant perturbation on the overall performance of the model.
- The episode reward standard deviation evaluation mechanism is the standard deviation of the reward calculated by the target agent after a number of episodes, which is mainly used to evaluate the stability of the influence of the constant perturbation strategy on the agent. The larger the standard deviation, the more unstable the constant perturbation strategy is.
- SR’s evaluation mechanism is the ratio of the number of times a constant perturbation can achieve the effect of an attack to the total number of attacks during operation. Therefore SR represents the average attack success rate of the counter perturbation, which is mainly used to detect and evaluate the effectiveness of constant perturbations.
- The evaluation mechanism of  $L_2$  refers to the concealment of constant perturbations.  $L_2$  represents the average 2-norm of the perturbation. The smaller  $L_2$  is, the more difficult it is to detect the perturbations with the naked eye.
- The evaluation mechanism of  $T$  refers to represent the average number of calculations perturbations.

### 4.3 Experimental results

To verify the effectiveness of the strategy in this paper, the results of 50 episodes for each experiment were taken for analysis. The experimental benchmarks (natural rewards), which do not interfere with the agent’s observations, are shown in Table 1, where **Reward $\pm$ std** in the rows denotes the mean reward value and standard deviation for 50 episodes; the number of **Frames** denotes the mean number of observations for 50 episodes; and these columns denote the three different models.

The results of the comparison experiments with the traditional single-sample attacks are shown in Table 2, which contains three different target models. We compared the rewards under three attacks, where  $T$  represents the average number of calculations perturbations. The results show that the traditional iteration-based single-sample attack method **PGD** plays a strong role in the DQN model, but has no attack effect in the PGD-DQN and SA-DQN models. Constant perturbation as an iterative perturbation achieves

a better perturbation effect in all three models. **cw** is currently a stronger single-sample perturbation method, but is more computationally intensive. The perturbation must be computed for each observation, while the constant perturbation only needs to be computed once to achieve a good perturbation effect, which greatly reduces the computational effort.

**Table 2: Comparison of the attack effect with PGD, CW and constant perturbation under the three models**

Model		PGD	CW	Ours
DQN	Reward±std	-20.8±0.37	-20.98±0.14	<b>-18.2±1.17</b>
	Frames	895.10	966.56	1081.60
	$L_2$	0.231	1.942	1.690
	SR	0.999	0.999	0.477
	$T$	895.10	966.56	<b>1</b>
PGD-DQN	Reward±std	21.0±0.0	-20.68±0.47	<b>-18.98±0.14</b>
	Frames	1637.46	794.78	918.3
	$L_2$	0.235	1.475	2.438
	SR	0.075	0.999	0.235
	$T$	1637.46	794.78	<b>1</b>
SA-DQN	Reward±std	21.0±0.0	-21.0±0.0	<b>-20.0±0.0</b>
	Frames	1634.26	765.66	839.64
	$L_2$	0.217	1.751	0.920
	SR	0.0	0.856	0.587
	$T$	1634.26	765.66	<b>1</b>

The results of the strategic time attack comparison experiments are shown in Table ?? . We compare the rewards under the two attack methods, and the table contains two different target models. From Table ?? , we can see that the strategy time attack is better for the DQN model, but not for the PGD–DQN model. We apply constant perturbation to the same strategy and the effect is significantly better, confirming that constant perturbation is easier to find the sensitive and vulnerable points of the model.

**Table 3: Comparison of the effect of the attack with the strategy time under the same strategy and the same models**

Model		Strategy Time	Ours
DQN	Reward±std	-17.58±2.41	<b>-18.9±4.10</b>
	Frames	1333.24	1000.36
	$L_2$	1.862	1.454
	SR	0.787	0.639
	$T$	19.78	<b>1</b>
PGD-DQN	Reward±std	0.34±13.23	<b>-19.82±0.38</b>
	Frames	2052.0	852.7
	$L_2$	2.579	1.401
	SR	0.681	0.324
	$T$	6.1	<b>1</b>

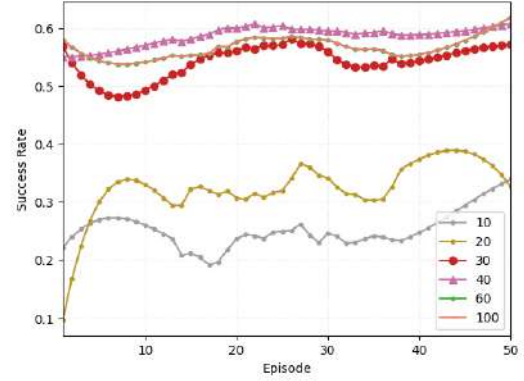
The constant perturbation strategy generalizes well to different models by comparing experiments. This phenomenon can be explained by the fact that if the agent is independent of its training and defense algorithms and learns continuous dependent decision

patterns, then adding constant perturbations to the observation is likely to disrupt the agent’s continuous decision making.

We also get a conclusion: The stability and accuracy of the constant perturbation depends largely on the quality of the  $Q$  function. If  $Q$  is poorly learned or has confusing gradients, it will give the wrong gradient direction and lead to the failure of the constant perturbation.

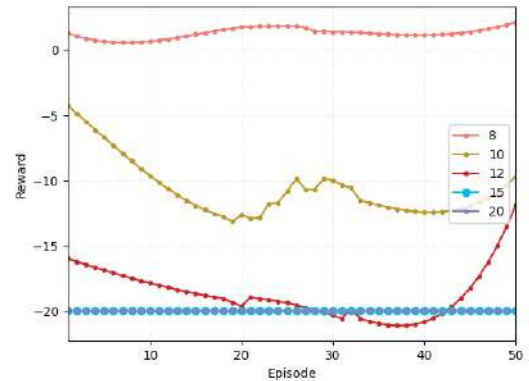
#### 4.4 Key hyperparameters

The algorithm in this paper involves several hyperparameters, and the values of these will have a certain impact on the experiments. Therefore, experiments are conducted on the hyperparameters, and the results of each experiment are taken for 50 episodes. We define  $e=1/255$ ,  $d=e/20$ .



**Figure 3: The success rate of a constant perturbation calculated with different numbers of samples.**

**4.4.1 The effect of  $N$  on the algorithm.** The hyperparameter  $N$  is the number of samples in the sample set, searching on more sample data can improve the SR of a constant perturbation. The



**Figure 4: Reward values after attacks with a constant perturbation, calculated for different  $\epsilon$  values.**

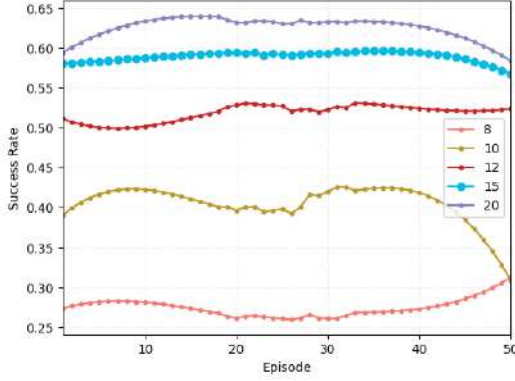


Figure 5: SR of attacking a constant perturbation calculated using different  $\epsilon$ .

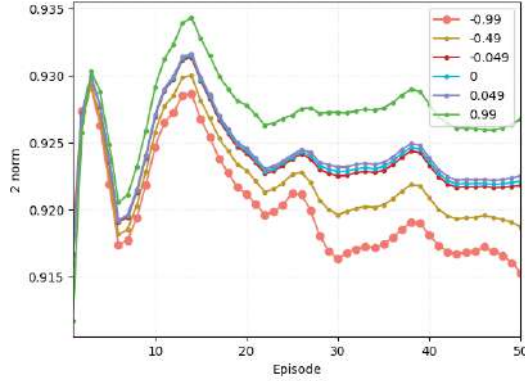


Figure 6: The  $L_2$  of constant perturbation calculated by different value of  $\beta$ .

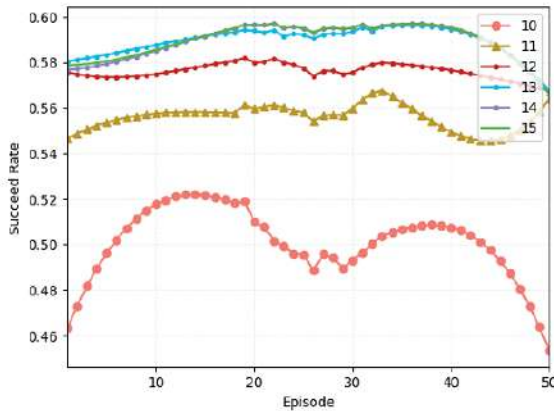


Figure 7: The SR of constant perturbation calculated by different  $\alpha$ .

larger parameter  $N$  is, the more sensitive pixels can be searched. The experimental result is shown in Fig.3, which shows the results of SR after the constant perturbation on the SA-DQN model. We set the  $\epsilon=15*e, \alpha=13*d, \beta=-0.099$  as a benchmark for experimental analysis.

The experiment shows that when  $N \in \{10, 20\}$ , the average SR is as  $\{0.25, 0.32\}$  and the average **reward±std** is as  $\{-9.54 \pm 11.4, -8.4 \pm 10.8\}$ . The effect of constant perturbation is obviously unstable. When  $N \in \{30, 40, 60, 100\}$ , the value of average SR is between 0.54 and 0.59, and the value of average **reward±std** is around  $-20 \pm 0.0$ . It proves that  $N$  has an important influence on the observation sensitive area. As  $N$  increases, the interference effect also increases.

**4.4.2 The influence of  $\epsilon$  on algorithm.** The hyperparameter  $\epsilon$  is the upper bound of the perturbation size, and searching over larger perturbations can improve the interference success rate of constant perturbations. We conducted a comparison experiment for the value of  $\epsilon$ , setting  $\epsilon = x * e$ , so the line markers in the figure indicate the value of  $x$ . The experimental results are shown in Fig.4 and Fig.5, which we use  $N=100, \alpha=13*d, \beta=-0.099$  as the benchmark for the experiment.

The experiment shows that when  $x \in \{8, 12\}$ , the average **reward±std** of 50 episodes corresponds to  $\{1.32 \pm 2.0, -10.64 \pm 10.3\}$ , and the SR is  $\{0.27, 0.40\}$ ; when  $x \in \{15, 20\}$ , the average **reward±std** is basically  $-20 \pm 0.0$ , and the average of SR is above 0.58, indicating that the constant perturbation at this time the effect of constant perturbation is strong. Overall, the experiments show that the size of  $\epsilon$  has an important effect on the constant perturbation, and the attack success rate increases as  $\epsilon$  increases.

**4.4.3 The influence of  $\beta$  on the algorithm.** We conducted comparative experiments on the value of the hyperparameter  $\beta$ , which is the constraint rate between neighboring perturbations. The experimental result is shown in Fig.6. We set the  $N=100, \epsilon=15*e, \alpha=13*d$  as a benchmark for experimental analysis.

According to the experiment, when  $\gamma \in \{-0.99, 0.99\}$ , the average **reward±std** in 50 episodes is basically  $-20 \pm 0.0$ . Fig. 6 shows that the use of constraint rates leads to smaller  $L_2$  to achieve a uniform effect, proving once again that the constant perturbation strategy is robust, efficient, and does not easily fall into local optima.

**4.4.4 The influence of  $\alpha$  on the algorithm.** The hyperparameter  $\alpha$  is the perturbation step size of each iteration. We compared the experimental values of  $\alpha$  and set  $\alpha = y * d$ , so the line markers in the figure indicate the values of  $y$ .

It can be seen from Fig.7 that when the perturbation step size is not appropriate, the induction effect may be unstable and in some cases the aggressiveness is not strong. We set the  $N=100, \epsilon=15*e, \gamma=-0.099$  as a benchmark for experimental analysis. Also, the experimental results show that SR takes higher value when  $y=13$ , average **reward±std** is  $-20.0 \pm 0.0$  and  $L_2=0.92$  is also relatively optimal.

## 5 CONCLUSION

The constant perturbation strategy in this paper differs from traditional attack algorithms in that it relies on the dependence of the DRL temporal order to verify whether an agent is safe and secure, and it can also verify an agent with security bounds. Compared

to traditional single-sample attack methods, the constant perturbation strategy is less computationally intensive and less costly than generative adversarial networks and attack methods based on imitation learning to construct alternative models. However, the method proposed in this paper has one drawback, the attack efficiency still needs to be improved, which can be used as the direction of improvement in the future.

## ACKNOWLEDGMENTS

In the completion of this paper, thanks to Teacher Fan's theoretical support and Chai Jixuan's analysis of experimental data. And this work was supported in part by the National Natural Science Foundation of China (61902260, 61972266).

## REFERENCES

- [1] Smruti Amarjyoti. 2017. Deep Reinforcement Learning for Robotic Manipulation - The state of the art. *CoRR* abs/1701.08878 (2017). arXiv:1701.08878 <http://arxiv.org/abs/1701.08878>
- [2] Vahid Behzadan and William H. Hsu. 2019. Analysis and Improvement of Adversarial Training in DQN Agents With Adversarially-Guided Exploration (AGE). *CoRR* abs/1906.01119 (2019). arXiv:1906.01119 <http://arxiv.org/abs/1906.01119>
- [3] Vahid Behzadan and Arslan Munir. 2017. Vulnerability of Deep Reinforcement Learning to Policy Induction Attacks. In *Machine Learning and Data Mining in Pattern Recognition - 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings (Lecture Notes in Computer Science, Vol. 10358)*, Petra Perner (Ed.). Springer, 262–275. [https://doi.org/10.1007/978-3-319-62416-7\\_19](https://doi.org/10.1007/978-3-319-62416-7_19)
- [4] Vahid Behzadan and Arslan Munir. 2017. Whatever Does Not Kill Deep Reinforcement Learning, Makes It Stronger. *CoRR* abs/1712.09344 (2017). arXiv:1712.09344 <http://arxiv.org/abs/1712.09344>
- [5] Vahid Behzadan and Arslan Munir. 2018. Mitigation of Policy Manipulation Attacks on Deep Q-Networks with Parameter-Space Noise. In *Computer Safety, Reliability, and Security - SAFECOMP 2018 Workshops, ASSURE, DECSOs, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11094)*, Barbara Gallina, Amund Skavhaug, Erwin Schoitsch, and Friedemann Bitsch (Eds.). Springer, 406–417. [https://doi.org/10.1007/978-3-319-99229-7\\_34](https://doi.org/10.1007/978-3-319-99229-7_34)
- [6] Christopher Berner, Greg Brockman, Brooke Chan, and Vicki Cheung. 2019. Dota 2 with Large Scale Deep Reinforcement Learning. *CoRR* abs/1912.06680 (2019). arXiv:1912.06680 <http://arxiv.org/abs/1912.06680>
- [7] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *CoRR* abs/1606.01540 (2016). arXiv:1606.01540 <http://arxiv.org/abs/1606.01540>
- [8] Juan C. Caicedo and Svetlana Lazebnik. 2015. Active Object Localization with Deep Reinforcement Learning. *CoRR* abs/1511.06015 (2015). arXiv:1511.06015 <http://arxiv.org/abs/1511.06015>
- [9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). <http://arxiv.org/abs/1412.6572>
- [10] Dan Horgan, John Quan, David Budden, Gabriel Barth-Maron, Matteo Hessel, Hado van Hasselt, and David Silver. 2018. Distributed Prioritized Experience Replay. *CoRR* abs/1803.00933 (2018). arXiv:1803.00933 <http://arxiv.org/abs/1803.00933>
- [11] Sandy H. Huang, Nicolas Papernot, Ian J. Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial Attacks on Neural Network Policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=ryv1RyBKL>
- [12] Chen Jinyin, Zhang Yan, Wang Xueke, Chai Hongbin, Wang Yu, and Ji Shoulin. 2022. A survey of attack defense and related security analysis for deep reinforcement learning. *Acta Automatica Sinica* 48, 1 (2022), 21–39. <http://doi.org/10.16383/j.aas.c200166>
- [13] Jernej Kos and Dawn Song. 2017. Delving into adversarial attacks on deep policies. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=BJcib5mFe>
- [14] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. 2017. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=r1Cy5yrKx>
- [15] Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. 2015. Human-level control through deep reinforcement learning. *Nat.* 518, 7540 (2015), 529–533. <https://doi.org/10.1038/nature14236>
- [16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *CoRR* abs/1312.5602 (2013).
- [17] Volodymyr Mnih, Adri Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous Methods for Deep Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings, Vol. 48)*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.). JMLR.org, 1928–1937. <http://proceedings.mlr.press/v48/mniha16.html>
- [18] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L. Lewis, and Satinder Singh. 2015. Action-Conditional Video Prediction using Deep Networks in Atari Games. *CoRR* abs/1507.08750 (2015). arXiv:1507.08750 <http://arxiv.org/abs/1507.08750>
- [19] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *CoRR* abs/1609.04747 (2016). arXiv:1609.04747 <http://arxiv.org/abs/1609.04747>
- [20] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2015. Trust Region Policy Optimization. *CoRR* abs/1502.05477 (2015). arXiv:1502.05477 <http://arxiv.org/abs/1502.05477>
- [21] David Silver, Aja Huang, Chris J. Maddison, and Arthur Guez. 2016. Mastering the game of Go with deep neural networks and tree search. *Nat.* 529, 7587 (2016), 484–489. <https://doi.org/10.1038/nature16961>
- [22] Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. 2020. Stealthy and Efficient Adversarial Attacks against Deep Reinforcement Learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 5883–5891. <https://ojs.aaai.org/index.php/AAAI/article/view/6047>
- [23] Hado van Hasselt, Arthur Guez, and David Silver. 2015. Deep Reinforcement Learning with Double Q-learning. *CoRR* abs/1509.06461 (2015). arXiv:1509.06461 <http://arxiv.org/abs/1509.06461>
- [24] Jingkan Wang, Yang Liu, and Bo Li. 2020. Reinforcement Learning with Perturbed Rewards. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 6202–6209. <https://ojs.aaai.org/index.php/AAAI/article/view/6086>
- [25] Ziyu Wang, Nando de Freitas, and Marc Lanctot. 2015. Dueling Network Architectures for Deep Reinforcement Learning. *CoRR* abs/1511.06581 (2015). arXiv:1511.06581 <http://arxiv.org/abs/1511.06581>
- [26] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane S. Boning, and Cho-Jui Hsieh. 2020. Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/f0eb6568ea114ba6e293f903c34d7488-Abstract.html>

# Twitter stance detection using deep learning model with FastText Embedding

Yongqing Deng

A graduate student in the School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, China  
dengyq\_397@163.com

Yongzhong Huang\*

A Full Professor with the School of Computer and Information Security, Guilin University of Electronic Technology, Guilin, China  
2389483289@qq.com

## ABSTRACT

The interactivity of social media platforms allows a large number of users to comment on different political or social issues to express their views, and identifying users' stances from online comment texts helps the government to monitor public opinion more effectively. The automatic recognition of stance information in comment text has become a new research hotspot in the field of natural language processing. Most of the existing text stance analysis corpus focuses on political topics in European and American countries, and high-quality stance analysis corpus research on political topics in Southeast Asian countries is relatively scarce. In order to stimulate this research direction, this paper provides a dataset about the 2022 Philippine presidential election, which annotates the stance information of the two popular presidential candidates and provides reliable data support for subsequent stance analysis model research. Next, we build a stance detection model of hybrid deep neural networks based on BiLSTM, CNN, and Attention, and we demonstrate its effectiveness on multiple datasets and obtain the best results on the SemEval-2016 dataset. In addition, we compare FastText and Word2Vec, two pre-trained word embeddings for word encoding, and discuss which word embedding is preferred in stance detection tasks. This result shows that the stance analysis model proposed in this paper can be effectively applied to Twitter text stance data.

## CCS CONCEPTS

• Computing methodologies; • Natural language processing; • Information systems; • Web and social media search; Sentiment analysis; • Machine learning; Language resources;

## KEYWORDS

Stance Detection, Deep Learning, Text Analysis, Natural Language Processing, Word Embeddings

### ACM Reference Format:

Yongqing Deng and Yongzhong Huang. 2023. Twitter stance detection using deep learning model with FastText Embedding. In *2023 2nd Asia Conference*

\*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590102>

on *Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590102>

## 1 INTRODUCTION

Nowadays, social media platforms have evolved into an essential part of personal social interactions. These social platforms are powerful tools for disseminating information; they allow users to have conversations, express opinions, share information, and exchange perspectives. At the same time, people rely on these platforms as their primary sources of news downtime in order to stay connected to the outside world and get real-time updates [1]. Users use these tools as a primary source of communication, and the tremendous dependence allows researchers to study different human behaviors through online means, such as public stances on society and politics.

Stance detection is an auto-classification process, which is interested in whether the author is likely to be opposed, favorable, or neutral to a claim or target in the text [2]. Stance detection techniques are especially important for understanding public perceptions of specific targets, often with the aim of determining how social media users feel broadly about certain controversial or specific areas, like candidates during elections.

Meanwhile, Stance detection is also a particularly difficult issue on Twitter. A great deal of this difficulty stems from the fact that Twitter content is brief, dynamic, constantly generates new hashtags and abbreviations, and deviates from the standard prose sentence structure [3]. In addition, implicit, explicit, ironic, metaphorical, and indeterminate expressions may bring a challenge to detecting stances because opinions are not always explicitly expressed [4].

Existing stance detection studies mainly adopt supervised learning methods, which require the use of predefined sets of tags to annotate the stance corpus. In recent several years, deep learning algorithms have been employed to research stance detection issues as artificial intelligence has grown more advanced. The work of Elfardy and Diab [5] used SVM models and stance classification in SemEval-2016 with lexical and semantic features with an F1 score that equaled 63.6%. A separate work by Wojatzki and Zesch [6] used stacked classifiers and syntactic features to classify stances in the SemEval-2016 stance dataset. With an F1 score of 62%, their model exhibits a marginal improvement in overall stance detection performance. Ghosh et al [7] summarize the performance of multiple supervised ML algorithms on the SemEval-2016 dataset, but only for NLP methods.

**Table 1: Sample of stance examples presented to labelers.**

Candidate	Statement	Stance
Bongbong	Let’s have a survey everyone is invited! Me first! My President is Bongbong Marcos Jr!	Favor
	We’ll never say, reverend chief robber of the country Bongbong Marcos! That’s not acceptable man! Let’s all use our brain cells!	Against
Leni	Can Bongbong Marcos make a difference?	Neutral
	Leni is our President and we are one PINK happy family!	Favor
	What do you expect to pinklawan? Their heart is full of hatred and their brain like Leni Robredo.	Against
	Leni’s ending statement gave me goosebumps.	Neutral

Moreover, the global vectors GloVe [8] and Word2Vec [9] are widely applied to catch the syntactic information and semantic relationships of word embeddings in numerous deep-learning models. These embedding models reveal valuable deep-learning methods for building word vector representation in terms of stance detection and are highly appreciated by researchers. These models were used in [10–13]. However, these representations are limited in solving certain problems, for example, it is not suitable for small corpora and requires large corpora for training and representing vector recognition of each word, as mentioned in [14]. Likewise, GloVe and Word2Vec, which were suggested in [15, 16], are not always the best options for researchers working with tiny datasets because of their time complexity, computational expense, and effectiveness.

Although various models with complex architecture have emerged as deep learning and pre-training have become increasingly popular, the computational issues that accompany complex architecture models cause them to take longer to train than simple models. Therefore, the aspiration of utilizing the stance in real-time applications is to use simpler model while balancing the performance of the model. This study intends to use another simple model architecture to deal with the stance detection problem using deep learning and word embedding. A novel stance detection model is applied in this study, which is based on the combination of Bi-LSTM and CNN deep neural network regulated by attention level. This study also compares the performance of the model when word2vec and FastText are used as different word embeddings, and the model is evaluated on three datasets.

## 2 DATASETS

In this section, we first introduce a dataset we collected for the 2022 Philippine presidential election, and provide a detailed description of the data source, collection method, labeling process, and distribution ratio of the labeled data. After that, we introduce the steps of data preprocessing and two well-known datasets of stance detection, which will be used for the experiments.

### 2.1 Data collection

For the study, we collected tweets about the 2022 Philippine presidential election. We collected nearly 800,000 tweets from May 1, 2021, to May 15, 2022. Our specific stance task is to identify the stances of two presidential candidates, Bongbong Marcos and Leni Robredo. For each candidate, we have three stance categories: favor, against, and none. We labeled two datasets with stance labels

individually for each of the candidates, Bongbong and Leni. Our data were labeled by three annotators who had no prior training. Therefore, we provide a group of examples for each stance class that they can refer to when implementing labeling tasks. Table 1 shows examples of the statements provided to the annotators.

To improve annotation yield, we require that tweets in the two datasets are mutually exclusive. Each tweet was labeled by three annotators, and the true label followed the majority principle. If each of the three annotators votes for three different classes, then we assume that the tweets are labeled neutrally because the stance is ambiguous. Our dataset contains a total of 2500 tweets with stance labels, 1250 for each candidate. The grammar and syntactic structure were preserved because we only used tweets with hashtags. The distributions of stance labels are shown in Table 2. with a roughly balanced distribution between the two candidates. Overall, the percentage of stance classes ranged from 29% to 37%.

### 2.2 Data preprocessing

Twitter text often contains informal text such as emoticons, URLs, usernames, and Twitter-specific hashtags. These informal texts present a formidable challenge to the task of stance detection. To eliminate unnecessary noise interference from experiments, this article pre-processed the Twitter text in advance:

- Remove non-text portions of the data, like punctuation marks and special characters;
- Converts full-width English characters from raw data to half-width English characters and performs word segmentation;
- Replace duplicates of multiple characters with a single character;
- Lemmatizes text and converts it to lowercase;
- Remove short links and @ tags from Twitter text.

### 2.3 SemEval-2016 Stance Dataset

In the supervised task, more than 4,000 tweets were labeled and categorized into five topics: “Atheism”, “Feminist Movement”, “Hillary Clinton”, “Climate Change is a Real Concern”, and “Legalization of Abortion”; in addition, the labeled tweets were sorted by the timestamp of each target, with a ratio of 8:2 between the training set and testing set [17]. Table 3 shows the distribution of training and testing sets for the five subjects and table 4 exemplifies some examples of tweets from the SemEval-2016 dataset.

**Table 2: Stance distribution for Bongbong and Leni.**

	% Favor	% Against	% None
Bongbong	36.32	29.44	34.24
Leni	34.56	33.52	32.00

**Table 3: Multiple examples of each target in the SemEval-2016 English dataset.**

Target	Train	Test	Total
Atheism	513	220	733
Climate Change is a Real Concern	395	169	564
Feminist Movement	664	285	949
Hillary Clinton	689	295	934
Legalization of Abortion	953	280	883
Total	2914	1249	4163

**Table 4: Examples of tweets from the SemEval-2016 dataset.**

Target	Statement	Stance
Feminist Movement	@PH4NT4M @MarcusChoOo @CheyenneWYN women. The term is women. Misogynist! #SemST	Favor
Atheism	American conservatism has everything to do with religion with all the good stuff taking out of it. #SemST	Against
Legalization of Abortion	Every time you respond to something that frustrates you, you let it steal away your time and happiness. #EasyWeightLoss #SemST	Neutral

**Table 5: Stance distribution for Biden and Trump.**

	% Favor	% Against	% None
Biden	31.3	39.0	29.8
Trump	27.3	39.9	32.8

## 2.4 US Presidential election 2020

The dataset is the first public Twitter dataset on the 2020 US Presidential Election. It has focused on the presidential and vice presidential candidates [3]. The release of this dataset is curated, documented, and updated weekly until the November 3, 2020 election and beyond. In addition, Kawintiranon and Singh constructed a dataset including the two presidential candidates, Joe Biden and Donald Trump. There are 2500 manually labeled tweets in total, with 1250 for each presidential candidate. The stance label distributions are shown in Table 5. Table 6 presents examples of statements submitted to MTurk workers.

## 3 MODEL AND EXPERIENCES

This section describes in detail the model architecture, experimental setup, and evaluation metrics proposed in this paper.

### 3.1 Model architecture

In this paper, a stance analysis model based on hybrid deep neural network FTEmb-BCA model is proposed (abbreviated as FBCA), this

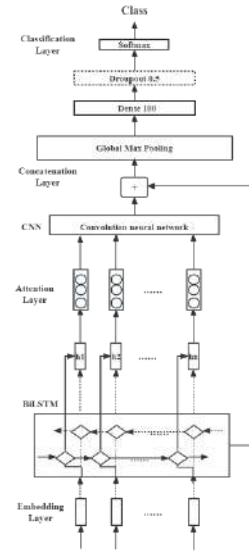
model adopts the idea of ensemble learning, and fuses the bidirectional long short-term memory network model (BiLSTM) and the convolutional neural network model (CNN). First, the word embedding model is used to convert each word of a sentence into a word vector as input to the BiLSTM model. The BiLSTM model can make full use of the context information to find the hidden vector for each word vector of the sentence, which contains the semantic information of the words in the sentence. Then, a simple and effective attention layer is introduced after the BiLSTM model to capture the long-distance association between words with different weights, so that the model can increase the weight of text keywords, better extract keywords, and improve model performance. After keyword weighting by the attention mechanism, it enters the CNN module, and uses multiple filters of different sizes to perform convolution on the input word vectors to achieve local feature extraction. The CNN layer further fuses and refines useful information by convoluting and maximizing pooling features, to help the model obtain better accuracy and achieve the purpose of stance classification. The complete stack of the proposed model for stance detection is shown

**Table 6: Examples of tweets from the US Presidential election 2020 dataset.**

Candidate	Statement	Stance
Biden	Biden will be a great president. I am voting for him in November.	Favor
	Biden has handled the pandemic poorly.	Against
	Biden spoke in Pennsylvania.	Neutral
Trump	Trump has been a great president. I am voting for him in November.	Favor
	Trump has handled the pandemic poorly.	Against
	Trump held a rally yesterday.	Neutral

in Figure 1. The component structure of the model is described in detail below.

- **Embedding layer:** map words to low-dimensional vectors. The same characters may have different meanings in different words and characters that are out of the semantic context of words may lead to classification errors. Therefore, we first preprocess the dataset and then classify the dataset to train the word embedding model. To get the neural network to compute them, we use the skip-gram algorithm in word2vec and a 300-dimensional FastText embedding to represent the input text.
- **BiLSTM layer:** use the BiLSTM network to obtain a deeper semantic vector representation of each word. This paper adopts an improved bi-directional long short-term memory network (BiLSTM) based on LSTM, that is, reading text data from front to back and back to front at the same time, which can further enhance the dependence of semantics on context. The hidden cell value is set to 100, and the dropout value is set to 0.5. We change the activation function used by the network and set it to a hyperbolic tangent function (tanh). The output of the layer is more centrally located at 0 thanks to this activation function’s S-shaped output, which ranges in value from -1 to 1. Furthermore, it generates a gradient bigger than the sigmoid function, which helps to hasten convergence.
- **Attention layer:** calculate the attention weight for each word. A layer of self-attention was added after BiLSTM, which is similar to the attention method suggested in [19]. The stance tendency of the sentence is not only related to the contextual information, but also highly relevant to the stance statement. Given a sentence, not all contextual words contribute equally to the semantics of the sentence. The attention mechanism is introduced in text stance detection to automatically learn the weight distribution, and assign different weight values to text features based on importance, which helps the classifier complete faster and more accurate classification.
- **CNN layer:** enter word vectors into the CNN network for operations such as convolution to extract hidden features. In our research, we apply the CNN layer to the results of the attention algorithm. This hidden level has a matrix form since the word embedding provides a vector representation over the tokens in the input. And more specifically, its precise shape is 80x400, allowing us to use a 1D Convolutional network with 400 filters and a 5x5 kernel. We use ReLU as the

**Figure 1: The structure of the stance classification model based on Bi-LSTM, Attention and CNN.**

activation function, which differs from hyperbolic tangent and computes faster.

- **Output layer:** the feature vector extracted by the final CNN is classified by Softmax. At the top of the CNN layer, we applied a Max Pooling function for a quadratic sampling of the obtained values, thus saving the computational effort and the number of parameters of the model. Especially, we add a small 2x2 kernel. At last, a further dense layer with a softmax activation function was used to estimate the probability distribution of each stance classification in the dataset. The model was trained for 30 periods using a categorical cross-entropy loss function [20] and an Adam optimizer, the best model was used in the classification phase.

### 3.2 Experiment setup

We take 80% of the raw data as the training set and 20% as the test set. After several iterations on the training set and the validation set, the model with the optimal convergence effect and classification accuracy was selected for testing on the test set. For all experimental models, the parameters of the experiment are shown in Table 7.

**Table 7: Experimental parameters.**

Parameter	Value
Embedding dimension	300
Dropout probability	0.5
Epochs	30
Optimizer	Adam
Decay rate	0.9
Learning rate	0.001
Batch size	64

**Table 8: Baseline models use for comparison.**

Baselines models
TF-IDF vectorization with an XGBoost classifier
TF-IDF vectorization with an SVM classifier
Fast-Text Embeddings with an SVM classifier
Fast-Text System

### 3.3 Evaluation Metrics

We evaluated the model with the metric provided by the organizers of the SemEval-2016 [21], which displays the  $F_1$  macro-average scores for two classes: “Favor” and “Against”. The test data also include members of the “None” class, but they are not included in the calculation:

$$F_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor} + R_{favor}} \quad (1)$$

$$F_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}} \quad (2)$$

where  $P$  indicates precision rate and  $R$  means recall rate. After that, the  $F_{avg}$  is calculated as:

$$F_{avg} = \frac{F_{favor} + F_{against}}{2} \quad (3)$$

## 4 RESULTS AND DISCUSSION

In this section, we conducted experiments on the proposed stance detection model and compare it with state-of-the-art models on two public datasets. After that, we performed ablation experiments to demonstrate the effectiveness of our method, and we carried out an error analysis and discussion of the experimental results.

### 4.1 Experiment results

In this subsection, we report the series of experiments we ran on our dataset utilizing various stance detection tasks, along with the outcomes we got using the baselines. We take the average of five runs after different initialization conditions as the final experimental results. See Table 8.

Tables 9, 10, and 11 report the results of the experiments carried out on the SemEval-2016 dataset, the Election 2020 Dataset, and the Philippine Presidential election 2022. The models in the first four rows, XGBoost using TF-IDF, SVM using TF-IDF, SVM with

averaged Fast-Text embeddings, and the Fast-Text system itself, are all based on linear classification.

For the SemEval-2016 dataset, the FastText system performed best overall among the linear classifiers. However, it is worth noticing that TF-IDF vectorization is better than FastText Embeddings when using the SVM [22] classifier. The results illustrate that we obtained state-of-the-art results using deep learning neural network models, performing improvements over any of the other methods presented in Section 1.

For the Election 2020 and Philippine 2022 datasets, interestingly, the FastText linear classifier combined with TF-IDF vectorization (TF-IDF + SVM) obtained better results than the other four neural networks. Moreover, the “Favor” class seems to be easier to learn than the “Against” class in Election 2020, and we find that the distribution of the “Favor” class is larger than the “Against” class by looking at the distribution of the data for both Biden and Trump, thus suggesting that the imbalance of the classes is responsible.

Several issues arise from the results in Tables 9, 10, and 11. First, there is consistency in the behavior of the model across the different datasets. Second, similar results were shown in three datasets: FastText is generally the most efficient word embedding model. One reason is that FastText can handle extra-lexical and unusual words as FastText embeddings are learned by combining location-dependent features, phrase representations, and sub-word information. Additionally, it was discovered that the supervised method with SVM worked well for various datasets. In terms of specific results, TF-IDF + SVM would outperform all other methods (in Election 2020 and Philippine 2022), which is unexpected.

### 4.2 Ablation experiments

For the proposed model, all components should be shown to have an impact on the final classification results. We do ablation experiments on the SemEval-2016 dataset and other parameters were kept constant, and all comparison experiments were evaluated using the 5-fold cross-validation method, and the average evaluation results were used as the final evaluation result. The comparison results are shown in Table 11.

From Table 12, it’s clear that the hybrid network and the attention mechanism have a big impact on classification accuracy. When CNN or BiLSTM is removed, compared to FBCA, we find that FBA and FCA have a relative improvement of 1.59% and 5.8% respectively. The much lower classification performance of the models suggests that hybrid networks The models’ significantly poorer classification performance show that hybrid networks can

**Table 9: Overall results for the SemEval-2016 Dataset.**

Model	$F1_{favor}$	$F1_{against}$	$F1_{avg}$
TF-IDF + XGBoost	49.31	65.17	57.24
TF-IDF + SVM	53.27	75.69	64.20
FTEmb + SVM	53.48	74.30	63.89
FTEmb + FastText	57.88	73.81	65.85
WBA (W2V + BiLSTM + Attention)	54.85	72.16	63.51
WBCA (W2V + BiLSTM + CNN + Attention)	55.94	72.74	64.34
FBA (FTEmb + BiLSTM + Attention)	76.43	67.24	71.83
FBCA (FTEmb + BiLSTM + CNN+ Attention) (our)	77.04	67.82	<u>72.43</u>

**Table 10: Results on the Election 2020 dataset.**

Model	$F1_{favor}$	$F1_{against}$	$F1_{avg}$
TF-IDF + XGBoost	50.98	63.86	57.42
TF-IDF + SVM	76.39	70.82	73.61
FTEmb + SVM	44.94	46.25	45.60
FTEmb + FastText	72.64	65.57	69.11
WBA (W2V + BiLSTM + Attention)	45.60	39.68	42.64
WBCA (W2V + BiLSTM + CNN + Attention)	47.29	39.04	43.17
FBA (FTEmb + BiLSTM + Attention)	72.70	64.32	68.51
FBCA (FTEmb + BiLSTM + CNN+ Attention) (our)	73.81	64.76	<u>69.29</u>

**Table 11: Results on the Philippine 2022 dataset.**

Model	$F1_{favor}$	$F1_{against}$	$F1_{avg}$
TF-IDF + XGBoost	73.64	62.78	68.21
TF-IDF + SVM	78.80	71.13	74.97
FTEmb + SVM	52.25	42.42	47.36
FTEmb + FastText	74.92	69.84	72.38
WBA (W2V + BiLSTM + Attention)	67.37	60.49	63.93
WBCA (W2V + BiLSTM + CNN + Attention)	68.74	62.25	65.50
FBA (FTEmb + BiLSTM + Attention)	73.06	66.88	69.97
FBCA (FTEmb + BiLSTM + CNN+ Attention) (our)	73.86	68.33	<u>71.10</u>

**Table 12: Classification performance of each component on SemEval-2016.**

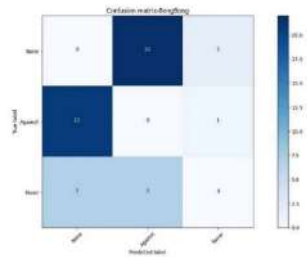
Model	$F1_{favor}$	$F1_{against}$	$F1_{avg}$
FB (FTEmb + BiLSTM)	72.06	63.59	67.83
FBC (FTEmb + BiLSTM + CNN)	74.84	65.23	70.04
FBA (FTEmb + BiLSTM + Attention)	76.43	65.24	70.84
FCA (FTEmb + CNN + Attention)	68.79	64.41	66.63
FBCA (FTEmb + BiLSTM + CNN + Attention)	77.04	67.82	72.43

more effectively make up for the shortcomings of individual deep learning models like CNN or BiLSTM, which can enhance the text classification performance. When the attention mechanism is removed, compared FB with FBA and compared FBC with FBCA, and we find that still have substantial improvement in classification performance, which contributes to increased classification accuracy. It is demonstrated that the classification performance can be

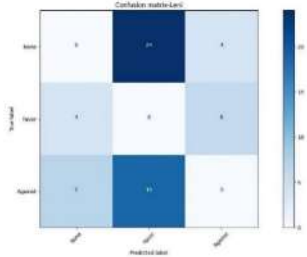
further improved by attention mechanism tuning. In conclusion, this experiment demonstrates that all the factors contribute to the classification performance of FBCA.

### 4.3 Error analysis

In the previous experience, we found that the model exhibits similar behavior in the Election 2020 and Philippine 2022 datasets—the



**Figure 2: Confusion matrix for Bongbong made with majority voting.**



**Figure 3: Confusion matrix for Leni made with majority voting.**

“Favor” class is easier to learn than the “Against” class, and through our analysis, we know that the Election 2020 results are due to class imbalance but haven’t explained the reason for Philippine 2022. Therefore, in this subsection, we provide a qualitative analysis to better understand the characteristics of the Philippines in 2022.

We performed an analysis of the prediction error. We scaled 240 tweets from this dataset, with 120 tweets each from Bongbong and Leni. We selected the three best-performing classifiers from Table 11 and identified 53 tweets that were flagged by at least two errors from these three classifiers, and their error types are shown in Figures 2 and 3. It can be observed that “Against” is frequently projected to be “Favor”, which is the most prevalent inaccuracy. In addition, the second most common source of error was Bongbong’s “Favor” being predicted as “Against” and Leni’s “None” being incorrectly predicted as “Favor”.

After the misclassified tweets were checked again manually, these were the most common sources of errors: First, the meaning of tweets is often unclear, which means that there is not sufficient context or background information to make an intelligent decision, to the point where there are just annotation errors; Second, a lot of Twitter text use metaphorical languages, such as irony, sarcasm, or the need for additional common sense or domain-specific expertise; Third, some tweets make indirect references to the topic without expressing a clear stance opinion on it.

## 5 CONCLUSION

The stance detection task plays an important role in public opinion analysis, and we can apply it to public event stance analysis to provide data reference for the government. To meet this requirement,

this paper proposes a stance detection method based on hybrid deep neural networks, which uses BiLSTM and CNN to extract features respectively, and is mediated by a certain degree of attention. The experimental results show the effectiveness of the proposed model on three different datasets, and the average accuracy score on the SemEval-2016 public dataset reached 72.43%, which was better than other deep learning methods. Furthermore, we attempt to measure the impact on the final result of the entire model by choosing a technique that encodes text input through word embeddings. It has been shown that for this architecture, using a pre-trained FastText embedding as a word representation can slightly improve the performance of a single neural network model, allowing obtaining the best results for stance detection. At the same time, this paper also provides a dataset of the 2022 Philippine presidential election for in-depth study by scholars in related fields. Recent efforts to apply transfer learning and unsupervised learning to stance detection have yielded encouraging results and are expected to be one of the major research directions in the field in the future, and using our dataset can help explore potential approaches to unsupervised stance detection techniques.

## ACKNOWLEDGMENTS

This research has received funding from the National Natural Science Foundation of China, Research on Knowledge Graph Construction and Inference Methods for Multilingual Events in ASEA. Grant no: 61866008. The authors appreciate the insightful comments from anonymous reviewers.

## REFERENCES

- [1] ALDayel A, Magdy W. Stance detection on social media: State of the art and trends[J]. *Information Processing & Management*, 2021, 58(4): 102597.
- [2] Al-Ghadir A I, Azmi A M, Hussain A. A novel approach to stance detection in social media tweets by fusing ranked lists and sentiments[J]. *Information Fusion*, 2021, 67: 29-40.
- [3] Kawintiranon K, Singh L. Knowledge enhanced masked language model for stance detection[C]//*Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*. 2021: 4725-4735.
- [4] Alkhalifa R, Zubiaga A. Capturing stance dynamics in social media: open challenges and research directions[J]. *International Journal of Digital Humanities*, 2022, 3(1-3): 115-135.
- [5] Elfardy H, Diab M. Cu-gwu perspective at semeval-2016 task 6: Ideological stance detection in informal text[C]//*Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 2016: 434-439.
- [6] Wojatzki M, Zesch T. ltl. uni-due at semeval-2016 task 6: Stance detection in social media using stacked classifiers[C]//*Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. 2016: 428-433.
- [7] Ghosh S, Singhania P, Singh S, *et al*. Stance detection in web and social media: a comparative study[C]//*Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*. Springer International Publishing, 2019: 75-87.
- [8] Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation[C]//*Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014: 1532-1543.
- [9] Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space[J]. *arXiv preprint arXiv:1301.3781*, 2013.
- [10] Tang D, Wei F, Yang N, *et al*. Learning sentiment-specific word embedding for twitter sentiment classification[C]//*ACL (1)*. 2014: 1555-1565.
- [11] Fu X, Liu W, Xu Y, *et al*. Combine HowNet lexicon to train phrase recursive autoencoder for sentence-level sentiment analysis[J]. *Neurocomputing*, 2017, 241: 18-27.
- [12] Araque O, Corcuera-Platas I, Sánchez-Rada J F, *et al*. Enhancing deep learning sentiment analysis with ensemble techniques in social applications[J]. *Expert Systems with Applications*, 2017, 77: 236-246.
- [13] Ren Y, Wang R, Ji D. A topic-enhanced word embedding for Twitter sentiment classification[J]. *Information Sciences*, 2016, 369: 188-198.

- [14] Giatsoğlu M, Vozalis M G, Diamantaras K, *et al.* Sentiment analysis leveraging emotions and word embeddings[J]. *Expert Systems with Applications*, 2017, 69: 214-224.
- [15] Wang Y, Huang M, Zhu X, *et al.* Attention-based LSTM for aspect-level sentiment classification[C]//*Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016: 606-615.
- [16] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. *arXiv preprint arXiv:1510.03820*, 2015.
- [17] Zotova E. Automatic stance detection on political discourse in Twitter[J]. 2019.
- [18] Rong X. word2vec parameter learning explained[J]. *arXiv preprint arXiv:1411.2738*, 2014.
- [19] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. *arXiv preprint arXiv:1409.0473*, 2014.
- [20] Goodfellow I, Bengio Y, Courville A. *Deep learning*[M]. MIT press, 2016.
- [21] Mohammad S, Kiritchenko S, Sobhani P, *et al.* Semeval-2016 task 6: Detecting stance in tweets[C]//*Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 2016: 31-41.
- [22] Cherkassky V, Ma Y. Practical selection of SVM parameters and noise estimation for SVM regression[J]. *Neural networks*, 2004, 17(1): 113-126.

# An Objective Reduction Evolutionary Multiobjective Algorithm using Adaptive Density-Based Clustering for Many-objective Optimization Problem

Mingjing Wang

Long Chen\*

wangmingjing@seu.edu.cn

chen\_long@seu.edu.cn

Southeast University, Nanjing

Nanjing, Jiangsu, China

Huiling Chen

Wenzhou University

wenzhou, China

chenhuiling.jlu@gmail.com

## ABSTRACT

Many-objective optimization problems (MaOPs), are the most difficult problems to solve when it comes to multiobjective optimization issues (MOPs). MaOPs provide formidable challenges to current multiobjective evolutionary methods such as selection operators, computational cost, visualization of the high-dimensional trade-off front. Removal of the reductant objectives from the original objective set, known as objective reduction, is one of the most significant approaches for MaOPs, which can tackle optimization problems with more than 15 objectives is made feasible by its ability to greatly overcome the challenges of existing multi-objective evolutionary computing techniques. In this study, an objective reduction evolutionary multiobjective algorithm using adaptive density-based clustering is presented for MaOPs. The parameters in the density-based clustering can be adaptively determined by depending on the data samples constructed. Based on the clustering result, the algorithm employs an adaptive strategy for objective aggregation that preserves the structure of the original Pareto front as much as feasible. Finally, the performance of the proposed multiobjective algorithms on benchmarks is thoroughly investigated. The numerical findings and comparisons demonstrate the efficacy and superiority of the suggested multiobjective algorithms and it may be treated as a potential tool for MaOPs.

## KEYWORDS

Objective Reduction, Evolutionary Multiobjective Algorithm, Density-Based Clustering, Many-objective Optimization Problem

### ACM Reference Format:

Mingjing Wang, Long Chen, and Huiling Chen. 2023. An Objective Reduction Evolutionary Multiobjective Algorithm using Adaptive Density-Based Clustering for Many-objective Optimization Problem. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3590003.3590103>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org). CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM. ACM ISBN 978-1-4503-9944-9/23/03...\$xxx  
<https://doi.org/10.1145/3590003.3590103>

## 1 INTRODUCTION

In practice, it is preferable for the majority of applications to evaluate as many objectives as feasible in order to meet diverse performance requirements more effectively [13, 17]. These optimization problems with more than conflicting three optimization objectives are called many-objective optimization problems (MaOPs). MaOPs are widespread in real-world application applications, referring to the optimization of more than three conflicting objectives simultaneously, such as air traffic control [9], groundwater monitoring [11], and molecular design [12]. MaOPs have recently sparked considerable attention in the evolutionary multiobjective optimization field since they provide formidable challenges to practically all extant multiobjective evolutionary algorithm categories (MOEAs) [21, 24]. The main challenges encountered by MOEAs in finding a representative set of Pareto optimal solutions for MaOPs are as follows: (i) The search process's capacity to converge to the Pareto front decreases as the objective space's dimensionality rises and increases in the number of non-dominated solutions. (ii) With the increase in the number of objectives, new requirements are put forward for the effectiveness of crossover and mutation operators of evolutionary algorithms. In a high-dimensional space, population members are far apart. Thus, two distant parent solutions generate two distant offspring who are unlike their parents. If so, recombination fails to produce promising progeny. (iii) The identification of surrounding solutions in a population becomes computationally costly in high-dimensional spaces when determining the degree of crowding of a solution in a population. (iiii) With the increase in the number of objectives, it is almost impossible to solve the MaOPs with the help of an evolutionary computation algorithm through the visualization of objective space. Therefore, the performance of the classical MOEAs deteriorates when tackling problems involving a larger number of conflicting objectives.

In recent years, a number of contributions have been made to overcome the limitations of current MOEAs in a variety of objective optimization applications. Some work is designed to strengthen dominance relations and new diversity promotion mechanisms in Pareto dominance-based MOEAs. Yang et al. [25] proposed a grid-based evolutionary algorithm (GrEA) to solve MaOPs and GrEA can strengthen the selection pressure toward the optimal direction while maintaining an extensive and uniform distribution among solutions. He et al. [8] designed a new fitness evaluation mechanism to continuously differentiate individuals into different degrees of optimality beyond the classification of the original Pareto

dominance and the concept of fuzzy logic is adapted to define a fuzzy Pareto domination relation. Yuan et al. [27] presented a new dominance relation-based evolutionary algorithm for MaOPs. Some work mainly focused on the design approximation of hypervolume values and more computationally efficient performance indicators in indicator-based MOEAs. Hisao et al. [10] propose the idea of using a scalarizing function-based hypervolume approximation method in IBEAs for MaOPs. Johannes et al. [2] used monte carlo simulation to approximate the exact hypervolume values for MaOPs. Cynthia et al. [18] proposed a performance assessment indicator  $\Delta_p$  for reducing the original expensive computing resources when faced with MaOPs. Moreover, some work designed novel updating strategies and the generation of more uniformly distributed weight vectors in decomposition-based MOEAs such as [1, 22, 28]. However, these recently suggested many-objective evolutionary algorithms may not be powerful enough to appropriately handle MaOPs with more than around 15 goals [4, 27], and their efficiency still has to be further investigated and tested on more real-world issues. Moreover, even with advancements in MOEAs, many-objective optimization's inherent difficulty in visualizing the high-dimensional Pareto front and picking favored solutions remains unabated [26].

A second major method for handling MaOPs is objective reduction, which, rather than focusing on making current MOEAs more scalable, aims to reduce the complexity of the issue by cutting down on the number of goals at either the decision-making or search stages. Recently, objective reduction approach has become one of the most essential techniques for MaOPs, as it may ease the challenges of selection pressure, computing expense, and human-computer interface visualization. The reason for this is that there is a minimal set of  $k$  ( $k < m$ ) conflicting objectives that can produce the same Pareto front as the original problem for many situations with  $m$  objectives in practice. These  $k$  objectives are frequently referred to as vital, while the others are considered redundant [19]. The objective reduction has attracted the attention of many academic researchers. Objective reduction can be roughly categorized into dominance structure-based approaches [19] and correlation-based approaches [20]. Most objective reduction algorithms return just one reduction goal set in a simulation and do not address the difference between approximate PF and actual PF or the competing criteria between error tolerance and the number of predicted objectives. The many-objective optimization approach immediately deletes unnecessary objectives, hence optimization results may vary from genuine solutions. In instance, other circumstances may affect objective connections, which may even change [5].

In this study, we designed an objective reduction evolutionary multiobjective algorithm using adaptive density-based clustering for MaOPs. In the proposed algorithm, the parameters in the density-based clustering can be adaptively determined by depending on the data samples constructed. And then based on the clustering outcome, the method utilizes an adaptive technique for objective aggregation that retains the original Pareto front's structure to the greatest extent possible. Ultimately, the performance of the suggested multiobjective algorithms on benchmarks is explored in depth. The numerical results and comparisons demonstrate the efficiency and superiority of the proposed multiobjective algorithms, which may be considered a promising tool for MaOPs.

## 2 MATERIALS AND METHODS

### 2.1 The definition of MaOP

In this section, some background can be given. A general MaOP can be mathematically formulated as:

$$\begin{aligned} \min \mathbf{f}(\mathbf{x}) &= (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_m(\mathbf{x}))^T \\ \text{subject to } g_i(\mathbf{x}) &\leq 0, i = 1, 2, \dots, u \\ h_j(\mathbf{x}) &= 0, j = 1, 2, \dots, v \\ \mathbf{x} &\in \Omega \end{aligned} \quad (1)$$

where  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  is a decision vector in the decision space  $\Omega$  with  $n$  dimensions,  $\mathbf{f} : \Omega \rightarrow \Theta \subseteq \mathbb{R}^m$  is an objective vector consisting of  $m$  objective functions, which maps  $n$  dimensions in decision space into  $m$  dimension objective space  $\Theta$ . And the  $\mathbf{f}(\mathbf{x})$  is called as MaOP when  $m > 3$ , if  $m \leq 3$   $\mathbf{f}(\mathbf{x})$  is called as multi-objective optimization.  $g_i(\mathbf{x}) \leq 0$  and  $h_j(\mathbf{x}) = 0$  mean inequality and equality constraints respectively.

### 2.2 The density-based clustering

Density-based clustering identifies clusters as zones containing densely packed items. The clusters are divided by zones of low population density [23]. DBSCAN [6] is a density based algorithm which discovers clusters with arbitrary shape by examining the density of objects inside a given location, density-based clustering may identify areas where similar items tend to congregate. Sparse areas divide one cluster from the next. In DBSCAN, there are two key input parameters radius of the cluster ( $Eps$ ) and the minimum required points inside the cluster ( $MinPts$ ). A detailed description of DBSCAN can be seen in Algorithm 1. In this DBSCAN, the parameters  $Eps$  and  $MinPts$  are difficult to set and seriously affect the performance of the algorithm.

---

#### Algorithm 1: DBSCAN

---

**Input:**  $D$ : data set;  
 $Eps$ : radius of the cluster;  
 $MinPts$ : minimum required points inside the cluster;  
**Output:** elitist cluster  
Set all objects in  $D$  unvisited;  
**for**  $i = 1$  **to**  $|D|$  **do**  
    **if**  $P_i$  is a cluster; **then**  
        | continue;  
    **else**  
        check  $NEps(p_i)$ ;  
        **if**  $NEps(p_i) < MinPts$  **then**  
            | Mark  $p_i$  as boundary point;  
        **else**  
            Mark  $p_i$  as the core point;  
            Add  $NEps(p_i)$  into new cluster  $C$ ;  
            **for each** unvisited  $q$  in  $NEps(p_i)$  **do**  
                check  $NEps(q)$ ;  
                **if**  $NEps(q) > MinPts$  **then**  
                    | Add nclustered  $NEps(q)$  into cluster  $C$ ;  
    **return** elitist cluster set  $C$ ;

---

## 2.3 The proposed algorithm

In this study, an objective reduction evolutionary multiobjective algorithm using adaptive DBSCAN (EDBSCANec) for MaOPs. In EDBSCANec, the MaOPs are optimized by NSGAIII [4] firstly. The  $Eps$  and  $MinPts$  are calculated with the current elitist solution  $P$  as [7] and then EDBSCAN is used to cluster current objectives. And then, the objectives in  $C$  are aggregated. For example, the  $K_i = \{M_1, M_2, \dots, M_m\}$  means the  $m$  objectives in  $i^{th}$  cluster  $C_i$ , and  $MK_i = \{C_1 * M_1 + C_2 * M_2 + \dots + C_m * M_m\}$  is the objectives aggregate function, where  $C_1 + C_2 + \dots + C_m = 1$ , the  $M_i$  is the  $i^{th}$  objective and  $C_i$  is the of weight coefficient of the  $i^{th}$  objective. Finally, calculate current elitist  $P$  on reduced objectives with NSGAIII.

---

### Algorithm 2: EDBSCANec

---

**Input:**  $P$ : population;  
 $N$ : population size;  
 $genmax$ : maximum iterations  
**Output:** elitist solution  $P$   
**while**  $t < genmax$  **do**  
  **if**  $t \bmod K_o$ ; **then**  
    Calculate current elitist  $P$  on all objectives with NSGAIII;  
     $t = t + 1$ ;  
  Calculate the  $Eps$  and  $MinPts$  with elitist  $P$ ;  
   $C \leftarrow \text{DBSCAN}(Eps, MinPts, P)$ ;  
  Aggregate each cluster in  $C$ ; **while**  $t \bmod K_o$  **do**  
    Calculate current elitist  $P$  on reduced objectives with NSGAIII;  
**return** elitist solution  $P$ ;

---

## 3 EXPERIMENT AND ANALYSIS

### 3.1 Experiment setup

In this part, the proposed EDBSCANec is verified on DTLZ benchmarks [15] and the EDBSCANec is compared with several classical and state-of-the-art algorithms such as NSGA-III [4], RVEA [3], HypE [14], aDECOR and fDECOR [16], NSGA-III is an NSGA-II-based classical approach for resolving MaOPs, HypE is useful when dealing with MaOPs when the PF is unknown. Both aDECOR and fDECOR are differential many-objective optimization algorithms that employ a clustering technique to reduce the number of objectives. Experiments are conducted on a machine equipped with 16GB RAM, an Intel(R) CORE CPU 7<sup>th</sup> Window system. Matlab2022b is used to do simulation verification. Each experimental group was done independently 30 times.

### 3.2 Experiment results

In this study, the hypervolume(HV) [24] and inverted generational distance (IGD) [21] are all used to measure the proposed method. The detailed experimental HV result on DTLZ1-DTLZ3 is recorded in Table 1 and the numbers of objectives ( $M$ ) are all set as 10, 20, and 30, and  $D$  means the number of decision variables. It can be observed from this Table 1 that the proposed EDBSCANec performs the best among all these comparison algorithms. Among all test problems, the performance of the algorithm proposed EDBSCANec in this

paper is only inferior to that of aDECOR and fDECOR in terms of HV. Among all the test problems, the algorithm performance is second only to the algorithm proposed EDBSCANec in this paper, which is aDECOR and fDECOR. In addition, the performance of the original multi-objective evolutionary algorithm without objective reduction is significantly lower than that of aDECOR, fDECOR and EDBSCANec for HV. The interesting phenomenon is that the performance of evolutionary algorithms based on objective reduction (aDECOR and fDECOR) is significantly better than that of ordinary multi-objective evolutionary algorithms such as NSGA-III, RVEA, and HypE.

Table 2 shows detailed experimental IDG results on DTLZ1-DTLZ3 of all algorithms. It can be observed from this Table 2 that EDBSCANec outperforms all other comparison methods for DTLZ1-DTLZ2 compression with different numbers of objectives and decision variables in terms of IDG value. The experimental phenomenon that occurs when considering the measure HV is also found when considering the IDG value. Among all the algorithms, the performance of the multi-objective evolutionary algorithm using the objectives reduction technology (EDBSCANec, aDECOR and fDECOR) is significantly better than that of the ordinary multi-objective evolutionary algorithms.

In addition, some statistical analysis results between the EDBSCANec and others on HV are shown in Table 3. The F-test is used to check if the variances of the two samples are equivalent. The T-test examines whether or not the sample means are identical. In addition, the P-value is determined using the significance test technique. It can be seen from this Table 3 that all the statistical results are less than 0.05 which shows the significant difference between our algorithm and others. Furthermore, the statistical analysis results between the proposed EDBSCANec and others on IGD are shown in Table 4. The same phenomenon can also be found in Table 4 that all the statistical results are all less than 0.05. Based on the above statistical results, a preliminary conclusion can come out that the proposed EDBSCANec may be treated as a potential tool for MaOPs.

## 4 CONCLUSION AND FUTURE WORK

Through objective reduction for MaOPs, the existing multi-objective evolutionary computation algorithm can enhance its ability to solve MaOPs as well as complex optimization problems with more than 15 objectives. In this study, we provide an evolutionary multiobjective method that makes use of adaptive density clustering (EDBSCANec) to reduce the number of objectives for MaOPs. In EDBSCANec, density-based clustering allows for the parameters to be adaptively chosen based on the data samples created. Using information gleaned via clustering, the algorithm applies a flexible method of objective aggregation that attempts to maintain the integrity of the initial Pareto front. Finally, the suggested multi-objective algorithms are tested extensively on benchmarks to see how well they perform. The recommended multiobjective algorithms EDBSCANec have been shown to be effective and superior in numerical discoveries and comparisons and as a result, they may be considered a useful tool for MaOPs in practical issues.

Although the test results of the algorithm EDBSCANec proposed in this paper on the standard data set show that it has great potential

**Table 1: Results of HV values on DTLZs**

Problem	M	D	NSGA-III	RVEA	HypE	aDECOR	fDECOR	EDBSCANec
DTLZ1	10	14	0.3698	0.3874	0.3589	0.8510	0.8052	0.9254
	20	24	0.2584	0.2302	0.2510	0.6921	0.7254	0.8274
	30	34	0.3214	0.1479	0.0125	0.7258	0.6874	0.8847
DTLZ2	10	19	0.0921	0.2154	0.2854	0.3485	0.3125	0.6584
	20	29	0.2584	0.0895	0.7421	0.7268	0.7485	0.7741
	30	39	0.5487	0.1589	1.5421	0.0090	0.1452	0.8521
DTLZ3	10	19	0.5428	0.3471	0.3658	0.6580	0.5863	0.8632
	20	29	0.6985	0.1258	0.5841	0.3695	0.6584	0.7885
	30	39	0.0025	0.1254	0.2584	0.4258	0.3695	0.6582

**Table 2: Results of IGD values on DTLZs**

Problem	M	D	NSGA-III	RVEA	HypE	aDECOR	fDECOR	EDBSCANec
DTLZ1	10	14	29.6584	18.6932	37.5482	17.6985	18.5632	15.3254
	20	24	63.3254	50.4897	39.5842	48.5695	50.3652	35.3215
	30	34	34.6985	69.3541	58.7855	36.9896	35.9854	30.2598
DTLZ2	10	19	2.6584	0.9956	0.9896	1.9854	1.6563	0.7584
	20	29	4.2514	9.5682	4.5821	3.6985	2.6985	1.6985
	30	39	8.5623	9.6352	11.2541	5.6987	4.5874	1.9856
DTLZ3	10	19	105.2594	112.6985	110.2584	98.5471	99.854	85.5695
	20	29	440.3652	115.6952	108.3698	99.3692	96.6854	95.2584
	30	39	445.6987	185.7452	165.6845	110.8745	112.6952	105.6954

**Table 3: Statistical analysis of results based on HV**

Metrcis	NSGA-III	RVEA	HypE	aDECOR	fDECOR
F-test	0.002354	0.000136	0.017808	0.000254	0.000362
T-test	0.016235	0.012541	0.032154	0.010685	0.041258
P-value	0.020125	0.020254	0.024180	0.022541	0.025840

**Table 4: Statistical analysis of results based on IGD**

Metrcis	NSGA-III	RVEA	HypE	aDECOR	fDECOR
F-test	0.003658	0.000145	0.017852	0.001474	0.002580
T-test	0.017452	0.026985	0.048542	0.018942	0.042512
P-value	0.022151	0.036582	0.025840	0.027413	0.014840

for MaOPs, especially for problems with more than 15 objectives, some issues are still worthy of further investigation in future work. First of all, the parallel mode of the algorithm proposed in this paper needs to be designed and proposed. Because the clustering learning process is constructed in this paper, it consumes more time than the original multi-objective evolutionary algorithm. The proposed parallel version can effectively improve the execution efficiency of the algorithm. In addition, there are few kinds of benchmark problems accessible for objective reduction in the literature. It is

very desired to create more benchmark problems with novel properties so that objective reduction methods EDBSCANec may be assessed in a more complete manner. Furthermore, the algorithm proposed in this paper can be applied to some practical problems with more than 15 objectives to further verify the performance of the algorithm proposed EDBSCANe in this study.

## ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of Jiangsu Province (BK20210204) and we would like to thank the anonymous reviewers for their helpful remarks.

## REFERENCES

- [1] Md Asafuddoula, Tapabrata Ray, and Ruhul Sarker. 2014. A decomposition-based evolutionary algorithm for many objective optimization. *IEEE Transactions on Evolutionary Computation* 19, 3 (2014), 445–460.
- [2] Johannes Bader and Eckart Zitzler. 2011. HypE: An algorithm for fast hypervolume-based many-objective optimization. *Evolutionary computation* 19, 1 (2011), 45–76.
- [3] Ran Cheng, Yaochu Jin, Markus Olhofer, and Bernhard Sendhoff. 2016. A reference vector guided evolutionary algorithm for many-objective optimization. *IEEE Transactions on Evolutionary Computation* 20, 5 (2016), 773–791.
- [4] Kalyanmoy Deb and Himanshu Jain. 2013. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints. *IEEE transactions on evolutionary computation* 18, 4 (2013), 577–601.
- [5] Rui Ding, Hong-bin Dong, Gui-sheng Yin, Jing Sun, Xiao-dong Yu, and Xian-bin Feng. 2021. An objective reduction method based on advanced clustering for many-objective optimization problems and its human-computer interaction visualization of pareto front. *Computers & Electrical Engineering* 93 (2021), 107266.
- [6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *kdd*, Vol. 96. 226–231.
- [7] Khalil Ibrahim Ghatwani and Athraa Jasim Mohammed. 2022. Intelligent Bat Algorithm for Finding Eps Parameter of DbScan Clustering Algorithm. *Iraqi Journal of Science* (2022), 5572–5580.
- [8] Zhenan He, Gary G Yen, and Jun Zhang. 2013. Fuzzy-based Pareto optimality for many-objective evolutionary algorithms. *IEEE Transactions on Evolutionary Computation* 18, 2 (2013), 269–285.
- [9] Jesús García Herrero, Antonio Berlanga, and José Manuel Molina López. 2008. Effective evolutionary algorithms for many-specifications attainment: Application to air traffic control tracking filters. *IEEE Transactions on Evolutionary Computation* 13, 1 (2008), 151–168.
- [10] Hisao Ishibuchi, Noritaka Tsukamoto, Yuji Sakane, and Yusuke Nojima. 2010. Indicator-based evolutionary algorithm with hypervolume approximation by achievement scalarizing functions. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*. 527–534.
- [11] Joshua Brian Kollat, Patrick M Reed, and RM Maxwell. 2011. Many-objective groundwater monitoring network design using bias-aware ensemble Kalman filtering, evolutionary optimization, and visual analytics. *Water Resources Research* 47, 2 (2011).
- [12] Johannes W Krusselbrink, Michael TM Emmerich, Thomas Bäck, Andreas Bender, Ad P IJzerman, and Eelke van der Horst. 2009. Combining aggregation with Pareto optimization: A case study in evolutionary molecular design. In *International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 453–467.
- [13] Hui Liu, Ye Li, Zhu Duan, and Chao Chen. 2020. A review on multi-objective optimization framework in wind energy forecasting techniques and applications. *Energy Conversion and Management* 224 (2020), 113324.
- [14] Adriana Menchaca-Mendez and Carlos A Coello Coello. 2017. An alternative hypervolume-based selection mechanism for multi-objective evolutionary algorithms. *Soft Computing* 21 (2017), 861–884.
- [15] Ivan Reinaldo Meneghini, Marcos Antonio Alves, Antônio Gaspar-Cunha, and Frederico Gadelha Guimaraes. 2020. Scalable and customizable benchmark problems for many-objective optimization. *Applied Soft Computing* 90 (2020), 106139.
- [16] Monalisa Pal, Sriparna Saha, and Sanghamitra Bandyopadhyay. 2018. DECOR: differential evolution using clustering based objective reduction for many-objective optimization. *Information Sciences* 423 (2018), 200–218.
- [17] Robin C Purshouse and Peter J Fleming. 2007. On the evolutionary optimization of many conflicting objectives. *IEEE transactions on evolutionary computation* 11, 6 (2007), 770–784.
- [18] Cynthia A Rodríguez Villalobos and Carlos A Coello Coello. 2012. A new multi-objective evolutionary algorithm based on a performance assessment indicator. In *Proceedings of the 14th annual conference on Genetic and evolutionary computation*. 505–512.
- [19] Dhish Kumar Saxena, Joao A Duro, Ashutosh Tiwari, Kalyanmoy Deb, and Qingfu Zhang. 2012. Objective reduction in many-objective optimization: Linear and nonlinear algorithms. *IEEE Transactions on Evolutionary Computation* 17, 1 (2012), 77–99.
- [20] Hemant Kumar Singh, Amitay Isaacs, and Tapabrata Ray. 2011. A Pareto corner search evolutionary algorithm and dimensionality reduction in many-objective optimization problems. *IEEE Transactions on Evolutionary Computation* 15, 4 (2011), 539–556.
- [21] Yanan Sun, Gary G Yen, and Zhang Yi. 2018. IGD indicator-based evolutionary algorithm for many-objective optimization problems. *IEEE Transactions on Evolutionary Computation* 23, 2 (2018), 173–187.
- [22] Yan-Yan Tan, Yong-Chang Jiao, Hong Li, and Xin-Kuan Wang. 2013. MOEA/D+uniform design: A new version of MOEA/D for optimization problems with many objectives. *Computers & Operations Research* 40, 6 (2013), 1648–1660.
- [23] Tran Manh Thang and Juntae Kim. 2011. The anomaly detection by using dbscan clustering with multiple parameters. In *2011 International Conference on Information Science and Applications*. IEEE, 1–5.
- [24] Ye Tian, Ran Cheng, Xingyi Zhang, Yansen Su, and Yaochu Jin. 2018. A strengthened dominance relation considering convergence and diversity for evolutionary many-objective optimization. *IEEE Transactions on Evolutionary Computation* 23, 2 (2018), 331–345.
- [25] Shengxiang Yang, Miqing Li, Xiaohui Liu, and Jinhua Zheng. 2013. A grid-based evolutionary algorithm for many-objective optimization. *IEEE Transactions on Evolutionary Computation* 17, 5 (2013), 721–736.
- [26] Yuan Yuan, Yew-Soon Ong, Abhishek Gupta, and Hua Xu. 2017. Objective reduction in many-objective optimization: evolutionary multiobjective approaches and comprehensive analysis. *IEEE Transactions on Evolutionary Computation* 22, 2 (2017), 189–210.
- [27] Yuan Yuan, Hua Xu, Bo Wang, and Xin Yao. 2015. A new dominance relation-based evolutionary algorithm for many-objective optimization. *IEEE Transactions on Evolutionary Computation* 20, 1 (2015), 16–37.
- [28] Yuan Yuan, Hua Xu, Bo Wang, Bo Zhang, and Xin Yao. 2015. Balancing convergence and diversity in decomposition-based many-objective optimizers. *IEEE Transactions on Evolutionary Computation* 20, 2 (2015), 180–198.

# Infrared small target detection based on the combination of single image super-resolution reconstruction and YOLOX

Wang Zhiyong\*

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
zhiyong-wang@foxmail.com

Xiang Xuefu

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
xxf20013011@163.com

Zeng Kan

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
58123515@qq.com

Zhang Zhenyu

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
177757276@qq.com

Li Yanan

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
liyanan@58suo.com

Song Dengpan

Automation Research Institute Co.,  
Ltd. Of China South Industries Group  
Corporation, Mianyang, China  
2841586820@qq.com

## ABSTRACT

For the infrared search and tracking system, it is necessary to increase the ability to detect small infrared targets against complex backgrounds. YOLOX is a high-performance detector, but its detection performance is constrained when it uses data from low-resolution infrared images with small targets. However, occasionally design constraints and budgetary restraints will prevent the optical system and sensor resolution from being increased enough to improve image quality. Real-ESRGAN is used to solve this issue by reconstructing a high-resolution infrared image from its low-resolution counterpart, which will be used as YOLOX-S's input. Also, the YOLOX-S training strategy is modified further to make it appropriate for the detection of infrared small targets, including the Mosaic and MixUp data augmentation and the size of ground-truth. The average precision achieved by the suggested method in this work increases from 63.70% to 77.19%, which shows a considerable improvement in infrared small target detection when compared with the original model by inputting original images.

## CCS CONCEPTS

• Computing methodologies; • Artificial intelligence; • Computer vision; • Computer vision problems; • Object detection;

## KEYWORDS

Infrared technology, Infrared small target detection, Deep learning, Super-resolution reconstruction, YOLOX

## ACM Reference Format:

Wang Zhiyong, Xiang Xuefu, Zeng Kan, Zhang Zhenyu, Li Yanan, and Song Dengpan. 2023. Infrared small target detection based on the combination of

single image super-resolution reconstruction and YOLOX. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590104>

## 1 INTRODUCTION

Infrared search and tracking system, with the benefits of all-weather imaging, has a wide range of applications in the field of remote sensing and monitoring [1]. Due to the target's lack of shape and texture information, low radiation intensity, and complex background, the infrared small target detection has proven to be a difficult problem [2].

Object detection algorithms based on deep learning have been widely used due to the good performance of the generalization capability and the ability to extract features from complex backgrounds [3]. The detection of small infrared targets against complicated backgrounds, however, is typically not optimal, particularly for low-resolution and low-quality infrared images. The low-resolution infrared image is readily lacking in high-frequency information of infrared small target in complex backgrounds, and frequently has issues like high noise, edge blur, and low contrast due to the influence of equipment, environment, bandwidth, and other factors [4]. The increase of image quality and resolution through improving the resolution of optical systems and sensors may be constrained by design restrictions and financial constraints.

Deep learning-based object detection algorithms now are divided into two categories: two-stage object detection algorithms [5, 6] and one-stage object detection algorithms [7–11]. Despite the fact that the former often have good detection accuracy, the low efficiency and slow speed cannot be disregarded [3]. The YOLO series is a common one-stage algorithm that can successfully accomplish the detection task fast and accurately, and it improves with time in terms of both accuracy and speed. A more modern version with an anchor-free technique is YOLOX [10]. Due to the complex backgrounds, low radiation intensity, and lack of target shape and texture information, the suggested training strategy of YOLOX is not totally suitable for the detection of infrared small targets in complex backgrounds.

To address the aforementioned issues, the Real-ESRGAN [12], which is a single image super-resolution (SR) algorithm based on

\*Corresponding Author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590104>

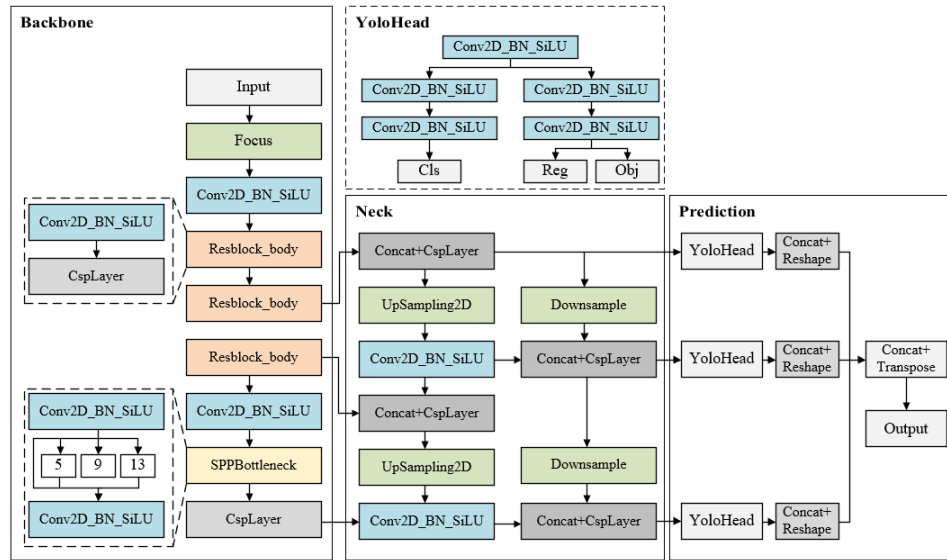


Figure 1: Model structure of YOLOX

generative adversarial network (GAN), is first used to reconstruct a high-resolution infrared image from its low-resolution counterpart, which will then be used as the input of the YOLOX detection framework. GAN is one of the key subfields of deep learning research and excellent for generating real world information. Real-ESRGAN's ability to retain the signals of small targets and enhance the details of infrared image can increase YOLOX's capacity for detection. Second, more optimization is made to the training strategies for YOLOX-based infrared small target detection. To enable YOLOX to properly learn the contextual features between the target and the backgrounds, the size of the ground truth is changed. Additionally, the Mosaic and MixUp data augmentation (MMDA) is re-optimized to make the YOLOX model more effective at detecting small targets in complex backgrounds.

## 2 METHOD

The initial step of our procedure involves using the Real-ESRGAN algorithm to reconstruct the high-resolution infrared image from its low-resolution counterpart. The target detection network receives the high-resolution infrared image in the second step, together with the training strategy optimization.

### 2.1 Infrared image SR reconstruction method based on GAN

By increasing the number of small target pixels that enter the convolutional layer, the SR reconstruction method based on GAN can lessen the feature loss of small targets caused on by recurrent downsampling in the target detection network [13].

Real-ESRGAN is an extension of the single image SR reconstruction algorithm ESRGAN [14], which models more real-world degradations by converting the traditional first-order degradation model in ESRGAN to a higher-order degradation process. Multiple

repeating degradation processes are used in higher-order degradation modeling, each of which is a classical degradation model with a unique set of parameters. Additionally, Real-ESRGAN replaces the VGG type discriminator in ESRGAN with a U-Net discriminator with skip connections and spectral normalization to enhance the discriminator's performance and stabilize training. More details are described by Wang X T et al.[12].

Several models are offered by Real-ESRGAN. All three versions of Real-ESRGAN—RealESRGAN\_x4plus, RealESRGAN\_x4plus\_anime\_6B, RealESRGAN\_x2plus—use the generator residual-in-residual dense block in ESRGAN. While the generator for the realesr-animevideo-v3 and realesr-general-x4v3 models is a VGG-style network without the batch normalizing layer. In our experiment, the realesr-generic-x4v3, a tiny little model for general scenes, is adopted to reconstruct a high-resolution infrared image from its low-resolution counterpart. This is done in accordance with the model size, application scene, and image reconstruction effect.

### 2.2 Infrared small target detection network

YOLOX-S is a lightweight model of YOLOX, which could make progress in the speed of defect detection. Its structure can be divided into three parts [3, 10] (as shown in Figure 1): Backbone (feature extraction network), Neck (feature fusion network), and Prediction (prediction network). Backbone performs convolution calculation on the input image and extracts sample features to input the feature fusion network Neck composed of FPN and PAN [3, 10]. Prediction's YoloHead is distinct from the heads of earlier YOLO series models. Utilizing a decoupled head for classification and localization solves the conflict caused by the coupling of the two tasks and improves the convergence speed.

The YOLOX model uses the MMDA strategy to improve detection performance, but it is closed for the last 15 epochs. The Mosaic data augmentation algorithm is an effective data augmentation

strategy proposed in ultralytics-YOLOv3 [7], and then widely used in YOLOv4 [8] and YOLOv5 [9], etc. It randomly selects 4 images from the train set and puts the contents of the 4 pictures into a synthetic image that is directly used for training. The MixUp data augmentation algorithm is originally designed for image classification task and later used for object detection task, which generates a weighted combination of random image pairs from the training data.

We close the MMDA for the last 140 epochs in order to enhance the YOLOX’s capacity to detect small targets in complex backgrounds. Additionally, the size of the ground-truth is adjusted to 40×40 pixels in order to prevent feature loss of small targets due to repeated downsampling in the convolutional layer and to enable the YOLOX model to properly learn the contextual features between the target and backgrounds.

### 2.3 Evaluation metrics

The performance of output results can be analyzed using a number of traditional evaluation metrics. Peak signal-to-noise ratio (PSNR), structural similarity index measurement (SSIM), and average precision (AP) are used to evaluate the effects of image SR reconstruction and target detection network performance, respectively.

PSNR is the ratio between the maximum possible power of a signal and the power of corrupted noise that affects the fidelity of its representation [15], which is one of the metrics widely used to evaluate the quality of image reconstruction between the original low-resolution and a restored high-resolution image. The higher the PSNR is, the better the quality of the reconstructed image is. The SSIM is a well-known quality metric used to measure the similarity between two images and decomposed into three factors: luminance, contrast and structure. AP is a metric of COCO-style.

## 3 EXPERIMENT AND ANALYSIS

### 3.1 Experimental details

We adopt a dataset for infrared detection and tracking of dim-small aircraft targets under ground/air background. It covers sky background and complex field background, includes 22 image sequences, 30 trajectories, 16177 frames and 16944 targets [16]. The target is an aerial fixed-wing UAV (fuel powered) with a fuselage length of 2.0m and a wingspan length of 2.6m. Each image has a resolution of 256 × 256 pixels.

We train YOLOX-S model for a total of 300 epochs with 5 epochs warmup, use stochastic gradient descent (SGD) for training. The training batch size is 64, the learning rate(lr) is  $\text{lr} \times \text{batch size} / 64$ , with an initial lr of 0.01 and the cosine lr schedule. The weight decay is 0.0005 and the SGD momentum is 0.9.

We remove any data that is invalid or does not contain a small target, setting 90% of the remaining data as the training set and the remaining 10% as the test set. Additionally, the YOLOX-S model requires images with an input size of 416416 pixels. The training process is accelerated by GPU.

The experimental platform’s operating system is Windows 10, the deep learning framework is Pytorch Real-ESRGAN and Pytorch YOLOX, the CPU is Intel Xeon Silver 4210R@ 2.40GHz, the GPU is an NVIDIA GeForce RTX 3090 with 64GB memory.

**Table 1: PSNR and SSIM results of image reconstruction**

Sequence	×1		×2	
	PSNR	SSIM	PSNR	SSIM
Data 1	40.219	0.982	37.685	0.929
Data 2	41.066	0.996	37.229	0.980
Data 3	42.054	0.997	41.209	0.996
Data 4	42.807	0.995	40.711	0.987
Data 5	39.381	0.980	38.407	0.973
Data 6	38.802	0.980	37.975	0.972
Data 7	39.498	0.984	38.620	0.978
Data 8	39.319	0.983	38.394	0.976
Data 9	39.571	0.984	38.700	0.978
Data 10	38.608	0.998	37.695	0.998
Data 11	39.193	0.983	38.283	0.976
Data 12	39.636	0.975	38.701	0.967
Data 13	39.757	0.982	38.857	0.975
Data 14	36.086	0.976	35.383	0.967
Data 15	36.710	0.967	36.189	0.958
Data 16	40.164	0.994	38.304	0.991
Data 17	37.141	0.988	36.222	0.982
Data 18	38.050	0.991	36.404	0.986
Data 19	39.969	0.994	38.714	0.991
Data 20	39.708	0.995	38.328	0.992
Data 21	38.203	0.995	36.948	0.992
Data 22	36.330	0.994	34.912	0.991

### 3.2 Contrast test of reconstruction image

The SR images of ×1 (256×256 pixels) and ×2 (512×512 pixels) are correspondingly reconstructed from the original images (256×256 pixels) to evaluate the impact of image reconstruction by PSNR and SSIM. When evaluating the quality of image reconstruction, the PSNR and SSIM are often determined by the reconstruction image and the real-world image. However, the PSNR and SSIM are calculated between the reconstructed image and the low-resolution original image, respectively, in order to evaluate the SR reconstruction effect with varied scale factors in this paper. As a result, the meaning of PSNR and SSIM’s values is diametrically opposed. Specifically, the higher the quality of the reconstruction image, the smaller the value.

The nearest-upsampled image with a scale factor of ×2 is used to calculate the PSNR and SSIM value for fair comparison (i.e., with the same resolution). By referring to [13] and contrasting the results of conventional interpolation techniques, it is discovered that nearest-neighbor interpolation generates small targets with shapes and highest gray values that are the closest to those in the original image.

The results of image reconstruction are shown in Table 1 for PSNR and SSIM, where each value corresponds to the mean value estimated from all images in the associated dataset sequence. The ×1 represents the results calculated between the SR image of ×1 and the original image, while the ×2 represents the results calculated between the SR image of ×2 and the nearest-upsampled image form the original image. As the table shows, PSNR and SSIM values of



**Figure 2: Comparison of image reconstruction of infrared small target (from data 6). (a) Origin 256×256;(b) Real-ESRGAN 256×256;(c) Nearest 512×512;(d) Real-ESRGAN 512×512**

×2 are all lower than those of ×1, which shows that the SR image with a larger scale factor can restore more information.

In Figure 2, image reconstruction of a small infrared target from data 6 is compared. (a) is the original image, (b) is the SR image of ×1, (c) is the nearest-upsampled image with a scale factor of ×2, and (d) is the SR image of ×2. The shape and greatest gray value of the small target in the SR image of 1 and the nearest-upsampled image are essentially unchanged, as can be seen from the local high-resolution image in the upper right corner of each sub-image. Although the nearest-upsampled image's small target size is 4 by 4 pixels instead of the original image's 2 by 2, there isn't much of an increase in the number of pixels with the greatest gray value. Due to the need of the same resolution, the nearest-upsampled image used to calculate PSNR and SSIM values is acceptable instead of the original image. The size and greatest gray value of the small target significantly increase for the SR image of ×2.

The shape and greatest gray value of the small target are almost same in the original image, the SR image of ×1, and the nearest-upsampled image, as shown in Figure 3's comparisons of image reconstruction of infrared small target from data 13. While the small target's shape for the SR image of ×2 is changed from a linear of two pixels to a rectangle of at least four pixels.

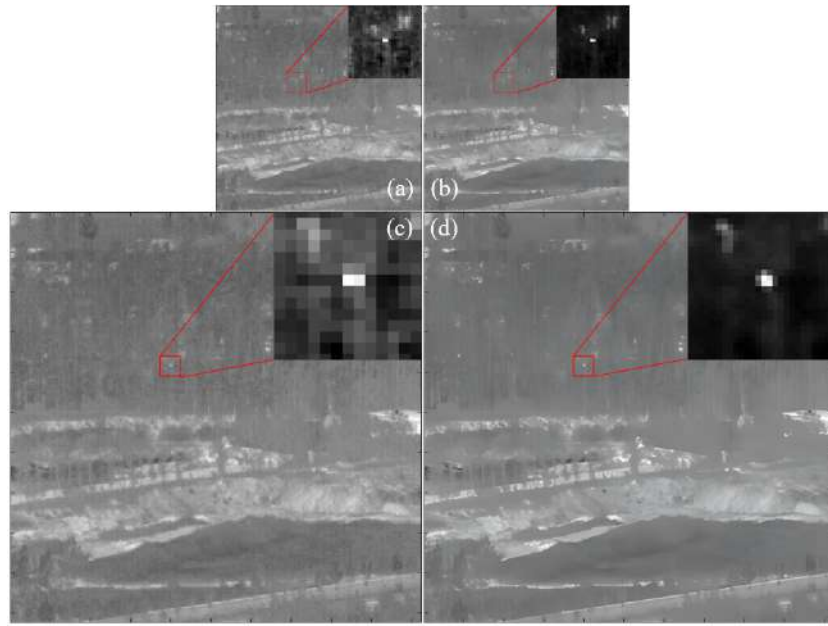
Similarly, Figure 3 shows comparisons of image reconstruction of infrared small target from data 13, and it can be seen that the shape and highest gray value of small target are almost the same in the original image, SR image of ×1 and nearest-upsampled image. While for the SR image of ×2, the shape of small target is changed from a linear of two pixels to a rectangle of at least 4 pixels, And the highest gray value has risen significantly from 177 to 212. Therefore, the Real-ESRGAN algorithm can retain the signals of small targets and enhance details of infrared image.

### 3.3 Contrast test of infrared small target detection

The original image and the SR image of ×1 and ×2 are utilized for the contrast test to verify the improvement of the YOLOX-S model detection accuracy by the SR image and the change of training strategies. We close MMDA for the last N epochs, N=180,140,100,60,15. For the original image, the SR image of ×1, and the SR image of ×2, the ground-truth size is set to 20×20, 20×20, and 40×40 pixels, respectively.

Figure 4 compares the AP results. The black triangles represent the achieved AP on the test set of original images, the best performance is approximately 63.7% AP when closing MMDA for the last 140, 100 and 60 epochs. The green circles represent the achieved AP on the test set of SR image of ×1, with the best performance of the detector even slightly declining to 63.1% AP. The red squares show the achieved AP on the test set of SR images of ×2, the best AP value is 77.19% when closing MMDA for the last 140 epochs, higher than closing MMDA for the last 15 epochs by 4.2% AP and the best performance achieved on the test set of original images by 13.49% AP, demonstrating the power of SR images with larger scale factors and closing MMDA at the right time.

The loss curves for red squares in Figure 4's results are shown in the figure 5(a). It demonstrates that when MMDA has been closed, all loss curves decreased quickly and the trends gradually became stable. The AP curves that correspond to the results of red squares in Figure 4 are depicted in Figure 5(b). As can be seen, once MMDA is closed, all AP values increase quickly, but the best detection performance improvement can be achieved when closing MMDA for the last 140 epoch.

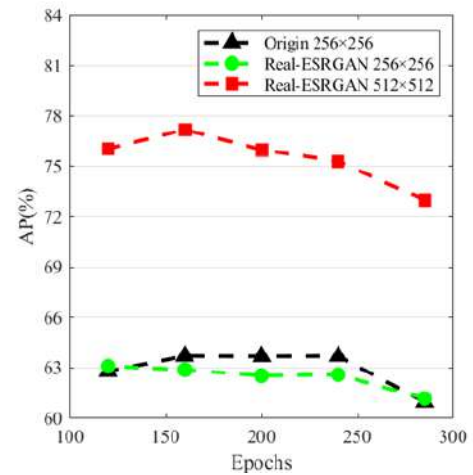


**Figure 3: Comparison of image reconstruction of infrared small target (from data 13). (a) Origin 256×256;(b) Real-ESRGAN 256×256;(c) Nearest 512×512;(d) Real-ESRGAN 512×512**

In fact, the training results are significantly impacted by the size of the ground-truth, which can have an impact on the detector’s ability to extract contextual features from both the small target and complex backgrounds. We train YOLOX-S model on the data of SR image of  $\times 2$  and close MMDA for the last 140 epochs. The detector achieves the performance of 68.77% AP, 72.60% AP, 74.90% AP, 77.19% AP and 78.62% AP when the size of ground-truth is set to  $28 \times 28$ ,  $32 \times 32$ ,  $36 \times 36$ ,  $40 \times 40$  and  $44 \times 44$  pixels respectively. We can see that the performance of the detector is greatly enhanced when the size of the ground-truth is increased from  $28 \times 28$  to  $40 \times 40$  pixels, and then somewhat improved when the size is increased from  $40 \times 40$  to  $44 \times 44$  pixels. We select  $40 \times 40$  pixels as an appropriate size for the ground-truth when the YOLOX-S model is trained and tested using the data from an SR image of  $\times 2$ . This decision is based on the performance of small target tracking in the future.

#### 4 CONCLUSIONS

In our study, the YOLOX-S model, which is based on single-image SR reconstruction and training strategy optimization, is used to enhance the capacity to detect infrared small targets in complex backgrounds, particularly for low-resolution and low-quality infrared images. Real-ESRGAN is used to reconstruct a high-resolution infrared image from its low-resolution counterpart, which will be taken as the input of YOLOX-S. It can retain the signals of small targets and enhance details of infrared image, which can improve the YOLOX-S’s detection ability. The training strategy of YOLOX-S is further optimized during the training process to make it more appropriate for the detection of infrared small targets, including closing MMDA for the last 140 epochs and altering the size of the ground-truth to  $40 \times 40$  pixels. The AP achieved by the suggested

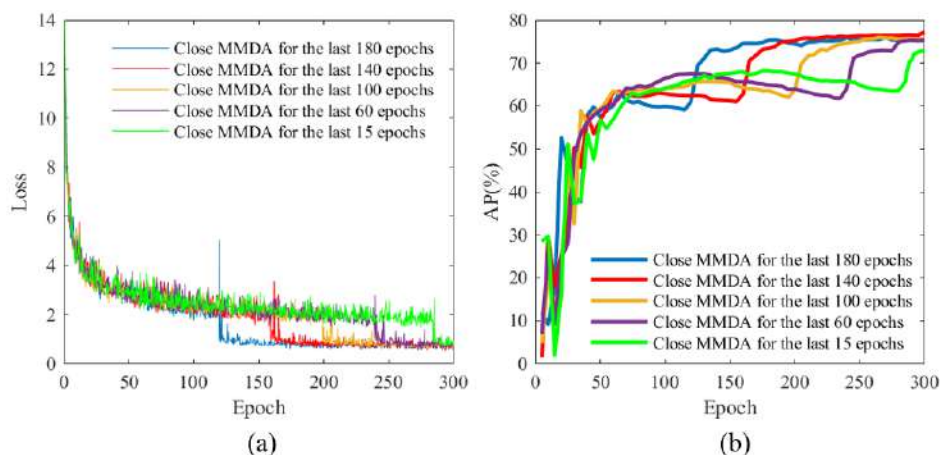


**Figure 4: Comparison of AP results. We close MMDA for the last N epochs,  $N=180,140,100,60,15$ . The black triangles, the green circles and the red squares represent the achieved AP on the test set of original images, the SR image of  $\times 1$  and  $\times 2$  respectively.**

method in this work increases from 63.70% to 77.19% after comprehensive experimental verification and comparison, demonstrating a notable improvement in infrared small target detection.

#### ACKNOWLEDGMENTS

This work is funded by the Fundamental Strengthening Technology Fund (Grant No. 2021-JCJQ-JJ-0739).



**Figure 5: (a) Loss curves; (b) AP curves. We train YOLOX-S model on the data of SR image of  $\times 2$  and close MMDA for the last  $N$  epochs,  $N=180,140,100,60,15$ .**

## REFERENCES

- [1] Hong Zhang, Lei Zhang, Ding Yuan, Hao Chen. 2017. Infrared small target detection based on local intensity and gradient properties. *Infrared Physics & Technology*, 89:88-96. <https://doi.org/10.1016/j.infrared.2017.12.018>
- [2] Wang X Y. 2018. Research on infrared dim and small target detection theory and methodology based on sparse dynamic inversion. Chengdu: University of Electronic Science and Technology of China.
- [3] Gujing Han, Tao Li, Qiang Li, Feng Zhao, Min Zhang, Ruijie Wang, Qiwei Yuan, Kaipei Liu and Liang Qin. 2022. Improved Algorithm for Insulator and Its Defect Detection Based on YOLOX. *Sensors*,22(16): 6186. <https://doi.org/10.3390/s22166186>
- [4] Zhang S Y. 2021. Research on Image SR Reconstruction Based on Generative Adversarial Network. China University of Mining and Technology. XuZhou: China University of Mining and Technology.
- [5] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*,39(6):1137-1149.<https://doi.org/10.1109/TPAMI.2016.2577031>
- [6] He, K, Gkioxari, G, Dollár, P, and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961-2969.
- [7] Redmon, Joseph, and Ali Farhadi. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- [8] Bochkovskiy A, Wang C Y, Liao H. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv preprint arXiv:2004.10934*, 2
- [9] Glenn Jocher *et al.* 2021. yolov5, <https://github.com/ultralytics-cs/yolov5>
- [10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun. 2021. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv: 2107. 08430*.
- [11] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu & Alexander C. Berg. 2016. Ssd: Single shot multibox detector. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing, 2016.
- [12] Xintao Wang, Liangbin Xie, Chao Dong, Ying Shan. 2021. Real-esrgan: Training real-world blind SR with pure synthetic data. 2021 *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, October 11-17, 2021, Montreal, BC, Canada, IEEE: 21442270.
- [13] X Zhou, L Jiang, C Hu, S Lei, T Zhang, X Mou. 2022. YOLO-SASE: An Improved YOLO Algorithm for the Small Targets Detection in Complex Backgrounds. *Sensors*, 22(12): 4600. <https://doi.org/10.3390/s22124600>
- [14] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. *Computer Vision – ECCV 2018 Workshops*, 11133: 63–79.
- [15] MA Yousuf, and MN Nobi. 2011. A new method to remove noise in magnetic resonance and ultrasound images. *Journal of scientific research*, 81-81. <https://doi.org/10.3329/jsr.v3i1.5544>
- [16] B Hui, Z Song, H Fan, P Zhong, We Hu, X Zhang, J Lin, H Su, W Jin, Y Zhang, Y Bai. 2019. A dataset for infrared image dim-small aircraft target detection and tracking under ground/air background. *Science Data Bank*. <https://doi.org/10.11922/sciencedb.902>. DOI:10.11922/sciencedb.902.

# An Encryption Scheme Using Multi-Scroll Memristive Chaotic System

Fan Wu, Musha Ji'e, Lidan Wang, Shukai Duan

College of Artificial Intelligence, Southwest University, Chongqing 400715, China

## ABSTRACT

In this paper, a novel multi-scroll memristive chaotic system is designed based on Chua's system via introducing a nonlinear memristor. The dynamics of this system is analyzed based on bifurcation diagrams, Lyapunov exponents and phase diagrams. Subsequently, an image encryption scheme based on this system is then proposed. First, the proposed chaotic system is used to generate continuously robust chaotic sequences, the hash values of plaintext images are embedded in the generation and selection of chaotic sequences and involved in each step of encryption to establish the coupling relationship between plaintext and ciphertext. Second, Knuth-Durstenfeld algorithm is used to scramble the high four-bit plane of the plain image twice, and the chaotic sequence is used as the index sequence, which greatly improves the efficiency and randomness of the permutation process. Finally, chaotic sequences are involved in DNA coding rules and pixel-level diffusion. The algorithm is highly sensitive to plain images, and it can realize adaptive encryption. Through performance analysis and comparison with recent literature, the proposed algorithm can cope with various attacks and show excellent performance.

## CCS CONCEPTS

• **Security and privacy** → Cryptography; Symmetric cryptography and hash functions; Cryptography; Symmetric cryptography and hash functions; Hash functions and message authentication codes.

## KEYWORDS

Multi-scroll attractor, Image encryption, Secure hash algorithm, Shuffling algorithm, DNA coding

## ACM Reference Format:

Fan Wu, Musha Ji'e, Lidan Wang, Shukai Duan. 2023. An Encryption Scheme Using Multi-Scroll Memristive Chaotic System. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590105>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590105>

## 1 INTRODUCTION

In recent years, memristive chaotic systems have attracted wide attention from scholars in the field of chaos. In 1984, Chua proposed Chua's system with simple structure, easy implementation and complex dynamics [1], and proved in [2] that this system can generate double-scroll attractors. Thereafter, Chua replaced diode in the Chua's system with a memristor model and constructed the memristive chaotic system first [3]. Chaotic systems are very sensitive to the initial state, pseudo-random and unpredictable, which is very suitable for image encryption. Fridrich first proposed image encryption based on chaotic systems in 1998 [4] and constructed the image encryption 'permutation-diffusion' structure. Since then, many image encryption schemes based on chaotic systems and permutation-diffusion have been proposed. In terms of image permutation, the classical Hilbert curve, Baker transformation, and Arnold transformation are used for permutation [5-7]. In terms of image diffusion, due to the low power consumption of DNA technology in biology, large-scale parallel computing can be realized, and chaotic sequence combined with DNA can effectively reduce the correlation of image pixels. Many scholars have carried out extensive research on chaotic image encryption based on DNA coding operation [8-10].

However, there are still many unsatisfactory schemes. For example, Yu et al [11] designed an image encryption scheme based on logistic chaotic mapping, due to the one-dimensional system used, the smaller key space cannot resist exhaustive attacks although the iteration speed is extremely fast. Refs. [12] proposed an encryption algorithm based on a six-dimensional hyperchaotic system, which consumes excessive resources despite the key space is large. Some of the current DNA coding-based image encryption also has shortcomings. An image encryption algorithm based on DNA and chaotic mapping was proposed in Ref. [13], the initial value of the chaotic system of this algorithm depends on pixel value of the plaintext image, which can be easily obtained by adversaries and has potential security risks. Ref. [14] proposed an image encryption algorithm based on DNA coding and spatio-temporal chaos, the algorithm scrambles the pixel positions of the image by DNA bases, the key is not involved in the diffusion part, so the computational complexity is low while the security of the encryption cannot be guaranteed.

Based on the inspiration discussed above, we construct a memristive multi-scroll chaotic system and design a scheme that can implement adaptive encryption by combining shuffling algorithm and DNA sequence. This paper has the following three main contributions: first, a 4D memristive chaotic system that can generate any number of multi-scroll is constructed by directly coupling a nonlinear flux-controlled memristor with the Chua's system without changing the remaining arbitrary terms. This system is capable

of generating a rich dynamical behavior by adjusting the memristive coupling strength and adjusting the initial values. Second, the efficiency of the algorithm was improved by scrambling the image twice by shuffling algorithm on the high four-bit plane, and the double diffusion scheme of DNA and pixel level was used to ensure the performance of the algorithm. Finally, the proposed encryption algorithm embeds the hash value of the plaintext into each link of permutation and diffusion by SHA-256. The algorithm not only makes reasonable use of the pseudo-random property of chaotic sequences, but also is highly sensitive to the plaintext images, which can well resist various attacks and realize adaptive encryption.

The structure of this paper is as follows: a multi-scroll memristive chaotic system is proposed and analyzed in Section 2. The proposed image encryption scheme was introduced in Section 3. Section 4 shows the simulation results and security analysis. Section 5 concludes this paper.

## 2 MEMRISTIVE CHAOTIC SYSTEM

### 2.1 The memristive multi-scroll chaotic system

Zhang et al. [15] proposed a nonlinear memristor model:

$$\begin{cases} \dot{i} = u(w)v \\ \dot{w} = av - bh(w) \\ u(w) = c + dh(w) \end{cases} \quad (1)$$

Coupling the memristor model (1) directly into the third equation of state of Chua's system [2] and our system is modeled and represented as:

$$\begin{cases} \dot{x} = \alpha(y - f(x)) \\ \dot{y} = x - y + z \\ \dot{z} = -\beta y - kz(u(w)) \\ \dot{w} = az - bh(w) \end{cases} \quad (2)$$

where  $f(x) = m_1x + \frac{1}{2}(m_2 - m_1)(|x + 1| - |x - 1|)$ .  $a \sim d$  are four control parameters.  $\alpha, \beta, m_1$ , and  $m_2$  are system parameters, and the parameter  $k$  denotes the memristive coupling strength. Depending on the number of multi-scroll attractors obtained,  $u(w)$  can be represented as:

$$\begin{cases} u_1(w) = c + dh_1(w) \\ u_2(w) = c + dh_2(w) \end{cases} \quad (3)$$

$h(w)$  can be written as [16]:

$$h(w) = \begin{cases} h_1(w) = \begin{cases} w, N = 0 \\ w - \sum_{i=1}^N (\text{sgn}(w - 1 + 2i) + \text{sgn}(w + 1 - 2i)) \\ N = 1, 2, 3, \dots \end{cases} \\ h_2(w) = \begin{cases} w - \text{sgn}(w), M = 0 \\ w - \text{sgn}(w) - \sum_{j=1}^M (\text{sgn}(w + 2j) + \text{sgn}(w - 2j)) \\ M = 1, 2, 3, \dots \end{cases} \end{cases} \quad (4)$$

$M$  and  $N$  in Eq. (4) are the controlled variable of the memristor model. Different odd or even multi-scroll attractors can be generated by selecting different  $N$  or  $M$  values. Figure 1 shows the phase portraits of different positions.

### 2.2 Dynamic behaviors of $k$ and $w$

The parameter is set to  $M = 2, \alpha = 8, \beta = 10, m_1 = 2/7, m_2 = -1/7, a = c = 1, b = 1.73, d = 0.03$  and the initial value is  $(0, 0, 0.1, 0)$ . When the memristive coupling parameter  $k$  varies between  $[0.1, 0.25]$ , the system exhibits complex dynamical behavior. The bifurcation diagram and the Lyapunov exponents of  $w$  are shown in Figure 2. The system experiences robust chaotic states with one positive Lyapunov exponents over the entire region. It can also be seen that there is a one-to-one correspondence between Lyapunov exponents and bifurcation diagrams.

When keeping the previous parameters unchanged and changing only  $b = 3$  and setting the parameter  $k = 0.2$ , the initial value is set to  $(0.1, 0, 0, w(0))$ . When  $w(0)$  varies between  $(-6, 6)$ , the bifurcation diagram and the Lyapunov exponents of  $w$  are shown in Figure 3. It can be seen that the bifurcation diagram has six long strips, and each strip is in a step ascending state. In addition, the Lyapunov exponents remains constant and has continuous and robust properties. In addition, when  $w(0)$  takes  $-5, -3, -1, 1, 3, 5$  in sequence, six coexisting chaotic attractors can be seen, and its phase portraits and corresponding time series are shown in Figure 4.

## 3 THE PROPOSED ENCRYPTION SCHEME

The framework of the proposed encryption scheme is shown in Figure 5.

### 3.1 Generation and selection of chaotic sequence

The plaintext image generates a 256-bit hash value by using SHA-256, it is processed into 32 decimal numbers and represented as follows:

$$K = \{k_1, k_2, \dots, k_{32}\} \quad (5)$$

$K$  is divided into two parts, as shown in Eq. (6). Eq. (7) is used to transform  $K1$  into  $w_0$  of the chaotic system, the initial value of chaotic system.  $K2$  is used to perform pre-iterations on the system to avoid transient effects and improve safety. First set the base value  $n_1$  for the pre-iteration value  $n$ , set  $n_1 = 1000, n_2 \in (0, 1000)$ , the pre-iterative value  $n$  is obtained by summing  $n_1$  and  $n_2$  by Eq. (8).

$$\begin{cases} K1 = (k_1 + k_3 + k_5 + \dots + k_{31}) \\ K2 = (k_2 + k_4 + k_6 + \dots + k_{32}) \end{cases} \quad (6)$$

$$w_0 = \text{mod}(K1, 7) \quad (7)$$

$$n = n_1 + n_2 = 1000 + \text{mod}(K2, 1000) \quad (8)$$

Set the plain image size to  $M \times N$ . Based on the pre-iteration  $n$  times, the 4D chaotic system iterates  $4 \times M \times N - 1$  times to obtain 4 chaotic sequences  $X, Y, Z$  and  $W$ , all with lengths of  $4 \times M \times N - 1$ .

### 3.2 Double bit-level permutation

To reduce the correlation between adjacent pixels of ciphertext, scrambling is carried out on the bit level of pixels. The gray image  $P$  is decomposed into 8 bit planes. In order to improve the encryption efficiency of the algorithm, the highest four bit planes carrying the most effective information are selected for two bit-level scrambling operations.

Firstly, the four selected bit planes are connected into a one-dimensional binary sequence  $P_1$  of size  $4 \times M \times N$ . To further enhance

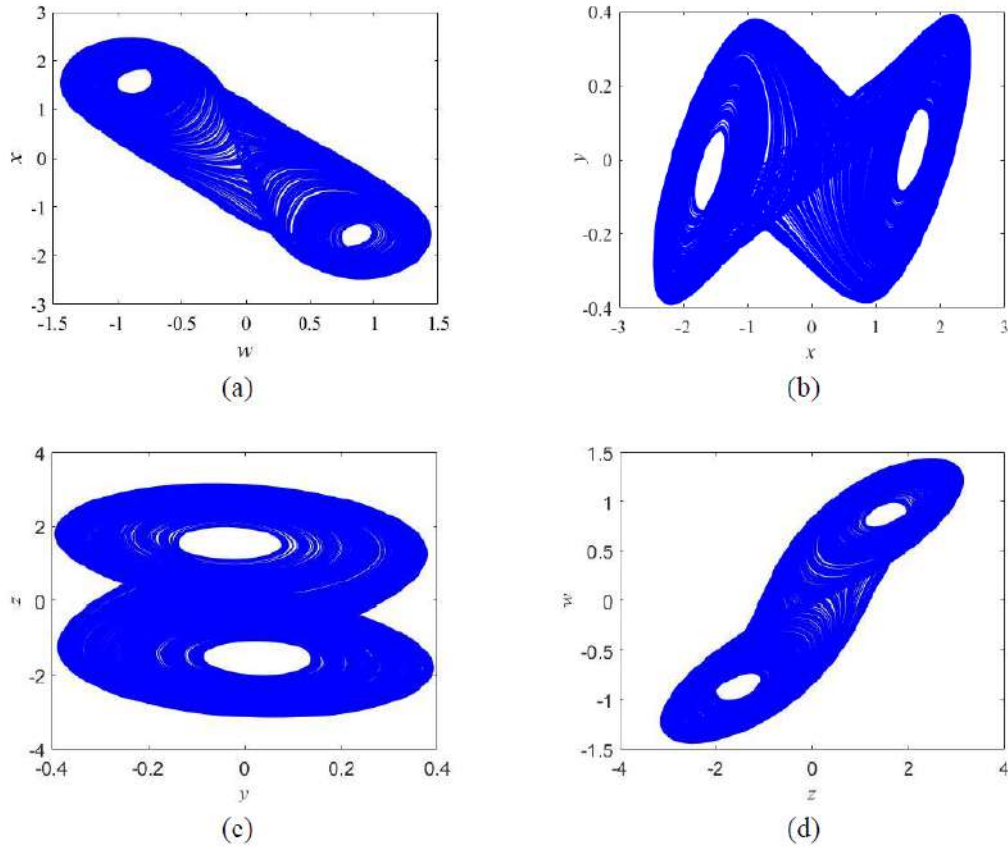


Figure 1: Phase portraits of the attractors at each position of the chaotic system (3). (a)  $w$ - $x$  plane; (b)  $x$ - $y$  plane; (c)  $y$ - $z$  plane; (d)  $z$ - $w$  plane.

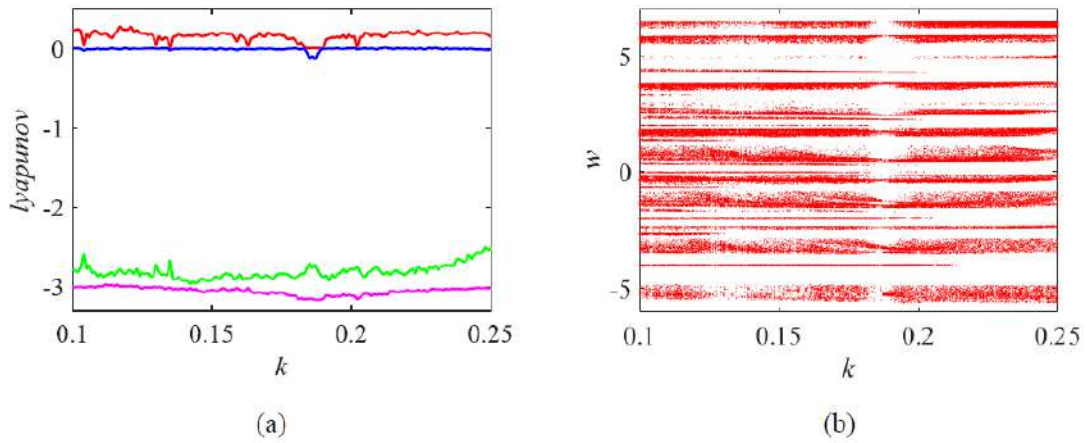


Figure 2: The dynamics of the memristor coupling strength  $k$ : (a) Lyapunov exponents (b) bifurcation diagram.

the coupling between the encryption algorithm and the plaintext image, Eq. (9) is used to define the index and Eq. (10) is used to select the desired chaotic sequence  $A$  and  $B$  for scrambling.

$$\text{index} = \text{mod}(\text{sum}(K), 6) + 1 \quad (9)$$

$$\begin{cases} \text{whenindex} = 1, \text{ then } A = X, B = Y \\ \text{whenindex} = 2, \text{ then } A = X, B = Z \\ \text{whenindex} = 3, \text{ then } A = X, B = W \\ \text{whenindex} = 4, \text{ then } A = Y, B = Z \\ \text{whenindex} = 5, \text{ then } A = Y, B = W \\ \text{whenindex} = 6, \text{ then } A = Z, B = W \end{cases} \quad (10)$$

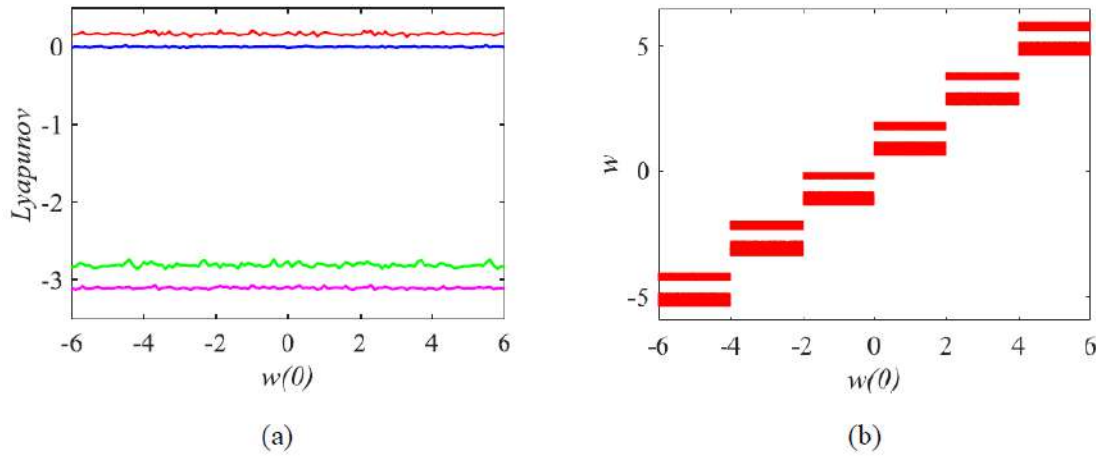


Figure 3: Dynamics of the system with initial value  $w(0)$ : (a) Lyapunov exponents (b) bifurcation diagram.

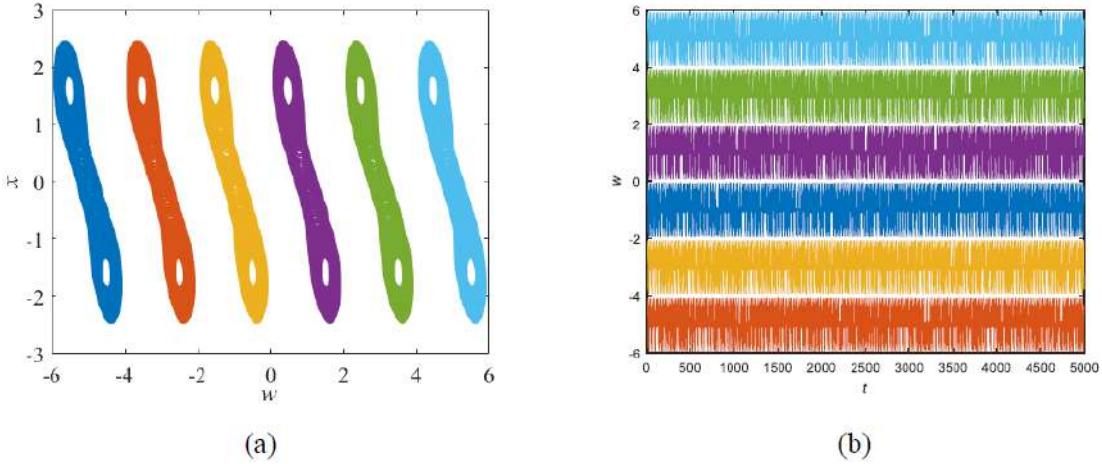


Figure 4: Phase portraits and time series of six coexisting chaotic attractors: (a) six coexisting attractors (b) time series of the state  $w$ .

The two selected chaotic sequences  $A$  and  $B$  are then processed into vector  $V1$  and vector  $V2$  by Eq. (11). both  $V1$  and  $V2$  are of length  $4 \times M \times N - 1$ .

$$\begin{cases} V1(i) = \text{mod}(\text{floor}((A(i) + 100) \times 10^{10}), 4 \times M \times N - i + 1) + 1 \\ V2(i) = \text{mod}(\text{floor}((B(i) + 100) \times 10^{10}), 4 \times M \times N - i + 1) + 1 \end{cases} \quad (11)$$

Then  $V1$  and  $V2$  perform two times bit-level permutations of  $P1$  according to the Knuth-Durstenfeld algorithm. The two permutation operations are shown in Eq. (12). Transform the vector  $P1$  into a matrix of  $4 \times M \times N$  after the first Knuth-Durstenfeld permutation, and transpose it. Then, the transpose matrix is expanded to obtain the one-dimensional vector  $P2$ , after which the second Knuth-Durstenfeld permutation operation is performed.

$$\begin{cases} P1(V1(i)) = P1(4 \times M \times N - i + 1) \\ P2(V2(i)) = P2(4 \times M \times N - i + 1) \end{cases} \quad (12)$$

Finally, the vector  $P2$  is transformed into 4 bit planes by Zigzag according to Eq. (13) after the scrambling is completed and merged with the bit planes not involved in scrambling to obtain the scrambled image  $Q1$ .

$$\begin{aligned} Data\{5\} &= \text{Zigzag}(P2(1 : M \times N)) \\ Data\{6\} &= \text{Zigzag}(P2(M \times N + 1 : 2 \times M \times N)) \\ Data\{7\} &= \text{Zigzag}(P2(2 \times M \times N + 1 : 3 \times M \times N)) \\ Data\{8\} &= \text{Zigzag}(P2(3 \times M \times N + 1 : 4 \times M \times N)) \end{aligned} \quad (13)$$

### 3.3 DNA and pixel-level diffusion

The scrambled image  $Q1$  performs DNA diffusion and pixel-level diffusion operation in this session. To further enhance the randomness of encryption, we determine the position of the chaotic sequence used for DNA diffusion with hash value  $K$  by Eq. (14) and Eq. (15) and obtain four chaotic sequences  $X1$ ,  $Y1$ ,  $Z1$ ,  $W1$  of

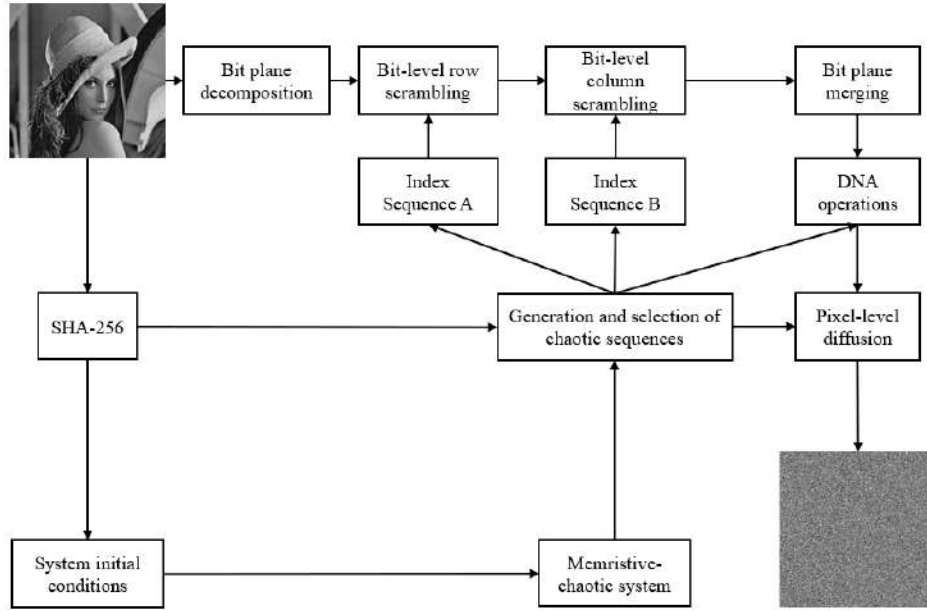


Figure 5: Framework diagram of the proposed scheme.

length  $M \times N$ ,  $s$  is an integer between  $[0, 2]$ .

$$s = \text{mod}(\text{sum}(K), 3) \quad (14)$$

$$\begin{cases} X1 = X(s \times M \times N + 1: (s+1) \times M \times N) \\ Y1 = Y(s \times M \times N + 1: (s+1) \times M \times N) \\ Z1 = Z(s \times M \times N + 1: (s+1) \times M \times N) \\ W1 = W(s \times M \times N + 1: (s+1) \times M \times N) \end{cases} \quad (15)$$

Perform the operation of Eq. (16) on the chaotic sequence selected by Eq. (15), sequences of rules for performing DNA operations can be obtained.

$$\begin{cases} Ex(i) = \text{floor}(\text{mod}(X1(i) \times 10^{10}, 8)) + 1 \\ Ey(i) = \text{floor}(\text{mod}(Y1(i) \times 10^{10}, 8)) + 1 \\ Ez(i) = \text{floor}(\text{mod}(Z1(i) \times 10^{10}, 8)) + 1 \\ E(i) = \text{floor}(\text{mod}(W1(i) \times 10^{10}, 256)) \end{cases} \quad (16)$$

Expanding  $Q1$  into a 1D vector  $R$ , according to the DNA coding rules of  $Ex$  and  $Ey$ , sequence  $R$  and  $E$  were encoded as DNA sequence  $R_{DNA}$  and  $E_{DNA}$  of size  $4 \times M \times N$ , respectively. The DNA sequence  $RE_{DNA}$  was then obtained by performing DNA XOR on the  $R_{DNA}$  and  $E_{DNA}$ . Finally, the  $RE_{DNA}$  is decoded and converted to an intermediate encrypted image  $Q2$  according to the DNA decoding rules of the sequence  $Ez$ .

In the pixel-level diffusion stage, select the chaotic sequence  $H$  from the sequence  $X$  from  $(s+1)(M \times N + 1)$  to  $(s+2)(M \times N)$ , and then processed as a discrete sequence by Eq. (17). The intermediate encrypted image is then converted into an encrypted image  $C$  of  $M \times N$  matrix by pixel-level diffusion through Eq. (18) and Eq. (19).

$$H(i) = \text{mod}(\lfloor (|h_i| - \lfloor |h_i| \rfloor) \times 10^{14} / 10^8 \rfloor, 256) \quad (17)$$

$$C^1 = Q_2^1 \oplus \text{mod}(\text{sum}(Q2), 256) \oplus H^1 \quad (18)$$

$$C^i = Q_2^i \oplus C^{i-1} \oplus H^i \quad (19)$$

The decryption process can be achieved by performing the inverse operation of the encryption algorithm.

## 4 EXPERIMENTAL RESULTS AND SECURITY ANALYSIS

The proposed encryption scheme uses standard Barbara (512×512), Lena (512×512), Peppers (512×512) images for the test. All experiments were implemented in MATLAB R2016a compiled environment.

### 4.1 Simulation results

Take Lena (512×512) as an example, the experimental parameters are  $x_0=0.1$ ,  $y_0=z_0=0$ ,  $w_0=6$ ,  $a=c=1$ ,  $b=1.73$ ,  $d=0.03$ ,  $k=0.2$ ,  $\alpha=8$ ,  $\beta=10$ ,  $N=1$ . The original image, cipher image and the decrypted image are shown in Figure 6.

### 4.2 Key space

For our encryption scheme proposed, one part of the key space is the 256-bit hash value with the key space is  $2^{256}$ , and the other part is the initial parameters of the chaotic system, whose key space is  $(10^{15})^4$ . The total key space is  $2^{256} \times (10^{15})^4 \approx 2^{455}$ , which is much larger than  $2^{100}$  [17], so all brute force attacks are invalid.

### 4.3 Histogram

Image histograms visualize information features by counting the distribution of image pixel values. Figure 7 statistics the pixel distribution of the two plain images and their encrypted images respectively. The encrypted image don't contain any statistical information related to the plain image. The encryption system in this paper has excellent performance in terms of histogram.

### 4.4 Correlation coefficient

Figure 8 takes Lena image as an example to test the correlation in the diagonal, vertical and horizontal directions. It can be clearly

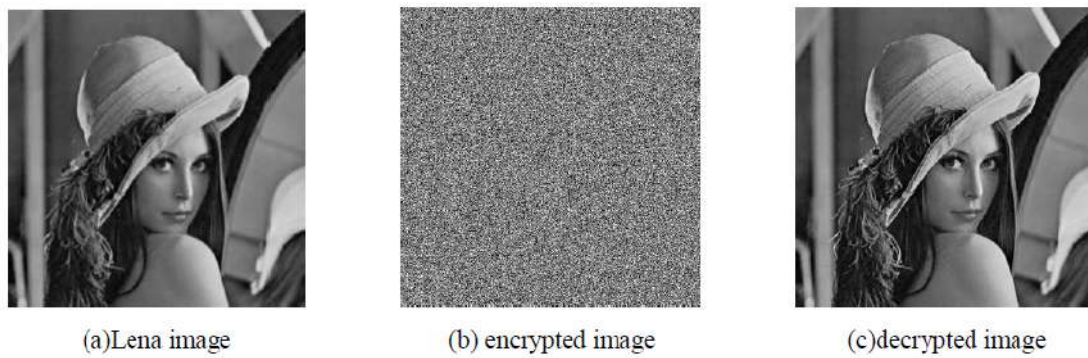


Figure 6: Experimental results of encryption and decryption.

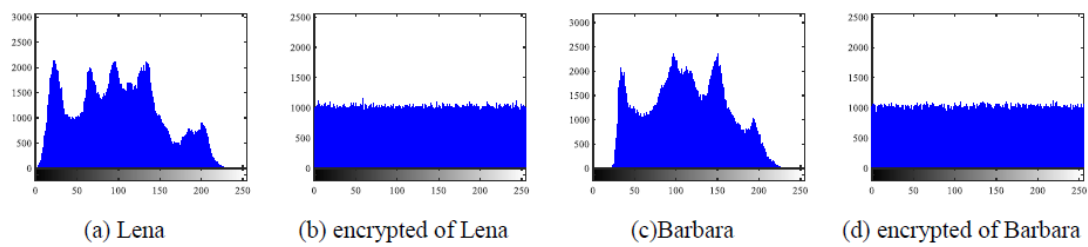


Figure 7: Histogram comparison of plaintext and encrypted image

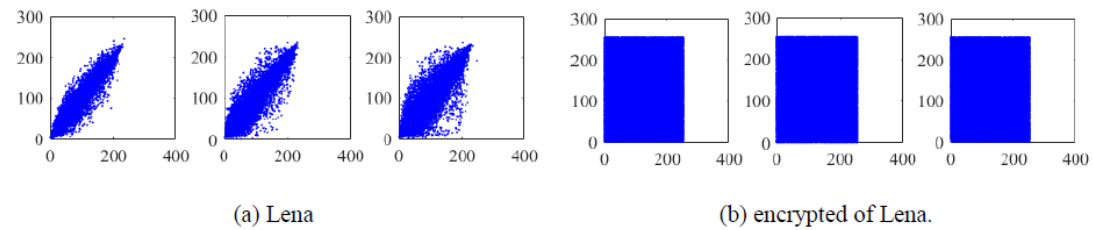


Figure 8: The correlation analysis in diagonal, vertical and horizontal directions.

Table 1: Correlation coefficients of plaintext and encrypted images.

Image	Plaintext image			Cipher images		
	Horizontal	Vertical	Diagonal	Horizontal	Vertical	Diagonal
Lena	0.9848	0.9739	0.9594	-7.9866e-05	-8.2037e-04	0.0050
Barbara	0.9590	0.8586	0.8415	-4.3810e-04	-7.8883e-04	-0.0017
Peppers	0.9836	0.9810	0.9703	-0.0020	0.0011	-0.0094

seen that the correlation of plaintext is strong, but weak for cipher images. Table 1 shows the quantified results of correlation coefficients of different images. The correlation coefficients of cipher images are close to 0.

#### 4.5 Information entropy

For image encryption, the information entropy reflects the randomness of pixel. The closer the entropy value of the cipher image is to

8, the better the encryption effect is. Table 2 lists the information entropy, which is close to 8 for the cipher image.

#### 4.6 Differential attack resistance

An attacker can break the encryption system by changing the plaintext to compare the encryption results. To resist differential attacks, a good encryption system should ensure that any small change in the plaintext image will result in a completely different ciphertext image. NPCR and UACI are commonly used to quantify resistance

**Table 2: Information entropy of encrypted images.**

Image	Information entropy	
	Plaintext image	Cipher image
Lena(512×512)	7.5929	7.9994
Barbara(512×512)	7.4664	7.9992
Peppers(512×512)	7.5715	7.9992

**Table 3: NPCR values of different images.**

Image name	Minimum(%)	Maximum(%)	Mean(%)
Lena	99.5819	99.6372	99.6077
Barbara	99.5842	99.6387	99.6079
Peppers	99.5831	99.6353	99.6109

**Table 4: UACI values of different images.**

Image name	Minimum(%)	Maximum(%)	Mean(%)
Lena	33.3568	33.6477	33.4693
Barbara	33.3772	33.5836	33.4665
Peppers	33.3354	33.5426	33.4552

**Table 5: Performance comparison with other schemes.**

Algorithms	Entropy	NPCR	UACI	Correlation coefficient		
				Horizontal	Vertical	Diagonal
Ref. [18]	7.9993	99.61	33.50	0.0037	-0.0004	-0.0378
Ref. [19]	7.9993	99.6084	33.4714	-0.21e-04	10e-04	-0.57e-04
Ref. [20]	7.9993	99.6069	33.4558	-0.0298	-0.0004	-0.0169
Ref. [21]	7.9993	99.6369	33.4335	1.4e-04	8e-04	6.7e-04
Ref. [22]	7.9994	99.6037	33.4463	3.4459e-04	-0.0064	0.0110
Our scheme	7.9994	99.6077	33.4693	-7.9866e-05	-8.2037e-04	-0.0050

to differential attacks. We perform 100 experiments on 3 images to eliminate randomness. The maximum, minimum and average values of NPCR and UACI are shown in Table 3 and Table 4 for each plaintext image by randomly changing 1-bit pixel value of one pixel and repeating 100 times. The theoretical values of NPCR and UACI are 99.6094% and 33.4635%, respectively. We can see that our experimental results are very close to it.

#### 4.7 Performance comparison with other algorithms

To eliminate the effects of other factors, we used Lena (512×512) for both the tested image and the references used for comparison. Table 5 compares several algorithms with excellent performance from 2018 to 2022 based on correlation coefficient, NPCR and UACI, information entropy. It can be seen that our scheme has a good performance.

## 5 CONCLUSION

In this paper, a multi-scroll memristive chaotic system is designed by coupling a memristor model with the Chua's system, and some brief dynamical analysis of the system is presented. On this basis, we propose a novel image encryption algorithm by combining secure hash algorithm, Knuth-Durstenfeld algorithm and DNA coding operation. The hash value of the ordinary image is embedded into the system, and the chaotic sequence is generated and selected by SHA-256 to establish the coupling relationship between plaintext and ciphertext. In the permutation stage, the permutation operation is performed at the bit level with the purpose of hiding the statistical characteristics of the original image and reducing the correlation of adjacent pixels. The chaotic sequence is used as the index sequence of the bit-level twice Knuth-Durstenfeld shuffle permutation. In the DNA coding operation phase, chaotic sequences are used as one-to-one encoding and decoding rules. Finally, the intermediate cipher image generated by DNA decoding and the chaotic sequence were diffused at pixel level to obtain the final encrypted image. The experimental results and analysis show that the proposed algorithm

can effectively resist common attacks. By comparing with various algorithms in recent years, the effectiveness and practicability of the proposed scheme are proved, and it is suitable for the field of image encryption.

## ACKNOWLEDGMENTS

Project supported by the National Key Research and Development Program of China (Grant No. 2018YFB1306600), the National Natural Science Foundation of China (Grant Nos. 62076207, 62076208, and U20A20227).

## REFERENCES

- [1] Matsumoto, T., *A chaotic attractor from Chua's circuit*. IEEE Transactions on Circuits and Systems, 1984. **31**(12): p. 1055-1058.
- [2] Chua, L., M. Komuro, and T. Matsumoto, *The double scroll family*. IEEE transactions on circuits and systems, 1986. **33**(11): p. 1072-1118.
- [3] Itoh, M. and L.O. Chua, *Memristor oscillators*. International journal of bifurcation and chaos, 2008. **18**(11): p. 3183-3206.
- [4] Fridrich, J. Image encryption based on chaotic maps. in 1997 IEEE international conference on systems, man, and cybernetics. Computational cybernetics and simulation. 1997. IEEE.
- [5] Zhou, N.R., et al., *Quantum image encryption based on generalized Arnold transform and double random-phase encoding*. Quantum Information Processing, 2015. **14**(4): p. 1193-1213.
- [6] Mao, Y., G. Chen, and S. Lian, *A novel fast image encryption scheme based on 3D chaotic baker maps*. International Journal of Bifurcation and chaos, 2004. **14**(10): p. 3613-3624.
- [7] Zhang, X., et al., *A chaos-based image encryption technique utilizing hilbert curves and H-fractals*. IEEE Access, 2019. **7**: p. 74734-74746.
- [8] Chai, X., Y. Chen, and L. Broyde, *A novel chaos-based image encryption algorithm using DNA sequence operations*. Optics and Lasers in engineering, 2017. **88**: p. 197-213.
- [9] Chai, X., et al., *An image encryption algorithm based on the memristive hyperchaotic system, cellular automata and DNA sequence operations*. Signal Processing: Image Communication, 2017. **52**: p. 6-19.
- [10] Liang, Z., et al., *Color image encryption algorithm based on four-dimensional multi-stable hyper chaotic system and DNA strand displacement*. Journal of Electrical Engineering & Technology, 2022: p. 1-21.
- [11] Yu, J., et al., *Image encryption algorithm by using the logistic map and discrete fractional angular transform*. Optica Applicata, 2017. **47**(1).
- [12] Wu, X., et al., *A novel lossless color image encryption scheme using 2D DWT and 6D hyperchaotic system*. Information Sciences, 2016. **349**: p. 137-153.
- [13] Dou, Y., et al., *Cryptanalysis of a DNA and chaos based image encryption algorithm*. Optik, 2017. **145**: p. 456-464.
- [14] Wen, H., S. Yu, and J. Lü, *Breaking an image encryption algorithm based on DNA encoding and spatiotemporal chaos*. Entropy, 2019. **21**(3): p. 246.
- [15] Zhang, S., et al., *Generating any number of initial offset-boosted coexisting chua's double-scroll attractors via piecewise-nonlinear memristor*. IEEE Transactions on Industrial Electronics, 2021. **69**(7): p. 7202-7212.
- [16] Zhang, S., et al., *Initial offset boosting coexisting attractors in memristive multi-double-scroll Hopfield neural network*. Nonlinear Dynamics, 2020. **102**(4): p. 2821-2841.
- [17] Alvarez, G. and S. Li, *Some basic cryptographic requirements for chaos-based cryptosystems*. International journal of bifurcation and chaos, 2006. **16**(08): p. 2129-2151.
- [18] Zefreh, E.Z., *An image encryption scheme based on a hybrid model of DNA computing, chaotic systems and hash functions*. Multimedia Tools and Applications, 2020. **79**(33): p. 24993-25022.
- [19] Erkan, U., et al., *2D  $e\pi$ -map for image encryption*. Information Sciences, 2022. **589**: p. 770-789.
- [20] Zhang, Y., *The image encryption algorithm based on chaos and DNA computing*. Multimedia Tools and Applications, 2018. **77**(16): p. 21589-21615.
- [21] Naskar, P.K., et al., *An efficient block-level image encryption scheme based on multi-chaotic maps with DNA encoding*. Nonlinear Dynamics, 2021. **105**(4): p. 3673-3698.
- [22] Chen, J., et al., *Exploiting self-adaptive permutation-diffusion and DNA random encoding for secure and efficient image encryption*. Signal Processing, 2018. **142**: p. 340-353.

# FlowTexNet: Fast Texture Synthesis for Massive Flow Field Visualization

Zijian Kang

Beijing Key Laboratory of Intelligent  
Information Technology, School of  
Computer Science & Technology,  
Beijing Institute of Technology,  
Beijing, China  
3120201034@bit.edu.cn

Wenyao Zhang\*

Beijing Key Laboratory of Intelligent  
Information Technology, School of  
Computer Science & Technology,  
Beijing Institute of Technology,  
Beijing, China  
zhwenyao@bit.edu.cn

Na Wang

Lenovo Research, Beijing, China  
wangna15@lenovo.com

## ABSTRACT

Flow field texture synthesis is a common and popular way to visualize flow fields. When massive flow fields are to be processed, existing algorithms based on line integral convolution (LIC) are not fast enough. In this paper, a new deep-learning-based method is proposed to synthesize flow textures for massive flow fields. Firstly, a deep neural network called FlowTexNet is built on the base of encoder-decoder architecture. Then the network is trained by flow textures generated by the original LIC algorithm. By this way, FlowTexNet can synthesize flow textures that have the same visualization effect as LIC textures. But FlowTexNet is much faster than the LIC algorithm. Test results show that the speedup of FlowTexNet is up to 450x when it is used to process massive flow fields and compared with the original LIC algorithm. Moreover, FlowTexNet can be applied to flow fields that are out of training, showing good generalization performance.

## CCS CONCEPTS

• Human-centered computing; • Visualization; • Visualization application domains;

## KEYWORDS

Deep learning, Flow field, Texture synthesis, LIC, Flow visualization

### ACM Reference Format:

Zijian Kang, Wenyao Zhang, and Na Wang. 2023. FlowTexNet: Fast Texture Synthesis for Massive Flow Field Visualization. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3590003.3590106>

## 1 INTRODUCTION

Flow field visualization plays an important role in the research of fluid mechanics, aerodynamics, meteorology and oceanography for its ability to disclose phenomena or laws contained therein. Flow

field texture synthesis is a popular way to visualize flow fields, whose goal is to convert flow field vector data obtained by simulation or observation into texture images that can show the underlying flow patterns. Currently, some texture synthesis algorithms have been developed for flow field visualization [1]. Line Integral Convolution (LIC) algorithm [2] is an excellent one of them. It can generate texture images that not only disclose the global structure of flow fields, but also provide plenty of local details without the visual confusion occurred in other visualization techniques such as streamlines and vector glyphs. For these reasons, LIC becomes a favorite method, and is widely used for flow field visualization. However, LIC has a problem of heavy computation, since line integral calculation is required by each pixel of textures. Therefore, variants of LIC, such as the Fast LIC [3] and various GPU versions of LIC [4], were proposed to speed up the computation. These improvements alleviate the problem to some extent. But they do not change the intrinsic way of synthesizing flow textures. When batches of massive flow fields are to be processed, these algorithms are still powerless, because the time spent on generating flow field textures is rather high. There should be a better way for flow field texture synthesis.

On the other hand, emerging deep learning technologies have presented very strong processing power in many areas. Lots of traditional problems are solved, improved, or processed more efficiently by the paradigm of deep learning. In flow field visualization, deep learning techniques have also been applied to feature extraction of flow field, for example, Eddy detection [5], Vortex detection [6], Shock detection [7], and other similar tasks [8, 9]. In literature, however, no deep learning work for flow field texture synthesis is reported.

To synthesize textures for flow fields as fast as possible, especially for massive flow fields, we explore the way of generating flow textures using deep learning, and propose a neural network model to achieve the goal. The proposed model is called FlowTexNet in this paper. It is a full convolutional encoder-decoder model. We train the model by textures synthesized by LIC algorithm, so it can generate LIC-style textures for flow fields. This is a new way rather than line integral integration to synthesize flow field textures. It can utilize powerful computation of modern deep learning platforms to accelerate the synthesizing. Our test results show that the speedup in massive batch processing can be up to hundreds of times, when it is compared with the original LIC algorithm.

The rest of this paper is organized as follows. Section 2 briefly reviews related work of texture synthesis for flow visualization and

\*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590106>

some deep learning applications in flow visualization. Section 3 details our newly proposed method for flow field texture synthesis, including the architecture of FlowTexNet and the computation of loss functions. Experiment and test results are given in Section 4, where FlowTexNet is trained, evaluated, and compared with the original LIC algorithm. Section 5 concludes the paper with some discussions and possible further work.

## 2 RELATED WORK

As a category of flow visualization method, flow field texture synthesis is aimed to generate texture images that can disclose the underlying flow patterns of flow fields. Currently, typical techniques for this goal include spot noise [10], LIC [2], texture advection [11], etc. These techniques can generate dense texture covering full field, and provide global and local information about flows. Here we only briefly introduce the LIC algorithm and its variants that are close to our current work.

The original LIC algorithm proposed by Cabral and Leedom [2] is an excellent and popular method for flow field texture synthesis. Its main idea is to convolve a noise texture using a kernel filter along streamlines. This feature enables it generate high quality of dense texture that can clearly disclose flow patterns. But the overhead of LIC is rather heavy, because the line integral convolution is applied for each pixel of the output texture. Therefore, Stalling et al. proposed the Fast LIC [3], where simple box filters are used for convolution. In performance, Fast LIC is about one order of magnitude faster than LIC. Zöckler [12] et al. proposed the parallel LIC algorithm, which further accelerates flow texture generation by parallel processing. Qin et al. proposed the GPU version of LIC [4] that obtained a speedup of 50 than the original LIC. Besides the acceleration, the original LIC algorithm has been extended and improved in other aspects. The detailed information can be referred to the survey about dense and texture-based flow visualization [1]. In this paper, the speed of synthesizing flow field textures is our main concerns. When massive flow fields are to be processed, we find that existing LIC algorithms are still less satisfactory in response time, though they work well in general analysis of individual flow field. In practice, there are tasks that need to analysis thousands of flow fields interactively. If LIC algorithms are used in such cases, the time cost on flow field texture synthesis will be far beyond the requirement of interactive analysis. We need to find a new way to speed up texture synthesis for massive flow fields.

Recently, deep learning techniques have gained great attention and achieved success in many areas such as computer vision [13] and natural language processing [14]. The strong power of deep learning may be feasible way for massive flow texture synthesis. Currently, to our best knowledge, no flow texture synthesis work based on deep learning is reported in literature. But some deep learning techniques have been applied to flow field feature extraction. Lguensat et al. [5] took the neural network model, U-Net, to detect ocean eddies in remote sensing ocean images. Duo et al. [15] introduced the ResNet50 model into ocean vortex detection, and got higher accuracy than traditional methods. Deng et al. [16] proposed the Vortex-Net for fast vortex detection. The performance of Vortex-Net is comparable to the classical instantaneous vorticity deviation method. But its speed is not fast enough. For this reason,

Wang et al. [17] proposed the Vortex-Seg-Net model where fully connected layers are removed to improve the speed of calculation. Deng et al. [6] proposed the Vortex-U-Net to further improve the accuracy and speed of vortex detection. Besides, a full convolution network named Shock-Net is proposed by Liu et al. [7] for shock wave detection in explosion fields.

The work mentioned above presents the potential of deep learning in flow field analysis. It could be possible to synthesize flow field texture using deep learning. With this motivation, we propose a neural network model to synthesize flow field textures in this paper.

## 3 METHOD

As mentioned before, existing LIC algorithms are not fast enough to generate massive flow field textures. To deal with this situation, we propose a deep-learning-based method for massive flow texture synthesis. In this method, a deep neural network model called FlowTexNet is first built, and then trained by flow field textures that are generated from LIC algorithms. When the model is trained and tuned fine enough, it is used to generate textures for massive flow fields. In this section, we will describe the architecture of FlowTexNet as well as the loss functions used to optimize the network.

### 3.1 Network Architecture

FlowTexNet is in general an encoder-decoder network. It takes the backbone of U-Net [18], but basic blocks and block connections are different from U-Net. As shown in Figure 1, there are two down-sampling layers and multiple convolution blocks in the encoder (i.e., the first half part of the network). Correspondingly, the decoder (i.e., the second half part) includes two up-sampling layers as well as convolution blocks similar to the encoder. Skip connections occurred in U-Net do not appear in our network model for two reasons. First, for the task of flow texture synthesis, we find out that low-level features from the encoder cannot help to improve the performance of the decoder. This is mainly due to the difference between vector fields and texture images. Second, connections that skip different level of feature maps increase the cost of computation and make the network more complex.

The input of FlowTexNet is two-dimensional flow field that contains vectors defined on rectangular grids. In practice, the difference between different input flow fields may be huge. For this reason, the encoder starts with two special convolution blocks. Each of them includes  $3 \times 3$  convolution (Conv  $3 \times 3$ ), instance normalization (IN) [19], and rectified linear unit (ReLU).

Besides the initial blocks, three other kinds of basic blocks are involved in our network model. They are named MBTNK1, MBTNK2, and MSA BottleNeck, respectively, and indicated by different color legends in Figure 1.

MBTNK1 is a modified version of Bottleneck1 (BTNK1) that was introduced in ResNet50 [20], while MBTNK2 is adapted from Bottleneck2 (BTNK2) of ResNet50. Figure 2 shows the structures of MBTNK1 and MBTNK2. The main difference is that batch normalization (BN) used for BTNK1 and BTNK12 are replaced by IN in MBTNK1 and MBTNK2. Both MBTNK1 and MBTNK2 are repeated

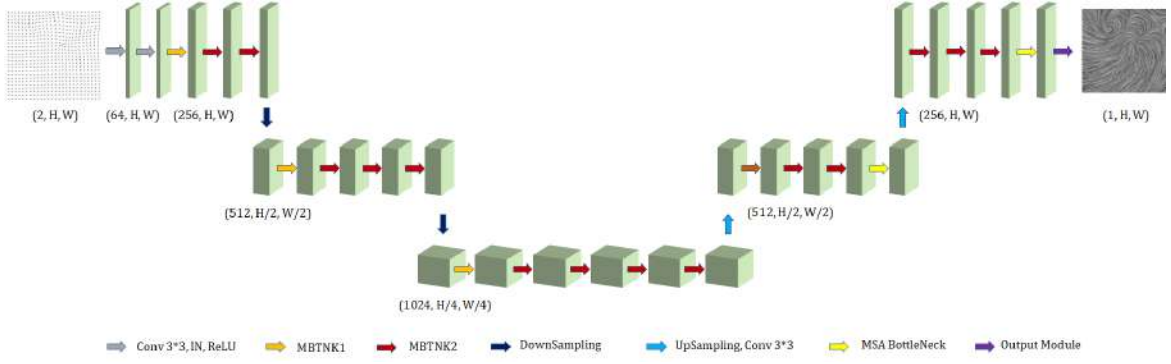


Figure 1: The architecture of FlowTexNet.

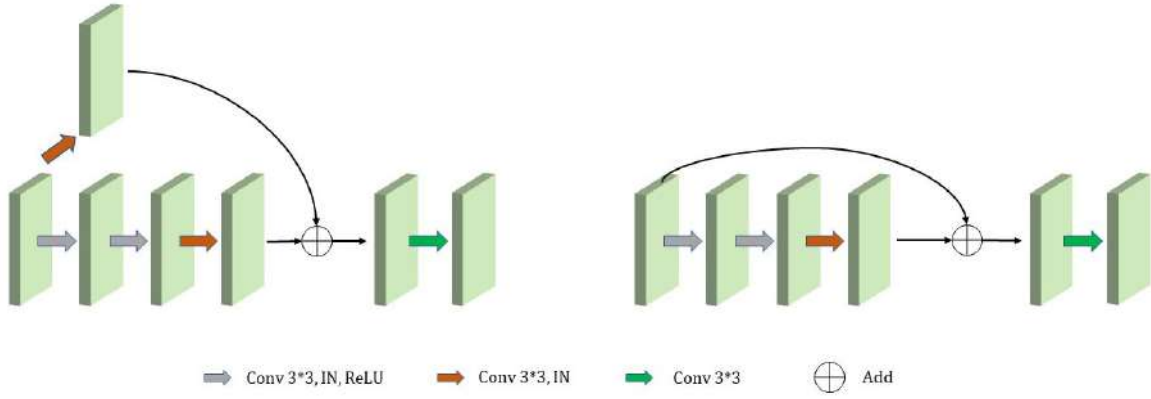


Figure 2: Structures of MBTNK1 (left) and MBTNK2 (right).

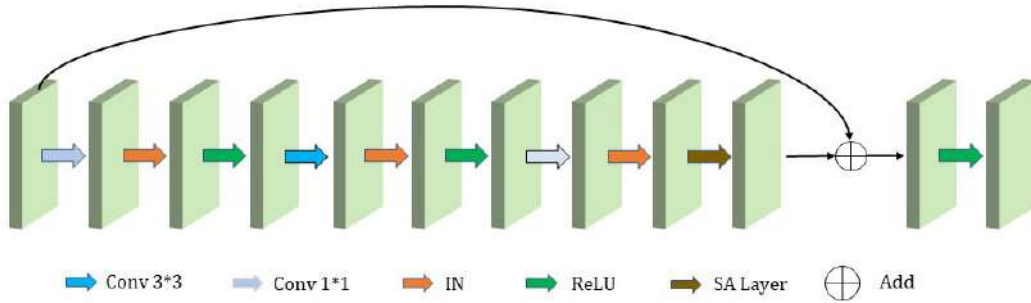


Figure 3: The structure of MSA BottleNeck.

multiple times in our network to extract flow features in different levels and help the prediction of output flow textures.

MSA BottleNeck is in fact an attention module used to strengthen the underlying useful flow features. Two MSA BottleNecks are included in our network. Both of them have the same structure as shown in Figure 3, where a shuffle attention (SA) layer is combined with several convolutions, instance normalizations, and rectified linear units. Here it should be noted that MSA BottleNeck is a

modified version of the shuffle attention bottleneck in [21]. The modification is that BN operation in shuffle attention bottleneck is changed to IN to fit our process of flow fields.

At the end of FlowTexNet, a specialized output module is designed to produce the final flow field textures. Figure 4 illustrates the output module, where 256 channels of feature maps from the decoder of FlowTexNet are gradually emerged into the final output

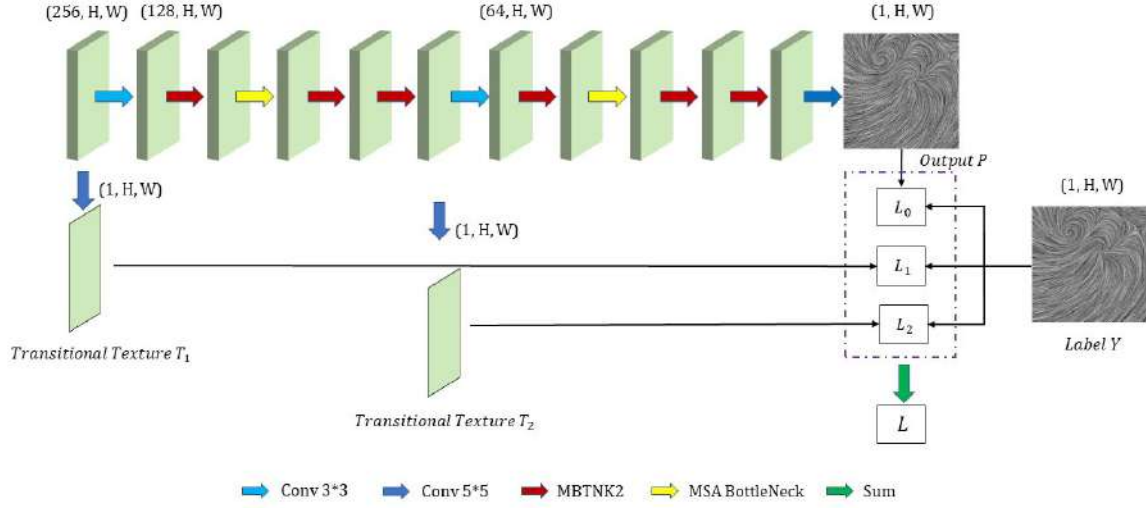


Figure 4: The structure of output module.

by concatenated processes of convolutions, MBTNK2 blocks, and MSA BottleNecks. By this means, flow textures are further refined.

In addition, FlowTexNet includes two pairs of downsampling and upsampling blocks. General convolution layers with  $3 \times 3$  kernels are used for downsampling. Bilinear interpolation rather than general deconvolution is used for upsampling to reduce artifacts occurred in recovering feature maps. To increase spatial continuity of feature maps, a general convolution layer with  $3 \times 3$  kernels is added for upsampling.

### 3.2 Loss Function

When the architecture of FlowTexNet is defined, we need to train it to obtain model parameters. The training dataset used for FlowTexNet consists of two parts: flow fields and corresponding texture images. The texture images are used as the ground truth labels. For each input flow field, the network will output a predicted texture image.

In general, predicted values and ground truth labels are taken to design a loss function to control model training. For FlowTexNet, however, two additional factors are taken into consideration. As shown in Figure 4, the initial and middle feature maps in the output module are merged by additional convolutions to produce two transitional textures,  $T_1$  and  $T_2$ , respectively. The transitional textures,  $T_1$  and  $T_2$ , are then combined with the output texture,  $P$ , and the ground truth label,  $Y$ , to design our loss function of network.

To be specific, the total loss function of FlowTexNet is defined as:

$$L = \lambda_0 L_0 + \lambda_1 L_1 + \lambda_2 L_2 \quad (1)$$

where  $\lambda_1, \lambda_2, \lambda_3$  are weighting coefficients for the three partial mean square errors (MSEs),  $L_0, L_1$ , and  $L_2$ . The three MSEs are defined as follows:

$$L_0 = \frac{1}{N \times H \times W} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W (P_{n,i,j} - Y_{n,i,j})^2 \quad (2)$$

$$L_1 = \frac{1}{N \times H \times W} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W (T_1_{n,i,j} - Y_{n,i,j})^2 \quad (3)$$

$$L_2 = \frac{1}{N \times H \times W} \sum_{n=1}^N \sum_{i=1}^H \sum_{j=1}^W (T_2_{n,i,j} - Y_{n,i,j})^2 \quad (4)$$

Where  $N$  is the batch size,  $H$  and  $W$  are the height and width of textures, respectively. FlowTexNet is trained by Adam optimizer with the above loss function.

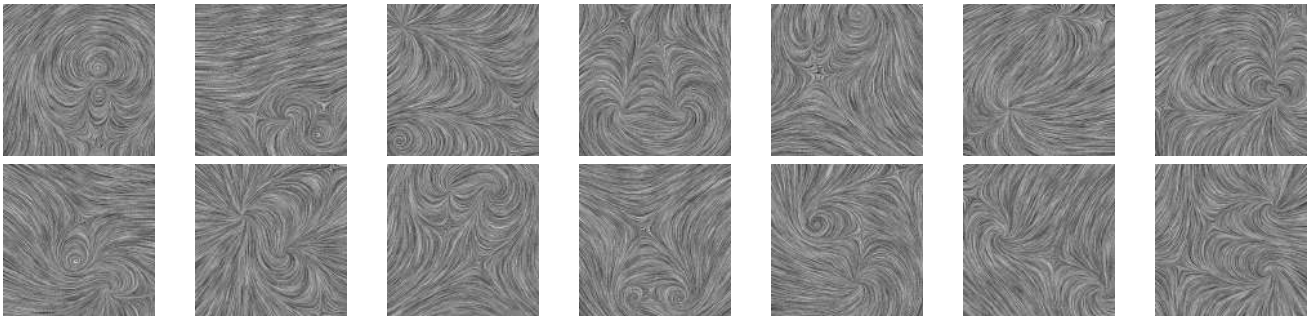
## 4 EXPERIMENT

### 4.1 Datasets

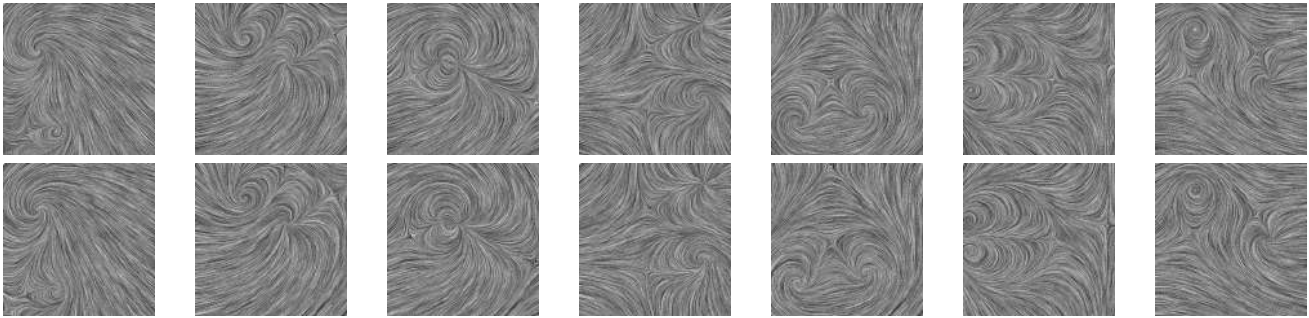
Currently, to our best knowledge, no public dataset is available for the training of FlowTexNet. Therefore, we built a dataset that includes 90,000 two-dimensional flow fields. Each field is characterized by four randomly distributed critical points. The involved critical points include saddle points and non-saddle points whose features are described in [22]. We made a tool to generate such flow fields at different sizes. For the convenience of network training, the field size of the dataset is fixed to  $128 \times 128$ . Moreover, a label texture, which is used as the ground truth of texture synthesis, is generated for each flow field by the original LIC algorithm. Some label texture examples are shown in Figure 5 where various flow patterns are presented.

### 4.2 Model Training

To train and test FlowTexNet, the dataset built by ourselves are randomly divided into a training set and a testing set. The training set contains 80,000 flow fields as well as associated labels, while the testing set has 10,000 flow fields. All training and testing tasks are performed on a deep learning platform that is equipped with python 3.7.1, pytorch 1.8.0, and NVIDIA GeForce RTX 3090 GPU device.



**Figure 5: Some examples of label textures in the dataset.**



**Figure 6: Comparisons of flow field textures. The first row is from FlowTexNet, and the second row is from the LIC algorithm.**

During the training, parameters of loss function,  $\lambda_1$ ,  $\lambda_2$ , and  $\lambda_3$ , are all set to 1.0. The training dataset is divided into mini-batches, each of which contains 5 flow fields. Adam optimizer is applied to update model parameters at a fixed learning rate 0.0002. After 16 epochs of training, model parameters of FlowTexNet reach to a stationary state where the total loss is about 48.0447. The total training time is about 32 hours.

### 4.3 Test Results

After the proper model training, we evaluated FlowTexNet with the prepared testing dataset. The evaluation includes the quality of synthesized textures, the performance of acceleration in massive processing, and the generalization of the model.

**4.3.1 Texture Quality.** The output textures of FlowTexNet are firstly compared in quality with the ground truth, i.e., label textures generated by the original LIC algorithm. Such a group of comparisons are given in Figure 6. We can see that flow patterns shown in textures obtained by FlowTexNet are almost indistinguishable from those generated by the LIC algorithm. In this sense, FlowTexNet achieves the same level of quality as the LIC algorithm. Here it should be noted that, as shown in Figure 6, the synthesized textures are in fact different from the ground truth in pixels, but they have the same visualization effect in terms of disclosing flow patterns.

**4.3.2 Processing Performance.** The main goal of FlowTexNet is to accelerate massive flow texture synthesis. To verify the acceleration performance, we tested FlowTexNet in batch processing with different batch sizes, and obtained the corresponding calculation

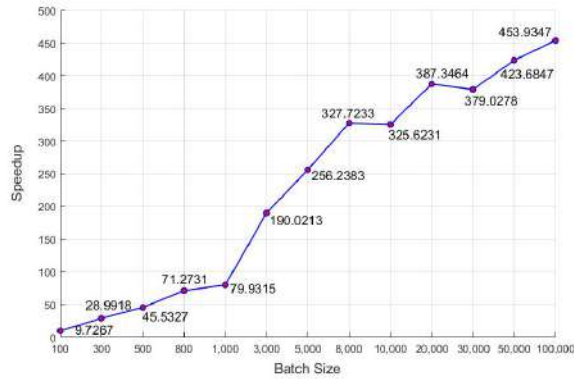
time spent on texture synthesizing. As a baseline, we also measured the time cost by the LIC algorithm to process the same amount of flow fields as FlowTexNet. All test results are given Table 1, from which we can see that speedup values obtained by FlowTexNet are rather high. The time cost by the LIC algorithm almost linearly increases with the amount of flow fields. When the batch size is up to a high level, the time cost is amazing. For FlowTexNet, however, the time cost maintains at pretty low levels. The increasing is very slow. Even if the amount of flow fields is up to 100,000, the time cost is only a little more than 42 seconds. This makes us achieve an excellent speedup up to 453.9374. The trend of speedup that varies with batch size is shown in Figure 7. It is clear that, the larger the scale of batch processing, the better the acceleration performance is. The superiority of FlowTexNet in processing massive flow fields is fully verified. In above tests, 100,000 flow fields are involved. Besides the dataset with 90,000 flow fields, additional 10,000 flow fields containing multiple critical points are generated and included in tests.

**4.3.3 Model Generalization.** FlowTexNet was trained by flow fields containing 4 critical points as described in Section 4.1. Previous tests show that the trained model can process similar flow fields successfully. Can the model trained with the designated dataset be applied to outside data? To explore this problem, we performed two additional experiments.

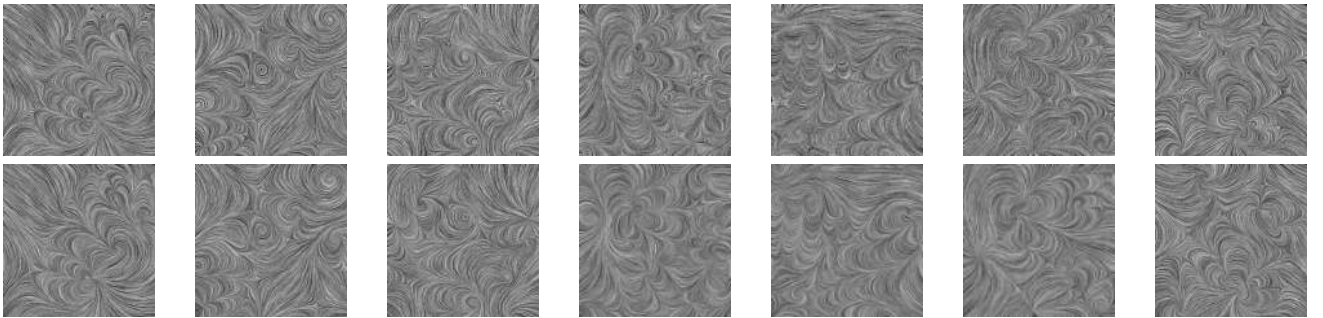
In the first experiment, we tested the model with new flow fields that are in the same size as the training data but have more or less than 4 critical points. These new flow fields are in different flow

**Table 1: Comparisons of time cost between FlowTexNet and the LIC algorithm**

Flow Field Size	Batch Size	Time cost by LIC (s)	Time cost by FlowTexNet (s)	Speedup
128*128	100	19.2364	1.9777	9.7267
128*128	300	58.8881	2.0312	28.9918
128*128	500	95.2727	2.0924	45.5327
128*128	800	161.1379	2.2608	71.2731
128*128	1,000	188.6124	2.3596	79.9315
128*128	3,000	567.0426	2.9841	190.0213
128*128	5,000	955.9914	3.7272	256.2383
128*128	8,000	1553.2450	4.7395	327.7233
128*128	10,000	1903.4954	5.8457	325.6231
128*128	20,000	3787.6668	9.7785	387.3464
128*128	30,000	5677.6085	14.9794	379.0278
128*128	50,000	9525.8302	22.4833	423.6847
128*128	100,000	19261.2225	42.4317	453.9347

**Figure 7: The trend of speedup obtained by FlowTexNet.**

patterns, because different number of critical points means different flow patterns. Some test results are shown in Figure 8, where flow textures generated by the LIC algorithm are also included as comparisons. We can see that flow field textures produced by FlowTexNet disclosed flow patterns correctly. FlowTexNet has the ability to accommodate various flow patterns.

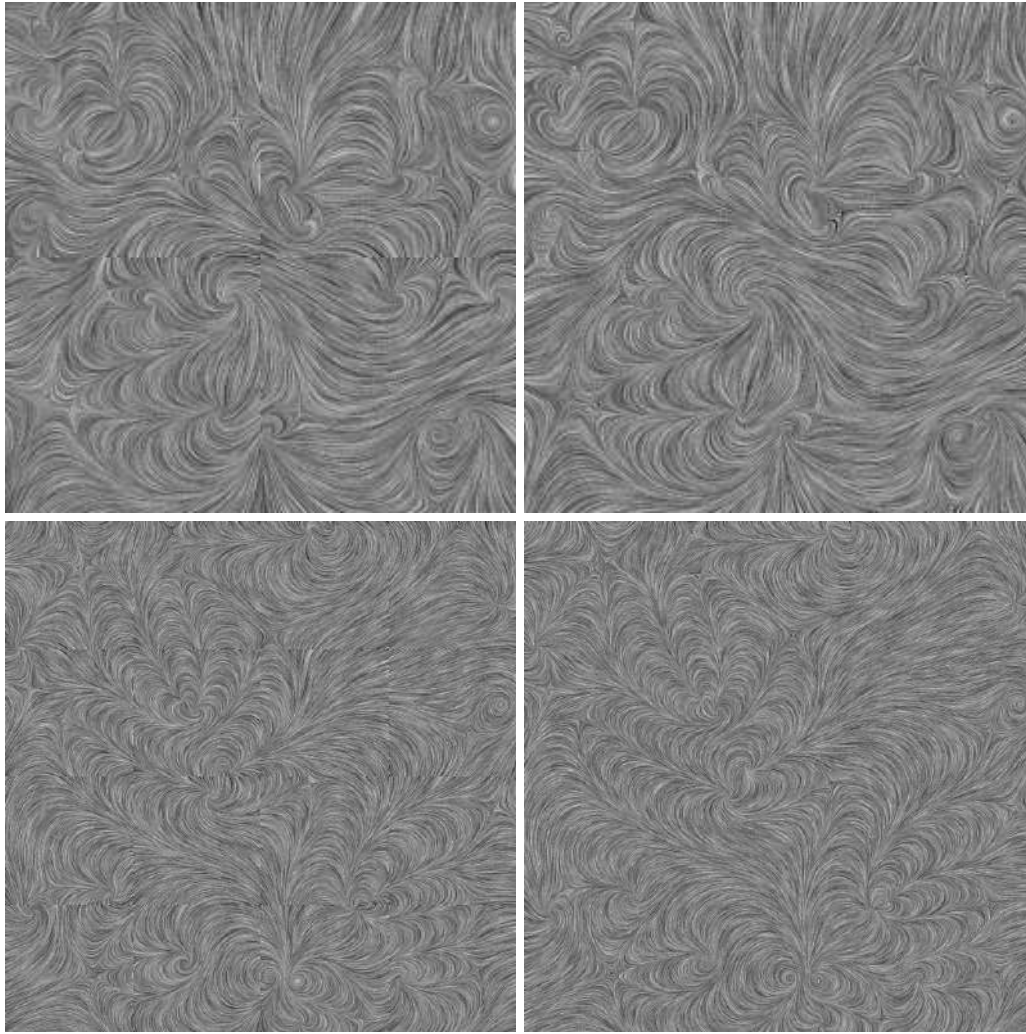
**Figure 8: Test results for flow fields with different number of critical points. The upper row is generated by LIC, and the lower row is produced by FlowTexNet.**

In the second experiment, we explored the ability of FlowTexNet to process flow fields in different sizes. The input size of FlowTexNet was initially set to 128\*128. And it was trained with the same size of flow fields. This means it cannot be directly applied for other cases. To process larger flow fields, there are two options. One is to retrain the model with new data. This is not feasible in practice. The other is to divide large field into small patches to fit the model. We tested FlowTexNet by this way. Two flow fields in sizes of 256\*256 and 512\*512 are cut into 128\*128 patches. When textures for these patches are produced by FlowTexNet, they are combined to recover the original full fields.

Test results are given in Figure 9. In both cases, FlowTexNet produced textures that showed global and local flow patterns correctly. But the split joint is not seamless. Weak discontinuous boundaries between patches can be perceived. Nonetheless, this does not affect the visualization of the entire flow field. Therefore, FlowTexNet can be applied for larger flow fields.

## 5 CONCLUSIONS

Flow texture synthesis is a basic way to visualize flow fields. When massive flow fields are to be processed, existing LIC algorithms are still less satisfactory in response time, though they work well



**Figure 9: Test results of two flow fields in different sizes. One is 256\*256 in the top line, and the other is 512\*512 in the bottom line. The left column is produced by FlowTexNet, while the right is by LIC.**

in general analysis of individual flow field. In this paper, we propose a deep-learning-based method for flow field texture synthesis. We build a deep neural network called FlowTexNet, and train the network with flow field textures generated by the LIC algorithm. When the network model is well trained, it can generate flow textures that are in the same level of quality as those obtained by LIC algorithm. But it does not follow the way of LIC. It in fact learns a new way to synthesize flow field textures, and utilizes powerful computation of modern deep learning platforms to accelerate the synthesizing. Test results show that the speedup can be up to hundreds of times when FlowTexNet is used to process massive flow fields and compared with the original LIC algorithm. Test results also indicate that FlowTexNet has good generalization performance. It can be applied to flow fields that are out of training and different in both sizes and flow patterns. Currently, however, large flow fields have to be divided into small patches to fit to the input size of FlowTexNet. This causes some weak discontinuities in final results.

Fortunately, visualization effect is not affected this phenomenon. In future, FlowTexNet will be extended to three-dimensional flow fields, and color encoding should also be taken into consideration to generate textures embedding more flow information.

## ACKNOWLEDGMENTS

This work was partly supported by State Key Laboratory of Computer Architecture (ICT, CAS) under Grant No. CARCH 202102.

## REFERENCES

- [1] Laramée, R. S., Hauser, H., Doleisch, H., Vrolijk, B., Post, F. H., and Weiskopf, D. 2004. The state of the art in flow visualization: Dense and texture-based techniques. *Computer Graphics Forum*, 23, 2, 203–221. <https://doi.org/10.1111/j.1467-8659.2004.00753.x>.
- [2] Brian Cabral and Leith Casey Leedom. 1993. Imaging vector fields using line integral convolution. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques (SIGGRAPH '93)*. Association for Computing Machinery, New York, NY, USA, 263–270. <https://doi.org/10.1145/166117.166151>.

- [3] Detlev Stalling and Hans-Christian Hege. 1995. Fast and resolution independent line integral convolution. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques (SIGGRAPH '95)*. Association for Computing Machinery, New York, NY, USA, 249–256. <https://doi.org/10.1145/218380.218448>.
- [4] Bo Qin, Zhanbin Wu, Fang Su, and Titi Pang. 2010. GPU-Based parallelization algorithm for 2d line integral convolution. In *Proceedings of the First international conference on Advances in Swarm Intelligence - Volume Part I (ICSI'10)*. Springer-Verlag, Berlin, Heidelberg, 397–404. [https://doi.org/10.1007/978-3-642-13495-1\\_49](https://doi.org/10.1007/978-3-642-13495-1_49).
- [5] Redouane Lguensat, Miao Sun, Ronan Fablet, Pierre Tandeo, Evan Mason and Ge Chen. 2018. EddyNet: A deep neural network for pixel-wise classification of oceanic eddies. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Valencia, Spain, 1764–1767. <https://doi.org/10.1109/IGARSS.2018.8518411>.
- [6] Liang Deng, Wenchun Bao, Yueqing Wang, Zhigong Yang, Dan Zhao, Fang Wang, Chongke Bi, and Yang Guo. 2022. Vortex-U-Net: An efficient and effective vortex detection approach based on U-Net structure. *Appl. Soft Computing*, 115, C (Jan 2022). <https://doi.org/10.1016/j.asoc.2021.108229>.
- [7] Yang Liu, Yutong Lu, Yueqing Wang, Dong Sun, Liang Deng, Fang Wang and Yan Lei. 2019. A CNN-based shock detection method in flow visualization. *Computers & Fluids*, 184(2019), 1–9. <https://doi.org/10.1016/j.compfluid.2019.03.022>.
- [8] Octavi Obiols-Sales, Abhinav Vishnu, Nicholas Malaya, and Aparna Chandramowlishwaran. 2020. CFDNet: a deep learning-based accelerator for fluid simulations. In *Proceedings of the 34th ACM International Conference on Supercomputing (ICS '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 1–12. <https://doi.org/10.1145/3392717.3392772>.
- [9] Pin Wu, Kaikai Pan, Lulu Ji, Siquan Gong, Weibing Feng, Wenyan Yuan, and Christopher Pain. 2022. Navier–stokes Generative Adversarial Network: a physics-informed deep learning model for fluid flow generation. *Neural Computing*, Appl. 34, 14 (Jul 2022), 11539–11552. <https://doi.org/10.1007/s00521-022-07042-6>.
- [10] Jarke J. van Wijk. 1991. Spot noise texture synthesis for data visualization. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques (SIGGRAPH '91)*. Association for Computing Machinery, New York, NY, USA, 309–318. <https://doi.org/10.1145/122718.122751>.
- [11] Nelson Max, Roger Crawfis, and Dean Williams. 1992. Visualizing wind velocities by advecting cloud textures. In *Proceedings of the 3rd conference on Visualization '92 (VIS '92)*. IEEE Computer Society Press, Washington, DC, USA, 179–184. <https://doi.org/10.1109/VISUAL.1992.235210>.
- [12] Malte Zöckler, Detlev Stalling, and Hans-Christian Hege. 1997. Parallel line integral convolution. *Parallel Computing*, 23, 7 (July 1997), 975–989. [https://doi.org/10.1016/S0167-8191\(97\)00039-2](https://doi.org/10.1016/S0167-8191(97)00039-2).
- [13] Rajat Kumar Sinha, Ruchi Pandey, Rohan Pattnaik. 2018. Deep learning for computer vision tasks: a review. *2017 International Conference on Intelligent Computing and Control (I2C2)*. <https://doi.org/10.48550/arXiv.1804.03928>.
- [14] Daniel W Otter, Julian R Medina and Jugal K Kalita. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32, 2 (February 2021) 604 – 624. <https://doi.org/10.1109/TNNLS.2020.2979670>.
- [15] Zijun Duo, Wenke Wang and Huizan Wang. 2019. Oceanic mesoscale eddy detection method based on deep learning. *Remote Sensing*, 11, 16(August 2019), 1921. <https://doi.org/10.3390/rs11161921>.
- [16] Liang Deng, Yueqing Wang, Yang Liu, Fang Wang, Sikun Li, and Jie Liu. 2019. A CNN-based vortex identification method. *J. Vis.* 22, 1 (February 2019), 65–78. <https://doi.org/10.1007/s12650-018-0523-1>.
- [17] Yueqing Wang, Liang Deng, Zhigong Yang, Dan Zhao, and Fang Wang. 2021. A rapid vortex identification method using fully convolutional segmentation network. *Vis. Comput.* 37, 2 (February 2021), 261–273. <https://doi.org/10.1007/s00371-020-01797-6>.
- [18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention. Lecture Notes in Computer Science*, 234–241. [https://doi.org/10.1007/978-3-319-24574-4\\_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [19] Dmitry Ulyanov and A. Vedaldi, V. Lempitsky. 2016. Instance normalization: The missing ingredient for fast stylization. <https://doi.org/10.48550/arXiv.1607.08022>.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Las Vegas, NV, USA. <https://doi.org/10.1109/CVPR.2016.90>.
- [21] Qing-Long Zhang and Yu-Bin Yang. 2021. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE Toronto, ON, Canada. 2235–2239. <https://doi.org/10.1109/ICASSP39728.2021.9414568>.
- [22] Perry, Anthony E. and B. D. Fairlie. 1979. Critical Points in Flow Patterns. *Advances in Geophysics*, 18 (1975), 299–315. [https://doi.org/10.1016/S0065-2687\(08\)60588-9](https://doi.org/10.1016/S0065-2687(08)60588-9).

# CIP-ES: Causal Input Perturbation for Explanation Surrogates

Sebastian Steindl

s.steindl@oth-aw.de

Innovation and Competence Center Artificial Intelligence,  
Technical University of Applied Sciences Amberg-Weiden  
Germany

Martin Sumner

Martin.Sumner@haw-landshut.de

Institute for Data and Process Science, Computer Science  
Department, Landshut University of Applied Sciences  
Germany

## ABSTRACT

With current advances in Machine Learning and its growing use in high-impact scenarios, the demand for interpretable and explainable models becomes crucial. Causality research tries to go beyond statistical correlations by focusing on causal relationships, which is fundamental for Interpretable and Explainable Artificial Intelligence. In this paper, we perturb the input for explanation surrogates based on causal graphs. We present an approach to combine surrogate-based explanations with causal knowledge. We apply the perturbed data to the Local Interpretable Model-agnostic Explanations (LIME) approach to showcase how causal graphs improve explanations of surrogate models. We thus integrate features from both domains by adding a causal component to local explanations. The proposed approach enables explanations that suit the expectations of the user by having the user define an appropriate causal graph. Accordingly, these expectations are true to the user. We demonstrate the suitability of our method using real world data.

## KEYWORDS

Explainable AI, Interpretability, Causality, Surrogate Models, Causability

### ACM Reference Format:

Sebastian Steindl and Martin Sumner. 2023. CIP-ES: Causal Input Perturbation for Explanation Surrogates. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590107>

## 1 INTRODUCTION

Machine Learning (ML) is increasingly used in critical applications such as healthcare [21], finance [1] and justice [27]. Hereby, the larger the impact of a model prediction in ML, the higher is the need for users to understand and trust the predictions. While deep learning methods benefit from their vast amount of parameters tuned on large data samples, this same property renders them less comprehensible for humans, provoking them to be regarded as *black-boxes*. This has led to increased attention to the research areas of Interpretable and Explainable Artificial Intelligence (XAI).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9944-9/23/03...\$15.00

<https://doi.org/10.1145/3590003.3590107>

Common XAI methods include SHAP (SHapley Additive exPlanations) [16] and LIME [22]. LIME uses a surrogate model trained on a local neighbourhood generated by perturbation to explain the black-box model.

For humans the notion of *Causality* is central. The way children learn and adults understand the environment is based on causal mechanisms [3]. The seminal work of Pearl [19] proposes a framework to formalize cause-effect-relationships and causal mechanisms. These get formalized as Directed Acyclical Graphs (DAG)  $G = (V, E)$  and Structural Causal Models (SCMs,  $\mathbb{C}$ ). While current ML techniques work well in the i.i.d. setting with observational data, there are problems with generalizing to real world scenarios that have led to increased research in integrating ML and Causality into Causal Machine Learning [15]. Since standard ML models and corresponding XAI methods are based solely on correlational statistics, a causal component can also improve their comprehensibility [26]. Accordingly, Holzinger et al. [13] define the term *Causability* as "the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding". Consequently, we propose to add a causal component to explanation surrogate techniques by performing the input perturbations based on a causal graph. There are parallels between Causality and XAI that are being explored in literature (cf. [18, 26]).

Chen et al. [5] propose to distinguish explanations that are *true to the data* from those that are *true to the model*. In this paper, we propose a method referred to as Causal Input Perturbation for Explanation Surrogates (CIP-ES), following a third novel perspective: *true to the user*. By having the user define a causal graph that embodies user expectations and knowledge and basing the explanations on this graph, we obtain explanations that are true to the users priors. Abstractly speaking, we use the human-in-the-loop as a generator or oracle for a world model in the form of a causal graph confined to the features of the dataset. The main contributions of this paper are *i)* a novel framework for causal input perturbation, i.e. CIP-ES, *ii)* the integration of CIP-ES and LIME on real datasets to finally, *iii)* introduce the notion of *true to the user* explanations and generate them with CIP-ES.

The remainder of this paper is structured as follows. In Section 2, we discuss related work. In Section 3, we motivate and detail our CIP-ES approach. We evaluate CIP-ES in Section 4 by applying it to real world data. Finally, we conclude our work and give an outlook of future development in Section 5.

## 2 RELATED WORK

In the following section we introduce relevant related work from both the XAI and the Causality domain.

**Shapley Value and causal adoptions.** The Shapley value from game theory interprets a feature as a participant of a game, forming coalitions with other features, to generate feature attributions [24]. Based on this approach, multiple extensions involving causal awareness have been proposed (cf. [11, 12, 14]).

**LIME.** In LIME [22], a local neighbourhood of synthetic datapoints  $\mathcal{Z}$  around a given  $x$ , that is to be explained, is used to fit an interpretable model. The predictions of the black-box model  $\hat{y}$  given by  $\mathcal{Z}$  act as labels for the perturbed datapoints. With the assumption that this interpretable model and the black-box model behave similar, the interpretable model gives insights into the local behaviour of the black-box model. As this is done only based on  $\mathcal{Z}$  and  $\hat{y}$ , this approach is model agnostic and hence not limited to a specific model class. Since the algorithm is limited to dependencies between features manifested as correlations in the dataset, we propose a causal adaption to this algorithm that enables having *true* to the user explanations.

**Causality and LIME.** Cinquini and Guidotti [7] show a way to integrate Causality and LIME for continuous data by applying causal discovery to recover the causal graph, which can be seen as *true* to the data.

**Personalized explanations.** Explanations and descriptions of an event might differ for various people in what has been coined as the *Rashomon effect* [4]. Schneider and Handali [23] argue that personalized explanations are favourable and might improve understandability. SimplEx is one example of a method for personalized explanations based on a user-chosen set of examples [8]. However it does not explicitly incorporate causal knowledge. In accordance with this concept of personalized explanations, our approach by design may lead to conflicting explanations, if different  $G_{man}$  are given as input.

## 3 MATERIALS AND METHOD

### 3.1 Materials

To test our method on different domains we use the Capital Bike-share dataset [10] and the MPG dataset from the UCI machine learning repository [9]. Both datasets contain regression tasks with the first one aiming to predict the number of rented bikes and the latter the miles-per-gallon (MPG) of a car. For the MPG data we used Min-Max scaling to preprocess the raw data. The relationships between the contained features of both datasets, e.g. the weather and the tendency to rent a bike are ordinary enough for anyone to apply common sense reasoning, using his or her existing knowledge and define a causal graph. We assume that there exist some physical mechanisms in nature that define a data-generating process. In the framework proposed by [19], this would be modeled by a SCM  $\mathbb{C}$ . Our training data  $(X, y)$  can then be seen as being generated by  $\mathbb{C}$  and entailing causal relationships matching those of  $G_{\mathbb{C}}$ .

### 3.2 Overview on the CIP-ES Approach

The generation of the neighbourhood  $\mathcal{Z}$  to explain  $x$  in standard LIME does not take causal relationships into account. We will differentiate between four causal graphs:  $G_{\mathcal{Z}}$  is used to generate the CIP-ES neighbourhood,  $G_{man}$  is defined by the user,  $G_X$  is the graph that would be algorithmically identified on the training data and  $G_{\mathbb{C}}$  is the true underlying causal graph that generated the data.

We propose to sample the causally motivated neighbourhood  $\tilde{\mathcal{Z}}$  by following the causal relationships described in  $G_{\mathcal{Z}}$  (see Fig. 1). With the Graph  $G_{man}$  being defined by the user and having  $G_{\mathcal{Z}} := G_{man}$ , this will lead to the explanations being aligned with the existing expectations and knowledge that led to  $G_{man}$ .

While the active research field of Causal Discovery has made a lot of progress in improving the identification of causal graphs on observational data, the problem remains NP-hard [6] and additional knowledge or assumptions are needed to make it feasible [19]. Therefore, the causal graph  $G_X$  that would be identified on the training data does not in general coincide with  $G_{\mathbb{C}}$ . Since  $G_{\mathbb{C}}$  is generally unknown, one would have to choose  $G_{\mathcal{Z}} := G_X$ . Because of this and the advantages mentioned above (e.g. having self-explanatory explanations), we propose to use  $G_{man}$  over using Causal Discovery. Moreover, in some situations we might have a priori knowledge about our data, that allows us to perform the perturbations more realistically than is practice by standard LIME.

For example in the Capital Bikeshare dataset [10], the features contain the month and season in which each datapoint is collected. The relationship between the features month and season is deterministic. Therefore we gain no new information by learning about the season, if we already know the month. Since in standard LIME all features are sampled independently, this coincides with the assumption that all features are independent from each other and existing, conflicting knowledge goes unused. Therefore, this can create corrupt data in the neighbourhood used to train the interpretable model.

From the causal perspective this is equivalent to a graph where every node has direct influence on  $y$  but is itself not influenced by anything else and hence has no parents. We call this a trivial causal graph  $G_{triv}$ . In contrast, CIP-ES generates  $\tilde{\mathcal{Z}}$  consistently and would not generate these faulty datapoints. There might be a few outlier scenarios where  $G_{triv}$  is sensible to assume, in which CIP-ES lets the user do this explicitly instead of implicitly. We summarise the main advantages of the CIP-ES adaption to LIME as

- avoiding the generation of corrupted data (cf. Sec. 3.1),
- giving self-explanatory explanations (cf. Sec 4.1) and
- producing explanations that are closer to reality, since we benefit of a human as the oracle for the causal graph.

### 3.3 The CIP-ES Algorithm

The main idea is to generate the neighbourhood in accordance to a causal graph  $G_{\mathcal{Z}}$  defined based on (expert) domain knowledge. Considering possible future advances in Causal Discovery, this might be combined with automatically discovered graphs, i.e.  $G_X$ .

We define  $PA_i$  to be the parents of node  $X_i$  within a DAG and as a special case  $PA_y$  to be the parents of  $y$ , the variable that is to be predicted. We first train  $f$  on the training data  $(X, y)$ . Then we define  $G_{\mathcal{Z}} := G_{man}$  by hand for real data. While this step of defining the causal relationship is not guaranteed to deliver the correct causal graph for the unknown  $\mathbb{C}$ , we argue that the benefit nonetheless lies in aligning  $G_{\mathcal{Z}}$  with  $G_{man}$  and therefore matching the explanations with the existing user expectation and knowledge.

To explain the prediction  $\hat{y} = f(x)$  we generate  $\tilde{\mathcal{Z}} \in \mathbb{R}^{N \times D}$  by three simple rules:

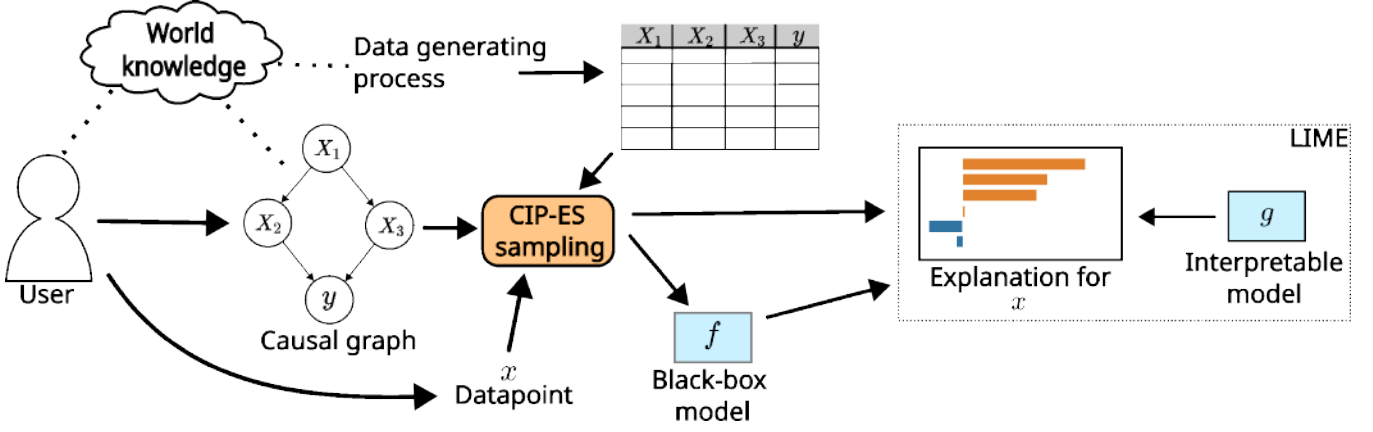


Figure 1: Visualization of the CIP-ES approach to explain the prediction  $f(x)$ .

- (1) Only perturb features with direct influence on  $y$ ,
- (2) sample those features taking into account the distribution of their parents and
- (3) if the parents contain a feature with direct influence on  $y$ , it has to be sampled first (Applied recursively)

We will now explain the rationale behind these rules. Firstly, we propose to only perturb those features that have a direct causal influence on  $y$ , i.e.  $X_i \rightarrow y$ . While in a causal chain structure like  $X_j \rightarrow X_i \rightarrow y$ ,  $X_j$  does have indirect influence on  $y$ , its effect is mediated by  $X_i$ . This means that learning about  $X_j$  while already knowing  $X_i$  does not change our estimate of  $y$ , even though  $X_j$  does have indirect effect on  $y$ . This is achieved by using the feature value  $x_j$  of the datapoint  $x$  that should be explained for all rows in  $\tilde{Z}$  with  $\tilde{X}_j = x_j \forall X_j \notin \text{PA}_y$ . This will lead to  $\text{Var}(X_j) = 0$ , thus providing no information for the interpretable model  $g$ . Consequently, we thereby force the explanation to state that  $X_j$  has no effect if  $X_j \notin \text{PA}_y$ .

Furthermore, from a causal perspective, the perturbation of  $X_i$  can be seen as an intervention, thus all incoming arrows would be ignored. CIP-ES explanations reveal this by stating that  $X_j$  has no influence, since all the information the user needs is already part of  $X_i$ .

We also propose to perform the perturbation itself according to  $G_{\tilde{Z}}$ . Looking again at the causal chain  $X_j \rightarrow X_i \rightarrow y$ , we sample  $\tilde{X}_i$  based on the distribution of  $X_j$ , to account for the indirect influence  $X_j$  has on  $y$ . According to rule 3 we need to check the parents of  $\text{PA}_{X_i}$ . If  $X_j \in \text{PA}_{X_i}$  and  $X_j \in \text{PA}_y$  we sample  $X_j$  first. This  $\tilde{X}_j$  is then used to sample  $\tilde{X}_i$ . For example with the MPG data, we first sample weight and then acceleration, since weight is a confounder between acceleration and our target variable (cf. Fig. 5). The sampling strategy uses conditional probabilities. We sample for continuous features  $\tilde{X}_i \sim \mathcal{N}(x_i, \sigma_{X_i}^2)$  where  $x_i$  is the value of the current feature for the datapoint that should be explained and with  $\sigma_{X_i}^2$  being the standard deviation of the selection of training data where  $X_j = x_j \forall X_j \in \text{PA}_i$  which reduces to  $\tilde{X}_i \sim \mathcal{N}(x_i, \sigma_{X_i}^2)$  if  $\text{PA}_i = \emptyset$ .

For categorical  $X_i$ , we define the function  $b_{P(X_i)}(X_i)$ , that samples the Bernoulli distribution given by  $(k; p(k))$  for all possible

values  $k$  in  $X_i$  and proceed analogously to the continuous case. With this  $\tilde{Z}$  our CIP-ES algorithm is completed and we then continue with standard LIME by generating  $\tilde{z}'$  and using K-LASSO to generate explanations.

## 4 RESULTS AND DISCUSSIONS

The following chapter first defines the methodology used to evaluate our approach, then shows the results on real data and finally evaluates and compares the CIP-ES fidelity and stability to LIME. While the true causal structure is unknown, this allows us to evaluate our approach both in a realistic situation and to conjecture about the validity of the explanations.

For all shown experiments the black-box model is a Multilayer Perceptron with one hidden layer consisting of 200 nodes and being trained for 2000 epochs. Otherwise, the default parameters from the MLPRegressor class of the scikit-learn library [20]. Note that for every experiment the datapoint used for the explanation is the same for each dataset, i.e., every comparison between LIME and CIP-ES shows the same datapoint, even if it might not look like it at first glance.

### 4.1 Influence of the Causal Graph

We found common Causal Discovery algorithms like FCI [25] or NOTEARS [28] fail to identify a realistic causal graph. For example, there would be obviously correlational or nonsensical relationships, like the humidity causing the month. This can be expected to happen in many real world scenarios, since the task is NP-hard [6] and additional knowledge or assumptions are needed to make it feasible [19]. This strengthens our proposal of defining a causal graph  $G_{man}$  manually and use this to perform causal input perturbation, i.e. setting  $G_{\tilde{Z}} := G_{man}$ . The explanations then are based on the knowledge and expectations about the causal structure that led to  $G_{man}$ .

The  $G_{man}$  that we defined for the used dataset is shown in Figure 4. While creating the graph, one has to sometimes make decisions about the causal structure that are not straightforward. Take for example  $season \leftarrow month \rightarrow weathersit$ . This case is

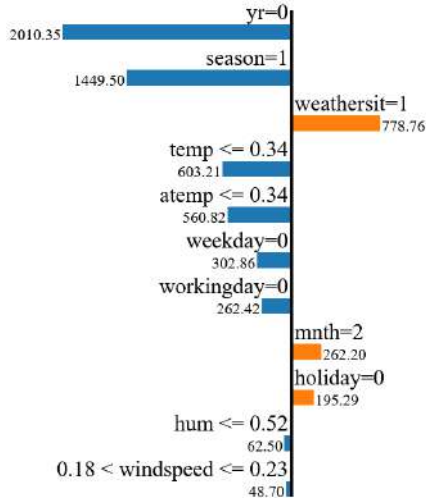


Figure 2: Explanation for a single  $x$  from real bike sharing data with standard LIME. Negative feature attribution is depicted in blue, positive in orange.

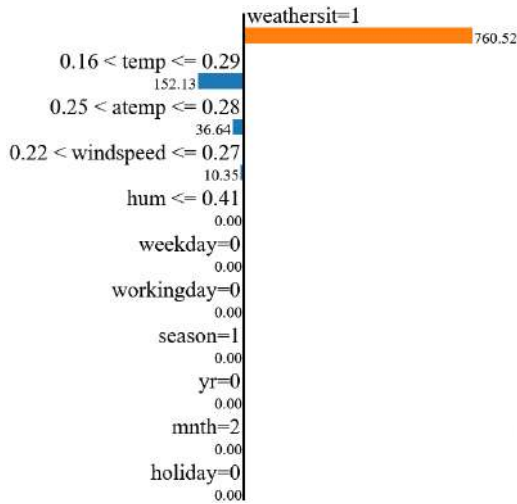


Figure 3: Explanation for a single  $x$  from real bike sharing data with CIP-ES. Negative feature attribution is depicted in blue, positive in orange.

special, because the month determines the season exactly. Therefore, we set  $season \leftarrow mnth$ . Now at first glance, one might assume that the season influences the weather situation and therefore propose  $mnth \rightarrow season \rightarrow weathersit$ . But we assert that since season can be seen as an more abstract version of the information contained in month, that month is actually a confounder between season and weather situation. Similarly the relationship  $holiday \rightarrow workingday \leftarrow weekday$  is special. To decide if a given day is a workingday, one needs both the information from holiday and weekday. We chose to construct  $G_{man}$  so that only workingday has direct influence on  $cnt$ , but different structures are conceivable. In a similar fashion, we designed  $G_{man}$  for the MPG dataset. The

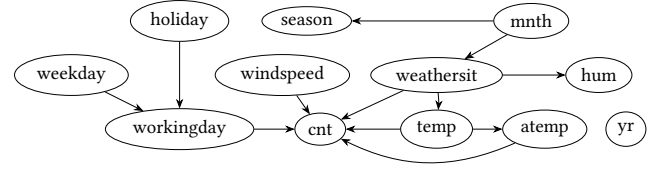


Figure 4: The hand-crafted, assumed causal structure of the bike share dataset. Each node represents a feature of the dataset. The outcome  $cnt$  is the number of rented bikes.

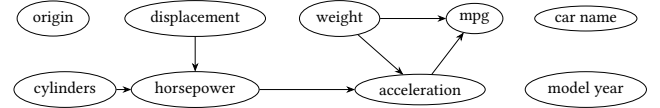


Figure 5: The hand-crafted, assumed causal structure of the MPG dataset. Each node represents a feature of the dataset. The outcome  $mpg$  is the miles per gallon.

main argument is that only acceleration and weight have direct influence on the MPG while e.g. the influence of horsepower exists only mediated by acceleration.

The explanations from standard LIME and CIP-ES are presented in Figures 2 and 3 for the bike share data and 6 and 7 for the MPG data, respectively. While according to standard LIME the year, season and weather situation have the strongest influence on the number of rented bikes, the CIP-ES explanation assigns influence only to four features in total. Our method ensures that only  $PA_y$  can have any influence. Furthermore, due to the sampling process, every  $\tilde{z}_{workingday}^{(i)}$  will have the same value for all  $i \in \{0 \dots N\}$ . Thus, even though  $workingday \in PA_y$ , it will be shown to have no influence in the explanation. This is an unfortunate side effect from the structure  $holiday \rightarrow workingday \leftarrow weekday$  and the way these concrete features are intertwined. This scenario remains an outlier.

Regarding the MPG data, we can see in Fig. 6 that standard LIME stresses the origin and model year. However, these are obviously confounded since building the same car with the same parts in a different country in a different year will not change its MPG. The CIP-ES explanation in Fig. 7 shows that only weight and acceleration have an impact. Note that the acceleration measures the time a car takes to reach a certain speed, therefore, a higher value for the acceleration feature means slower acceleration.

We argue that the assumptions about how the SCM  $\mathcal{C}$  influences the data generation and that are implicitly encoded into the graph  $G_{man}$  improve the explanations by achieving a higher level of Causability with regards to the expectations and knowledge of the user. For example according to standard LIME, the year in which the datapoint was collected has the highest influence. This explanation is dissatisfying without additional information (e.g. new competitor) on why the year would have any influence on the amount of rented bikes, let alone one this strong. In  $G_{man}$  this is reflected by the missing arrow from  $yr$  to  $cnt$ . The effect from e.g. month on count itself is not direct either. It is mediated by the weather situation, i.e. someone would not rent a bike just because it is summer, but

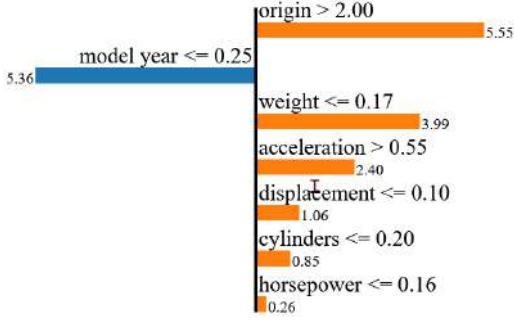


Figure 6: Explanation for a single  $x$  from real MPG data with standard LIME Negative feature attribution is depicted in blue, positive in orange.

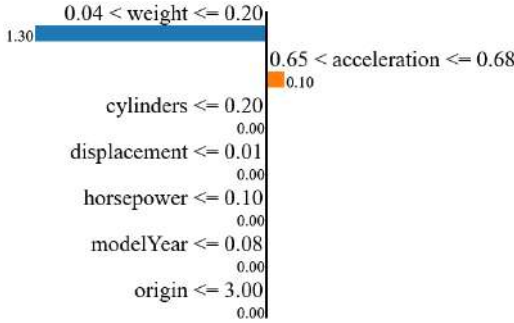


Figure 7: Explanation for a single  $x$  from real MPG data with CIP-ES. Negative feature attribution is depicted in blue, positive in orange.

because the weather situation is well suited to ride a bike, which generally happens more often in summer than e.g. winter.

Those assumptions made by  $G_{man}$  show in the CIP-ES explanation (Figure 3). The focus gets heavily put on the weather situation and to a lesser extent on temperature. Besides those two, only wind speed and *atemp* have any influence. Compared with the vanilla LIME, the CIP-ES explanation is self-explanatory for the user who generated  $G_{man}$ .

The explanations from CIP-ES are therefore sparse compared to those by standard LIME. This is due to  $G_{\tilde{Z}}$  being more complex than  $G_{trio}$  and therefore having nodes that are forced to have no influence. Thus, this behaviour is expected in general and it can be an advantage, since good explanations should focus on a few selected causes [17].

## 4.2 Evaluation of Fidelity and Stability

In line with Cinquini and Guidotti [7], we evaluate CIP-ES according to the fidelity and stability of the generated explanations and compare them with LIME. Fidelity is evaluated as coefficient of determination, also called  $R^2$ -score, of the interpretable model trained on the neighbourhood of  $x$  with respect to the output of the black-box-model, which represents the ground truth. The score quantifies the ability of the interpretable model to mimic the black-box model

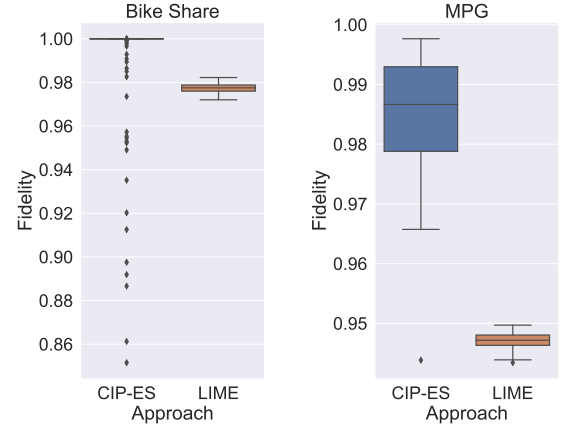


Figure 8: Fidelity of CIP-ES (blue) and LIME (orange) for the Bike Share (left) and MPG (right) dataset.

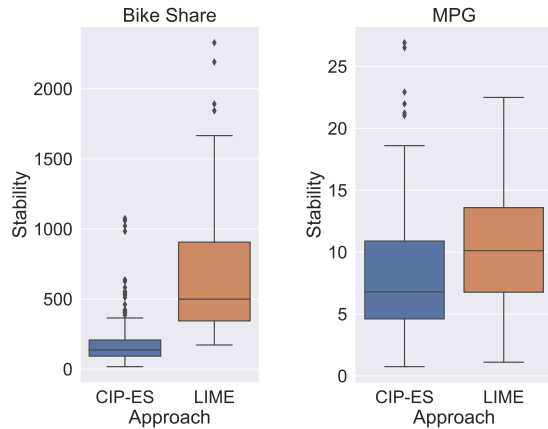
and can be seen as a measure of fidelity of the approach. The maximum value is 1.0. We generate a neighbourhood for each data point in the validation set to assess the fidelity.

The results shown in Fig 8 indicate overall a higher fidelity for the CIP-ES approach compared to standard LIME. The mean fidelity for CIP-ES on the bike dataset is close to 1.0, and for both datasets higher than that of vanilla LIME. This means that the interpretable model  $g(x)$  and the black-box  $f(x)$  are more similar on the CIP-ES neighbourhood  $\tilde{Z}$  than the LIME neighbourhood  $Z$ . This is important since the explanations are given based on  $g(x)$ . Therefore, for the explanations to be relevant, having similar  $g(x)$  and  $f(x)$  is crucial. We argue that this is due to CIP-ES aligning  $\tilde{Z}$  (i.e. the training data for the interpretable model) with the training data of the black-box model.

To evaluate the stability of the approach, we are using a local Lipschitz estimation [2]:  $LLE_x = \text{avg}_{x_i \in \mathcal{N}_x^k} (||e_i - e||_2 / ||x_i - x||_2)$  over a neighbourhood  $\mathcal{N}_x^k$  of a datapoint  $x$  with  $k$  neighbours. This estimation quantifies average proportionality of the difference similar data points  $x_i$  (neighbours) to  $x$  and their explanations  $e$ . A lower value for  $LLE_x$  means higher stability. The neighbourhood is created via an embedding, which is trained on the training dataset. We select a neighbourhood of  $k = 5$  closest neighbours of  $x$  via the Euclidean distance of the embedding vectors. The stability evaluation is conducted for each data point in the validation set. Fig. 9 shows the result for CIP-ES and LIME for both datasets, where CIP-ES has a lower mean  $LLE$  in each dataset. In the bike dataset, CIP-ES has a lower variance compared to LIME.

## 4.3 Limitations of Our Approach

While the integration of causality into methods from XAI is promising, as substantiated by our proposed CIP-ES approach, additional research efforts are required. We evaluated CIP-ES with LIME [22] on tabular data only. While a key advantage of LIME is that it is not only model-agnostic, but also works on multiple data modalities. For other modalities such as images, it is more challenging to



**Figure 9: Stability of CIP-ES (blue) and LIME (orange) for the Bike Share (left) and MPG (right) dataset.**

generate a causal graph and sample according to it, which we aim to address in future work.

## 5 CONCLUSIONS

We present CIP-ES, an approach to perform input perturbation based on a causal graph. We propose to have this graph be given by the user, therefore matching the generated explanations from the surrogate model to the users expectations and knowledge. We argue that CIP-ES is therefore *true to the user* and generates explanations that are self-explanatory to the user. We sample the input for the surrogate model according to the graph and therefore ensure that the explanations follow the graph.

Since the quality of an explanation is heavily subjective and is itself also interpreted by the user, we argue that this *true to the user* explanation can be useful. A possible scenario might be that multiple users or stakeholders agree upon one causal graph first and then will also agree on following explanations.

Future work will focus on extending the approach to different modalities. In addition, we aim at integrating CIP-ES with other explanation methods besides LIME that are based on input perturbation and a surrogate model.

## REFERENCES

- [1] Peter Martey Addo, Dominique Guegan, and Bertrand Hassani. 2018. Credit Risk Analysis Using Machine and Deep Learning Models. 6, 2 (2018), 38. Issue 2. <https://doi.org/10.3390/risks6020038>
- [2] David Alvarez Melis and Tommi Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc.
- [3] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. 2022. On Pearl’s Hierarchy and the Foundations of Causal Inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl* (1 ed.). Association for Computing Machinery, 507–556.
- [4] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statist. Sci.* 16, 3 (2001), 199–231. <https://doi.org/10.1214/ss/1009213726>
- [5] Hugh Chen, Joseph D. Janizek, Scott Lundberg, and Su-In Lee. 2020. True to the Model or True to the Data? *arXiv:2006.16234* [cs, stat]
- [6] David Maxwell Chickering, David Heckerman, and Christopher Meek. 2004. Large-Sample Learning of Bayesian Networks Is NP-Hard. *Journal of Machine Learning Research* 5 (Dec. 2004), 1287–1330.
- [7] Martina Cinquini and Riccardo Guidotti. 2022. CALIME: Causality-Aware Local Interpretable Model-Agnostic Explanations. *arXiv e-prints*, Article arXiv:2212.05256 (Dec. 2022), arXiv:2212.05256 pages. [arXiv:2212.05256](https://arxiv.org/abs/2212.05256) [cs.AI]
- [8] Jonathan Crabbe, Zhaozhi Qian, Fergus Imrie, and Mihaela van der Schaar. 2021. Explaining Latent Representations with a Corpus of Examples. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., 12154–12166.
- [9] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [10] Hadi Fanaee-T and Joao Gama. 2013. Event Labeling Combining Ensemble Detectors and Background Knowledge. *Progress in Artificial Intelligence* 2 (2013), 1–15. <https://doi.org/10.1007/s13748-013-0040-3>
- [11] Christopher Frye, Colin Rowat, and Ilya Feige. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 1229–1239.
- [12] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 4778–4789.
- [13] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery* 9, 4 (2019), e1312. <https://doi.org/10.1002/widm.1312> PMID:32089788
- [14] Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. 2022. On Measuring Causal Contributions via do-interventions. In *International Conference on Machine Learning*. PMLR, 10476–10501.
- [15] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. *arXiv:2206.15475* [cs, stat]
- [16] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (2017), Vol. 30. Curran Associates, Inc.
- [17] Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (second ed.).
- [18] Raha Moraffah, Mansoor Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal Interpretability for Machine Learning - Problems, Methods and Evaluation. *ACM SIGKDD Explorations Newsletter* 22, 1 (May 2020), 18–33. <https://doi.org/10.1145/3400051.3400058>
- [19] Judea Pearl. 2009. *Causality* (2 ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511803161>
- [20] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [21] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2021. Secure and Robust Machine Learning for Healthcare: A Survey. *IEEE Reviews in Biomedical Engineering* 14 (2021), 156–180. <https://doi.org/10.1109/RBME.2020.3013489>
- [22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- [23] Johannes Schneider and Joshua PETER Handali. 2019. PERSONALIZED EXPLANATION FOR MACHINE LEARNING: A CONCEPTUALIZATION. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Stockholm & Uppsala, Sweden.
- [24] Lloyd S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games II*, Harold W. Kuhn and Albert W. Tucker (Eds.). Princeton University Press, Princeton, 307–317.
- [25] Peter Spirtes, Clark N. Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*. MIT press.
- [26] Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu, and Xianzhi Wang. 2021. Causality Learning: A New Perspective for Interpretable Machine Learning. *arXiv:2006.16789* [cs, stat]
- [27] Aleš Završnik. 2020. Criminal Justice, Artificial Intelligence Systems, and Human Rights. *ERA Forum* 20, 4 (2020), 567–583. <https://doi.org/10.1007/s12027-020-00602-0>
- [28] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. 2018. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

# Research on Identification Method of Gap Nonlinear Vibration

Jialiang, S, and Sun\*

The Fifth Institute of Electronics, Ministry of Industry and  
Information Technology  
sunjialiangsy22@163.com

Jingying,L, and Liu

The Fifth Institute of Electronics, Ministry of Industry and  
Information Technology  
1099098464@qq.com

## ABSTRACT

The research of linear vibration has been very mature, but the gap nonlinear element widely exists in the actual structure and vibration system. The general linear vibration theory can not meet the needs of solving the nonlinear dynamic problems of clearance, and the research on the nonlinear vibration of clearance is essential. Based on the description function method, the forced response calculation of linear and nonlinear systems and the nonlinear detection method of clearance systems are studied in this paper. The forced response calculation of linear and clearance nonlinear systems mainly depends on the digital filtering method to realize the corresponding simulation of the system with clearance nonlinear. Based on the definition of linear system, a concrete scheme for judging whether the system contains the nonlinear characteristics of clearance is given. The position identification of the clearance nonlinear system based on the description function and the polynomial fitting inversion of the description function is studied and verified by an example.

## CCS CONCEPTS

• :: • **Theory of computation** → Models of computation;

## KEYWORDS

Gap nonlinearity, Description function method, Polynomial fitting inversion, Vibration

## ACM Reference Format:

Jialiang, S, and Sun and Jingying,L, and Liu. 2023. Research on Identification Method of Gap Nonlinear Vibration. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023)*, March 17–19, 2023, Shanghai, China. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590138>

## 1 INTRODUCTION

With the development of science and technology, mechanical vibration has become an important issue in various engineering fields. In the past, people have been paying attention to the research of linear vibration, and its theory and method have been improved and can be used for modeling, simulating and measuring linear

\*ialiang Sun(1997-),male, assistant engineer, Master degree in reliability, engaged in reliability and systems engineering research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
CACML 2023, March 17–19, 2023, Shanghai, China

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9944-9/23/03...\$15.00  
<https://doi.org/10.1145/3590003.3590138>

systems. However, there are various nonlinear factors in the actual mechanical system.

It is well known that the actual system in engineering almost always contains various nonlinear factors, such as the clearance, dry friction, bearing oil film in mechanical system, large deformation of structural system, nonlinear material constitutive relationship, nonlinear control strategy of control system, etc. [1]The linear system model is a simplified model of the system that requires low accuracy for the convenience of analysis or the system nonlinearity has little impact on the system performance. Generally, the linear system model can well approximate the dynamic behavior of the actual system; However, in recent years, with the development and progress of science and technology and the continuous improvement of system performance requirements, this linear approximation is not always reliable, and the neglected nonlinear factors sometimes cause unacceptable errors in analysis and calculation; Moreover, more and more nonlinear phenomena in engineering have also attracted people's attention, and nonlinear problems have become one of the hot issues in current research. Therefore, it is necessary to carry out nonlinear research on nonlinear systems and reveal the essence of nonlinear systems, which is of great significance for the analysis and design of nonlinear systems. For this reason, many mathematical theories and methods have been developed to model, solve and analyze nonlinear systems.

## 1.1 Background and significance

In essence, all mechanical problems in engineering are nonlinear. Some classical mechanical theories are based on the simplified treatment of practical problems based on some assumptions, such as small deformation assumption, linear elasticity assumption, and the assumption that the boundary conditions remain unchanged. If any of the above assumptions is not satisfied, a nonlinear phenomenon will occur, which corresponds to geometric nonlinearity, material nonlinearity, and boundary nonlinearity respectively.

There are two main reasons to promote the development of nonlinear recognition. First, in the vibration experiment, the nonlinear phenomenon of the gap has attracted more and more attention of engineers. For example, in the linear modal analysis of Airbus A400M and A350XWB, the gap nonlinear phenomenon was found in the elastic mounting device and hydraulic cylinder, landing gear, and fuselage tail cone auxiliary power unit. Secondly, the emphasis on environmental protection and other aspects has promoted the development of nonlinear identification laterally. For example, a report by the European Aviation Research Advanced Group claimed that<sup>[2]</sup> the target of reducing the emissions of carbon dioxide and nitrogen oxide per kilometer per passenger by 75% and 90% respectively must be achieved by 2050. In order to reduce the energy consumption of aviation equipment, it is a simple and effective

method to use emerging composite materials to reduce the equipment mass, but the reduction of equipment mass inevitably makes the gap nonlinear behavior in the system more obvious [3].

## 1.2 Accessibility

Gap nonlinear structure identification is mainly divided into three steps, namely nonlinear detection, position identification and type identification. The so-called nonlinear detection mainly solves the problem of judging whether the gap nonlinear system contains nonlinear, that is, judging whether the movement of structural members shows nonlinear characteristics. Nonlinear position identification is mainly used to find and determine the position where the gap nonlinearity exists. The main problem of nonlinear feature description is to determine the type of nonlinearity (gap nonlinearity) and describe the function form of nonlinearity<sup>[4]</sup>.

The nonlinear system identification method based on the description function method is a relatively intuitive and effective method. In the existing methods and processes, it is very difficult to identify the system with gap nonlinear elements and other nonlinear combinations. And the method can not give qualitative or quantitative criteria for the effectiveness of recognition results. Therefore, it is necessary to improve the existing methods and processes based on description function recognition, and provide criteria for the effectiveness of recognition results.[5]

Simulation tools are a great help in the research of theoretical methods. Accurate and effective simulation tools can, on the one hand [6], avoid the influence of uncontrolled or unpredictable factors existing in the actual project on the research of identification methods. On the other hand, the parameter information of the nonlinear system built by the simulation tools can be known and can be used to verify whether the identification method is effective. Therefore, the establishment and improvement of nonlinear system simulation tools is also one of the research work of this paper.[7]

## 2 MODELING AND DETECTION OF GAP NONLINEAR DYNAMIC SYSTEM

In the study of nonlinear dynamic systems, modeling and simulation are two important links. The research of nonlinear theory and simulation benefits from the development of natural science and computer technology since the 20th century.[8] In this chapter, some main methods are analyzed and examples are applied to calculate, laying a theoretical foundation for the key problems that may be involved, and providing examples for the experimental verification of the gap nonlinear system.[9]

### 2.1 Modeling of nonlinear dynamics

The simplest dynamic system is a single-degree-of-freedom system. The single-degree-of-freedom MCK (mass, damping, stiffness) system as shown in the figure is investigated.[10] Under the excitation force  $f(t)$ , the mass  $M$  will vibrate. The differential equation of motion of the system is established by using the principle of d'Alembert. Perform force analysis on mass block  $M$ , as shown in Figure 1

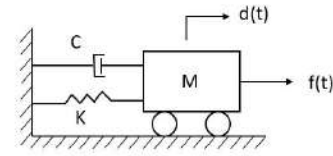


Figure 1: MCK system with single degree of freedom

The force balance equation of mass  $M$  in the direction of motion can be obtained.

$$f_C(t) + f_K(t) + f_I(t) = f(t) \quad (1)$$

After obtaining the differential equation of motion of the single-degree-of-freedom MCK system, it is not difficult to extend it to the discrete system with  $N$  degrees of freedom. [11]For such a multi-degree-of-freedom system, a set of force balance equations can be obtained:

$$M\ddot{d}(t) + C\dot{d}(t) + Kd(t) = f(t) \quad (2)$$

$[M]$ ,  $[C]$ ,  $[K]$  are  $N \times N$  matrix, which is the mass, damping and stiffness matrix of the system.

### 2.2 Gap nonlinear dynamic systeming

The common nonlinear types in dynamic tests include nonlinear damping, polynomial stiffness, saturation, clearance, Coulomb friction, etc. These nonlinear phenomena are often related to displacement and velocity. According to the relationship with displacement or velocity, nonlinearity can be divided into stiffness nonlinearity and damping nonlinearity. A system that contains both stiffness nonlinearity and damping nonlinearity is called hysteresis nonlinearity. [12]The nonlinear characteristics are complex. In nonlinear research and analysis, researchers usually simplify and express the nonlinear characteristics as the characteristics of force and displacement or velocity. Table 1 lists the types of clearance nonlinearity and their expressions and diagrams

The dynamic equation of the linear system is known. The motion equation of the multi-degree-of-freedom nonlinear system is the combination of the linear partial dynamic equation of the system and the function of the nonlinear element:

$$[M] \{\ddot{x}\} + [C] \{\dot{x}\} + [K] \{x\} + f_n(x, \dot{x}) = \{f\} \quad (3)$$


Where,  $f_n(x, \dot{x})$  is the nonlinear force of the system, which is a function of displacement or velocity;  $\{f\}$  is external incentive.

### 2.3 System nonlinear detection

The so-called nonlinear detection of the system is to detect whether the gap system has nonlinear characteristics. The simplest way to detect system nonlinearity is to use the definition of linear system.

The so-called linear system means that the input and output of the system meet the superposition principle. That is, if the response of the system to an arbitrary excitation force is  $y_1(t)$ , [13]and the

**Table 1:** Expression and diagram of gap clearance

Nonlinear type	Expression	FIG
Gap Nonlinear	$\begin{cases} \dot{x}_1 = v_1 \\ \dot{v}_1 = -\frac{1}{m_1}k_1x_1 - \frac{1}{m_1}k_2x_2 \\ \dot{x}_2 = v_2 \\ \dot{v}_2 = -\frac{1}{m_2}k_2x_1 - \frac{1}{m_2}k_3x_2 \end{cases}$	

response to another independent external excitation force  $X_2^{(t)}$  is  $y_2^{(t)}$ , then for any constant a and b, the response of the system when subjected to the superimposed external force  $ax_1(t) + bx_2(t)$  should be the sum of the responses obtained by the corresponding independent excitation forces, that is, the response of the system should be  $ay_1(t) + by_2(t)$ . Therefore, it is only necessary to use two different levels of excitation signals to excite the system, and analyze the two groups of excitation and response signals of the obtained system to calculate whether it meets the superposition principle. [14] If it meets the requirements, it is a linear system. If it does not, it is proved that the system contains nonlinear characteristics. The specific judgment basis can be calculated by formula (4).

$$err(t) = \frac{\frac{y_2(t)}{x_2(t)} - \frac{y_1(t)}{x_1(t)}}{\frac{y_1(t)}{x_1(t)}} \times 100\% \quad (4)$$

Where  $err(t)$  is the defined "relative error parameter", which is a time domain function. According to the principle of superposition, if the system is a linear system, the value of  $err(t)$  should be infinitely close to zero at any time; If the system has clearance nonlinearity,  $err(t)$  will appear relatively large.

Using the digital filter method introduced earlier, a linear system and a nonlinear system are constructed respectively. The linear parts of the two systems are the same, expressed as MCK matrix:

$$M = [1000; 200; 30; 5];$$

$$C = [10 \ -5 \ 0 \ 0; -5 \ 10 \ -5 \ 0; 0 \ -5 \ 10 \ -5; 0 \ 0 \ -5 \ 10];$$

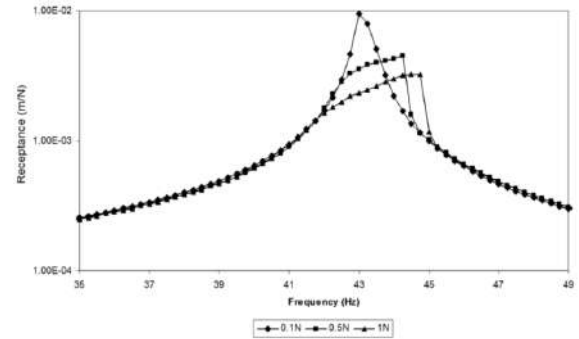
$$K = [1000 \ -500 \ 0 \ 0; -500 \ 1000 \ -500 \ 0; 0 \ -500 \ 1000 \ -500; 0 \ 0 \ -500 \ 1000];$$

The constructed nonlinear system has a gap nonlinearity between 1 degree of freedom and the ground. The expression of nonlinear force is:

$$nlf = \begin{cases} 0 & |x_1| < 0.005 \\ 100 \cdot (x_1 - 0.005 \cdot \text{sign}(x_1)) & |x_1| > 0.005 \end{cases} \quad (5)$$

The two systems are respectively given a step signal with amplitude of 0.1N and 1N as the excitation signal. For linear systems, the order of magnitude of the relative error parameter  $err(t)$  is about 10 to the negative 12th power, infinitely close to 0, indicating that the input and output signals of the system meet the superposition principle, and the system is linear; [15] For nonlinear systems, the relative error parameter  $err(t)$  has a numerical order of about 10. It is obvious that the input and output of the system do not meet the superposition principle, and the system contains nonlinear characteristics.

In the frequency domain, we can also detect the gap nonlinear characteristics of the system through the frequency response function curve of the output ratio of the system under different excitation. Similar to the definition in the time domain to determine whether the system contains nonlinear elements, if the system is a linear system, the frequency response function representing the transfer characteristics between the input and output of the system should not change with the change of the system excitation, that

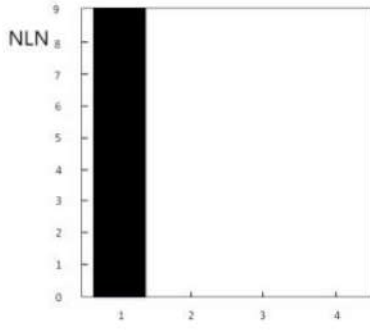
**Figure 2:** FRF of gap nonlinear system under different excitation levels

is, the frequency response function curves of the input and output of the system under different excitation should be coincident. Using the above system model, if the system is a gap nonlinear system, the frequency response function of the system's input and output is not only a function of frequency, but also affected by the excitation signal. That is to say, the input and output frequency response functions of the system will be different under different excitation, not strictly coincident, as shown in Figure 2. Similarly, the relative error of the frequency response function under different levels of excitation can be calculated by imitating the "definition of the relative error parameter" in the time domain. The "relative error" obtained by the frequency response function of the same linear system without/with gap nonlinear elements is shown in the dotted line and the solid line in the figure, which can also clearly determine the system linearity and gap nonlinearity.

### 3 GAP NONLINEAR POSITION IDENTIFICATION BASED ON DESCRIPTION FUNCTION METHOD

Gap nonlinear structure identification is mainly divided into three steps, namely nonlinear detection, position identification, and parameter estimation. The so-called clearance nonlinear detection mainly solves the problem of judging whether the clearance system contains nonlinearity, that is, judging whether the vibration of structural members shows nonlinear characteristics. The nonlinear position identification of clearance is mainly to find and determine the position of clearance. The nonlinear parameter estimation mainly determines the parameters of the clearance nonlinear model. This chapter mainly introduces the position identification and parameter identification of the system gap nonlinear element based on the description function method.[16]





**Figure 3: Position identification of multi-degree-of-freedom system with only one clearance nonlinear element**

Then the "nonlinear index (NLN)" for nonlinear position recognition can be written as:

$$NLN_r = [\Delta_r] \left\{ H_i^N \right\} = \Delta_{r1} H_{1i}^N + \Delta_{r2} H_{2i}^N + \Delta_{r3} H_{3i}^N + \cdots + \Delta_m H_{ni}^N \quad (20)$$

In the equation, if there is nonlinearity in the degree of freedom  $r$ ,  $[\Delta_r]$  is not 0, and  $NLN_r$  is not a value of 0, which indicates that there is nonlinearity in the degree of freedom  $r$ .  $NLN_r$  can be calculated from the right side of the equation

$$NLN_r = \delta_{ri} - Z_{r1} \cdot H_{1i}^N - Z_{r2} \cdot H_{2i}^N - \cdots - Z_{rn} \cdot H_{ni}^N \quad (21)$$

Among them, the linear frequency response function  $[H^L]$  can be approximated by the response obtained by small magnitude excitation of the system. Therefore, the nonlinear position  $NLN_r$  can be identified by summing all calculated frequencies.

After identifying the nonlinear position  $NLN$ , the nonlinear matrix can be calculated through the equation, and then the nonlinear description function of the gap can be obtained. The nonlinear type and parameters can be obtained by fitting the nonlinear description function curve.

### 3.3 Simulation case

In order to verify the effectiveness of the position recognition method, a four-degree-of-freedom nonlinear system with clearance is now constructed as shown in the figure. The parameters of the linear part of the system are:

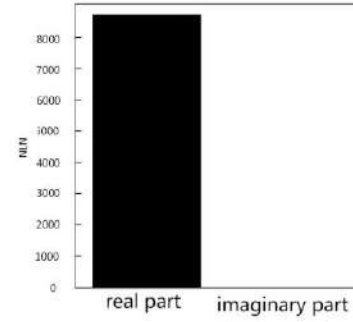
$$M = [1 \ 0 \ 0 \ 0; 0 \ 2 \ 0 \ 0; 0 \ 0 \ 3 \ 0; 0 \ 0 \ 0 \ 5];$$

$$C = [10 \ -5 \ 0 \ 0; -5 \ 10 \ -5 \ 0; 0 \ -5 \ 10 \ -5; 0 \ 0 \ -5 \ 10];$$

$$K = [1000 \ -500 \ 0 \ 0; -500 \ 1000 \ -500 \ 0; 0 \ -500 \ 1000 \ -500; 0 \ 0 \ -500 \ 1000];$$

On the basis of the above linear system, there is stiffness nonlinearity with clearance between the first degree of freedom of the multi-degree of freedom system and the ground:

At m1, sinusoidal sweep excitation with amplitude level of  $N=0.1N$  and  $N=10N$  is respectively given, and the response from m1 to point m4 is measured. By calculating the nonlinear index  $NLN_r$  of each degree of freedom. The results are shown in Figure 3 and Figure 4



**Figure 4: Virtual and real distribution of NLN with 1 degree of freedom**

### 3.4 Summary of this chapter

This chapter introduces the position identification of nonlinear elements of the system based on the description function method. First, the theoretical derivation of this method is systematically introduced. Through the simulation case, it is proved that this method can give a relatively accurate determination of the position of the nonlinear elements of the system in the gap nonlinear system with only one nonlinear element. At the same time, how to identify the types and parameters of nonlinear elements by describing the function and describing the inverse of the function is introduced in detail. In order to improve the accuracy of parameter identification, the unconstrained linear optimization function `fminsearch.m` in MATLAB is used for parameter estimation. Finally, this chapter has carried out the simulation verification on the clearance nonlinear identification. The simulation results fully show that the description function and polynomial inversion method have good recognition effect on the clearance nonlinear system.

## 4 SUMMARY AND PROSPECT

According to the definition of nonlinear characteristic matrix of clearance nonlinear system and the calculation characteristics of nonlinear frequency response function matrix, combined with the definition method of Sherman Mason method, this paper combines the characteristics of nonlinear frequency response function matrix with the results of nonlinear position. The characteristics of the function matrix and the results of nonlinear position identification are solved by using the effective column in the matrix to represent the nonlinear description function of the system. Finally, the correctness of the algorithm is proved by a concrete example. In this paper, the polynomial fitting inversion method is used to transform the description function representing nonlinear characteristics into the form of restoring force. By observing the calculation formula of the description function, it is concluded that the solution of the description function and the inversion process of the corresponding description function meet the superposition principle. Through the analysis of the conversion coefficient between the first several orders of nonlinear force and its corresponding description function, it is found that the conversion coefficient of any fixed order nonlinear force is fixed and can be calculated by general formula. The examples in this paper prove the feasibility of nonlinear polynomial

inversion for clearance nonlinear system, but there are still several aspects to be improved:

1. For the function method described in this article, only a few simple attempts have been made. For other more complex types, such as the identification of nonlinear systems with multiple non-linear elements, further research efforts are needed.

2. Calculation of nonlinear description function. Although different calculation methods are given according to different system types in this paper, the linear frequency response function and clearance nonlinear frequency response function of the system are obtained by different excitation levels. It is assumed that the non-linearity contained in the system is the response displacement and velocity, and the response functions between different non-linear elements are independent of each other. It does not consider the calculation of complex nonlinear description function when there is coupling between nonlinear components, nor does it consider the influence of external environment on the dynamic characteristics of the system, especially the nonlinear characteristics.

It is hoped that the follow-up work can achieve more applications of nonlinear types, deal with more and more complex coupling nonlinear types, and further improve and verify the theory in combination with practical engineering experiments.

## REFERENCES

- [1] J.P. Noël, L. Renson, G. Kerschen, Complex dynamics of a nonlinear aerospace structure: experimental identification and modal interactions, *J. Sound Vib.* 333 (2014) 2588–2607
- [2] High Level Group on Aviation Research, ACARE Flightpath 2050 – Europe’s vision for aviation.
- [3] [http://www.acare4europe.com/sites/acare4europe.org/files/document/Flightpath2050\\_Final.pdf](http://www.acare4europe.com/sites/acare4europe.org/files/document/Flightpath2050_Final.pdf), Visited on 1 October 2015
- [4] Y.S. Lee, A.F. Vakakis, D.M. McFarland, L.A. Bergman, Non-linear system identification of the dynamics of aeroelastic instability suppression based on targeted energy transfers, *Aeronaut. J.* 114 (2010) 61–82
- [5] J.P. Noël, S. Marchesiello, G. Kerschen, Subspace-based identification of a nonlinear spacecraft in the time and frequency domains, *Mech. Syst. Signal Process.* 43 (2014) 217–236
- [6] C.M. Richards, R. Singh, Identification of multi-degree-of-freedom non-linear systems under random excitations by the “reverse path” spectral method, *J. Sound Vib.* 213 (1998) 673–708
- [7] S.M. Spottiswood, R.J. Allemang, Identification of nonlinear parameters for reduced order models, *J. Sound Vib.* 295 (2006) 226–245
- [8] M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, John Wiley & Sons, New York, NY, USA, 1980
- [9] J.P. Noël, G. Kerschen, Frequency-domain subspace identification for nonlinear mechanical systems, *Mech. Syst. Signal Process.* 40 (2013) 701–717
- [10] J.P. Noël, G. Kerschen, E. Foltête, S. Cogan, Grey-box identification of a nonlinear solar array structure using cubic splines, *Int. J. Non-linear Mech.* 67 (2014) 106–119
- [11] D.J. Ewins, A future for experimental structural dynamics, in: *Proceedings of the International Conference on Noise and Vibration Engineering (ISMA)*, Leuven, Belgium, 2006
- [12] S.W. Shaw, C. Pierre, Normal modes for non-linear vibratory systems, *J. Sound Vib.* 164 (1) (1993) 85–124
- [13] J.R. Wright, M.F. Platten, J.E. Cooper, M. Sarmast, Identification of multi-degree-of-freedom weakly non-linear systems using a model based in modal space, in: *Proceedings of the International Conference on Structural System Identification*, Kassel, Germany, 2001
- [14] M.F. Platten, J.R. Wright, J.E. Cooper, G. Dimitriadis, Identification of a nonlinear wing structure using an extended modal model, *AIAA J. Aircr.* 46 (5) (2009) 1614–1626
- [15] M. Peeters, G. Kerschen, J.C. Golinval, Dynamic testing of nonlinear vibrating structures using nonlinear normal modes, *J. Sound Vib.* 330 (2011) 486–509
- [16] M. Haroon, D.E. Adams, A modified H2 algorithm for improved frequency response function and nonlinear parameter estimation, *J. Sound Vib.* 320 (2009) 822–837
- [17] Azenha A., Machado J. On the describing function method and the prediction of limit cycles in nonlinear dynamical systems[J]. *Systems Analysis Modelling Simulation*, 1998, 33(3):307–320.

# Real-time Emulation of MASQUE-based QUIC Proxying in LTE Networks using ns-3

Donát Scharnitzky  
dscharnitzky@edu.bme.hu

Department of Telecommunications and Media  
Informatics, Budapest University of Technology and  
Economics  
Budapest, Hungary

Sándor Molnár  
molnar@tmit.bme.hu

Department of Telecommunications and Media  
Informatics, Budapest University of Technology and  
Economics  
Budapest, Hungary

Zsolt Krämer  
zsolt.kramer@ericsson.com  
Ericsson  
Budapest, Hungary

Attila Mihály  
attila.mihaly@ericsson.com  
Ericsson  
Budapest, Hungary

## ABSTRACT

Tools for real-time emulation of mobile networks are valuable for researchers due to the high amount of time and resources it allows to save compared to carrying out measurements in live networks. In this paper we present the rationale, design and prototype implementation of a novel net device in the ns-3 open source network simulator that allows for end-to-end real-time emulation of LTE networks with real endpoints. We then show the performance evaluation of a QUIC proxy built on MASQUE using our emulated LTE setup. Our results confirm the intended behavior of the implementation, however, we also show the limitations of the real-time capabilities of ns-3.

## CCS CONCEPTS

• **Networks** → **Network simulations**; *Transport protocols*; • **Software and its engineering** → *Software implementation planning*.

## KEYWORDS

ns-3, QUIC, MASQUE, real-time emulation

### ACM Reference Format:

Donát Scharnitzky, Zsolt Krämer, Sándor Molnár, and Attila Mihály. 2023. Real-time Emulation of MASQUE-based QUIC Proxying in LTE Networks using ns-3. In *2023 2nd Asia Conference on Algorithms, Computing and Machine Learning (CACML 2023), March 17–19, 2023, Shanghai, China*. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3590003.3590995>

## 1 INTRODUCTION

ns-3 is a discrete-time network simulator written in C++ [15]. It is widely used in networking for studies related to routing, transport

layer, testing prototypes, and a variety of other scenarios. It is also capable of simulating mobile networks including LTE and has an external mmWave module [10]. ns-3 has a real-time simulation mode, which can be used to work with real servers when emulating mobile networks, however, a real client is not supported (at the writing of this paper) and this limits the possible measurement scenarios.

QUIC is a novel transport protocol created by Google and under standardization by IETF [9]. In the design of QUIC a core concept is encryption, which makes it impossible to manage it from middleboxes. This feature increases privacy and security at the cost of functionality (e.g., performance enhancing proxies are hard to design and deploy). To overcome this limitation, MASQUE was proposed in IETF, which enables the creation of cooperative QUIC proxies, without compromising the encryption [6].

The main motivation of our work is to extend the emulation capabilities of the ns-3 simulator with real-time, end-to-end LTE emulation. For this purpose we designed and implemented a new net device module for the popular ns-3 simulator (a net device represents a NIC in ns-3). This enables a real client to communicate with a real server via an emulated LTE network. Such an environment provides an appropriate platform to easily perform performance evaluation of new protocols, such as QUIC, where the protocol is rapidly evolving and the continuous evaluation and testing is especially important.

The contribution of this paper is twofold. First, we present the design and implementation of our novel net device module, which enables LTE UE emulation with a real client node. We also show its working mechanism and how it can be integrated in a ns-3 environment. Second, we demonstrate its use by an important case study of a QUIC proxy built on MASQUE protocol using our emulated LTE setup. We also show our first performance evaluation results of this study focusing on metrics like the completion time and the round-trip time. For comparison purposes the study includes three cases, i.e., the IP forwarding case and also two modes when the MASQUE proxy is used (stream and data forwarding modes). For the sake of the comparison we also consider both low throughput and high throughput cases.



This work is licensed under a Creative Commons Attribution International 4.0 License.

CACML 2023, March 17–19, 2023, Shanghai, China  
© 2023 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-9944-9/23/03.  
<https://doi.org/10.1145/3590003.3590995>

The paper is organized as follows. We discuss related works regarding ns-3, proxying and QUIC in Section II. We present the implementation highlights in Section III. Our net device and the measurement environment are presented In Section IV. The measurement results is demonstrated in Section V. Finally, Section VI concludes our paper.

## 2 RELATED WORK

### 2.1 Evaluating the performance of QUIC in simulations

An implementation of the QUIC protocol in ns-3 has been presented in [3]. The authors describe the design and implementation in detail and validate the behavior of the protocol with various congestion control algorithms. The code has been made public, and has been used by the community in further research projects such as [1], where the authors study the performance of QUIC with interactive, low-latency traffic such as cloud gaming, 4K streaming and online gaming. This implementation has also been used in [8], which compares TCP and QUIC performance in LTE networks. The module since then has been extended in [12] where they added the BBR (Bottleneck Bandwidth and RTT) congestion control algorithm to the protocol and provided interfaces for further extensions. While the proper behavior of QUIC is validated and the module is immensely valuable for researchers, it takes considerable effort to keep the implementation aligned with the latest version of the protocol.

### 2.2 Extending the emulation capabilities of ns-3

The ns-3 simulator[15] and its models have proven to behave realistically and are generally considered to be capable of producing measurement results that accurately represent the behavior of real networks. However, increasing the fidelity of the simulations by using real implementations of protocols and algorithms instead of the ns-3 models remains a crucial area of research. Besides the increased accuracy of the findings, another advantage of these approaches is that it can mitigate the duplicate efforts of implementing both the real-world protocols and the simulation models. One important achievement in this field is the Direct Code Execution cradle for the ns-3 simulator [19], which enables the use of real Linux kernel network stacks on the endpoints with simulated networks. Another direction towards similar goals is the extension of the emulation capabilities of the simulator, such as by the authors of [2] and [4]. These works proposed enhancements to the real-time emulation mode of ns-3, making it easier to use and configure real network interfaces for the simulations.

The authors of [13] showed that it is possible to extend the emulation support based on the EmuFdNetDevice class of the simulator, by designing a novel, DPDK-based file descriptor net device. The measurements presented in the paper showed that DpdkNetDevice was capable of achieving much higher data rates with significantly lower CPU load. Moreover, the detailed description of the design and implementation of the new net device class is also available in the paper. An interesting use case of such an emulation-based testbed can be seen in [18] where a framework is presented for QUIC interoperability testing.

Real-time emulation of LTE networks with ns-3 has been studied by [5] and [16]. Both are complex solutions, with [5] using a modified LTE stack integrated with LabView and [16] creating an integrated approach by combining ns-3 and the CORE simulator in order to achieve the result.

### 2.3 MASQUE and Cooperative Proxying of QUIC Traffic

Multiplexed Application Substrate over QUIC Encryption (MASQUE) [17] is a new protocol for using QUIC as tunnel for IP and UDP traffic. It extends the HTTP CONNECT method, thus it requires explicit user request in order to work. The use case for such a protocol is similar to VPNs (Virtual Private Networks), a proxy is requesting objects from servers on behalf of the user, which can improve privacy depending on threat level (shifts the trust to the proxy provider from the web service provider in the context of the Internet). MASQUE can be used to tunnel a QUIC connection if the client supports it as well. The two different modes of MASQUE are datagram mode and stream mode. In datagram mode, the tunneling connection (also referred to as outer connection) does not acknowledge, and in stream mode, it does. Building on MASQUE makes it possible to implement and deploy middlebox functionalities with explicit consent and without compromising the integrity of the privacy or security context of the connection [6].

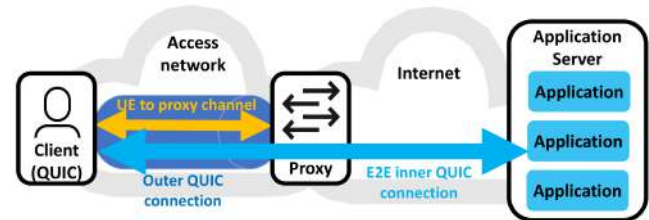


Figure 1: Proxy functionality with QUIC tunnelling

The authors of [7] studied the performance implication of using MASQUE with different protocols (including QUIC). The throughput is a few percent less compared to traffic without tunneling, and in some cases MASQUE has better performance, e.g. when the client-proxy link is noisy but has low delay.

## 3 IMPLEMENTATION

In this section we describe our additions and modification to ns-3, focusing on the novel net device that enables LTE UE emulation with a real client node.

### 3.1 LteUeFdNetDevice

We aimed for a setup where we have a real client and a real server with a real proxy in front of it, with the proxy and the client being connected through ns-3 with LTE emulation, however this is not supported by ns-3. The main reason is that NAT is not implemented in the UE of LTE network and the LteUeNetDevice (C++ class of the net device that a host uses to connect to the LTE network in ns-3) does not support using file descriptors to communicate with the outside environment. The latter limitation could be potentially

overcome by using two NetDevices (the C++ class of ns-3 net device), but we opted for creating a new NetDevice that is a mix of LteUeNetDevice and FdNetDevice (C++ class of the net device in ns-3 that supports file descriptors) to have finer control over it, called LteUeFdNetDevice. We use this new net device to mitigate the first limitation by implementing a simplified, limited NAT (with the help of another new net device on the other end of the LTE network).

This net device uses file descriptors in downlink and LTE in up-link directions, which enables the use of a real client to connect to it and then use the simulated LTE to connect to the proxy. LteUeFdNetDevice inherits from LteUeNetDevice, thus ns-3 can use it as the net device of the UE. The file descriptor capability was implemented based on FdNetDevice, however we do not inherit from it to reduce complications originating from the fact that both of these classes have the same base class (NetDevice) and implement (some of) its methods as shown in Figure 2. The classes with green background are added by us, they are based on existing implementations.

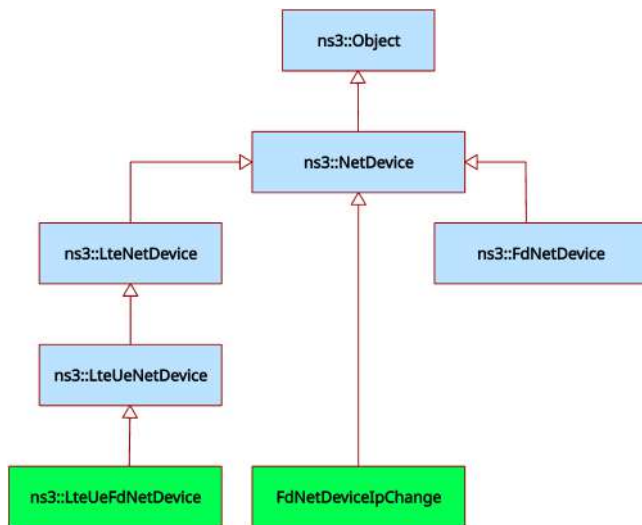


Figure 2: Inheritance diagram of net devices. Our additions are shown with green background.

### 3.2 Packet journey

When a packet coming from the client arrives at the LteUeFdNetDevice, it will be handled by the ForwardUp C++ function at some point, where we call the LTE send function. This behavior basically routes the packet from the file descriptor to the LTE network. In the LTE network the eNodeB gets the packet via radio link, and sends it to the SGW node, which sends it to the PGW node, and then finally it is sent to the RightNode. This node is in ns-3 with a modified FdNetDevice (we call FdNetDeviceIpChange), and its task is to implement the communication between ns-3 and the proxy. It sends the packet to the proxy unmodified, the proxy handles it and replies. The packets coming back are handled by the ForwardUp C++ function, which implements the IP changing in FdNetDeviceIpChange, to change the destination address to that of the UEs and also adds a GTPU (GPRS Tunnelling Protocol, [11]) header

with source address of the PGW node. This ensures that the LTE network will handle the packets correctly and the UE will receive them (to the SGW these packets appear as if the PGW sent them). On the side of the UE, in LteUeFdNetDevice when the packet is received, the destination address is changed back to the address of the client. Since the UDP checksum in the GTPU header is set to 0, UdpL4Protocol (the C++ class in ns-3 that represents the UDP layer) needed a modification to not drop packets with 0 checksums (this behavior is allowed by the standard, see [14]).

## 4 MEASUREMENT ENVIRONMENT

The setup is realized in Docker containers as shown in Figure 3. We are using a modified version of the setup from [7], which was created based on [18]. The client container contains the QUIC client, the sim container contains the simulated LTE environment implemented in ns-3[15], the proxy and server containers contain the MASQUE proxy and QUIC server, respectively. The sim container has a pair of virtual network devices since the EmuEpc (the module in ns-3 that implements the LTE EPC network) requires them for the SGW-eNodeB communication. The direct communication between the two interfaces of the sim container is blocked via iptables firewall, thus the packets are forced to go through ns-3.

### 4.1 ns-3 scenario

We have implemented a scenario in ns-3 that creates the above setup. It is required to create routes between the RightNode and the PGW and the RightNode and the SGW since the RightNode will send packets to the SGW, pretending to be coming from the PGW. The RightNode also needs routes to the proxy, since it acts as a router between ns-3 and the proxy.

The eNodeB, the UE and the RightNode have mobility models (ConstantPositionMobilityModel class in ns-3). Neither of them are moving and they constitute a rectangular triangle. The distance between the eNodeB and UE is 99 meters, which is a typical distance in LTE networks (the position of the RightNode is actually irrelevant).

The scenario configures the LTE and the simulation parameters as shown in Table 1. The modules are the following:

- Global:** sets the simulation type to real time and the checksum calculation to be off.
- RealtimeSimulatorImpl:** sets the hard limit parameter of the simulator. This is the maximum seconds that the simulator can lag behind the wallclock.
- LteUePhy:** sets the transmit power and noise level of all UEs.
- LteEnbPhy:** sets the transmit power and noise level of all eNodeBs.
- LteSpectrumPhy:** turns off error models of the LTE spectrum layer.
- LteHelper:** controls parameters of the LTE helper, which is used to set up the different parts of the LTE network.
- LteEnbRrc:** sets the default transmission mode of eNodeBs to SISO.
- DropTailQueue<Packet>:** default values for the drop tail queues. The max queue size is set to a low value to induce packet loss.

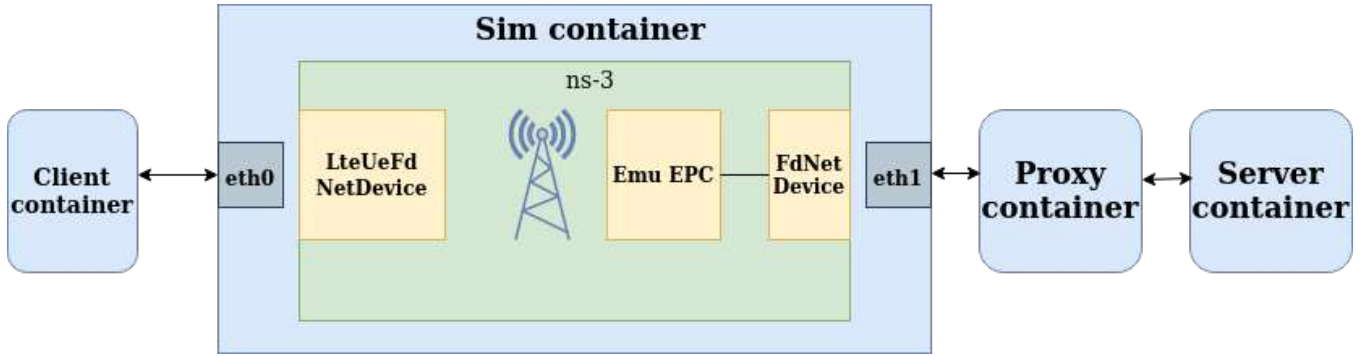


Figure 3: Setup topology

**LteAttribute:** configures the scheduler type, the loss- and fading modes used in the simulations.

**FadingModelAttribute:** sets the file name to use and the number of RBs (the default is 100 also).

**EnbDevice:** configures the parameters of the net devices of the eNodeBs. The first 2 are the number of resource blocks allocated in downlink and uplink, respectively. The last 2 are the EARFC of the channel in downlink and uplink, respectively. The number of resource blocks are changed depending the measurement.

## 5 MEASUREMENTS

The measurements were done using the setup described above. We differentiate between lower and higher LTE throughput measurements, which we can control by setting the number of resource blocks (DLBandwidth and ULBandwidth) to either lower or higher values. The process consists of downloading a 10MB sized file. The RTT between the client and proxy containers is around 38ms as reported by the ping program (when no file download is happening). The one way delay between the RightNode and PGW is configured to be 10ms.

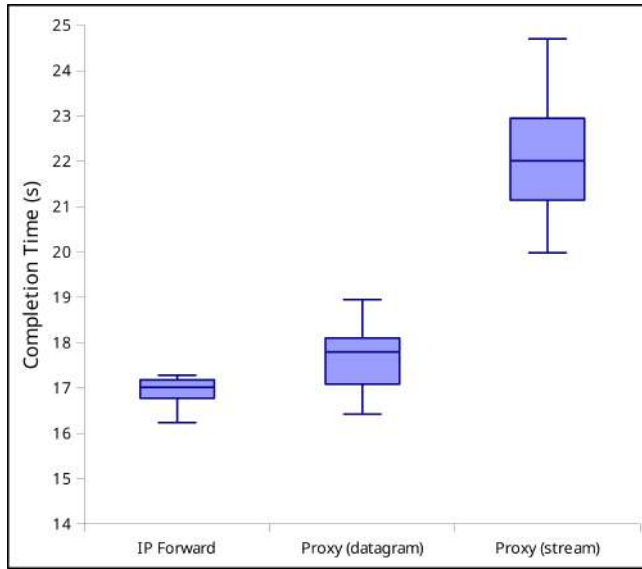
### 5.1 Lower performance

We set the LTE parameters to values that causes low throughput. To achieve this, the uplink and downlink resource blocks are each given the value 25.

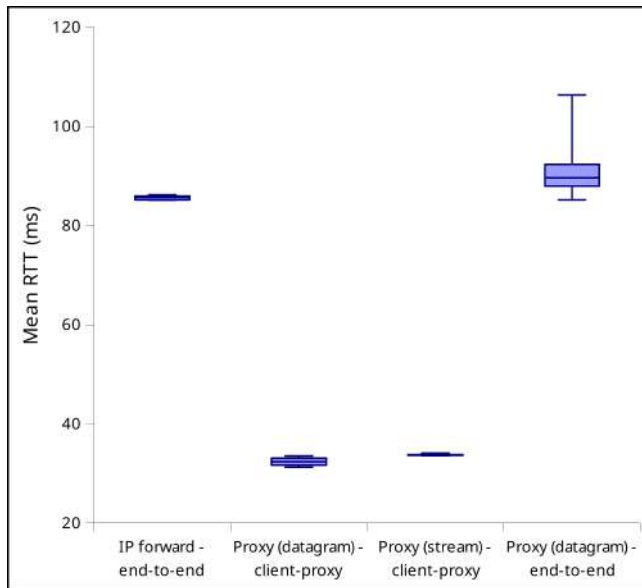
Figure 4 shows the results of measurements with 3 different set of parameters where each of them ran 10 times. The first uses basic IP forwarding and the last two uses the MASQUE proxy. IP forwarding is accomplished via setting up DNAT in the proxy container and turning off proxy mode in the client and server (and not starting the MASQUE proxy in the proxy container). In proxy datagram mode, the outer QUIC layer does not acknowledge the packets, instead this is done in the tunneled connection (which also uses the QUIC protocol). In stream mode, the acknowledgement is done in the proxy (outer connection), and as can be seen in the figure, it increases the completion time significantly more (30.9% on average), than datagram mode (4.5% on average). This is the expected behaviour that results from the implemented infinite buffer size of the stream mode. Comparing this result to the prior work of [7] one can see the same effect.

Table 1: The table shows the parameters and their values for the different modules in ns-3.

Global	
Simulator ImplementationType	RealtimeSimulatorImpl
ChecksumEnabled	False
RealtimeSimulatorImpl	
SynchronizationMode	HardLimit
HardLimit	0.2s
LteUePhy	
TxPower	10.0
NoiseFigure	7.0
LteEnbPhy	
TxPower	30.0
NoiseFigure	5.0
LteSpectrumPhy	
DataErrorModelEnabled	true
CtrlErrorModelEnabled	true
LteHelper	
UsePdschForCqiGeneration	false
UseIdealRrc	true
LteEnbRrc	
EpsBearerToRlcMapping	RLC_UM_ALWAYS
DefaultTransmissionMode	0
DropTailQueue<Packet>	
MaxSize	16
LteAttribute	
SchedulerType	PfFfMacScheduler
PathlossModel	FriisPropagation LossModel
FadingModel	TraceFadingLossModel
FadingModelAttribute	
TraceFilename	fading_trace_EPA_3kmph.fad
RbNum	100
EnbDevice	
DLBandwidth	{25,100}
ULBandwidth	{25,100}
DLEarfcn	9895
ULEarfcn	27785



**Figure 4: Completion time shown for IP forward (left), proxy with datagram mode (middle) and proxy with stream mode (right) setups with lower bandwidth limit**



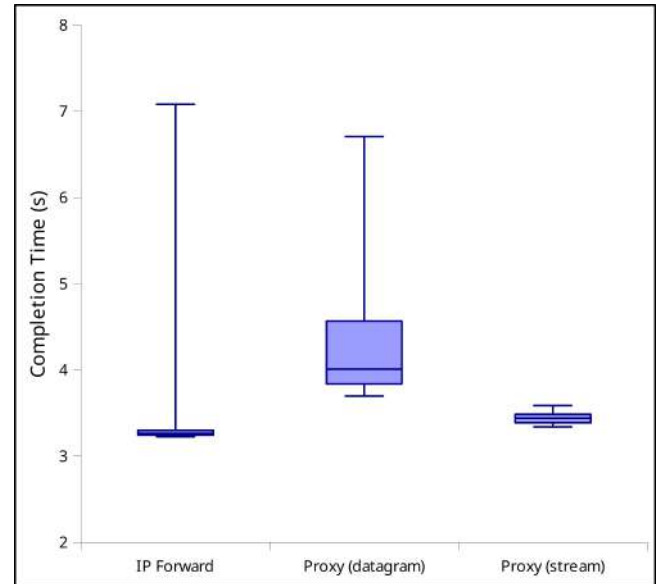
**Figure 5: From the perspective of the client, the RTT shown for (from left to right): 1. IP forward between the client and server, 2. Proxy (datagram mode) between the client and the proxy, 3. Proxy (stream mode) between the client and proxy, 4. Proxy (datagram mode) between the client and server.**

Figure 5 shows the RTT for the different set of measurements. As can be seen, the delay between the client and the proxy (in proxy mode) in either datagram mode or stream mode is the same. The RTT is higher in IP forward mode and in proxy (datagram) mode (between the client and server), because of the 25ms one way delay

between the proxy and the server, and the difference between them is small, the latter being slightly higher. In these configurations the simulator's lag behind real time adds to the overall RTT, which causes the delay to be higher than expected (10ms one way delay between the RightNode and PGW, and the LTE networks delay should be less than 35ms). The proxy stream mode client to server RTT is not shown in this figure because the results were erroneous.

## 5.2 Higher performance

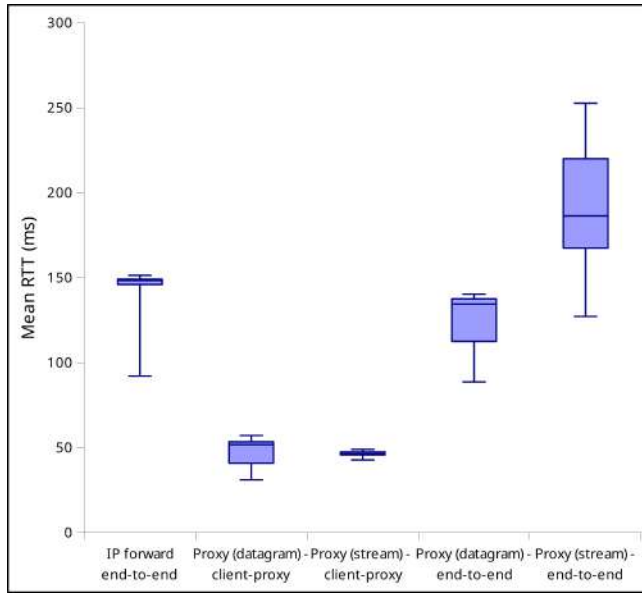
We set the LTE parameters to values that causes higher throughput to simulate adequate reception conditions. To achieve this, the uplink and downlink resource blocks are each given the value 100.



**Figure 6: Completion time shown for IP forward (left), proxy with datagram mode (middle) and proxy with stream mode (right) setups with higher bandwidth limit**

Figure 6 shows the results of measurements with the same setup as above, with the only difference of higher resource block resulting in higher bandwidth limit. In this case, the stability of the IP forward and datagram proxy mode is degraded (as indicated by the outlier higher completion time). The average completion time in datagram proxy mode is 19.3% higher compared to IP forward mode. The completion time in stream proxy mode is 5.8% less than in IP forward mode on average, which is caused by the individual high value in the latter, as the median is actually 5.4% higher.

Figure 7 shows the RTT for the different set of measurements. Comparing with the previous results (Figure 5) the RTT increased overall, which can be explain by the higher throughput, thus higher computation requirements by the simulator, even though the completion times are much lower. The delay, however, not increased equally. In case of IP forwarding, where previously it was slightly lower than proxy in datagram mode (client-server delay), now it is slightly higher. We can also see in this figure that the proxy in stream mode has the highest RTT. The increase is 57ms for IP



**Figure 7: From the perspective of the client, the RTT shown for (from left to right): 1. IP forward between the client and server, 2. Proxy (datagram mode) between the client and the proxy, 3. Proxy (stream mode) between the client and proxy, 4. Proxy (datagram mode) between the client and server, 5. Proxy (stream mode) between the client and the server.**

forward mode, and 15ms for proxy datagram mode (client-proxy delay), which hints that the added delay by ns-3 increased, but it is not the sole cause of this, and the RTT between the proxy and server is higher as well. A possible cause for the latter is higher processing requirement in the server, necessitated by the higher throughput.

## 6 CONCLUSIONS

Adding a new net device to ns-3 that enables a real client to communicate with a real server via an emulated LTE network has two main benefits. It makes it easier to test new protocols in a simulated network environment, since there is no need to implement it in the simulator and a real LTE environment is not needed.

The implementation is a proof of concept, there are multiple ways to enhance it. Future works could expand on it to enable full NAT support, which in turn would enable multiple real clients to connect through one LTE UE node. A study could assess the feasibility of a similar modification to the mmWave module. This would further increase the testing capabilities on ns-3.

It is clear that simulating an LTE network have limitations caused by the resource (memory, CPU speed) hungry nature of the process. In case of ns-3 it manifests as additional delay. For our low throughput setups this is not pronounced, however, in high throughput scenarios this can be significant. It is for further study to understand the achievable cellular throughput in real time using more powerful computing resources. If real-time ns-3 emulation is not strictly required, our implementation provides a valuable tool since it does not require any radio specific equipment.

## ACKNOWLEDGMENTS

The authors are thankful for the fruitful discussions and guidance from Mirja Kühlewind, Marcus Ihlar, Magnus Westerlund and Zaheduzzaman Sarker (Ericsson).

## REFERENCES

- [1] Armir Bujari, Claudio E Palazzi, Giacomo Quadrio, and Daniele Ronzani. 2020. Emerging interactive applications over quic. In *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 1–4.
- [2] Gustavo Carneiro, Helder Fontes, and Manuel Ricardo. 2011. Fast prototyping of network protocols through ns-3 simulation model reuse. *Simulation modelling practice and theory* 19, 9 (2011), 2063–2075.
- [3] Alvise De Biasio, Federico Chiariotti, Michele Polese, Andrea Zanella, and Michele Zorzi. 2019. A QUIC implementation for ns-3. In *Proceedings of the 2019 Workshop on ns-3*. 1–8.
- [4] Helder Fontes, Rui Campos, and Manuel Ricardo. 2016. Improving ns-3 emulation support in real-world networking scenarios. *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 3, 9 (2016), e5–e5.
- [5] Rohit Gupta, Bjoern Bachmann, Russell Ford, Sundeep Rangan, Nikhil Kundargi, Amal Ekbal, Karamvir Rathi, Maria Isabel Sanchez, Antonio De La Oliva, and Arianna Morelli. 2015. Ns-3-based real-time emulation of LTE testbed using LabVIEW platform for software defined networking (SDN) in CROWD project. In *Proceedings of the 2015 Workshop on ns-3*. 91–97.
- [6] Zsolt Krämer, Mirja Kühlewind, Marcus Ihlar, and Attila Mihály. 2021. Cooperative performance enhancement using QUIC tunneling in 5G cellular networks. In *Proceedings of the Applied Networking Research Workshop*. 49–51.
- [7] Mirja Kühlewind, Matias Carlander-Reuterfelt, Marcus Ihlar, and Magnus Westerlund. 2021. Evaluation of QUIC-based MASQUE proxying. In *Proceedings of the 2021 Workshop on Evolution, Performance and Interoperability of QUIC*. 29–34.
- [8] Apostolos I Kyrtziz and Panayotis G Cottis. 2021. QUIC vs TCP: A Performance Evaluation over LTE with NS-3. *Communications and Network* 14, 1 (2021), 12–22.
- [9] Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasnic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, et al. 2017. The quic transport protocol: Design and internet-scale deployment. In *Proceedings of the conference of the ACM special interest group on data communication*. 183–196.
- [10] Marco Mezzavilla, Sourjya Dutta, Menglei Zhang, Mustafa Riza Akdeniz, and Sundeep Rangan. 2015. 5G mmWave module for the ns-3 network simulator. In *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. 283–290.
- [11] Evolved Universal Terrestrial Radio Access Network. 2011. S1 Application Protocol (S1AP)(Release 10). *Technical Specification* 36 (2011).
- [12] Umberto Paro, Federico Chiariotti, Anay Ajit Deshpande, Michele Polese, Andrea Zanella, and Michele Zorzi. 2020. Extending the ns-3 QUIC Module. In *Proceedings of the 23rd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. 19–26.
- [13] Harsh Patel, Hrishikesh Hiraskar, and Mohit P Tahiliani. 2019. Extending network emulation support in ns-3 using DPDK. In *Proceedings of the 2019 Workshop on ns-3*. 17–24.
- [14] Jon Postel et al. 1980. User datagram protocol. (1980).
- [15] George F Riley and Thomas R Henderson. 2010. The ns-3 network simulator. In *Modeling and tools for network simulation*. Springer, 15–34.
- [16] Ayman Sabbah, Abdallah Jarwan, Ismael Al-Shiab, Mohamed Ibnkahla, and Maoyu Wang. 2018. Emulation of large-scale lte networks in ns-3 and core: A distributed approach. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*. IEEE, 1–6.
- [17] David Schinazi. 2022. *Proxying UDP in HTTP*. Internet-Draft draft-ietf-masque-connect-udp-11. Internet Engineering Task Force. <https://datatracker.ietf.org/doc/html/draft-ietf-masque-connect-udp-11> Work in Progress.
- [18] Marten Seemann and Jana Iyengar. 2020. Automating QUIC Interoperability Testing. In *Proceedings of the Workshop on the Evolution, Performance, and Interoperability of QUIC*. 8–13.
- [19] Hajime Tazaki, Frédéric Urbani, and Thierry Turlatti. 2013. DCE Cradle: Simulate network protocols with real stacks. In *Workshop on NS3 (WNS3)*.

Received 23 February 2023; accepted 5 March 2023



ISBN: 978-1-4503-9944-9