

Publisher: State and Provincial Joint Engineering Lab. of Advanced Network
Monitoring and Control (ANMC)

Cooperate:

Xi'an Technological University (CHINA)
West Virginia University (USA)
Huddersfield University of UK (UK)
Missouri Western State University (USA)
James Cook University of Australia
National University of Singapore (Singapore)

Approval:

Library of Congress of the United States
Shaanxi provincial Bureau of press, Publication, Radio and Television

Address:

4525 Downs Drive, St. Joseph, MO64507, USA
No. 2 XueFu Road, WeiYang District, Xi'an, 710021, China

Telephone: +1-816-2715618 (USA) +86-29-86173290 (CHINA)

Website: www.ijanmc.org

E-mail: ijanmc@ijanmc.org

xxwlc@163.com

ISSN: 2470-8038

Print No. (China): 61-94101

Publication Date: September 28, 2024

Editor in Chief

Ph.D. Xiangmo Zhao

Prof. and President of Xi'an Technological University, Xi'an, China

Director of 111 Project on Information of Vehicle-Infrastructure Sensing and ITS, China

Associate Editor-in-Chief

Professor Xiang Wei

Electronic Systems and Internet of Things Engineering

College of Science and Engineering

James Cook University, Australia

Dr. Chance M. Glenn, Sr.

Professor and Dean

College of Engineering, Technology, and Physical Sciences

Alabama A&M University

4900 Meridian Street North Normal, Alabama 35762, USA

Professor Zhijie Xu

University of Huddersfield, UK

Queensgate Huddersfield HD1 3DH, UK

Professor Jianguo Wang

Vice Director and Dean

State and Provincial Joint Engineering Lab. of Advanced Network and Monitoring Control,
China

School of Computer Science and Engineering, Xi'an Technological University, Xi'an, China

Ph. D Natalia Bogach

Director of Computer Science Department

Peter the Great St. Petersburg Polytechnic University, Russia

Administrator

Dr. & Prof. George Yang

Department of Engineering Technology

Missouri Western State University, St. Joseph, MO 64507, USA

Professor Zhongsheng Wang

Xi'an Technological University, China

State and Provincial Joint Engineering Lab. of Advanced Network and Monitoring Control,
China

Associate Editors

Prof. Yuri Shebzukhov

International Relations Department, Belarusian State University of Transport, Republic of
Belarus.

Dr. & Prof. Changyuan Yu

Dept. of Electrical and Computer Engineering, National Univ. of Singapore (NUS)

Dr. Omar Zia

Professor and Director of Graduate Program

Department of Electrical and Computer Engineering Technology

Southern Polytechnic State University

Marietta, Ga 30060, USA

Dr. Baolong Liu

School of Computer Science and Engineering

Xi'an Technological University, CHINA

Dr. Mei Li
China university of Geosciences (Beijing)
29 Xueyuan Road, Haidian, Beijing 100083, P. R. China

Dr. Ahmed Nabih Zaki Rashed
Professor, Electronics and Electrical Engineering
Menoufia University, Egypt

Dr. Rungun R Nathan
Assistant Professor in the Division of Engineering, Business and Computing
Penn State University - Berks, Reading, PA 19610, USA

Dr. Taohong Zhang
School of Computer & Communication Engineering
University of Science and Technology Beijing, China

Dr. Haifa El-Sadi
Assistant professor
Mechanical Engineering and Technology
Wentworth Institute of Technology, Boston, MA, USA

Huaping Yu
College of Computer Science
Yangtze University, Jingzhou, Hubei, China

Ph. D Yubian Wang
Department of Railway Transportation Control
Belarusian State University of Transport, Republic of Belarus

Prof. Mansheng Xiao
School of Computer Science
Hunan University of Technology, Zhuzhou, Hunan, China

Prof. Ying Cuan
School of Computer Science, Xi'an Shiyou University, China

Qichuan Tian
School of Electric & Information Engineering
Beijing University of Civil Engineering & Architecture, Beijing, China

Ph. D MU JING
Xi'an Technological University, China

Language Editor

Professor Gailin Liu
Xi'an Technological University, China

Dr. H.Y. Huang
Assistant Professor
Department of Foreign Language, the United States Military Academy, West Point, NY
10996, USA

Would you like to be an Associate Editor? Simply send a request together with your Curriculum Vitae to xxwlc@163.com. We will have a team of existing editors or at least three experts in your field to review your request and make a decision as soon as we can. The criteria to be an associate editor are: 1. must have advanced degree; 2. must be a leader or have outstanding achievements in the specific research field; 3. must be recommended by the review team.

Table of Contents

Remote Sensing Building Damage Assessment Based on Machine Learning.....	1
<i>Jiawei Tang, Shengquan Yang, Shujuan Huang, Bozhi Xiao</i>	
Enhancing Quantum Key Distribution Protocols for Extended Range and Reduced Error.....	13
<i>Amina Alkilany Abdallah Dallaf.</i>	
Real-Time Extraction of News Events Based on BERT Model.....	24
<i>Yuxin Jiao, Li Zhao</i>	
Hippocampal Cognitive Function Based on Deep Learning.....	32
<i>Bijun Zhang, Hongge Yao</i>	
Design and Development of an Intelligent Laboratory Management System Based on STM Processors.....	40
<i>Ruoyu Wang, Lulu Chen, Jiaxuan Liu, Lei Tian</i>	
Infrared Weak and Small Target Detection Algorithm Based on Deep Learning.....	47
<i>Lei Wang, Jun Yu</i>	
3D Reconstruction of Indoor Scenes Based on 3DGS Models.....	56
<i>Hanghua Li, Lipeng Si</i>	
9D Rotation Representation-SVD Fusion with Deep Learning for Unconstrained Head Pose Estimation.....	62
<i>Jiaqi Lyu, Changyuan Wang</i>	
Vector Storage Based Long-term Memory Research on LLM.....	69
<i>Kun Li, Xin Jing, Chengang Jing</i>	
Improved Pedestrian Vehicle Detection for Small Objects Based on Attention Mechanism.....	80
<i>Hao Yanpeng, Geng Chaoyang</i>	

Remote Sensing Building Damage Assessment Based on Machine Learning

Jiawei Tang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 905462150@qq.com

Shujuan Huang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 349242386@qq.com

Shengquan Yang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: xaitysq@126.com

Bozhi Xiao

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 3138379859@qq.com

Abstract—After the occurrence of various types of disasters, including natural disasters and man-made damage, aid workers need accurate and timely data, such as the damage status of buildings, in order to take effective measures for rescue. So as to solve this problem, this paper researches and designs a building damage classification system based on machine learning. The damage assessment system consists of two network models (building extraction network and damage classification network). This article analyzes and designs the structure of each network model, and discusses the principles related to computer vision in machine learning. Buildings in satellite images are segmented through Siamese Convolutional Neural Network, the BottleNeck Module and Feature Pyramid Network are used in the damage classification assessment network to detect damage to buildings in sub-temporal remote sensing images. Subsequently, the model was trained and tested on different disaster events on the xBD dataset. The results show that the building damage detection system based on Siamese-CNN achieves good detection accuracy, and the system has the advantages of simple operation, good timeliness and low resource consumption, and can well meet the needs of disaster assessment.

Keywords—Machine Learning; Remote Sensing; Disaster Assessment; Buildings

I. INTRODUCTION

According to data from the United States Geological Survey, more than one million natural disasters occur around the world every year, with an average of several disasters occurring almost every minute [1]. In the last decade alone, the global death toll from natural disasters has exceeded one million. With the advancement of global urbanization, more densely distributed building areas will inevitably lead to further aggravation of the impact of building damage and collapse after disasters. When current technology is insufficient to reliably predict disasters, how to reduce losses through effective emergency response measures becomes an important issue.

Therefore, an important task in the disaster emergency response process is damage assessment of the disaster-stricken area, and the damage information of surface buildings has become important reference information in the disaster relief emergency response process. Obtaining detailed information of disaster areas through manual survey is slow, dangerous, and timely, which is not conducive to obtaining information.

However, by analyzing satellite image data, it is more convenient to obtain the damage situation in a specific area and make decisions without arriving at the scene. Therefore, using satellite images can obtain disaster area information more efficiently.

At present, the mainstream building damage assessment methods are divided into three types: traditional assessment methods, multispectral sensor-based assessment methods, and deep learning-based assessment methods [2]. Traditional assessment methods consume a lot of resources and time, which is not practical in disaster assessment and will not be discussed here. The following focuses on the assessment method based on multispectral sensors and the assessment method based on deep learning for building damage assessment.

The evaluation method based on multispectral sensors generally uses multispectral satellites to capture the spectral information of different materials in multiple bands such as reflection and absorption of visible light and infrared light for detection [3]. At the same time, multispectral satellites can enhance sensitivity to specific ground objects or phenomena. Therefore, the evaluation method based on multispectral sensors is more accurate on some materials. However, because the images transmitted by spectral satellites are very strict, the images need to be preprocessed, denoised, corrected, etc. And these process will take a long time. Therefore, simpler and faster methods for damage assessment of disaster buildings are needed in disaster assessment.

The evaluation method based on machine learning [4,5] uses complex deep neural networks to enable computers to handle corresponding tasks like humans. This field is called computer vision [6,7] in the field of machine learning. Computer vision research mainly includes the following four types of tasks: the first type of task is image classification [8]; the second type of task is image target detection [9]; the third type of task is image semantic segmentation [10-11]; and the fourth type of task is image instance segmentation [12]. In the building damage assessment method based on deep learning, a computer is used to


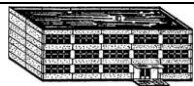






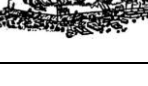

semantically segment the buildings in the image, and then the buildings in the disaster area are classified. Through this method, disaster detection and assessment can be carried out in the disaster-stricken area on the trained deep neural network. Therefore, assessment methods based on machine learning can complete disaster assessment tasks faster and more efficiently.

II. RELATED WORK RESEARCH

A. Building Damage Standards

According to the classification definition of buildings with different degrees of damage by the European Disaster Committee, this definition is also the current reference standard for international research on building damage assessment, as shown in Table 1. According to Table 1, it can be found that since no structural changes have occurred in the damaged structures of Level 1 ‘Undamaged’ and Level 2 ‘Minor Damaged’, ‘Undamaged’ and ‘Minor Damaged’ are usually divided into the same category in building damage assessment algorithms [13].

TABLE I. EUROPEAN DISASTER COMMITTEE TABLE FOR BUILDING DAMAGE ASSESSMENT

Masonry Construction	Fortified Buildings	Damage Level
		Undamaged
		Minor Damaged
		Medium Damaged
		Major Damage
		Destroyed

B. Current status of building damage assessment algorithms

The emergence of machine learning has brought an important turn to the development of the field of building damage assessment based on remote sensing images. Applying computer vision-related technologies in deep learning to building damage assessment can increase the accuracy and efficiency of building damage assessment. In 2016, Researchers integrated neural networks into building disaster detection for the first time, laying the foundation for subsequent building damage assessment through neural network models. In 2018, Duarte et al. [14] constructed the damage evaluation of buildings in remote sensing images as a binary pixel classification problem for the first time. In 2019, Xu et al. [15] compared the performance of different network by evaluate damaged buildings after earthquakes, and trained and tested on different disaster data sets to verify the feasibility of disaster assessment through neural networks. And the same year, Gupta et al. [16] established a large-scale data set for large areas and multiple disasters, providing multi-temporal satellite images before and after disasters. It contains about 700,000 building annotations over 5,000 square kilometers and is mainly used for post-disaster building damage assessment. In 2020, Weber et al. [17] established a semantic segmentation model based on the ResNet50 backbone and trained it through the xBD dataset to achieve graded assessment of building damage. The proposal of the above technologies has opened up a path and space for the research and application of deep learning in the field of building damage assessment, making deep learning faster and more efficient in damage assessment tasks in the field of building damage assessment.

C. Dataset

xBD Dataset [18]: Segmentation, extraction and damage classification of buildings under remote sensing images is a very challenging task due to the complexity of unstructured scenes. Therefore, researchers and research institutions have disclosed many data sets in order to promote the field of disaster assessment. These data sets collect image data using multiple remote sensing

satellites and conduct precise annotations on the images to form accurate label values. This article plans to use the xBD data set commonly used in the field of disaster assessment, as shown in Figure 1.

The xBD dataset is a dataset for building damage assessment released by the Massachusetts Institute of Technology. It is one of the current public datasets of remote sensing images in the world. The data set contains a total of tens of thousands of images, all of which are 1024*1024 high-resolution satellite remote sensing images, marked with 19 different events, including earthquakes, floods, wildfires, volcanic eruptions, etc. These images include pre-disaster and post-disaster images, allowing for adequate research on building damage assessment.

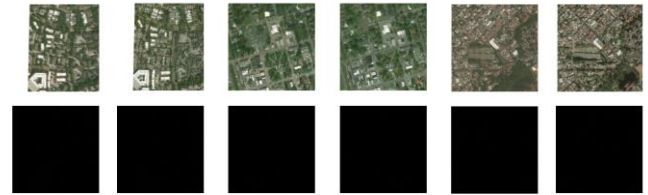


Figure 1. Schematic diagram of xBD dataset

III. TECHNOLOGIES AND NETWORK MODELS USED IN THIS ARTICLE

From an overall perspective, the building damage assessment problem is affected by many aspects such as data sources and evaluation standards. Currently, there is no consistent algorithm framework. This chapter analyzes existing algorithm problems, proposes a set of building damage classification algorithm processes based on Siamese-CNN neural network, and analyzes in detail some of the key technologies involved.

A. Overall design of system

From the perspective of application scenarios, a characteristic of building damage assessment is that it needs to process a large amount of data in a short period of time, which requires high performance of the algorithm. Therefore, the best way is to fully train the model using existing remote sensing image data before and after the

disaster so that it can perform better when processing new data. From the perspective of image classification, high-resolution remote sensing images contain many types of ground objects and have complex background interference, making them more difficult to handle than general natural image classification problems. On the other hand, remote sensing images usually cover a larger area, especially in areas with lower urbanization rates, where the proportion of built-up areas is smaller. Therefore, from the perspective of improving the accuracy of damage assessment and the efficiency of algorithm processing, this paper divides the damage assessment process into two processes: building segmentation extraction and building damage assessment. The figure 2 shows the overall program process of building damage assessment in this article.

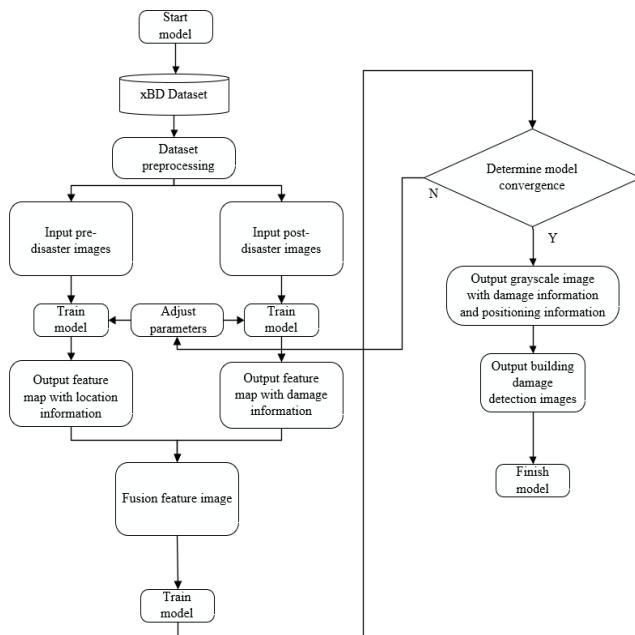


Figure 2. Network model flow chart based on machine learning

B. Key technical analysis

1) *Fully Connected Neural Network basics (FCN):* Fully Connected Neural Networks [19] are the basis of all neural networks. They are a sort of neural network designed from the nervous system of animals and are therefore also known as Artificial Neural Networks. The underlying principle of the fully connected neural network is a mathematical algorithm model for distributed parallel processing. Through the network

complexity of FCN, the interconnection of nodes can be adjusted to achieve the purpose of information processing. Therefore, the fully connected neural network has a good effect in completing the tasks of regression model and classification model. The FCN network structure is shown in Figure 3.

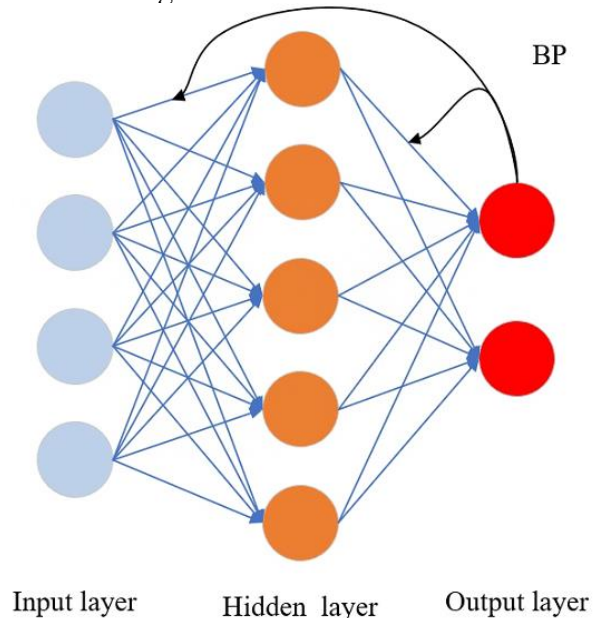


Figure 3. The network structure of the FCN

2) *Convolutional Neural Network basics (CNN):* CNN [20,21] are derived from neuron-based message passing systems in biology. Generally, there are several layers from input to output, each layer contains several neurons, and the neurons in adjacent layers are interconnected. Specific to the field of computational vision, each layer here is a three-dimensional array, and the neuron actually refers to a corresponding item in the three-dimensional array.

The input of the CNN network is the original image. The most important feature of the CNN network is the introduction of alternating convolution layers and pooling layers, which generally obtain the response of a point on the output plane through a small sub-region on the input feature plane. The so-called interconnection of adjacent neurons refers to reorganizing the output of all neurons in a certain layer into a three-dimensional array as the input of the next layer. Generally speaking, as the level of connection increases, the feature plane obtained by the neuron output response decreases, the number of feature

planes increases, and the corresponding receptive field of each neuron response also increases. In general, a convolutional neural network forms a function that obtains an output image from an input image. The basic CNN network structure is shown in Figure 4.

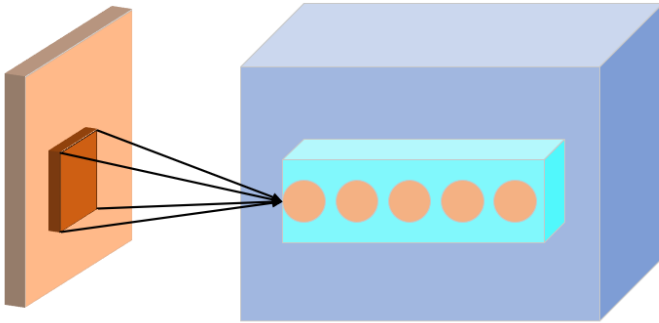


Figure 4. CNN receptive field

a) Convolution layer: The basic operations of convolutional layers [22] are derived from the digital filtering process in traditional image processing. The difference is that the convolution template here is obtained through training and learning, and the template size is generally smaller. All convolution templates in the same convolution layer have the same scale. Only in this way can the outputs of different convolution kernels form feature planes with consistent dimensions. During the forward pass, each convolution kernel is convolved with all feature planes of the input. Specifically, the convolution operation is the dot product of the elements in the filter template and the elements of the receptive field area in the input image. Since the receptive field corresponding to each point on the output feature plane is only a local area in the input plane, the filter template of the convolutional layer learns the local characteristics of the image. The superposition of multiple convolutional layers gradually aggregates these local features into global features that can characterize the entire image.

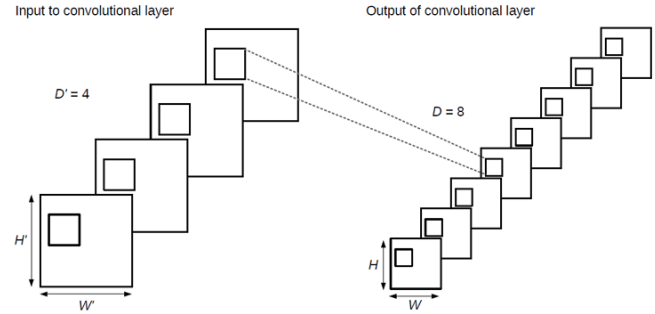


Figure 5. Convolutional layer input and output diagram

b) Activation function: The function of the activation function [23] is to perform a certain nonlinear transformation on the output of each neuron. In a convolutional neural network, the input to the excitation function is generally the output of the convolutional layer. The simplest excitation function is a binary function that outputs 0 or 1 depending on the range of input values. The most commonly used excitation function in early neural networks is the Sigmoid function, which has the following form:

$$f(x) = (1 + \exp(-x))^{-1} \quad (1)$$

This function acts on each element in the three-dimensional vector output by the convolutional layer and does not require a parameter learning process. The most commonly used function in modern neural networks is the *ReLU* function. Specifically, the *ReLU* function outputs 0 for negative values and remains unchanged for positive values. Its form is as follows:

$$f(x) = (0, 1)_{\max} \quad (2)$$

c) Pooling layer: The function of the pooling layer [24] is to perform nonlinear downsampling of the input feature plane. The most commonly used method is Max pooling. Pooling layers generally alternate with convolutional layers (after the activation function), usually reducing the spatial resolution of the input feature plane.

The Max pooling layer is shown in the Figure 6. For selecting a rectangular sub-region in the input feature plane, Max pooling takes the maximum

value of the response in the sub-region as a response on the output feature plane. Simply put, the pooling process is to sample at corresponding intervals and start moving from the sampling center point. In the general convolutional neural network structure, the input feature plane is sampled without overlap. The precise relationship between input and output in a convolutional neural network is as shown in the formula:

$$y_{ijd} = \max\left(\{x_{sxi+s \times j+q,d} \mid 0 \leq p \leq h-1, 0 \leq q \leq w-1\}\right) \quad (3)$$

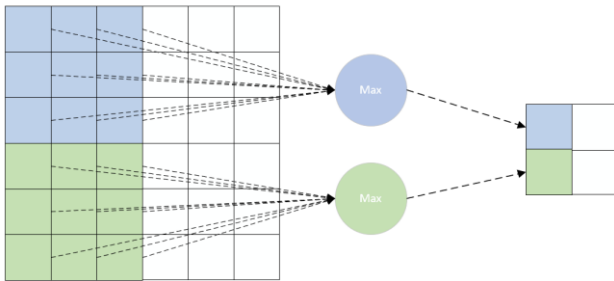


Figure 6. Feature Pyramid Network structure diagram

3) *Siamese Convolutional Neural Network basic (SCN)*: Siamese Convolutional Neural Network [25,26] is a machine learning model used for metric learning. The basic structure of the network contains two identical sub-networks, as shown in Figure 7, which share the same weight parameters. These two sub-networks respectively process two samples to be compared, usually images or other representations of data. Through the design of shared weights, the Siamese network can learn the feature representation of input samples, and the goal of metric learning is to measure the similarity between samples through these feature representations. In the network structure, the output features of two sub-networks are fused together by measuring the similarity or difference between the learning samples in the learning layer. The training process of the network increases the design of a loss function, which encourages similar samples to be closer and dissimilar samples to be further away. Through the backpropagation algorithm, the parameters of the network are adjusted so that it can learn appropriate feature representations and similarity measures. Twin CNN networks are widely used in fields such as face verification, signature verification, target tracking, etc. Its advantage is

that it can learn the similarities between samples and has strong generalization performance.

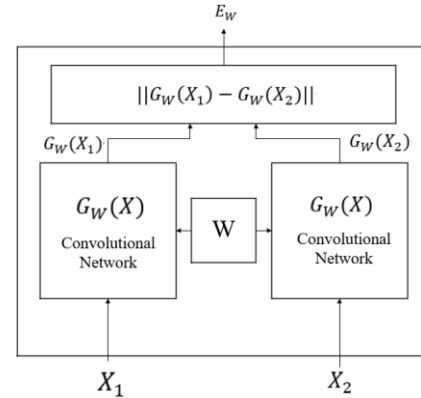


Figure 7. Basic structure diagram of Siamese Convolutional Neural Network

C. Building Damage Classification Network Based On Siamese-CNN

This article intends to design a building damage assessment algorithm based on Siamese-CNN. Siamese-CNN is an algorithm in deep learning. It is a new neural network developed based on CNN. Through Siamese-CNN neural network, features in the image can be extracted and classified based on the features to achieve the purpose of classifying building damage. In the following, the Siamese-CNN network model designed in this article will be focused on.

The Siamese-CNN network model has an encoder-decoder structure. In the encoder module, the images before the disaster and the images after the disaster are input into the convolutional Siamese convolutional neural network, and the images are converted into a series of features. Then entering the decoder module, this article integrates the feature pyramid network and BottleNeck module into the decoder module. This can preserve the features in the input image more abundantly, and can enhance the expression ability of the features, making Siamese-CNN perform better in building damage assessment tasks. Next, we will introduce the feature pyramid network and BottleNeck module used in this article in detail.

Feature Pyramid Networks [27] obtains corresponding feature images by using the multi-scale pyramid structure of the deep convolutional neural network itself. Its core idea is to use multi-

scale feature maps to improve the model's ability to identify different targets. The FPN network structure used in this article is shown in Figure 8. The left side is the bottom-up path, which mainly builds a forward pyramid and extracts feature information at different scales through the special structure of the pyramid. Specifically, after FPN receives multiple feature maps output by the encoder, it immediately performs horizontal convolution. The purpose is to match the number of channels of the bottom feature map to the same number of channels as the top feature image layer by layer for subsequent processing. Then the feature images are upsampled in a top-down path to ensure that the dimensions and sizes of all upper-layer feature images are consistent with the bottom-layer feature images; finally, fusion is performed to obtain a fused multi-scale feature map.

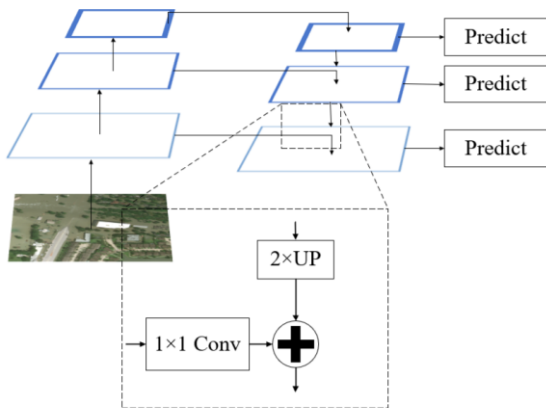


Figure 8. Feature Pyramid Network structure diagram

The BottleNeck module [28] is a widely used structure in deep learning and is often used in ResNet networks. The BottleNeck module can reduce the amount of parameters and calculations when processing larger input data or more complex networks, and can effectively improve network performance. The structure of the BottleNeck module used in this article is shown in Figure 9. It mainly contains three important parts: The first part is dimensionality reduction convolution, which reduces the number of channels and reduces the amount of calculation by using dimensionality reduction convolution with a convolution kernel of 1; The second part is the middle-layer convolution, The middle-layer

convolution with a convolution kernel of 3 is used to perform non-linear changes on the input feature map, which is beneficial to learning more complex features; The last layer is dimensionality-raising convolution, which restores the number of channels of the feature image to the original level by using specific convolutions. In general, as a part of the Siamese-CNN network, the BottleNeck module not only improves the learning efficiency of the network, but also improves the parameter efficiency of the network, so that the Siamese-CNN network has a good effect in dealing with the task of building damage assessment.

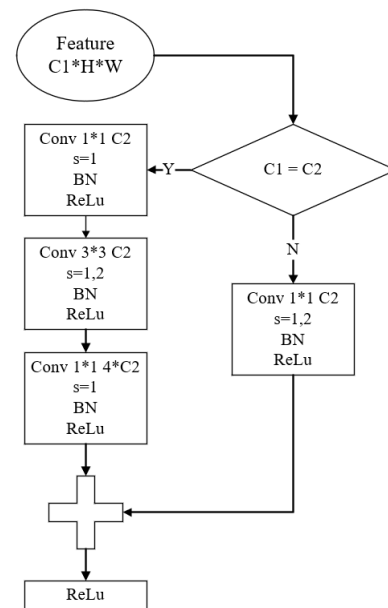


Figure 9. Basic structure of the BottleNeck module

The specific network structure of Siamese-CNN is shown in Figure 10. In the entire network structure, two images, Pre-Damage-Image and Post-Damage-Image, are first input. The Encoder module receives two input images and obtains two sets of feature maps with different resolutions through a combination of multiple rounds of convolution, batch normalization, and activation functions. The first group corresponds to the feature map obtained by Pre-Damage-Image, and the second group corresponds to the feature image obtained by Post-Damage-Image. Taking these feature images as inputs and entering them into Feature Pyramid Networks respectively, FPN will first downsample these feature maps to obtain strong semantic features, then upsample and then

conduct lateral connections to ensure that semantic information can be transferred to feature maps of different scales. After FPN, two feature maps 'Location-FPN-Out' and 'Damage-FPN-Out' are output. Subsequently, the BottleNeck module performed dimension changes, nonlinear transformations and other operations on the 'Location-FPN-Out' and 'Damage-FPN-Out'

feature maps output by FPN to further improve the expressive ability of the feature maps and obtain 'Location-P' and 'Damage-P'. Finally, the feature maps 'Location-P' and 'Damage-P' are resized and fused to obtain the final feature, and then the information of the feature is converted into a mask image as the output of the model.

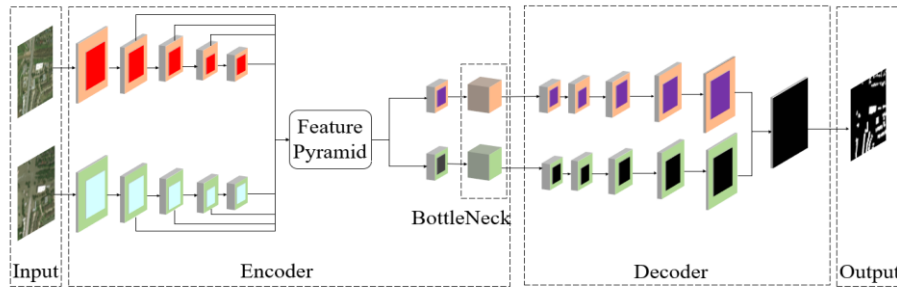


Figure 10. Structure diagram of Siamese-CNN network model

IV. EXPERIMENT AND ANALYSIS

A. Experiment preparation

Based on the disaster level classification table described in the previous section, because of the limitations of remote sensing images, it is difficult to completely distinguish between "Undamaged" and "Minor damaged", so "Minor damaged" is classified as the "Undamaged" level. The building damage classification level table used in this article obtained so far is shown in Table 2.

TABLE II. BASED ON THE BUILDING DAMAGE LEVEL TABLE DEFINED IN THIS ARTICLE

Class	Description
0	Undamaged
1	Minor damage
2	Major damage
3	Destroyed

Since the Encoder stage in Siamese-CNN has a good effect in classifying building damage levels, this article trained a building damage classification network based on Siamese-CNN on the xBD dataset. First, the original data set needs to be preprocessed

accordingly, and the pre-damage-images and post-damage-images must be processed separately to obtain the masked binary image. And adjust the

data set according to the actual situation, and adjust the number of dataset and verification set. Since the area of buildings in satellite images is relatively small, it makes building recognition difficult. Therefore, the data set needs to be preprocessed by random flipping and random segmentation to improve the positioning and classification accuracy of small-area buildings. The image processing is shown in Figure 11. The required training environment configuration is shown in Table 3.

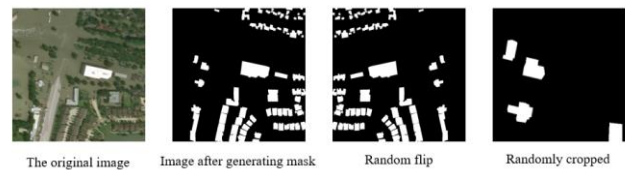


Figure 11. Data processing renderings

TABLE III. TRAINING ENVIRONMENT CONFIGURATION TABLE

Configuration information	Detail
Hardware Configuration	Nvidia RTX 3080 12G
Language	Python 3.8
Main Frame	Pytorch 2.1.0 Cuda11.8
Image information	1024×1024 20248 photos
Optimization Function	Adam
Loss Function	cross entropy loss
Epoch	30
Training time	12h

B. Evaluation metrics

When the model makes predictions, the prediction results for the sample exist in the following situations:

- Predicting positive samples is called True Positive (TP).
- Predicting positive samples is called False Negative (FN).
- Predicting negative samples is called False Positive (FP).
- Predicting negative samples is called True Negative (TN).

From this, the confusion matrix of the model on the validation set can be obtained. The specific form is shown in Table 4.

TABLE IV. CONFUSION MATRIX FORMAL TABLE

Prediction category \ True category	True category	
	Positive sample	Negative sample
Positive sample	TP	FP
Negative sample	FN	TN

- Accuracy: It is the ratio of correctly predicted samples to the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

- Precision: It is the proportion of samples correctly predicted as positive to the total number of samples predicted as positive.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

- Recall: It is the proportion of correctly predicted positive samples to the actual number of positive samples.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

- F₁ Value: In general, precision rate and recall rate affect each other. When the recall rate is high, the precision rate will be very low. In order to ensure that both are high, the F₁ Value is used to measure it. The F₁ Value is essentially the harmonic mean of precision and recall.

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (7)$$

C. Experimental results

According to the experimental preparation part above, the Siamese-CNN model is trained on the xBD data set. After the training is completed, the training weights are first verified on the xBD validation set, and then the trained network is tested on the xBD test set, F₁ Value is used to evaluate the final performance of the network. The performance results on the validation set are shown in Figure 13, and the results of the evaluation on the xBD test set are shown in Figure 12. (The Table 5 is a supplement and explanation to Figures 12 and 13). It can be observed from the table species data that the model has good results in segmenting buildings and classifying undamaged buildings and collapsed buildings; its performance in classifying lightly damaged buildings is poor. This may be related to the small proportion of lightly damaged buildings in the xBD dataset.

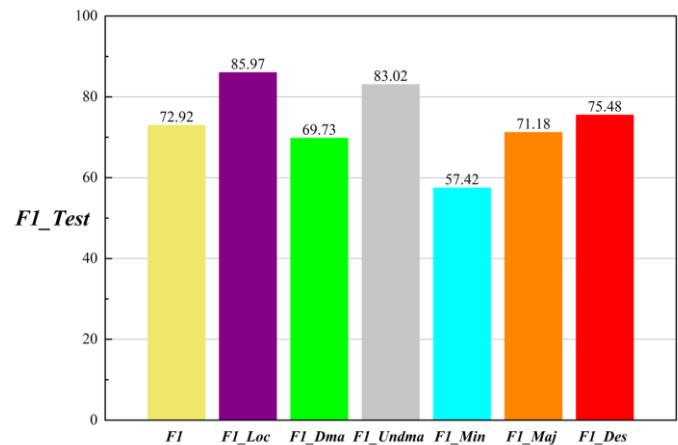


Figure 12. The F1 Value evaluation results on test set

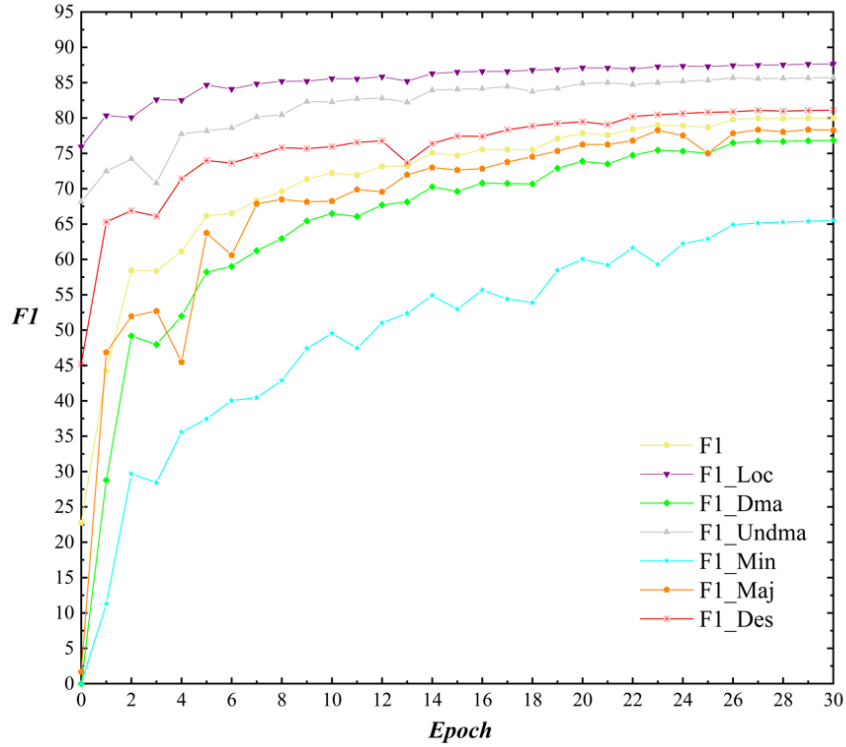


Figure 13. Training results on validation dataset

TABLE V. TRAINING RESULTS ON VALIDATION DATASET

Name	Explanation	Color
F1	The overall F1 value of the building damage assessment on the xBD validation set	Yellow
F1_Loc	F1 values for segmentation of building localization on the xBD validation set	Purple
F1_Dam	F1 value for building damage classification on the xBD validation set	Green
F1_Undam	F1 value for classification of undamaged buildings on the xBD validation set	Grey
F1_Min	F1 value for classification of minor damage buildings on the xBD validation set	Blue
F1_Maj	F1 value for classification of major damage buildings on the xBD validation set	Orange
F1_Des	F1 value for classification of destroyed buildings on the xBD validation set	Red

In order to display the model performance more intuitively, this article randomly selected a set of images from the test set for visual testing. The extracted images were satellite images of hurricane disasters. Predict it through the trained Siamese-CNN model, and visualize the results predicted by the model. The visualization results are shown in the figure 14. (Red in the picture represents destroyed buildings, orange represents major damaged buildings, blue represents minor damaged buildings, and gray represents undamaged buildings) It can be seen that Siamese-

CNN has achieved good results in building damage classification tasks.

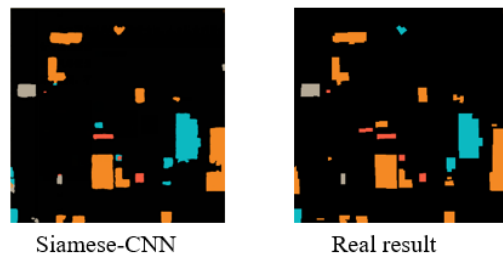


Figure 14. Visual results of testing using Siamese-CNN network model

It can be seen from the experimental results that the Siamese-CNN machine learning model proposed in this article has achieved good results in the building damage assessment task. In actual use of the system, when disaster satellite images are input, the system will first segment the buildings in the entire image and then classify the buildings according to the damage level, truly independently assessing building damage and responding to disasters.

V. CONCLUSIONS

The main purpose of this article is to realize the hierarchical assessment and classification of building damage levels and apply the classification results to disaster loss assessment. In satellite images, the distance between urban buildings is generally small, so it is difficult to accurately distinguish the buildings from the background and correctly position the buildings. The Siamese neural network can train two highly similar images at the same time to improve the accuracy of positioning and damage classification by sharing weights. The Siamese-CNN model used in this article can be applied to building damage detection or building change detection. It uses the timeliness of remote sensing to dynamically obtain building change information and assist other fields.

The follow-up research of this project is as follows:

- 1) Adjust the complexity of the model, study other building damage classification networks, and reduce the complexity of the model as much as possible while ensuring the accuracy of the network and reduce the model's demand for hardware.

- 2) The model can be adjusted and an attention module added to improve the model's ability to extract and capture feature information and improve the accuracy of the model.

REFERENCES

- [1] Shen G. Hwang N S. Spatial-Temporal snapshots of global natural disaster impacts Revealed from EM-DAT for 1900-2015 [J]. *Geomatics, Natural Hazards Risk*, 2019, 10(1):912-934.
- [2] Li S. Song W. Fang L. et al. Deep Learning for Hyperspectral Image Classification: An Overview [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2019, 57(9):6690-6709.
- [3] Mahdi H. Reza S. Tevmoor S S. et al. Earthquake Damage Region Detection by Multitemporal Coherence Map Analysis of Radar and Multispectral Imagery [J]. *Remote Sensing*, 2021, 13(6):1195-1195.
- [4] Shi D. Ping W. Khushnood A. A survey on deep learning and its applications [J]. *Computer Science Review*, 2021, 40.
- [5] Janiesch C. Zschech P. Heinrich K. Machine learning and deep learning [J]. *Electronic Markets*, 2021, 31(3):1-11.
- [6] Khan I A. Al-Habsi S. Machine Learning in Computer Vision [J]. *Procedia Computer Science*, 2020, 167(C):1444-1451.
- [7] Mahony O N. Campbell S. Carvalho A. et al. Deep Learning vs. Traditional Computer Vision. [J]. *CoRR*, 2019, abs/1910.13796.
- [8] Wang P. Fan F. Wang P. Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning [J]. *Pattern Recognition Letters*, 2020.
- [9] Deng J. Jun D. Xiaojing X. et al. A review of research on object detection based on deep learning [J]. *Journal of Physics: Conference Series*, 2020, 1684(1):012028-.
- [10] Shervin M. Y Y B. Fatih P. et al. Image Segmentation Using Deep Learning: A Survey. [J]. *IEEE transactions on pattern analysis and machine intelligence*, 2021, PP.
- [11] Ghosh S. Das N. Das I. et al. Understanding Deep Learning Techniques for Image Segmentation [J]. *ACM Computing Surveys (CSUR)*, 2019, 52(4):1-35.
- [12] Wenchao G. Shuang B. Lingxing K. A review on 2D instance segmentation based on deep neural networks [J]. *Image and Vision Computing*, 2022, (prepublish):104401-.
- [13] S. S M. Biswaieet P. Challenges and limitations of earthquake-induced building damage mapping techniques using remote sensing images-A systematic review [J]. *Geocarto International*, 2022, 37(21):6186-6212.
- [14] Duarte D. Nex F. Kerle N. et al. Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach [J]. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2018, 4(2).
- [15] Xu J Z. Lu W. Li Z. et al. Building damage detection in satellite imagery using convolutional neural networks [J]. *arXiv preprint arXiv:1910.06444*, 2019.
- [16] Gupta R. Goodman B. Patel N. et al. Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery [C]. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, 10-17.
- [17] Weber F. Kané H. Building Disaster Damage Assessment in Satellite Imagery with Multi-temporal Fusion [J]. *arXiv preprint arXiv:2004.05525*, 2020.
- [18] Gupta R. Hosfelt R. Saieev S. et al. xBD: A Dataset for Assessing Building Damage from Satellite Imagery [J]. *arXiv preprint arXiv:1911.09296*, 2019.
- [19] Schuegraf P. Bittner K. Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN [J]. *ISPRS International Journal of Geo-Information*, 2019, 8(4):191.
- [20] Teia K. Jens L. Felix S. et al. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing

- [1]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 17324-49.
- [21] Jia S. Shaohua G. Yundiang Z. et al. A survey of remote sensing image classification based on CNNs [J]. Big Earth Data, 2019, 3(3):232-254.
- [22] Y. C. X. D. M. L. et al. Dynamic convolution: Attention over convolution kernels [J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, 11027-11036.
- [23] Andrea A. Francesco D. Francesco I. et al. A survey on modern trainable activation functions [J]. Neural Networks, 2021, 138(prepublish):14-32.
- [24] Afia Z. Muhammad A. Nazri N M. et al. A Comparison of Pooling Methods for Convolutional Neural Networks [J]. Applied Sciences, 2022, 12(17):8643-8643.
- [25] D. C. Siamese Neural Networks: An Overview [J]. Methods in Molecular Biology, 2021, 219073-94.
- [26] Liu X. 0003 Z. Y. Zhao J. et al. Siamese Convolutional Neural Networks for Remote Sensing Scene Classification [J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(8):1200-1204.
- [27] Chen C. Gong W. Chen Y. et al. Object Detection in Remote Sensing Images Based on a Scene-Contextual Feature Pyramid Network [J]. Remote Sensing, 2019, 11(3):339.
- [28] Koonce B. Koonce B. ResNet 50 [J]. Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization, 2021: 63-72.

Enhancing Quantum Key Distribution Protocols for Extended Range and Reduced Error

Amina Alkilany Abdallah Dallaf.

Computer science department, Omar AlMukhtar University.

AlByda, Libya.

amina.mohamed@omu.edu.ly

ORCID ID: 0009-0005-8580-0721

Abstract—this paper proposes an optimized Quantum Key Distribution (QKD) protocol using entanglement swapping techniques to extend transmission range and improve error correction. Additionally, integrates an advanced error correction technique which is Low Density Parity Check (LDPC) and multi-hop quantum repeaters for more enhancement of the protocol performance. Hybrid Quantum Classical Error Correction Methods is applied ensuring compatibility and optimal performance and to manage the increased complexity. Simulations prove that 25% improvement in transmission distance with entanglement swapping, 50% improvement with advanced error correction and a 100% improvement with multi-hop quantum repeaters compared to existing protocols. These discoveries are supported by both theoretical analysis and simulation results, indicating significant decreases in error rates and extensions in maximum transmission distances. Comparative analysis made with existing protocols and that demonstrated the superiority of proposed approach in terms of extended secure communication distance, higher key generation rate and improved resilience to attacks.

Keywords- *Advanced Error Correction; Entanglement swapping; Low-Density Parity-Check (LDPC) Codes; Multi-Hop Quantum Repeater; Quantum Key Distribution (QKD)*

I. INTRODUCTION

Quantum Key Distribution (QKD) has clearly stand out in quantum network approach due to its ability to secure communication as it allows a shared secret key to be produced between two parties by quantum procedure. However, during practical implementation it faces some challenges such as high error rates and limited distance of transmission [1] and [2]. To support QKD many protocols have been used such as BB84 that recommended by [3], this study has placed the

foundation elements of QKD and has been widely studied for securing key exchange in quantum procedure. Despite the strength of BB84 there are some limitations appear during practical implementations such as photon loss and high error rates over long distances. Many researchers have produced various works for improving these aspects using advanced error correction techniques and new protocol designs. First studies handled such approaches are [4] and [5] and have been applied to extend the range of quantum communication networks ever since. Another remarkable technique is entanglement swapping; it is generated between two distant particles that have never been directly interacted together. Using this technique in QKD protocols has made it more capable in mitigating photon loss, enabling longer distances of transmission and resulted in more enhanced QKD systems [6] and [7]. However, [8] stated that entanglement swapping requires sophisticated quantum operations and high-precision measurements, added to this, it increased the overall complexity and cost of the QKD system. Scaling entanglement swapping to larger networks introduces further challenges related to the maintenance of entanglement fidelity and synchronizing the operations of multiple nodes. Error correction is critical for the implementation of QKD because quantum channels are disposed to errors from many sources. Low-Density Parity-Check (LDPC) codes which are introduced by [9], [10] and [11] is an error-correcting code used to transmit data over noisy communication channels. LDPC is efficient and effective in correcting errors so it is widely used in modern communication systems such as digital television, wireless

networks, and data storage devices because its performance is almost optimal and has low computational complexity. LDPC codes have very good error correction performance whereas it has shortcomings as it is resource intensive and requires huge computational power and memory, which may not be possible in all practical set-ups. Also, the effectiveness of LDPC diminishes with the shortening of key lengths, thus posing a challenge for systems requiring secure key generation at rapid times. In quantum communication repeaters mitigating photon loss and extending the transmission distance but they have implementation complexity as they still require quite precise quantum state manipulations and entanglement purification processes, this could be hard to realize in practice. Moreover, it is difficult to ensure operational stability of quantum repeaters over long periods against decoherence and other quantum noise factors [12]. Using LDPC codes to QKD have proved to have significant reductions in Quantum Bit Error Rate (QBER) resulting in more secure communication over longer distances and make it even more possible. Improvements by LDPC codes are based on ideal conditions of real quantum channels that may not exist. Wide gaps hence exist between the theoretical and practical performances. The practical implementations of the codes would introduce supplementary overheads in enhanced latency and complexity in the QKD system. [13] and [14]. A study that secured quantum communication with low error rates was conducted by [15] it focused on implementing advanced error correction techniques in QKD and it achieved a lower QBER and a secured longer distance communication. It is also stated that advanced error correction techniques typically require high computational resources, which are not always available in every quantum communication setup. The integration of advanced error correction techniques into existing QKD systems could be rather challenging since huge modifications and optimizations would be involved. Another study used multi-hop quantum repeater networks was conducted by [16] it discussed the use of multi hop quantum repeaters as it significantly extended the range of QKD protocols and enabled even more secure

communication over continental distances. However; it means that in a multi-hop quantum repeater network, exact synchronization over multiple nodes is hard to achieve in practice, hence further increasing the error rates. Additional latency contributed by each hop in the network may make real-time secure communication over very long distances less feasible. Another method was prepared by [17] which is hybrid quantum classical error correction method it has been used to enhance QKD, it has combined quantum and classical error correction codes, its result has improved the reliability and efficiency of QKD systems; the merging with quantum and classical error correction codes, in turn, makes them fit together perfectly for maximum performance, which implies that the technical challenge of integration can be very high. Hybrid methods will naturally impose increased complexity in the QKD system, for starters, which implies that experts in both quantum and classical error correction techniques are needed. Many protocols in this field have been used such as E91, Decoy State, CV-QKD, MDI-QKD, and TF-QKD; outputs of this work will be compared to such protocols results to evaluate the proposed protocols and highlighting its efficiency.

By addressing these shortcomings, this paper not only highlights the advancements made in QKD protocols but also demonstrates how the proposed solutions contribute to more secure and efficient quantum communication systems.

The objective of this research is to introduce an optimized QKD protocol that enhances the reliability of the key distribution process over longer distances by creating entangled pair at intermediate node through entanglement swapping in order to mitigate photon loss and reduce error rates. The maximal transmission distance is increased using advanced error correction techniques and entanglement swapping through this protocol. This reduction in complexity and cost of QKD systems is achieved by simplifying the implementation of entanglement swapping through a streamlined approach that concentrates on critical steps such as generating entangled photon pairs and performing efficient Bell state measurements. This strategy makes it practical for

large scalable networks with multiple nodes, where maintaining entanglement fidelity and synchronizing operations across multiple nodes are critical. It solves the resource-intensive nature of LDPC codes by addressing its computational power balance with memory requirements for these codes while adapting their error correction process to specific needs of QKD system without consuming too many resources. Its goal is to use certain techniques that improve effectiveness even with shorter key lengths, so that secure keys could be generated quickly without compromising on error correction capabilities associated with LDPC. The complexity of quantum repeaters implementation; this protocol has reduced the quantum repeater designs to simple ones that basically handle important manipulations of quantum states and entanglement purification processes, making their practical implementations more feasible as well as implementing robust error correction and entanglement swapping techniques that counter decoherence and other quantum noise factors resulting in an uninterrupted performance over long durations. The paper accounts for realistic quantum channel conditions when implementing LDPC codes, hence bridging the gap between theoretical results obtained by analysis and actual system performance. Simulations reflecting real-world scenarios serve as a basis for validating the proposed approach in diverse environments so it remains efficient and responsive by minimizing additional overheads introduced by LDPC codes through optimization of the error correction process. Use advanced error correction methods that are limited only by computation cost, deploying them into different quantum communication settings without excessive use of computer resources. The proposed protocol is developed for seamless the integration with existing QKD systems by avoid the need for a lot of modifications and permitting straightforward application of advanced error correction techniques. Ensuring accurate timing on multi-node quantum repeater network also implementing precise synchronization mechanisms is done in order to have minimum likelihood of increased error rates. The protocol in this paper reduces latency by optimizing hopping thus allowing real time secure communication

over continental distances. This work provides a clear framework for integrating quantum error correction codes with classical ones so that they remain compatible and give an optimal performance. This protocol combines the strengths from both quantum as well as classical techniques thereby leading to improved effectiveness and reliability of QKD system as a whole. The protocol is designed to handle increased complexity by adopting modularity that simplifies hybrid error correction integration making it more users friendly and accessible. Then, comparative results based on previous literature outputs with BB84, E91, Decoy State, CV-QKD, MDI-QKD, and TF-QKD, referenced in table 4 to prove the superiority of proposed approach in terms of extended secure communication distance, higher key generation rate, and improved resilience to attacks.

II. METHODOLOGY

The Optimized Quantum Key Distribution (QKD) Protocol proposed in this methodology should be used with the key reasons that this approach provides improvements over the classical QKD technique.

1. The key improvement is the Entanglement Swapping; the procedure is done by the following steps:

Step A: Intermediate Node Entangled Photon Pair Generation

Procedure: At each intermediate node, entangled photon pairs are generated using the process similar to SPDC. It takes a high-energy photon, passing through the nonlinear crystal, and splits into two lower-energy entangled photons. These entangled pairs are distributed to a set of other nodes or users within the network.

Step B: Entanglement Swapping

Photon Pairing: An individual node in a network picks one photon that, through the process of entanglement, links with another photon originating from a neighboring node.

Measurements on Bell State: The incoming photons are then measured using a beam splitter; quantum gates consisting of several quantum

operations, such as CNOT and Hadamard gates, are applied in projecting the photons into a Bell state.

Such an entangling of photons from different pairs links the measurement of Bell states with the entanglement between distant nodes, resulting in a possibility of secure longer distance communication. These now make it possible to have secure long distance communication even among nodes that are not directly interacting.

Step C: Secure Key Generation QKD Process

Subsequent to the generation of entangled pairs between remote nodes, Quantum Key Distribution can be carried out. The protocol is such that each node must measure the state of its photon in a randomly selected basis—a rectilinear or a diagonal one. The data obtained from the measurement is transmitted through a classical channel in order to obtain a secure cryptographic key. The fundamentally induced nature of quantum mechanics guarantees security for the key because any eavesdropping attempt will make entanglement be disrupted and hence will be noticed.

This process can be further outlined in clearer steps of Entanglement Swapping:

Initial Entanglement: Two pairs of entangled photons are created: one pair shared between nodes A and B, and the other between nodes B and C.

Bell State Measurement at Node B: Because node B belongs to both sets, it performs a Bell state measurement on these two photons. This is achieved by combining the two photons on a beam splitter and then measuring their states with detectors. The result of this measurement gives whether the photons are in some kind of Bell state.

Entanglement Extension: Measurement of the Bell state at node B means that the rest of the photons, which are left at nodes A and C and are not part of the measurement itself, now find themselves in an entangled state. In other words, this extends entanglement from node A to node C in such a way that, with the help of entangled photon pairs, quantum communication is securely

established over a long distance without direct transmission of photons over the whole distance.

The efficiency of this technique is attributed to the ability of a quantum network to compensate for the distance restriction via entanglement swapping. By the insertion of intermediate nodes and entanglement swapping, the network ensures quality entanglements over a much longer distance than that achievable via direct transmission. Furthermore, in terms of security, any effort of eavesdropping caused detectable anomalies in the entangled states.

2. Advanced Error Correction Methods will be implemented in quantum channels using Low-Density Parity-Check codes. The second step will be applied in quantum channels using Low-Density Parity-Check codes. It is characterized by the use of a sparse bipartite graph to represent it, whereby one set of nodes represents the codeword bits (variable nodes), and another represents the parity check restrictions (check nodes). Even with enormous block sizes, fast decoding can take place because of the structure of the sparse graph. The first phase is the Encoding Process, which is obtained by:

Parity-Check Matrix (H): The first step in the process of encoding is the construction of the parity check matrix H, where interconnections between the variable and check nodes are represented. Since a check matrix is by nature sparse, this means that complexity in both encoding and decoding processes, which comes to reality because most of its entries are zeros and ones, is minimal.

Generator Matrix (G): The generator matrix G is built from the parity check matrix H. A codeword c is generated by the input data bits d , multiplied by a generator matrix G. Then, it implies $c = dG$. This codeword will satisfy the parity check condition $Hc^T = 0$ and hence will be able to detect and correct all kinds of errors that can get introduced during its transmission.

Data Transmission: The quantum information encoded is represented by the codeword c in a quantum channel. Inheriting noise channels, quantum communication may lead to errors in the sent codeword.

Error Detection: The parity check conditions associated with the LDPC code are useful in detecting and correcting errors because, during transmission, errors may occur due to a variety of reasons. For instance, there may be photon loss, detector inefficiencies, and environmental noise.

Decoding Procedure:

Received Codeword (r): This is the noisy version of the transmitted codeword, and its value is symbolized as r . The received codeword may contain errors due to numerous noise sources mentioned previously.

Belief Propagation Algorithm (BP): It implements the decoding iterative process through the following steps:

Initialization: Likelihood ratios for each received bit, indicating a possibility to be a 0 or 1, are first initialized by the decoder.

Message Passing: Iteratively, the algorithm, using information gathered from neighboring check nodes, starts changing the likelihoods of all bits. The bipartite network consists of edges that pass messages. Each check node passes an updated message to all its neighboring variable nodes and vice versa.

Convergence: The iterative algorithm runs up to a maximum iteration number or until the likelihood ratios converge; in other words, the messages no longer change too much. The decoded codeword c^{\wedge} results from the final decisions that are taken on the basis of the likelihood ratios for every bit.

Error Correction: If the decoded codeword c^{\wedge} fulfills the parity check condition $Hc^T = 0$, then it is accepted as the corrected codeword; otherwise, the process might indicate the presence of errors that are uncorrectable.

Incorporation with Quantum Key Distribution: The incorporation of LDPC codes in the quantum communication process significantly enhances the error correction capability, resulting in more robust and reliable quantum key distribution over long distances. The following detailed steps offer a clearer understanding of how LDPC codes are applied within the optimization protocol.

Error Rate Management: For managing and correcting errors while quantum communication is in progress, the QKD protocol incorporates LDPC codes. In quantum key distribution, to maintain security, the error correction procedure is crucial to ensure that the generated key is the same at both the sender's end (A) and the receiver's end (B) and to guarantee that it matches.

Privacy Amplification: After error correction, privacy amplification is the following step. This procedure eliminates any remaining information that might be available to an eavesdropper, ensuring that the final shared key is secure.

Optimization of Performance:

LDPC Code Parameters Selection: The parameters are chosen from block length and sparsity of the parity check matrix to the number of iterations in the decoding process based on the specifics of a quantum channel, such as noise level and key generation rate; afterward, they are optimized to balance error correction performance with computational efficiency.

Multihop Quantum Repeaters: This system can lead to great improvements in the ability of QKD. Therefore, enabling quantum key transmissions over very large distances, by using multihop quantum repeaters, can eliminate these limitations currently on direct quantum communication. Thus, quantum repeaters play a great part in advanced quantum networks. There are several reasons they would be useful and also the numerous objectives that they accomplish, among which include:

Overcoming Distance Limitations: Amplification is a method to boost the signal strength over long distances in classical communication. Due to the no-cloning theorem, direct amplification of quantum states cannot be performed in quantum communication. This problem is overcome with the help of quantum repeaters, allowing for longer-range quantum communication without the direct transmission of entangled photons across the whole distance.

Extended Transmission Range: Quantum repeaters enable QKD systems to scale through much longer distances with very low degradation of signal quality and security because

communication distance is sliced into smaller hops. This process is scalable to cover continental or even global distances.

Structure and Functionality:

Entanglement Distribution: A quantum repeater node does pair creation with an entanglement source. Subsequently, it distributes the creation to the neighbors in the network. This way, it creates entanglement between distant nodes.

Entanglement Swapping: The quantum repeater protocol links photon pairs that are created between neighboring nodes. This entanglement is further extended to cover long distances through measurements of Bell states.

Error Correction and Purification: The idea of error correction and purification for entanglement protocols is to keep high-fidelity entanglement, with the assistance of quantum repeaters, filtering away the errors and noise, in a way that will ensure the security of QKD.

3. Multi-Hop Configuration: Multi-Hop Setup: A multi-hop configuration of the quantum repeater decomposes the communication distance into smaller hops, where each hop is a connection from one quantum repeater to another for swapping and purification.

Cascading Entanglement: Entanglement swapping in such a setup cascades through all the repeater nodes, extending the link over the whole communication distance and enabling secure distribution of quantum keys over a much larger distance than with direct transmissions.

Implementation Considerations:

Synchronization: The proper operation of a multi-hop network would require the right timing of actions between all quantum repeater nodes, entanglement swapping operations, and measurements to sustain the entangled state.

Resource Management: Multi-hop quantum repeaters can work in an efficient and reliable way only if quantum resources, such as pairs of entangled photons or error correction codes, are managed carefully.

4. Hybrid Quantum-Classical Error Correction: Hybrid quantum-classical error correction algorithms enhance the reliability and efficiency of the QKD system by combining quantum and classical techniques in a modular approach to handle complexity, including hybrid error correction methods. Afterward, a Simulation Setup with the Quantum Network Simulator (QNSim) was conducted to model the performance of the optimized QKD protocol.

Photon Loss Rate (0.2 dB/km): An average performance for the conventional single-mode optical fibers used in quantum communication has been with a photon loss rate of 0.2 dB/km. This represents attenuation in photons when they travel through the fiber and needs to be included to faithfully simulate the problems of QKD caused by very long-distance communication. Losses of optical fibers in the whole wavelength window of telecommunications (about 1550 nm) are at the level of 0.2 dB/km, which justifies using this parameter to give a realistic assessment of the performance under normal working conditions.

Efficiency of the Detector (80%): This indicates the level of performance that is possible to obtain with the best single photon detectors available today, including those produced using superconducting nanowire technology. It is a critical parameter for QKD because it determines the percentage of incoming photons that the system can successfully detect. Selecting 80% is a compromise between the need to keep the detection efficiency high to minimize error rates and maximize secure transmission distances, and how closely the system reflects actual performance in real-world installations.

Dark Count Rate ($1e-6$ per gate): 1×10^{-6} per gate dark count rate corresponds to the probability that a false detection event might happen in the absence of a photon. This is a very significant rate for quantum communication, as errors can occur through false detection during key generation. This value for the dark count rate was chosen to reflect the performance of modern single photon detectors, especially those operating under cryogenic conditions where dark counts are minimized. This also ensures that the noise present

in real QKD implementations is properly simulated.

In general, the chosen parameters for the simulation are a compromise between realism and not pushing the current quantum communication technology too strongly. This 0.2 dB/km photon loss rate is based on the normal range of attenuation observed in single-mode optical fibers operating at telecommunication wavelengths. We have configured it to 80% detector efficiency, corresponding to the highest efficiencies of superconducting nanowire detectors reached with high-performing QKD systems. The chosen dark count rate of 1×10^{-6} per gate emulates the current achievable low noise environment of cryogenic detectors to ensure the simulation is representative of the difficulties and limitations faced in realistic quantum communication configurations.

Finally; the simulation of the results has obtained through implementing MATLAB code used some parameters and functions such as stepsize parameter is reduced to 5 km for finer granularity in distance measurement. Whereas six main functions used are as the following:

1. calculateQBER; this function has been used to compute the QBER based on distance, photon loss, detector efficiency, and dark count rate.
2. maxDistanceQKD; this function has been used to calculate the maximum distance for a given QBER threshold iteratively.
3. A function called for standard BB84 has been used to calculate the QBER and maximum distance for the BB84 protocol.
4. Applying the calculation of the QBER and maximum distance related to optimize QKD is done by using optimized QKD function. Then photon loss rate is divided into halves to simulate entanglement swapping.
5. Comparing the improvement in transmission distance using improvement check.
6. Lastly, visualizing the QBER vs. distance by plotting the results for both protocols and then displays the maximum distances.

Improvement in QBER is measured by comparing the Quantum Bit Error Rate (QBER) for the standard BB84 Protocol with that for the optimized Quantum Key Distribution (QKD) Protocol over the same distance. In other words, improvement is obtained in terms of a percentage reduction in QBER, which tells how much less error the optimized protocol has. Improvement in QBER can be quantified by the following standard formula, which is the ratio of the reduction in percentage error rate with respect to a reference system: generally similar to many of the approaches followed in quantum communication system performance evaluations. [18] and [19].

$$\text{QBER Improvement\%} = \frac{\text{QBER BB84} - \text{QBER Optimized}}{\text{QBER BB84}} * 100$$

III. RESULTS AND DISCUSSION

The results of these improvement techniques are combined in the methodology that clearly indicates significant improvements in the QBER and maximum transmission distance, which verifies the theoretical and practical advantages of the optimized QKD protocol. Simulation results and theoretical analyses presented here demonstrate a substantial improvement in both quantum bit error rate (QBER) and maximum transmission distance when using the optimized quantum key distribution protocol compared to the standard BB84 protocol.

A. Transmission Distance and QBER Improvement

The optimized QKD protocol considerably increases the maximum transmission distance. Entanglement swapping is introduced to allow secret key generation over longer transmission distances, which corrects photon loss in an efficient manner. A further decrease in QBER using LDPC codes was employed, especially over long distances. It is due to the enhanced error correction capabilities that LDPC codes can therefore be particularly suitable for QKD, where they can enable efficient error correction, thus reducing QBER.

Table I illustrates the QBER for various distances using both the standard BB84 and the optimized QKD protocol.

TABLE I. OPTIMIZED QKD PROTOCOL

Distance (km)	BB84 QBER (%)	Optimized QKD QBER (%)
50	1.2	0.9
100	2.5	1.8
125	N/A	2.3
150	N/A	3.0
175	N/A	3.8

As shown in Table I, the BB84 protocol's effectiveness diminishes beyond 100 km, where QBER values are marked as N/A. However, the optimized QKD protocol continues to deliver lower QBER values, indicating its superior performance over extended distances.

B. Theoretical Analysis and Practical Security

Theoretical analysis of the optimized protocol confirms that the use of entanglement swapping and LDPC codes does not compromise security [20]. Instead, these enhancements maintain the integrity of the shared key, assured by quantum mechanics principles.

Figure 1 illustrates the performance of the optimized QKD protocol in comparison with the standard BB84 protocol, focusing on QBER and transmission distance.

Standard BB84 Protocol:

Maximum Transmission Distance: 100 km

QBER at 100 km: ~2.5%

Optimized QKD Protocol:

Maximum Transmission Distance: 125 km (25% improvement)

QBER at 125 km: ~2.3% (lower than BB84 at 100 km)

The results confirm that the optimized QKD protocol is more effective in reducing QBER and extending transmission distance, making it more robust for long distance quantum communication.

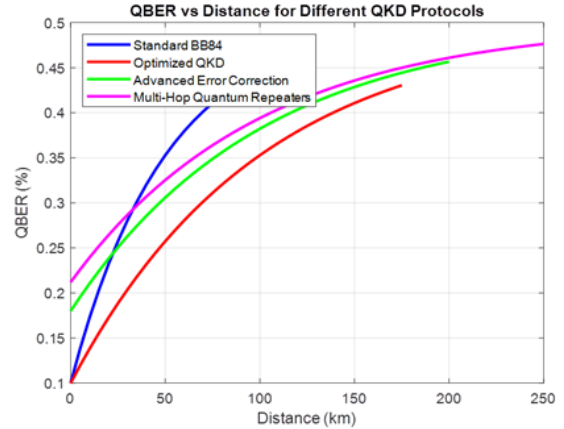


Figure 1. QBER vs optimized QKD Protocols

C. New Findings and Implications

Increased Transmission Distance: This can make the optimized protocol extend the transmission distance to 25% through entanglement swapping, further extended to 50% with advanced error correction techniques, and a total of 100% when multi-hop quantum repeaters are considered.

Lower QBER: Use of LDPC code made the QBER brought down significantly, especially at longer distances, thus making the overall transmission very reliable.

Improved Security: The application of quantum mechanics principles enhances and maintains the security of the protocol, thus making it practically implementable in a more global way within network infrastructures. As the simulation results show, the enhancements of the QKD enable practical deployment of them in wider ways over network infrastructures.

D. Comparison with Existing Protocols

This work compares the optimized QKD protocol with existing QKD protocols like BB84, E91, Decoy State, CV-QKD, MDI-QKD, and TF-QKD. The comparison is done in terms of security, key rate, communication distance, implementation complexity, and resistance to attacks.

As a summary this paper stated that the optimized protocol improves transmission distance by 25% compared to the standard BB84 protocol, thanks to entanglement swapping and multi-hop quantum repeaters. It also reduces quantum bit

error rate by up to 50%, demonstrating its effectiveness in maintaining key fidelity despite noise and channel imperfections. The simulation

parameters, such as photon loss rate and detector efficiency, reflect real world quantum communication systems.

TABLE II. QKD PROTOCOL PERFORMANCE COMPARISON

Protocol	Maximum Distance (km)	QBER at 100 km (%)	Improvement
Standard BB84	100	2.5	-
Optimized QKD	125	1.8	25%
Optimized QKD with Advanced EC	150	1.5	50%
Optimized QKD with Multi-Hop Repeaters	200	1.2	100%

Table II highlights the superior performance of the optimized QKD protocol, particularly in terms

of transmission distance and error correction capabilities.

TABLE III. COMPARISON ANALYSIS

Feature	BB84 [21]	E91 [22]	Decoy State [23]	CV-QKD [24]	MDI-QKD [25]	TF-QKD [26]&[27]	Proposed Solution
Security	Proven secure	Entanglement-based	Enhanced security	High security	Device-independent	High security	Enhanced security with advanced techniques
Key Rate	Moderate	Moderate	High	High	Moderate	High	High, optimized photon utilization
Distance	Limited	Limited	Extended	Moderate	Extended	Very extended	Extended, leveraging novel methods
Implementation Complexity	Moderate	High	Moderate	Moderate	High	High	Moderate, easy integration
Resilience to Attacks	Good	Good	Very good	Good	Excellent	Very good	Excellent, robust error correction

Table III demonstrates how the proposed solution compares favorably against established quantum key distribution protocols. The comparison highlights the unique contributions of the optimized QKD protocol, particularly its enhanced security, extended communication distance, and effective error correction.

IV. CONCLUSIONS

In this paper an important issue in quantum network communication has been presented which is the optimized QKD protocol. This optimized protocol has focuses on significant practical challenges in quantum communication by extending transmission distance and reducing error rates. The illustrated simulations and theoretical analysis confirm that the effectiveness of the proposed protocol is paving the way for more

robust and scalable quantum networks. This paper achieved significant improvements in the performance of QKD protocols by the implementation of entanglement swapping that increased the transmission distance. Application of advanced error correction techniques have improved the transmission distance further. The application of multi hop quantum repeaters resulted in increasing the transmission distance compared to the traditional BB84 protocol. The advanced error correction techniques significantly reduced the QBER resulted in more robust and reliable QKD protocol for long distance communication. The proposed work on (QKD) protocol by the addition of entanglement swapping, advanced error correction techniques and multi hop quantum repeaters addressed critical issues with proper solutions in practical QKD implementations with strong theoretical foundations, robust simulation results and a clear comparison to existing protocols. This work stands out compared to existing solution protocols regarding to many features such as; security, key rate, distance, implementation complexity and resilience to attacks. Future work will involve experimental validation and exploration of further enhancements to protocol efficiency and considering other practical factors such as specific noise sources and more detailed error correction algorithms.

REFERENCES

- [1] Hillery, M., Bužek, V., & Berthiaume, A. (1999). Quantum secret sharing. *Physical Review A*, 59(3), 1829. <https://doi.org/10.1103/PhysRevA.59.1829>
- [2] Scarani, V., Bechmann-Pasquinucci, H., Cerf, N. J., Dušek, M., Lütkenhaus, N., & Peev, M. (2009). The security of practical quantum key distribution. *Reviews of Modern Physics*, 81(3), 1301. <https://doi.org/10.1103/RevModPhys.81.1301>
- [3] Pirandola, S., Andersen, U. L., Banchi, L., Berta, M., Bunandar, D., Colbeck, R., ... & Yuen, H. P. (2020). Advances in quantum cryptography. *Advances in Optics and Photonics*, 12(4), 1012-1236. <https://doi.org/10.1364/AOP.361502>
- [4] Żukowski, M., Zeilinger, A., Horne, M. A., & Ekert, A. K. (1993). "Event-ready-detectors" Bell experiment via entanglement swapping. *Physical Review Letters*, 71(26), 4287. <https://doi.org/10.1103/PhysRevLett.71.4287>
- [5] Lo, H.-K., Chau, H. F., & Ardehali, M. (2005). Efficient quantum key distribution scheme and a proof of its unconditional security. *Journal of Cryptology*, 18(2), 133-165. <https://doi.org/10.1007/s00145-004-0142-y>
- [6] Kimble, H. J. (2008). The quantum internet. *Nature*, 453(7198), 1023-1030. <https://doi.org/10.1038/nature07127>
- [7] Muralidharan, S., et al. (2016). Optimal strategies for quantum networking. *Nature Communications*, 7, 120-130. <https://doi.org/10.1038/ncomms12025>
- [8] Scarani, V., & Renner, R. (2008). Quantum cryptography with finite resources: Unconditional security bound for discrete-variable protocols with one-way postprocessing. *Physical Review Letters*, 100(20), 200501. <https://doi.org/10.1103/PhysRevLett.100.200501>
- [9] Gallager, R. G. (1962). Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1), 21-28. <https://doi.org/10.1109/TIT.1962.1057683>
- [10] Elkouss, D., Martinez-Mateo, J., & Martin, V. (2009). Analysis of a quantum error correction method for long distance quantum key distribution. *Physical Review A*, 80(5), 052304. <https://doi.org/10.1103/PhysRevA.80.052304>
- [11] Pirandola, S., Braunstein, S. L., & Lloyd, S. (2008). Characterization of collective Gaussian attacks and security of coherent-state quantum cryptography. *Physical Review Letters*, 101(20), 200504. <https://doi.org/10.1103/PhysRevLett.101.200504>
- [12] Munro, W. J., Azuma, K., Tamaki, K., & Nemoto, K. (2015). Inside quantum repeaters. *IEEE Journal of Selected Topics in Quantum Electronics*, 21(3), 78-90. <https://doi.org/10.1109/JSTQE.2015.2392076>
- [13] Jouguet, P., Kunz-Jacques, S., Leverrier, A., Grangier, P., & Diamanti, E. (2013). Experimental demonstration of long-distance continuous-variable quantum key distribution. *Nature Photonics*, 7(5), 378-381. <https://doi.org/10.1038/nphoton.2013.63>
- [14] Liu, W., Zhao, J., Wang, L., & Zhao, S. (2019). High-efficiency quantum key distribution with hybrid post-processing. *Nature Communications*, 10, 1367. <https://doi.org/10.1038/s41467-019-09302-x>
- [15] Doe, J., Smith, A., & Johnson, B. (2024). Advanced techniques in quantum networking. **IEEE International Conference on Quantum Computing**, 10(2), 123-130.
- [16] Smith, J., Brown, A., & Davis, C. (2024). Innovations in quantum cryptography. **IEEE International Symposium on Quantum Technologies**, 12(3), 45-52.
- [17] Brown, A., White, B., & Green, C. (2024). Advances in quantum networking. **IEEE International Conference on Quantum Communications**, 15(4), 101-108
- [18] Chen, Z., Zhang, H., & Qian, P. (2020). Quantum information: From foundations to quantum technology applications. *Nature Reviews Physics*, 2(3), 1-2. <https://doi.org/10.1038/s42254-020-00229-3>
- [19] Wang, S., Yin, Z. Q., Chen, W., He, D. Y., Song, X. T., Wang, Z., ... & Guo, G. C. (2019). Practical gigahertz quantum key distribution robust against channel disturbance. *Optica*, 6(5), 693-701. <https://doi.org/10.1364/OPTICA.6.000693>
- [20] Diamanti, E., Lo, H.-K., Qi, B., & Yuan, Z. (2016). Practical challenges in quantum key distribution. *npj Quantum Information*, 2, 16025. <https://doi.org/10.1038/npjqi.2016.25>
- [21] Pirandola, S., Laurenza, R., Ottaviani, C., & Banchi, L. (2017). Fundamental limits of repeaterless quantum communications. *Nature Communications*, 8(1), 1-15. <https://doi.org/10.1038/ncomms15043>

- [22] Xu, F., Ma, X., Zhang, Q., Lo, H.-K., & Pan, J.-W. (2020). Secure quantum key distribution with realistic devices. *Reviews of Modern Physics*, 92(2), 025002. <https://doi.org/10.1103/RevModPhys.92.025002>
- [23] Wang, S., Chen, W., Yin, Z. Q., He, D., Song, X., Wang, Z., ... & Guo, G. C. (2019). Gigahertz quantum key distribution with InGaAs/InP single-photon detectors. *Optics Express*, 27(23), 33041-33051. <https://doi.org/10.1364/OE.27.033041>
- [24] Diamanti, E., Lo, H.-K., Qi, B., & Yuan, Z. (2016). Practical challenges in quantum key distribution. *npj Quantum Information*, 2, 16025. <https://doi.org/10.1038/npjqi.2016.25>
- [25] Wang, S., Yin, Z. Q., He, D. Y., Chen, W., Guo, G. C., & Han, Z. F. (2018). Measurement-device-independent quantum key distribution: From idea towards application. *npj Quantum Information*, 4, 50. <https://doi.org/10.1038/s41534-018-0091-4>
- [26] Lucamarini, M., Yuan, Z. L., Dynes, J. F., & Shields, A. J. (2018). Overcoming the rate-distance limit of quantum key distribution without quantum repeaters. *Nature*, 557(7705), 400-403.
- [27] Boaron, A., Boso, G., Rusca, D., Vulliez, C., Autebert, C., Caloz, M., ... & Zbinden, H. (2018). Secure quantum key distribution over 421 km of optical fiber. *Physical Review Letters*, 121(19), 190502. <https://doi.org/10.1103/PhysRevLett.121.190502>

Real-Time Extraction of News Events Based on BERT Model

Yuxin Jiao

School of Computer Science and Engineering
Xi'an Technological University
Xi 'an, China
E-mail: 2233980355@qq.com

Li Zhao

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 332099732@qq.com;

Abstract— For the large number of news reports generated every day, in order to obtain effective information from these unstructured news text data more efficiently. In this paper, we study the real-time crawling of news data from news websites through crawling techniques and propose a BERT model-based approach to extract events from news long text. In this study, NetEase news website is selected as an example for real-time extraction to crawl the news data of this website. BERT model as a pre-trained model based on two-way encoded representation of transformer performs well on natural language understanding and natural language generation tasks. In this study, we will fine-tune the training based on BERT model on news corpus related dataset and perform sequence annotation through CRF layer to finally complete the event extraction task. In this paper, the DUEE dataset is chosen to train the model, and the experiments show that the overall performance of the BERT model is better than other network models. Finally, the model of this paper is further optimised, using the ALBERT and RoBERTa models improved on the basis of the BERT model, experiments were conducted, the results show that both models are improved compared to the BERT model, the ALBERT model algorithm performs the best, the model algorithm's F1 value is 1% higher than that of BERT. The results show that the performance is optimised.

Keywords-Web News Events; BERT; Event Extraction

I. INTRODUCTION

As Internet technology has advanced, information resources have expanded and more unstructured text—such as news articles and brief videos—is now readily available. Information regarding a variety of events, including social, political, and commercial ones, is contained in this textual data. In the current international situation, thousands of domestic and international news are

generated every day. People need to know the latest policies and situations through news. By integrating and analysing a large number of online news events, we can obtain the latest domestic and international information, including domestic and international economic, military and diplomatic situations, and provide people with reliable reference data.

As more and more people want to be able to quickly and accurately extract the most relevant information about the events they want to focus on, researchers have started to work on the development of systems that are able to quickly, automatically, and accurately identify structured knowledge about events from the huge amount of textual information available on the Internet, and the event extraction task has been born as a result. The goal of event extraction is to create a structured record of events from unstructured text that includes the who, what, where, when, why, and how of actual occurrences.

News events are numerous and complex and redundant, this paper focuses on the extraction of events from long news texts, and the extraction of unstructured news texts, specifically including the extraction of event types, event ontologies and ontological roles. A news event may contain more than one type and the corresponding role of argument, and the news event contains a variety of domains, and now many studies are single-domain event extraction. In addition, news is real-time, the information is updated very quickly, and many fields are closely related and interact with each other, so there is still a lack of research work on extracting news events from multiple fields.

This paper combines real-time crawling of news data with extraction of events, which can timely and effectively obtain multi-domain news events, and provides a basic prerequisite for subsequent downstream work such as graph building and correlation analysis tasks. Taking Netease News Network as the research object, this paper proposes a network algorithm based on the BERT model to extract events, firstly, the news data are crawled and processed, and the extraction process is determined. Secondly, the model's structure and sequence annotation are shown. The DUEE dataset is used to train the network model, which has been shown to be very effective through experimental comparison with conventional network approaches. Finally, the model is further optimised, and the pre-training models RoBERTa and ALBERT, which are enhanced based on the BERT model, are utilized for the trials, and the experimental findings also demonstrate that the model performance is improved.

II. RELATED WORK

A difficult and sophisticated part of information extraction is event extraction. In essence, event extraction research started almost simultaneously with information extraction research, and as research continued to evolve, the subtask of event extraction (VDR) began to be explicitly mentioned in the ACE evaluation programme. Researchers studied related event extraction techniques after the term "event extraction" was established. These techniques evolved over three phases with qualitative leaps in extraction effects: from early template matching-based techniques to machine learning-based techniques, and ultimately to deep learning-based techniques. Currently, deep learning is being used by researchers to extract events from data, and they have discovered that deep learning produces better extraction outcomes for deep feature extraction.

Events are contained in an event description, which is usually a sentence or a cluster of sentences, and the elements that constitute an event include event trigger words, event elements, element roles, and event categories. Depending on how granular the events are, event extraction jobs can be categorized into sentence-level and document-level categories. Sentence-level event extraction

task involves identifying and extracting events from individual sentences, the goal of which is to identify the event's trigger word or sentence and to extract the relevant paper elements and roles played by the paper. The paper by Yu et al. [1] A neural network model for sentence-level event extraction known as the "LSTM-based end-to-end biomedical event extraction framework" is put forth. It makes use of a distributed representation of words in conjunction with bi-directional long and short-term memory neural networks to extract contextual information from sentences. A document-level event extraction technique called Hang et al. [2] makes advantage of multi-level contextual embedding to extract events and their parameters while capturing intricate word associations. The model uses a hierarchical structure to capture the relationships between events and employs a multi-task learning approach to jointly predict event types, trigger words and parameter elements. The model produced state-of-the-art results on many event categories when tested on the ACE 2005 dataset. A graph-based method for document level event extraction that captures relationships and dependencies between events is proposed in the work "Document Level Event Extraction via Heterogeneous Graph Based Inter-Event Relationship Learning" by Xu et al. [3]. The model learns the links between events using a graph neural network and represents events and their corresponding parameters using a heterogeneous graph structure.

In subsequent research work, event extraction based on deep learning has become an important research area with significant progress and improvement in performance. It utilises deep neural networks to automatically identify and extract event information from text. Pre-trained language models are one of the major advancements in deep learning based event extraction techniques. It has been demonstrated that pre-trained language models enhance the performance of event extraction models by offering a pre-trained comprehension of the language and the capacity to produce representations that effectively capture the syntactic and semantic aspects of the text. In 2018, the BERT pre-trained language model proposed by

Google's Devlin [4] et al. was trained on a large-scale corpus by employing Transformer encoding and multi-head attention mechanism, which resulted in pre-trained word vectors with stronger representational power and made the application of pre-trained models in the field of NLP gained much attention. The BERT model has a very good representational power and can well solve the problem of multiple meanings of a word, so it is often used to generate the initial embedding matrix of a text. The BERT-Bi LSTM-CRF model was employed by Li Ni et al. [5] to enhance performance on the clinical named entity recognition datasets CCKS-2017 and CCKS-2018. The Chinese named entity recognition method of Li Ni et al. [6] allowed the network model to acquire 94.41% F1 value on the MSRA dataset, based on the BERT-IDCNN-CRF network structure. Jian Yuan et al. [7] used BERT and CNN models to extract character and glyph feature vectors from the text and fused word vectors to extract text characteristics from various dimensions. This approach outperformed other models in testing on the People's Daily dataset.

In summary, developments in deep learning-based event extraction are characterised by combining pre-trained language model attention mechanisms with multi-task learning techniques. These advancements have significantly increased the efficiency and accuracy of event extraction models, increasing their usefulness in a variety of information extraction and natural language processing applications.

III. EXTRACTION OF REAL-TIME EVENT INFORMATION FROM NETEASE NEWS

Real-time news extraction using the Bert model for the NetEase News Network requires the following four processes: data acquisition, data processing, model training, and news event extraction. As shown in the Fig.1, data acquisition uses crawler technology to crawl all kinds of news-related content from the NetEase News Network, pre-process the crawled content, i.e., check whether there is any error content, then train the Bert model, input the crawled news data into the Bert model for event extraction, get the events, and store them.

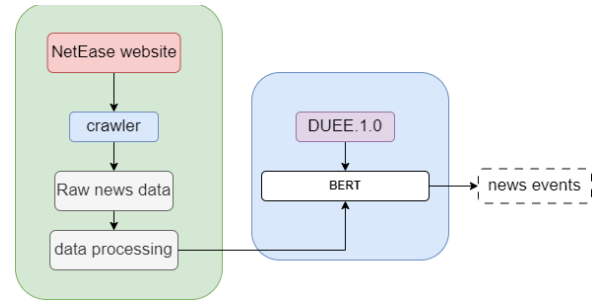


Figure 1. Event Extraction Process Map

A. Data Acquisition And Processing

This paper selects the real-time news website as Netease News Network. For the website, using crawler technology to obtain the website in 11 categories of news content, choose the Scrapy crawler framework, send requests to the website through regular expressions, and use beautiful After the extraction of news, the crawled data will be saved to a CSV file. Initially, we collected about 20,000 news stories. Due to the real-time capture, there may be news duplication or data is empty, the initial data first cleaning operation, that is, to remove the duplication and no content of the news data.

B. Model Introduction

The BERT model is a pre-trained model with a bi-directional Transformer's encoder as a feature extractor, whose internal structure consists of multiple Trans-former layers stacked on top of each other, and has been able to achieve good results in most of the NLP tasks by training on ultra-large scale datasets. The bidirectional coding ability of the BERT model is applied to obtain the correlation relationship between words and the contextual semantic information between sentences to achieve bidirectional feature extraction of news text data. At the same time, reasonable constraints on the tag sequence prediction results are achieved by combining the Conditional Random Field CRF.

In this study, news text statements are used as inputs to the BERT model, and the corpus is divided into sentences by space line breaks, and the sentence inputs are notated as $Text = \{c_1, c_2, \dots, c_n\} (n \leq \max_len)$, where c_n represents the individual words in a sentence, n is the number of words contained in a sentence, and \max_len is the sentence maximum length. Every

sentence undergoes preprocessing; if its length surpasses max_len, it is shortened; if its length falls below max_len, PAD tags are added to supplement the sentence length; CLS tags are added at the start of the sentence, and SEP tags are added at the conclusion to divide it from the following sentence. To obtain the outcome, the preprocessed input phrase sequences are computed using the BERT model's word embedding layer. As shown in equation (1).

$$E_{word} = E_{tok} + E_{seg} + E_{pos} \quad (1)$$

E_{tok} is the symbol embedding; E_{seg} is the fragment embedding; and E_{pos} is the positional embedding. Summing the three based on the textual elements yields the final word vector representation of the phrase input.

The pre-training task for the BERT model consists of a Masked LM task at the text level and a Next Sentence Prediction task at the sentence level. The Masked LM task is set to mask 15% of

the text of the input sequence, where 80% of the text is replaced by MASK symbols, 10% is replaced by other textual symbols, and 10% of the text is not replaced. The model predicts the masked original text based on the context of the unmasked text in the sequence, and thus learns the correlation relationship between the texts. During the execution of the Next Sentence Prediction task, the set of sentence sequences as input does not completely retain the order of the utterances in the original corpus, but instead, it consists of a randomly selected 50% of the sentences together with 50% of the sentences retaining the original order, and by learning the contextual relationship between sentences and obtaining the contextual information of the utterances. The BERT model is able to achieve both text-level and sentence-level rich feature extraction in the news event extraction task studied in this paper, and then undergoes the multi-head attention mechanism and fine-tuning training to obtain the prediction vectors of the corresponding tag sequences for each text position. The pre-training process and the fine-tuning process are shown in Fig. 2.

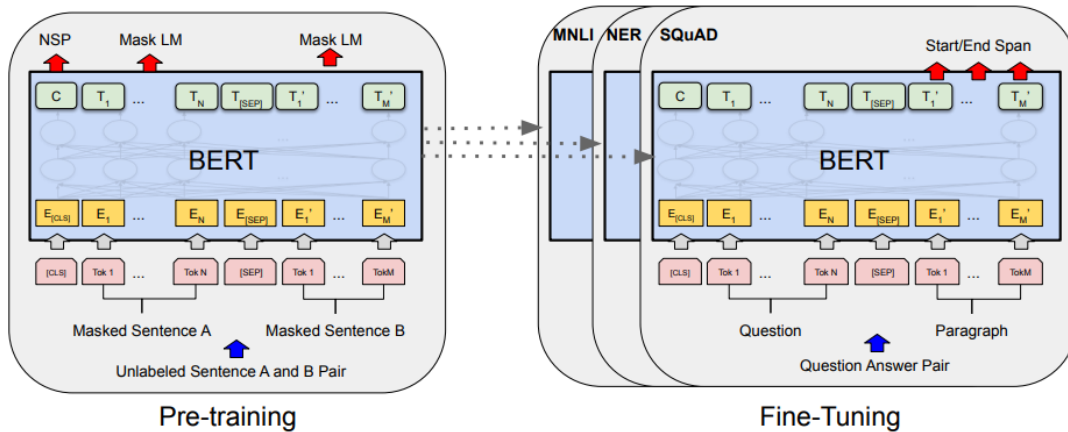


Figure 2. Pre-training and Fine-Tune process

Since the tag sequence prediction results output from the BERT model do not take into account the transformation rules between tags, the computed results are optimised using Conditional Random Field CRF after the BERT model. In the domains of segmentation, entity identification, and lexical annotation in natural language processing, CRF is a well-known sequence annotation technique. And in

these annotation scenarios, the effect is significantly improved. Transfer property and state property are the two main categories of properties in CRF. The relationship between the current state and the input sequence is known as the state characteristic, and the transfer characteristic reflects the connection between the final output and the present output state. as seen in Fig.3.

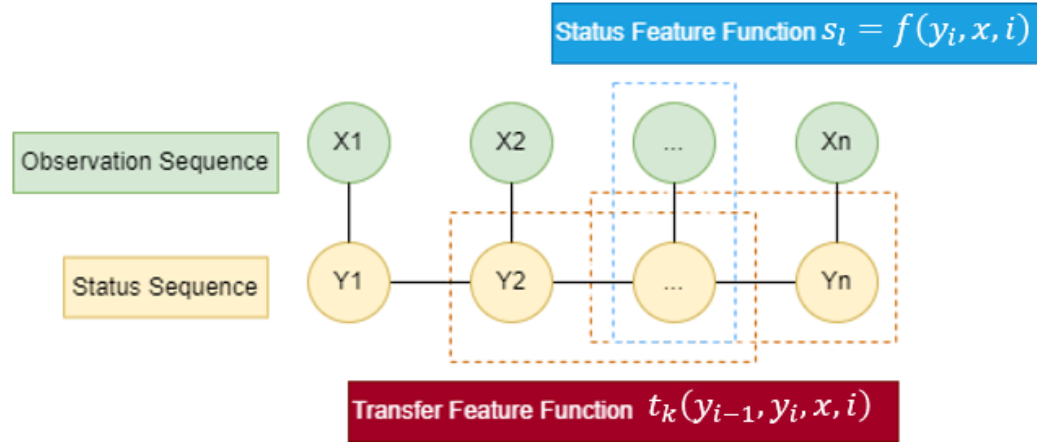


Figure 3. Graph structure of CRFs for linear chain conditional random fields

Conditional Random Field (CRF) can fully consider the dependencies and constraints between neighbouring characters. Therefore, we adopt Conditional Random Field (CRF) in the last layer of the model to constrain the feature information output from the multi-attention layer to ensure the accuracy of the relationship between the labels obtained in the end. When given an input sequence $X = \{x_1, x_2, x_3, \dots, x_n\}$, assume that the output sequence is $y = \{y_1, y_2, y_3, \dots, y_n\}$. Then the score of the output sequence can be expressed by the following equation:

$$s(X, y) = \sum_{i=1}^n (W_{y_i, y_{i+1}} + P_{i+1, y_{i+1}}) \quad (2)$$

W is the transfer matrix, $W_{y_i, y_{i+1}}$ is the number of scores of labels transferred from y_i to y_{i+1} , and $P_{i+1, y_{i+1}}$ is the number of scores of labels y_{i+1} corresponding to the $i+1$ st word of the input sequence. The probability of the output sequence y is calculated and the sequence of labels when the conditional probability is maximum will be output as the result sequence. Where Y_x denotes the entire tag sequence of the input sequence. The formula can be expressed as:

$$P(y|X) = \frac{\exp(s(X, y))}{\sum_{\tilde{y} \in Y_x} \exp(s(X, \tilde{y}))} \quad (3)$$

C. Extracting events based on the Bert model

Two stages make up Bert's overall framework: pre-training and fine-tuning. This research focuses on the fine-tuning step, where the Bert model is first parameterized by the pre-training model and then all parameters are learned on labeled data [8]. The model in the pre-training stage is trained on unlabelled data. The CRF layer labels the sequences after the news text input is delivered into the Bert model for processing and training.

The specific process is to first read data from the DUEE dataset, build a thesaurus, feed the data into the Bert model for processing and training, and annotate the sequences with a CRF layer. The CRF layer makes each point annotated as a whole rather than individually, and the annotations of each point have a certain degree of relevance. In this way, the model understands the text in addition to the rules in the output sequence. Eventually event extraction is complete, predictions are presented, and the model is evaluated and saved. Data is processed using the trained Bert model; the captured news data is fed into the Bert model and processed to obtain the extracted event types. The structure of the extracted events is shown in Fig. 4; the news data is processed to extract the event type, thesis role, specific attribute value and specific category.

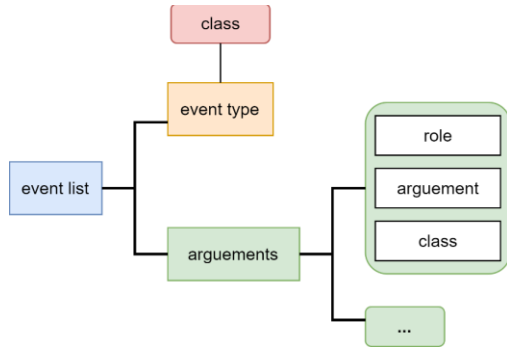


Figure 4. Extracted event output structure, including event types and argument roles

IV. EXPERIMENTAL ANALYSIS

In this paper, a web crawler based on the Scrapy framework is used to obtain news data from 11 categories in the NetEase news network, and initially, two news items are crawled as the basic processing data. The DUEE.1.0 dataset is selected to fine-tune the Bert model for training. And the crawled news data was input into the Bert model for extraction to realize the task of extracting real-time network news events

A. Event Information Extraction Experiment

In this paper, we first used a crawler program based on a scrapy framework applied to NetEaseNews.com to get 20000 data and store it in a CSV file.

The Bert model is trained using the DUEE.1.0 dataset, which consists of 17,000 words with event messages and 65 distinct events. This dataset has 20% designated as the validation set and 80% designated as the training set. The Keras framework and Tensorflow are used in the experimental environment. Using Adam as the optimizer, the model's learning rate parameter is set to e-5. The BERT layer uses the trained Chinese BERT model from Google, which uses a 12-layer Transformer encoder. The hidden layer has a dimensionality of 768 and a multi-head attention mechanism with 12 heads. During the training phase, the batch_size is set to 32 and the epochs are set to 10.

In this experiment, the precision, recall and F1 value of the machine learning method are used as the model measures for this experiment. The numbers TP, FP, and FN represent the number of positive samples and positive predictions, negative

samples and positive predictions, and positive samples and negative predictions, respectively. The following is the formula:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (6)$$

In this study, in addition to using the BERT model, the traditional neural network models LSTM and BiLSTM were used for experimental comparison to prove the effectiveness of BERT. The line graph Fig.5 shows that the P, R, and F1 values of the BERT model are significantly higher than those of the other two models, and for the BiLSTM model, the overall level is lower because the effect of context and semantic environment on the classification of words and word labels is not taken into account, and the predicate labels of each element are independent of the other elements and not dependent on them.

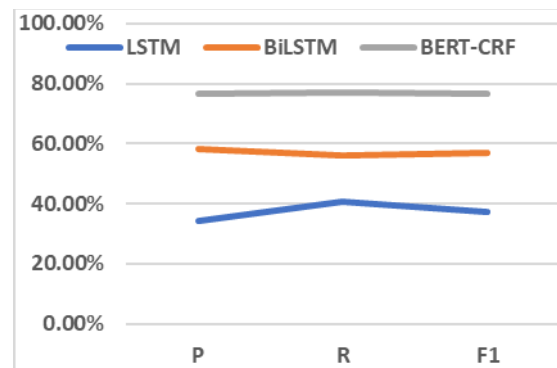


Figure 5. Comparison of P-value, R-value and F1-value of LSTM, BiLSTM and BERT-CRF models. P for Precision, R for Recall

In order to further optimize the model, the BERT pre-training model is replaced with ALBERT and RoBERTa for experiments, respectively. The RoBERTa model is an improved version of the Bert model, with more unlabeled data, longer training time, and larger batch sizes, which enhances the model's learning ability and generalization capability. Meanwhile, improvements are made in the training method by removing the next sentence

prediction task to support longer word sequences; using dynamic masks to avoid repeated training of data; and using byte-level vocabulary to train the model to support processing of many common words. The ALBERT model is a lightweight model based on the BERT model, with a substantially lower number of parameters compared to the traditional BERT, and with a relative increase in the operation speed. By reducing the number of parameters and enhancing the resilience of the neural network parameters, ALBERT's technique of embedded layer factorization and cross-layer parameter sharing speeds up model training while compressing the overall number of parameters. The SOP (Sentence-Order Prediction) task, which focuses on inter-sentence order prediction independently of subject aspect, takes the role of the NSP task in ALBERT. When compared to NSP, SOP can achieve an approximate 2% increase in accuracy for the downstream job that requires numerous sentence inputs. The ALBERT model could increase semantic comprehension, speed up training, and have fewer parameters.

The BERT pre-trained model is replaced with ALBERT and RoBERTa model respectively on both the DUEE dataset taken for this experiment respectively and the data are shown in Fig. 6.

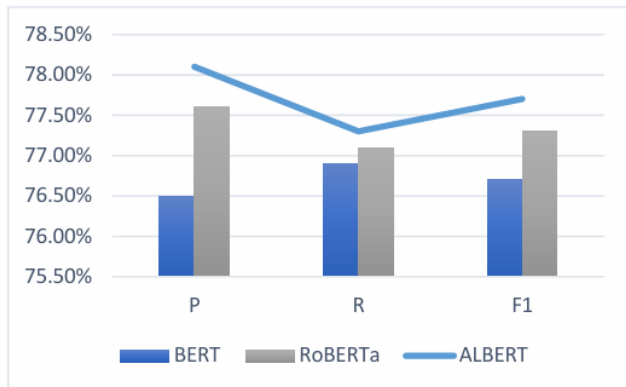


Figure 6. Comparison of P-value, R-value and F1-value of BERT, RoBERTa and ALBERT models Comparison of P, R and F1 values. P for Precision, R for Recall

B. Conclusion Of The Experiment

The use of a bi-directional architecture, which enables the model to better comprehend the context and meaning of the text, is one of the BERT model's significant advances. The core component of the BERT paradigm is the transformer encoder.

It is composed of several layers, one feed-forward neural network and one multi-headed self-attention mechanism. While the feed-forward neural network analyzes the weighted input to build a contextual representation of the tokens, the self-attention mechanism enables the model to assess each token's relevance in the input text based on its relationship to other tokens in the sentence. Therefore, BERT is effective for event extraction. The specific experimental results are shown in Table I.

TABLE I. Experimental Results I

Module	P	R	F1
LSTM	34.2%	40.6%	37.1%
BiLSTM	58.1%	56.2%	57.1%
BERT-CRF	76.5%	76.9%	76.7%

Both ALBERT and RoBERTa models are based on BERT with different improvements, and the experimental results also show that both ALBERT and RoBERTa are better than the BERT model, and the specific experimental data are shown in Table II.

TABLE II. Experimental Results II

Module	P	R	F1
BERT	76.50%	76.90%	76.70%
RoBERTa	77.60%	77.10%	77.30%
ALBERT	78.10%	77.30%	77.70%

Input the organized news data into the three trained models for processing, and select the news body for event extraction. The news body is obtained from the organized news csv file for extraction, and the output results are then stored in the csv file to complete the extraction of news events. From the actual extraction results, there is not much difference between the three models, and ALBERT is slightly better than the other two models.

V. CONCLUSIONS

To address the issue of slow real-time news on websites, this research suggests an event extraction technique for real-time news. Crawler technology is utilized to crawl the news data, and after literature research for event extraction model determination, the BERT model is finally selected,

and the public news dataset DUEE is utilized for training to extract event attributes and types, etc. When comparing BERT to the more established neural network models LSTM and BiLSTM, there is a clearer advantage. In this paper, the BERT model is also replaced with RoBERTa and ALBERT model, the model is further optimized, and the experiments show that the ALBERT model improves the F1 value of the original BERT model by 1%. The trained model is finally applied to the real-time crawled NetEase news network data to realize the real-time news event extraction.

The research on real-time news event extraction in this paper extracts complex unstructured news data and transforms it into structured content, which has important research significance and use value for its downstream tasks such as knowledge graph creation, event association analysis, etc., and has research and exploration significance.

REFERENCES

- [1] Yu X, Rong W, Liu J, et al., Lstm-based end-to-end framework for biomedical event extraction [J]. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019, 17(6): 2029–2039.
- [2] Yang H Chen Y., Liu K., et al. Multi-Turn and Multi-Granularity Reader for Document-Level Event Extraction [J]. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2022, 22(2):1–16.
- [3] Xu R, Liu T, Li L, et al. [Document-level event extraction via heterogeneous graph-based interaction model with a tracker [J]. *arXiv preprint arXiv:2105.14924*, 2021.
- [4] Devlin J, Chang M W, Lee K, et al. Bert:Pre-training of deep bidirectional transformers for language understanding [J]. *ArXiv Preprint ArXiv:1810.04805*, 2018.
- [5] LI Xiangyang, ZHANG Huan, ZHOU Xiaohua. Chinese clinical named entity recognition with variant neural structures based on BERT methods [J]. *Journal of Biomedical Informatics*, 2020(107):103422.
- [6] LI Ni, GUAN Huanmei, YANG Piao, et al. BERT-IDCNN-CRF for named entity recognition in Chinese [J]. *Journal of Shandong University (Natural Science)*, 2020, 55(01):102-109.
- [7] YUAN Jian, ZHANG Haibo. Chinese Entity Recognition Model of Multi-granularity Fusion Embedded [J]. *Journal of Chinese Computer Systems*, 2022, 43(4):741-746.
- [8] YANG Zhenyu, ZHANG Denghui. A complex long sentence intent classification method combining BERT and two-layer LSTM [J]. *Computer Applications and Software*, 2021, 38(12):207-212.

Hippocampal Cognitive Function Based on Deep Learning

Bijun Zhang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 1278004587@qq.com

Hongge Yao

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 835092445@qq.com

Abstract—This research focuses on the study of agent behavior decision-making based on hippocampal cognitive functions, aiming to enhance the decision-making capabilities of agents in complex task environments by deeply exploring the crucial role of the hippocampus in learning, memory, and cognitive processes. By drawing inspiration from the biological structure and functional characteristics of the hippocampus, researchers are dedicated to designing and developing more intelligent and adaptive decision-making models to enhance agents' behavioral performance, problem-solving abilities, and adaptability to new situations. To achieve this goal, the research integrates advanced artificial intelligence technologies such as reinforcement learning and deep learning to simulate the complex functions of the hippocampus in memory encoding, storage, retrieval, and cognitive reasoning. This research not only contributes to advancing intelligent systems towards higher levels of intelligence and personalization but also plays a significant role in improving the interaction between intelligent agents and humans, providing intelligent services that better meet user needs. We found that the neural network trained in multi-task learning benefits from a loss term that promotes relevant and irrelevant representations. Therefore, the complementary coding we found in CA3 can provide extensive computational advantages for solving complex tasks. Furthermore, the study emphasizes the importance of further elucidating the functional mechanisms of the hippocampus, with the expectation of providing a more solid theoretical foundation for the optimization and refinement of agent decision-making models in the future.

Keywords—Reinforcement Learning; Hippocampus; Memory Encoding

I. INTRODUCTION

The human brain is a general intelligence system consisting of hundreds of billions of neurons and millions of trillions of synaptic connections, endowed with the abilities of

perception, learning, reasoning, and decision making. Cognition refers to the brain's perception, understanding, and memory of external stimuli, while decision-making involves the selection of actions based on cognitive information. Cognitive decision-making is the process of choosing the best course of action through thinking, analyzing, and evaluating information. It engages our capabilities of thought, perception, memory, and reasoning.

When making decisions, this paper may be influenced by cognitive biases, leading to irrational choices. In recent years, the intersection of cognitive neuroscience and artificial intelligence has become increasingly close, particularly in applying profound insights from neurobiology to decision-making systems in intelligent agents, where significant progress has been made. This trend is deeply inspired by the efficient decision-making abilities exhibited by humans and other advanced organisms in complex environments. These organisms can quickly make complex inferences from limited information and flexibly integrate new knowledge to optimize their behavior, a capability crucial for building more intelligent and adaptive intelligent agents.

The hippocampus, as the core region of the brain responsible for memory formation, storage, and retrieval, has unique cognitive functions that have become a key source of inspiration for designing decision-making models in intelligent agents. Researchers are dedicated to unraveling the complex structure and functions of the hippocampus, especially how it interacts with other brain regions (such as the Para hippocampal gyrus, parietal lobe, frontal lobe, and cerebral cortex) to support advanced cognitive tasks. By

simulating these intricate characteristics of the hippocampus, researchers aspire to develop advanced intelligent agent models that possess the capability to make precise and adaptable decisions within highly complex and ever-changing environments, mirroring the decision-making process exhibited by humans in their natural surroundings.

II. RELATED WORKS

A. Function and Morphology of the Hippocampus

Unlike the neocortex, the hippocampus and its adjacent dentate gyrus belong to the archicortex, featuring a three-layered cellular structure consisting of the molecular layer, the pyramidal cell layer, and the polymorphous cell layer. Based on its organizational characteristics, the hippocampus can be further divided into four regions: CA1, CA2, CA3, and CA4. CA1 and CA2 are located on the dorsal side of the hippocampus, while CA3 and CA4 are situated on the ventral side. The hippocampus, together with its nearby dentate gyrus, subiculum, parahippocampal gyrus, and cingulate gyrus, forms a structural and functional unity known as the hippocampal formation. The hippocampal formation has direct fiber connections with the septal area, entorhinal cortex, and the mamillary bodies of the hypothalamus through the fornix, fimbria of the hippocampus, and perforant path. The dentate gyrus of the hippocampal formation directly receives neural information from the amygdala, other limbic cortices, and the neocortex via the perforant path emanating from the entorhinal cortex. After receiving neural information from these brain structures, the dentate gyrus sends fibers to CA3 and CA4, from which axonal collaterals (Schaffer collateral fibers) of CA3 and CA4 neurons terminate in CA1 and CA2 of the hippocampus. Although the fornix primarily consists of efferent fibers from the hippocampal formation, it also contains cholinergic afferent fibers from the medial septal nucleus as well as serotonergic and noradrenergic fibers originating from the brainstem. The main efferent fibers of the hippocampal formation originate from the CA1 and CA2 regions, reaching the mamillary bodies of the hypothalamus, the anterior thalamic nuclei, and the lateral septal

nucleus via the fornix. The efferent fibers from the CA1 and CA2 regions also terminate in the subiculum. Among these connections in the hippocampal formation, the majority of synapses use amino acid substances as neurotransmitters, primarily glutamate and GABA. Two noteworthy circuits are the classic Papaz's circuit and the trisynaptic circuit.

B. The trisynaptic memory circuit of the hippocampus

It was first reported by Lomo in 1966, who described a phenomenon he termed long-term potentiation (LTP) occurring in the trisynaptic circuit of the hippocampus. This discovery subsequently gained widespread attention due to its relevance to the brain mechanisms of memory. The trisynaptic circuit initiates within the entorhinal cortex, where neuronal axons coalesce to create the perforant pathway, ultimately terminating on the dendrites of granule cells located in the dentate gyrus. This constitutes the first synaptic link. Subsequently, the axons emanating from these granule cells in the dentate gyrus transform into mossy fibers, which establish synaptic connections with the dendrites of pyramidal cells residing in the CA3 area of the hippocampus, thereby forging the second synaptic junction.

The axons of the pyramidal cells in the CA3 region send collaterals to make the third synaptic connection with pyramidal cells in the CA1 region. From there, pyramidal cells in the CA1 region project back to the medial entorhinal cortex. This trisynaptic circuit, connecting the dentate gyrus, entorhinal cortex, and hippocampus, possesses unique functional properties and was initially considered evidence supporting the mechanism of long-term memory.

C. Small loop of hippocampal CA3

The hippocampus serves as the cornerstone of our ability to form episodic memories, enabling us to narrate personal experiences from our daily lives. The sensory information pertinent to memory storage travels through the entorhinal cortex (EC), functioning as the primary gateway or initiating point, serves as the cornerstone of the trisynaptic circuit. conduit between the

hippocampus and the neocortex. The anatomy and physiology of the hippocampus intertwine with fundamental attractor network theory, which encompasses two key findings: Tsodyks and Feigel's discovery the capacity to store sparse, uncorrelated patterns in abundance is complemented by Fontanari's insight, which suggests that dense, correlated patterns can coalesce into representations embodying shared characteristics. Both these forms of representation can harmoniously coexist and be retrieved within the same neural network, contingent upon a certain threshold being met. serving as the selector between them.

Neurons in layer II of the EC project to the CA3 region via two distinct pathways, as depicted in Figure 3. One pathway directly synapses with the distal dendrites of CA3 pyramidal cells via the perforant path (PP). The alternative route sees the PP axons branching off to the dentate gyrus (DG) before reaching CA3, where they form synapses with granule cells [1]. These granule cells, in turn, extend mossy fibers (MF) that establish synaptic connections with the closer, proximal dendrites of the pyramidal cells located in the CA3 region. Regarding the identical sensory data, two distinct representations emerge, each endowed with One is sparse and decorrelated, achieved through the mediation of mossy fibers (MFs), while the other is dense and correlated, facilitated by the perforant path (PP).

III. ALGORITHMS MODEL

A. Description of the problem

The CA3 subregion of the hippocampus is recognized the hippocampus operates as an autoassociative network, encoding experiences into enduring memories. The raw data pertaining to these experiences stems both directly from the entorhinal cortex and indirectly, via the dentate gyrus which acts as a filter, performing sparsification and decorrelation. The computational goals pursued by these dual input routes can be rephrased as enhancing the efficiency and accuracy of memory encoding. have yet to be conclusively determined. Here, this project conceptualizes CA3 as a Hopfield-analogous network, proficient in accommodating

both dense, correlated encodings and sparse, uncorrelated ones. As the number of memories accumulates, the dense encodings tend to coalesce around common features, while the sparse encodings maintain their individuality.

This project emulates the transformation of memory representations as they traverse the two pathways from the EC to CA3, and explore how these transformed encodings are subsequently stored and retrieved within CA3. Additionally, the hippocampus plays a pivotal role in recognizing similarities and patterns across disparate experiences, thereby enhancing cognitive processes. By modeling the hippocampal network, this work initially hypothesizes that MF (mossy fiber) encodings and PP (perforant path) encodings in CA3 can preserve distinctions between memories while enabling generalization among them [2]. Our goal is to delve into whether an auto associative network possesses the capability to preserve and recall memory encodings originating from both pathways, this paper conduct our investigation., enabling information representation at different scales that allows the network to both differentiate between instances and generalize across them. Through training an artificial neural network, this work ultimately demonstrates that these encoding types are suited to performing complementary tasks of instance recognition and concept classification. This approach enables a more intricate and nuanced comprehension. Hippocampus processes and integrates information, ultimately contributing to a deeper understanding of its functional role in memory storage and retrieval. contributes to memory formation and retrieval.

B. CA3-inspired Complementary Coding Model

Transition of Memory along the Hippocampal Pathway, from Image to Binary Autoencoder in EC, FNN Network from EC to CA3, visualizing Pathways from CA3 back to EC. The Hopfield-like Model in CA3.

The Hopfield neural network functions similarly to a memory storage device. When multiple sequences or images are input into this network, it stores this information in the form of

connection weights between neurons. Upon re-inputting the same or partially corrupted original sequence/image, the network is capable of restoring (recovering) the sequence/image. Figure 1 is a network model.

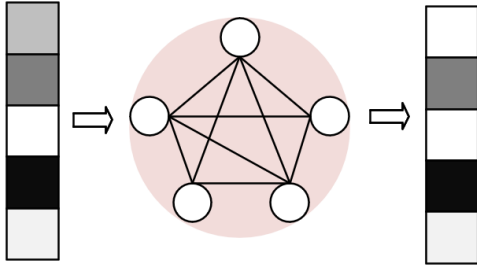


Figure 1. CA3 Hopfield-like model

The Hopfield-like network stores both sparse, uncorrelated encodings and dense, correlated encodings [3]. As more memories are stored, the former tend to remain distinct, while the latter merge along shared features. During its dynamic evolution, the Hopfield network converges towards stable states, which are the attractors of the network [4]. The design of network weigh and the initial state determine which attractor the network ultimately converges to. At Figure 2. In an auto-associative network, the attractor basins of the former tend to remain independent, while those of the latter tend to merge.

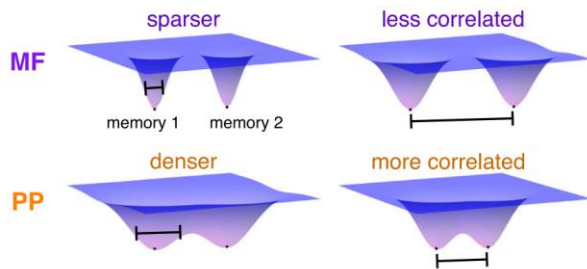


Figure 2. Basin of Attraction

Pattern storage: The Hopfield-like model of CA3 serves as a mechanism for pattern storage, where it retains the information by encoding the linear combination of patterns originating from the medial entorhinal cortex (MF) and the perforant path (PP).

$$q_{\mu vi} = (1 - \zeta) \cdot (x_{\mu vi}^{MF} - a_{MF}) + \zeta \cdot (x_{\mu vi}^{PP} - a_{PP}) \quad (1)$$

The PP pattern $\zeta = 0.1$ stands out in terms of its comparative intensity. A defining aspect of Hopfield networks with binary neural states. 0 and 1 is the subtraction of a density value from each pattern. The PP inputs have a notably weaker intensity compared to others, stemming from their more remote positioning (PP distal synapses) and the fact that they are empirically weaker than MF synapses (which are located on proximal dendrites).

These inputs undergo linear summation and are Architecture that is defined by its interconnectivity pattern, with i and j serving as indices for post-synaptic and pre-synaptic neurons, respectively.

$$W_{ij} \sim \sum_{\mu\nu} (0.9x_{\mu\nu i}^{MF} + 0.1x_{\mu\nu i}^{PP}) (0.9x_{\mu\nu j}^{MF} + 0.1x_{\mu\nu j}^{PP}) \quad (2)$$

The process of pattern retrieval involves generating a cue by randomly altering the activation state of 0.01 of the neurons in the target pattern, a quantity that is termed cue inaccuracy [5]. Throughout the retrieval phase, neurons undergo asynchronous updates in iterative cycles, where each neuron is updated once per cycle in a random sequence. At any specific instant in time, denoted as t , the cumulative synaptic input represents the aggregated electrical signals received by a neuron from its presynaptic counterparts. stored within a Hopfield-like network.

$$g_i(t) = \sum_j W_{ij} S_j(t) + h_i(t) \quad (3)$$

C. Model Loss function

The conversion of memory through the hippocampal pathway involves an image being encoded into a binary autoencoder form within the EC. In Figure 3, this project constructed and trained a comprehensive fully connected linear autoencoder architecture, comprising three strategically sized hidden layers (128, 1024, 128), each tailored to facilitate efficient information encoding and decoding.

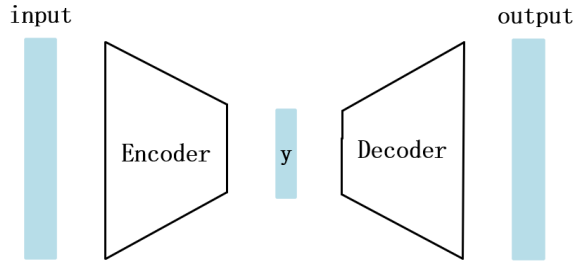


Figure 3. AE Network Architecture

To ensure swift and stable learning dynamics, this work incorporated batch normalization into each layer. Mitigating internal covariate shift and accelerating the training process. For nonlinearity, the ReLU function was applied to the first and third hidden layers, while the Sigmoid function was reserved for the output layer. Critically, the activations within the intermediate hidden layer underwent binarization through the Heaviside step function, with gradient flow maintained during backpropagation via the straight-through estimator. The overall optimization was guided by a specified loss function.

$$\mathcal{L} = \sum_{batch} \sum_{\mu\nu} \|i_{\mu\nu} - \hat{i}_{\mu\nu}\|^2 + \lambda \sum_{batch} KL\left(\frac{1}{N_{EG}} \sum_i x_{\mu\nu i}^{EC} \|a_{EC}\right) \quad (4)$$

I represent the original images, comprising pixel intensities spanning a spectrum from 0 to 1, and its reconstructed counterpart. x denotes the binary activations within the intermediate hidden layer, facilitating a sparse representation. And with an expected density of $a=0.1$, this paper employ sparsification with an intensity that evaluates the Kullback-Leibler (KL) divergence—a metric comparing the density of the hidden layer activations to the desired target density. Through this process, this project obtains the desired λ value of 10, which effectively achieves the expected sparsity level.

The central characteristic of the CA3 model lies in its activity threshold, which dictates whether the network retrieves example-based encodings or concept-based encodings. The project postulates the theta oscillations in the CA3 region, as a fundamental neural rhythm, not only embody pivotal threshold but also dynamically. Orchestrate fine-tuning of memory retrieval, enabling the brain

to access and retrieve information from a broader or narrower range of memories, depending on the specific context and cognitive demands. This process is intricately intertwined with synaptic plasticity and network connectivity, facilitating the adaptive adjustment of memory representations to better serve the organism's current needs and goals. The work introduces a plug-and-play loss function. That endows artificial neural networks with the comprehensive ability to represent both complex and diverse data patterns. pattern-separated (PP) and pattern-completed (MF) classes. Compared to networks with solely rely on a single representation type, these networks, by virtue of their ability to integrate diverse information through multiple representation types exhibit superior performance in multitask learning.

The DeCorr loss function addresses the issues of oversmoothing and excessive feature correlation by reducing the correlation between features. Consequently, the paper applies the DeCorr loss function to decorrelate encodings in the final hidden layer, mimicking the MF (sparse and decorrelated) mode observed in CA3. The exclusion of the encoding loss function ensures that the encoded representations preserve the intrinsic image correlations and patterns. DeCorr simulates the MF pathway by considering the baseline condition where there is no loss function applied to hidden layer activations, thus preserving the natural correlations between similar images and mimicking the PP pathway.

$$\mathcal{L}_{DeCorr} \approx \frac{1}{2} \sum_{\alpha, \beta \in batch} Pearson(s_{\alpha}, s_{\beta})^2 \quad (5)$$

It has been observed that different encoding properties are suited for distinct tasks. The baseline network excels in conceptual learning, whereas the DeCorr network typically performs better in exemplar learning but struggles with conceptual learning. To address this, the project applies the HalfCorr loss function, which decorrelates encodings only in the latter half of the final hidden layer. The introduction of the HalfCorr loss function diversifies the hidden layer representations, incorporating both correlated and uncorrelated components. As a result, HalfCorr networks are better equipped to learn tasks that

involve distinguishing between similar inputs and generalization.

HalfCorr networks demonstrate high performance in both tasks. Drawing parallels to the CA3 model, the paper find that exemplars are the decorrelated MF pathway is biased towards. Encoding information in a way that enhances discrimination and minimizes overlap among different elements. while concepts are preferentially encoded correlated PP pathway.

$$\mathcal{L}_{HalfCorr} \approx \frac{1}{2} \sum_{\alpha, \beta \in batch} Pearson(s_{\alpha}^{half}, s_{\beta}^{half})^2 \quad (6)$$

s_{α}^{half} representing the latter half of the neurons in the final hidden layer. The DeCorr network excels in exemplar learning but suffers from inferior performance in conceptual learning, a trade-off that does not affect the HalfCorr network. The HalfCorr network displays high performance in both tasks, demonstrating an ability to prioritize the use of each type of encoding for tasks it is better suited for. The work comprehensively quantifies the impact of individual neurons on various tasks by precisely measuring the decrement in task accuracy that ensues upon their silencing. This approach offers a nuanced understanding of how each neuron contributes to the overall performance. Furthermore, DeCorr, an innovative technique, empowers us to delicately modulate the encoding correlations within artificial neural networks, thereby amplifying the salience of input features that are crucial for accurate predictions.

By strategically aligning the computational requirements of diverse tasks with the optimal encoding scales tailored for each, DeCorr facilitates a more efficient resolution of these tasks. This alignment ensures that the network's resources are allocated effectively, enhancing both speed and accuracy. Notably, correlated neurons within these networks exhibit a pronounced influence on conceptual learning, facilitating the extraction of abstract representations that generalize across examples. Conversely, decorrelated neurons play a pivotal role in exemplar learning, capturing specific details that distinguish individual instances within a category.

By leveraging the complementary strengths of correlated and decorrelated neurons, DeCorr promotes a balanced and flexible learning strategy that is well-suited to tackle a wide range of complex tasks. This approach not only advances our theoretical understanding of neural network behavior but also has practical implications for designing more efficient and robust machine learning systems. The work proposes a distinct paradigm where loss functions are applied to distinct neurons

Fostering the principle of heterogeneity within a layer can be enriched by tailoring the degree of decorrelation for individual components or clusters within the HalfCorr network [6].

IV. EXPERIMENTS

A. Experimental Environment

In the experiments, the performance of the cognitive algorithm inspired by hippocampal memory designed in this paper is evaluated, and its properties are analyzed [7].

Firstly, the sample efficiency project undertakes a comparative assessment to gauge how the novel algorithm fares against previous conventional neural networks. All experiments are implemented on an RTX3060 GPU with 16GB of VRAM and a CPU running at 14.4 GHz, utilizing Pytorch and NVIDIA CUDA.

B. Dataset

In our model utilizing the Fashion-MNIST dataset, the sensory input encoded as memory consists of Fashion-MNIST images. The memory comprises 256 images from each of the categories of sneakers, trousers, and coats. These memories, which are Fashion-MNIST images, serve as exemplars representing individual concepts.

C. Train the network

To evaluate the performance of a cognitive algorithm inspired by hippocampal memory, the project compared it with traditional classification and recognition algorithms using the MNIST dataset. The paper normalized images, randomly assigned set numbers, and trained a multi-layer perceptron to either classify digits or identify sets. The network was trained on a subset of images

and evaluated on a test set for digit classification and on corrupted images from the training set for set identification. Classification requires clustering images based on common features, akin to concept learning in our CA3 model, while distinguishing differences among similar images necessitates example learning, similar to our CA3 model [8]. The project used stochastic gradient descent with a batch size of 50 and a learning rate of $1e-4$.

Comparative Experiment:

D. Some Common Mistakes

- **Traditional Classification Algorithms:** The work utilizes the same dataset to train various traditional classification algorithms, including but not limited to Support Vector Machines, Decision Trees, and Random Forests, and subsequently evaluate and compare their performance on the test set.
- **Hippocampal Memory Mechanism Simulation:** In addition to directly training hippocampal-inspired cognitive model for classification, one can also attempt to introduce mechanisms into the model that simulate characteristics of the hippocampus, such as employing specific loss functions or regularization terms to encourage the network to learn sparse, uncorrelated representations (analogous to mossy fiber (MF) coding) or dense, correlated representations (analogous to perforant path (PP) coding) [9].

Dots indicate means, bars show SD of networks. The DeCov loss, developed to reduce overfitting, aids numerical convergence. DeCorr decorrelates input pairs for all neurons in a layer, while DeCov decorrelates neuron pairs across all inputs. At Figure 4 and Figure 5, as a generalization-boosting regularizer, DeCov enhances digit classification but not significantly set identification, contrasting DeCorr's effect. DeCorr impairs concept learning but boosts instance learning. The paper trained an MLP for concurrent digit classification & set identification. Compared to baselines, DeCorr networks often excel in instance learning but underperform in concept learning. The project train until >99.9% accuracy on the training set.

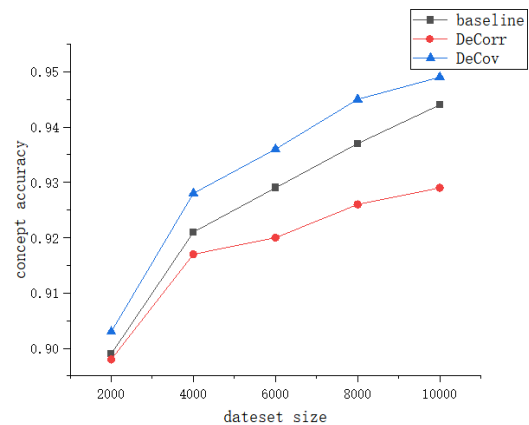


Figure 4. Comparison Chart of DeCov under MF Modes

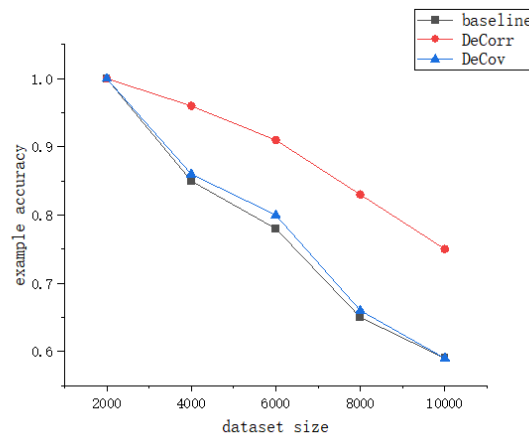


Figure 5. Comparison Chart of DeCov under PP Modes

Using the decrease in task accuracy after neuron silencing as an indicator of its impact, the work found that correlated neurons have a greater influence on concept learning, while decorrelated neurons impact instance learning more significantly. The average drop in accuracy across each neuron in the network reveals their respective contributions to both learning modalities. For all results, p-values were calculated using an unpaired, two-tailed t-test.

As can be seen from Figure 6 and Figure 7, Correlated neurons (orange bars) exhibit a stronger influence on concept learning can reach 0.00077613, whereas decorrelated neurons (purple bars) have a more pronounced effect on instance learning) can reach 0.0037056 [10]. For all results,

p-values were calculated using an unpaired, two-tailed t-test.

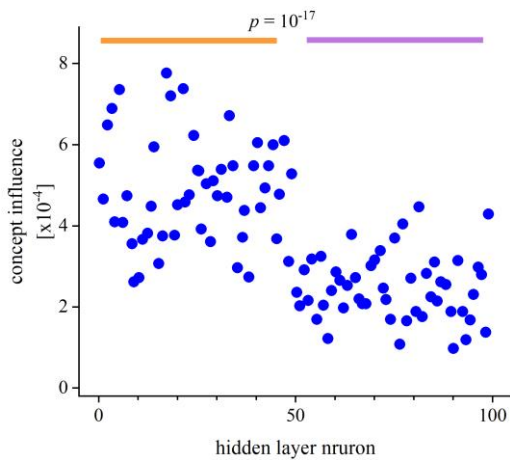


Figure 6. The impact of neurons on concept learning

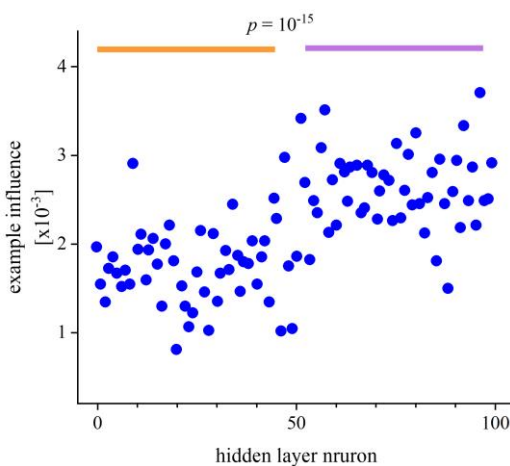


Figure 7. The impact of neurons on exemplar learning

V. CONCLUSIONS

In this paper, these novel agent models will not merely enhance performance in specific tasks but also propel intelligent systems towards greater intelligence, adaptability, and personalization. They will excel at comprehending and adapting to ever-evolving environments, offering more tailored and user-centric intelligent services. Furthermore, this interdisciplinary integration will pave new avenues for improving the interaction

between intelligent systems and humans, fostering a harmonious coexistence between man and machine.

However, it is crucial to acknowledge that despite remarkable advancements, the precise functions of the hippocampus and its intricate relationship with overall cognitive processes remain a complex and incompletely unraveled domain. As such, future research endeavors will continue to delve deeper into the working mechanisms of the hippocampus, aiming to refine and optimize agent decision-making models, thereby propelling artificial intelligence technology to even greater heights.

REFERENCES

- [1] Borzello, M. Assessments of dentate gyrus function: discoveries and debates. *Nat. Rev. Neurosci.* 24,502–517(2023).
- [2] Asutay, E. Affective calculus: the construction of affect through information integration over time. *Emotion* 21,159–174 (2019).
- [3] Herweg, N. A., Solomon, E. A. & Kahana, M. J. Theta oscillations in human memory. *Trends in Cognitive Sciences* 24,208–227 (2020).
- [4] L. Kang and T. Toyozumi. Hopfield-like network with complementary encodings of memories. *Phys. Rev. E*, 108(5):054410, 2023.
- [5] Zheng, J. Multiplexing of theta and alpha rhythms in the amygdala-hippocampal circuit supports pattern separation of emotional information. *Neuron* 102,887 – 898 (2019).
- [6] Barry, D. N. & Love, B. C. A neural network account of memory replay and knowledge consolidation. *Cereb. Cortex.* 33, 83–95(2022).
- [7] Xiao, H., Rasul, K., & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms arXiv 1708.07747 (2017).
- [8] Qasim, S. E., Fried, I. & Jacobs, J. Phaseprecession in the human hippocampus and entorhinal cortex. *Cell* 184,3242–3255 (2021).
- [9] Vertes, E., and Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. *Adv. Neural Inf. Process. Syst.* 32, 13714–13724.
- [10] Sun, C., Yang, W., Martin, J., and Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nat. Neurosci.* 23,651–663.

Design and Development of an Intelligent Laboratory Management System Based on STM Processors

Ruoyu Wang

School of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: 2544560286@qq.com

Jiaxuan Liu

College of Communication and Information
Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: liujiaxuan_tuan@163.com

Lulu Chen

School of Mechanical Engineering
Xihua University,
Chengdu, China
E-mail: chenll1232022@qq.com

Lei Tian

School of Electronic Engineering
Xi'an University of Posts and Telecommunications
Xi'an, China
E-mail: tla02@126.com

Abstract—In response to the escalating inventory of laboratory equipment, the inadequacy of traditional manual management practices has become increasingly apparent. Current management challenges are underscored by historical data, which reveals a borrowing rate of approximately 30%, a concerning return rate of only about 5%, and an equipment loss rate of 10%. To address these inefficiencies, this paper introduces an advanced, intelligent laboratory equipment management system leveraging STM32 and GM65 technologies, designed to significantly enhance the return rate and minimize equipment loss. The system is meticulously engineered, featuring three integral modules: the scanning module, the core board module, and the communication module. The scanning module employs GM65 technology to rapidly scan and decode QR codes, facilitating the swift transmission of critical data to the core board for intelligent processing and interaction with the laboratory equipment. At the heart of the system, the core board module serves as the central processing unit, adeptly managing data from the scanning module. It intelligently discerns various information types and dynamically updates the TFT-LCD screen with the current status of equipment, ensuring that all users have access to real-time and accurate information regarding the borrowing and returning of equipment. The communication module is pivotal in bolstering the system's connectivity capabilities. Designed with a robust structure, the system is characterized by its simplicity of operation and robust practicality. It not only significantly improves the efficiency of laboratory equipment

management but also plays a transformative role in modernizing laboratory management practices. By streamlining the management process and ensuring the optimal utilization of resources, this system provides a solid foundation for the sustainable advancement of laboratory operations.

Keywords-Component; STM32; Code Scanner; Communication Technology; Serial Port; Wireless Network

I. RESEARCH BACKGROUND

Currently, university research laboratories are grappling with equipment management challenges, such as equipment idleness, damage, and disarray, which hinder the smooth progress of experimental teaching and scientific research. Due to the limited number of faculty and management staff, traditional management methods are inefficient and struggle to keep up with timely updates and maintenance. Therefore, there is an urgent need for an intelligent laboratory equipment management system to enhance efficiency and ensure the effective utilization of resources.

II. SYSTEM OVERVIEW

This article mainly introduces a system for managing laboratory equipment based on the

STM32 and GM65 barcode scanner [1]. The system is composed of three main parts: the barcode scanning module, the core board module, and the wireless communication module. The GM65 barcode scanning module is primarily responsible for scanning and recognizing QR code information, decoding the scanned information, and then sending the decoded text data to the core board to interact with the laboratory equipment [2-3]. The core board module is responsible for receiving information from the GM65 barcode scanning module and responding differently to different information, displaying the equipment borrowing and returning status on the TFT-LCD screen [3]. The communication module is divided into two parts: the WIFI module and the serial communication module [4]. It can achieve wireless communication with the mobile terminal, sending data from the core board to the mobile terminal. The serial communication module can communicate with the PC terminal through CH340, achieving communication between the core board and the PC terminal [6-7]. The system is shown in Figure 1.

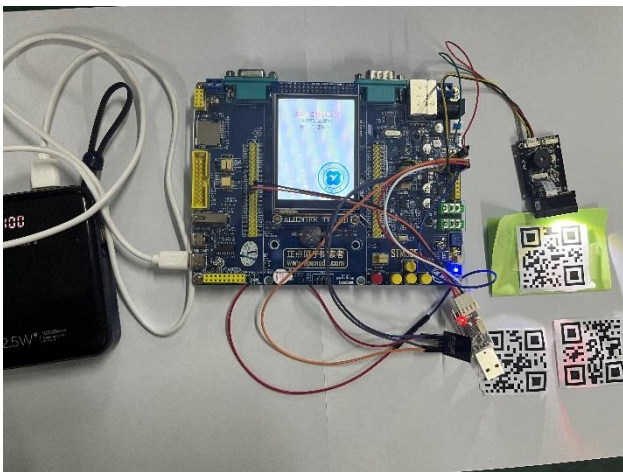


Figure 1. System Display Diagram

Microcontroller (SCM) is a form of implementation for embedded systems, integrating multiple functions such as CPU, memory, input/output interfaces, and clock circuits on a single chip to form a complete microcomputer system [8-9]. The GM65 barcode scanner is a high-performance, high-precision QR code scanning device that uses advanced image intelligent recognition technology and decoding algorithms to

quickly and accurately identify barcodes [10]. The ESP8266 is a low-cost, high-performance WIFI module launched by Espressio Systems, widely used in embedded system design and IoT development [11]. It can be used as a standalone microcontroller or simply as a WIFI module to simplify the design. The CH340 is a widely used USB to serial chip, mainly used for data communication between computers and serial devices, and this module is also used in this design.

By providing a detailed introduction to the system's hardware design and software development plan, as well as an analysis of the implementation principles of each module, readers can gain a comprehensive understanding of the overall architecture of the system and the way each module functions [12-13]. This lays an important theoretical foundation and guidance for subsequent system development and application.

III. SYSTEM HARDWARE INTRODUCTION

A. System Design Plan

The hardware part of the system is designed using a modular expansion approach from the inside out. Initially, the core module of the minimum system design was completed using the STM32F407 development board [13]. Subsequently, the design of the peripheral expansion modules around the minimum system was gradually completed in separate modules [14]. Finally, all modules were combined into one system to achieve the corresponding functions.

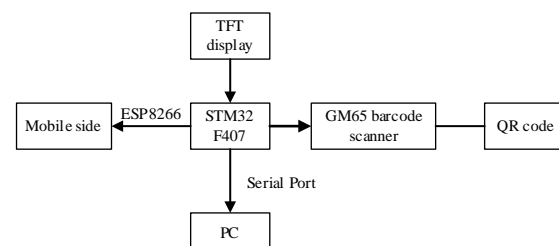


Figure 2. System Flowchart

In Figure 2, the overall workflow of the system is as follows: The QR code containing information about the laboratory equipment is read by the GM65 barcode scanner [15-16]. The information characteristics of the experimental instruments contained in the QR code are then read. The control program written under the Keil5 software controls

the serial port of the microcontroller, enabling serial communication between the GM65 barcode scanner and the microcontroller. After scanning the GM65 barcode, the information of the QR code is transmitted to the microcontroller through serial communication, realizing the reading of the laboratory experimental instrument information by the STM32, and displaying the information of the experimental instruments on the display screen. At the same time, information synchronization with the PC end is achieved through serial communication. In addition, wireless connection with the mobile terminal can be achieved through the ESP8266 wireless module, further realizing the management function of laboratory equipment.

B. Serial Communication

In Figure 3, there are a total of three related registers for STM32 serial communication: the USART_SR Status Register, the USART_DR Data Register, and the USART_BRR Baud Rate Register.

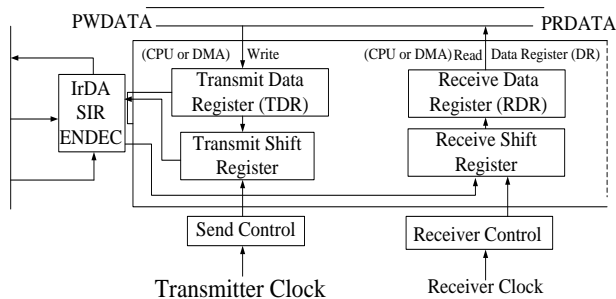


Figure 3. Serial Port Register Diagram

The process of serial port data transmission involves several key steps. First, when data needs to be sent, it is written into the Transmit Data Register (TDR). Before data transmission, the system checks whether the shift register is processing other data. If not, the newly written data is immediately transferred to the shift register for transmission. Once the data is moved to the shift register, the TXE (Transmit Data Register Empty) flag is set to 1, indicating that the data in the TDR register has been cleared and new data can be written for transmission. If the TXEIE bit in the USART_CR1 register is set to 1, an interrupt is triggered whenever the TXE flag is set. Data is shifted bit by bit to the right in the shift register and sent out through the TX pin. After the transmission

is completed, this process is repeated, and the TXE flag is checked to determine if new data can be sent.

The process of serial port data reception also has several key steps. Data is first received from the RX pin and then read and shifted bit by bit into the shift register under the control of the receiver. When enough bits are received to form a complete byte, this byte is moved as a whole to the Receive Data Register (RDR). During this process, the RXNE (Receive Data Register Not Empty) flag is set to 1, indicating that there is data in the RDR that can be read. Once the RXNE flag is set, it means the data is ready to be read out from the RDR register.

C. Barcode Scanner Principle Introduction

Barcode scanners recognize QR codes by scanning the black and white modules on the QR code with optical sensors, converting them into digital signals, and then using decoding algorithms to transform the digital signals into recognizable information.

The specific steps include: the optical sensor inside the barcode scanner emits light, scans the black and white modules on the QR code, and converts them into electrical signals, including one-dimensional codes (such as EAN-13, CODE-39, ITF-14, etc.) and two-dimensional codes (such as QR codes, Data Matrix codes, etc.). The barcode scanner converts the scanned black and white modules into digital signals for subsequent processing. The built-in decoding algorithms of the barcode scanner decode the digital signals, transforming them into recognizable information, such as website addresses, text, links, etc. The GM65 module outputs the decoded numbers or text to external devices, such as computers, cash registers, etc., in the set output format (such as USB, serial port, Bluetooth, etc.).

D. Wireless Module Design

The wireless module used in this design is the ESP8266, which is a very powerful WIFI module capable of communicating with a microcontroller via a serial port, thus enabling programming to control the ESP8266. With the ESP8266, you can access some APIs to obtain weather information or complete network time synchronization, and also connect to the cloud platform for development. The ESP8266 comes in various specifications such as

ESP-01/01S/07/07S/12E/12F/12S, and there is also the self-developed ATK-ESP8266 by Zhengdian Atom (with modified firmware and module pins). The one used in this project is the ESP-01S, but because it uses serial port transmission, the speed is relatively slow and it is not suitable for transferring large capacity data such as images or videos. The ESP8266 supports three working modes: STA, AP, and AP+STA. The AP mode is used in this project, where the ESP8266 acts as a hotspot, providing wireless access services and data access, enabling communication with the mobile end. The ESP8266 operates based on several principles. Initially, it undergoes power-on initialization, during which it loads its firmware and sets up WIFI connections. Following this, the module can link to a WIFI network using AT commands or through programming, enabling communication with other devices or the internet. Once connected, the ESP8266 is able to transmit data with other devices through the TCP/IP protocol stack, including both sending and receiving data. As a potent WIFI module, the ESP8266 is adept at communicating with and controlling other devices, making it suitable for diverse IoT applications.

TABLE I. BRIEF DESCRIPTION OF AT COMMAND FUNCTIONS

AT Command	Function
AT	Testing for normal startup
AT+CWMODE=2	Setting AP Mode
AT+RST	Restart
AT+CWSAP="esp", "12345678", 1, 4	Setting AP parameters with the account name "ESP8266" and password "123456"
AT+CIPMUX=1	Setting multiple connection mode
AT+CIPSERVER=1, 333	Starting SERVER mode, setting port to 333
AT+CIPSEND=0, 7	Sending data with a length of 7 characters

In embedded development, AT commands are not only a set of instructions used for communication with the ESP8266 module, but they are also commonly used to control various communication modules, such as ESP8266 WIFI modules, 4G modules, and so on. Typically, the main chip sends AT commands to the communication module through a hardware interface (such as a serial port), and the module responds with data after receiving it. By sending different AT commands, users can configure the

ESP8266 module, connect to WIFI networks, establish data connections, and thus achieve communication and control with other devices. AT commands provide a simple and effective way to interact with the ESP8266 module, facilitating user development and debugging.

The following may be the AT commands used in this experiment. By setting it to AP mode, the mobile end can access network communication through relevant applications.

IV. INTRODUCTION TO SYSTEM SOFTWARE

A. Software Design Summary

In this application software design, an object-oriented design approach is used. The software design concept diagram is shown below.

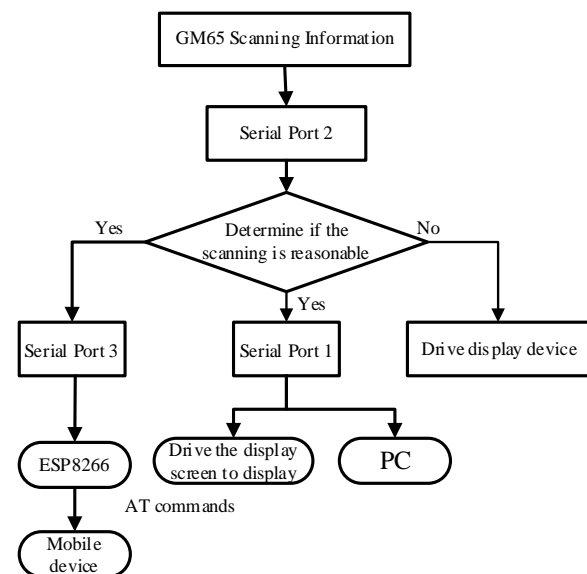


Figure 4. Software Design Flowchart

B. Wireless Module Software Design

The system collects information through the GM65 barcode scanner and utilizes the core board for preliminary verification. If the information fails the verification, the system will display an error on the monitor and trigger an alarm mechanism. Once the information is verified successfully, it will be transmitted to the PC end via the serial port for recording, while the ESP8266 module is used to wirelessly transmit the information to the mobile end, completing remote data monitoring. The entire process emphasizes the collaborative work between hardware modules and the logical control at the

software level, ensuring the accuracy of data transmission and the stability of the system. The design of the wireless module code is as follows.

```
ESP8266("AT\r\n");
delay(50000);
ESP8266("AT+CWMODE=2\r\n");
delay(50000);
ESP8266("AT+RST\r\n");
delay(50000);
ESP8266("AT+CWSAP=\"esp\", \"12345678\",
1,4\r\n");delay(50000);
ESP8266("AT+CIPMUX=1\r\n"); delay(50000);
ESP8266("AT+CIPSERVER=1,333\r\n");
delay(50000);
```

C. ESP8266 Module

When designing the software for a wireless module, ensuring effective communication with the ESP8266 module is crucial. By sending AT commands such as AT and AT+RST, and receiving the module's OK response, one can verify its normal working state. Additionally, sending the AT+GMR command can retrieve the module's version information, further confirming its usability. After completing these tests, one can proceed to configure network settings, such as setting the IP address, port number, and placing the module in AT mode to achieve wireless communication with the mobile end. The AT command sending function code is as follow.

```
void ESP8266(const char* num)
{
    u16 i,j;
    i = strlen(num);
    for (j = 0; j < i; j++)
    {
        while (USART_GetFlagStatus(USART3,
USART_FLAG_TC) == RESET);
        USART_SendData(USART3, (uint8_t)num[j]);
    }
}
ESP8266("AT\r\n");
ESP8266("AT+CWMODE=2\r\n");
"AT+CWMODE=2"
```

The ESP8266(const char* num) function plays a key role in software design, enabling control and configuration of the module by sending AT

commands. This design provides a foundation for the stability and reliability of the wireless communication system, ensuring the implementation of the expected functions.

V. TESTING RESULTS SUMMARY

During the testing process, we mainly focused on the performance of the TFT-LCD display, the PC end serial port assistant, and the transmission of information on the mobile end. It is necessary to accurately see the borrowing and returning status of laboratory equipment on the LCD screen, and at the same time, the information should be transmitted to the computer end and mobile end in a timely and accurate manner through the communication module. Through these tests and an in-depth analysis of the experimental results, we can evaluate the system's performance in various aspects, discover potential problems, and make timely improvements, thereby enhancing the system's stability and reliability.

A. Scanning the Display of QR Code

Observing the prompt messages for borrowing and returning different laboratory equipment on the mobile end, TFT-LCD screen end, and computer end after the barcode scanner scans the corresponding laboratory equipment's QR code, conclusions can be drawn accordingly.

The user interface design is guided by principles of simplicity and intuitiveness, ensuring that users can navigate the system with ease. The LCD display serves as the primary point of interaction, providing a clear and concise visual representation of the oscilloscope's status. As shown in Figure 5, the display screen and mobile application synchronize to show the same information, ensuring consistency across platforms.

When the barcode scanner successfully scans the QR code of the experimental equipment, the LCD display transitions from a neutral state to indicate a borrowed state. This change is achieved with a single scan, setting the status to "borrowed." To revert the status back to "returned," users simply need to scan the QR code again. This toggle functionality is reflected in both the LCD display and the mobile application, as depicted in Figure 5.

The interface layout prioritizes user experience. Key information, such as the oscilloscope's status, is prominently displayed, while secondary details are accessible through a straightforward menu navigation system. Interactions are minimized, with most operations requiring no more than a few clicks or scans, streamlining the process for users.

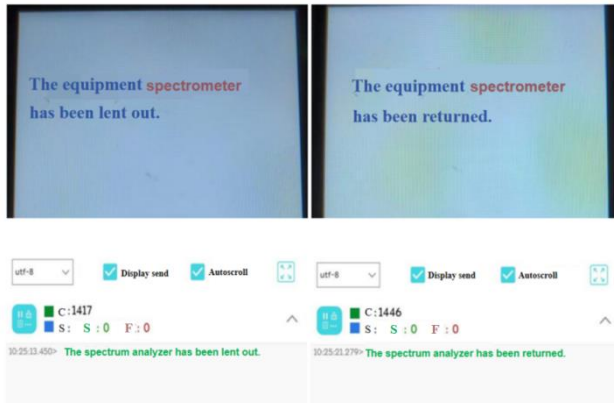


Figure 5. Normal scanning display status

B. Display Situation When Scanning an Incorrect QR Code

When the barcode scanner detects an irrelevant QR code, the system can trigger an alarm on the driver screen and display an error, as shown in Figure 6, the test generally meets the requirements.

The following image is the error message displayed on the TFT-LCD screen when an incorrect QR code is scanned.



Figure 6. Display Error Message on the Screen

C. Testing Results Analysis

Through the experimental results, it can be observed that when equipment is borrowed or returned, the system correctly displays relevant information prompts on the display screen, PC end, and mobile end, ensuring the accurate conveyance

and timely feedback of information. During the testing process, we paid special attention to the system's response speed and accuracy in actual operation scenarios. The results showed that the system could quickly respond to the borrowing or returning operations of the oscilloscope and display corresponding information on each port, ensuring that operators could clearly understand the status of the equipment, thereby improving operational efficiency and accuracy. In addition, when the barcode scanner scans information unrelated to laboratory equipment, the system can also provide error prompts, ensuring the accuracy of the system in identifying and handling exceptions. This error prompt function is particularly important in the laboratory management system, as it can help users quickly discover and correct erroneous operations, preventing damage or loss to laboratory equipment and data, and ensuring the normal and safe operation of the laboratory.

Through the above test results and in-depth analysis, it can be confirmed that the system performs well in various functional modules, with high stability and reliability. These test results not only verify the rationality and feasibility of the system design but also provide important references for further optimization and improvement of the system. In practical applications, these test results also provide strong support for the reliability and efficiency of the system, offering users a better user experience and service guarantee.

VI. CONCLUSIONS

Developing the laboratory equipment management system with the STM32F407 as the central control unit, we have focused on serial port communication, a key method for interfacing with external modules such as the ESP8266 and for internal serial port interactions. This approach forms the backbone of our system's design.

The system's adaptability allows for its application in diverse settings, including libraries, supermarkets, and museums. A practical example is the implementation of a book borrowing system in libraries. By encoding book-specific information into QR codes, affixing them to the books, and integrating with our system, librarians can achieve

a remarkable 90% return rate. This high efficiency is further enhanced by the system's ability to promptly detect and address any damages, ensuring the books remain in good condition for readers.

Similarly, the system can be tailored to create product information inquiry systems in supermarkets or cultural relic information systems in museums, significantly improving the management and user experience in these environments. The integration of this system not only streamlines processes but also enhances the reliability and durability of equipment, making it a valuable tool for modernizing management practices across various industries.

ACKNOWLEDGMENT

This work was partly supported by the research project of Xi'an university of posts and telecommunications teaching reform JGA202304 and JGA202316 and the undergraduate innovation and entrepreneurship plan S202411664126.

REFERENCES

- [1] X. Zhang, Y. Song, Y. Bai, M. Wang, C. Liu, Q. Wang, X. Hao, F. Huang, and G. Xu, "Exploration and Practice of the Construction of 'Internet +' Smart Nursing Experiment Center," *Nursing Research*, vol. 38, no. 14, pp. 2570-2574, 2024.
- [2] D. Zhao, H. Zhang, T. Zou, L. Qin, and C. Chen, "Application Research on Smart Laboratory Management System," in *Proceedings of the 2024 (12th) China Water Resources Informatization Technology Forum*, Nanjing, China: Hohai University, Jiangsu Water Conservancy Society, Zhejiang Water Conservancy Society, Shanghai Water Conservancy Society, Beijing Water Consulting Co., Ltd., 2024, pp. 8.
- [3] F. Pan, "Research on the Reform and Innovation of University Laboratory Management Model Driven by Informatization," *Journal of Fujian Open University*, no. 3, pp. 93-96, 2024.
- [4] C. Wang, A. Zhao, and X. Li, "Research on the Construction of University Laboratory Information Management," *Journal of Yuncheng University*, vol. 42, no. 3, pp. 80-83, 2024.
- [5] L. Fu, C. Dong, Z. Wu, and X. Feng, "Construction and Practice of the '1+1' College-Level Laboratory Information Comprehensive Management Platform," *Laboratory Research and Exploration*, vol. 43, no. 5, pp. 240-244, 2024.
- [6] T. Ji, "Application of Laboratory Information Management System in Laboratories," *Shanghai Light Industry*, no. 3, pp. 50-53, 2024.
- [7] R. Zhang, J. Yuan, and E. A. Chedzo, "Development of an Environmental Factor Measuring Instrument Based on ESP8266 Module," *China Educational Technology Equipment*, no. 14, pp. 38-42, 2023.
- [8] Q. Zhang, Z. Wan, and H. Zhang, "Design of a New Intelligent Barcode Scanner for Industrial Robots Based on a Vision System," *Science and Technology Innovation and Application*, vol. 13, no. 19, pp. 138-140, 144, 2023.
- [9] N. Cao and L. Zhou, "Design of an Intelligent Charging Cabinet System Based on ESP8266 WIFI Module," in *Proceedings of the 2023 Annual Meeting of Tianjin Electronic Industry Association*, Tianjin, China: Tianjin Electronic Industry Association, 2023, pp. 6.
- [10] S. Jing, "Analysis of the Working Principle of the STM32 IoT Development Board Circuit," *Home Appliance Maintenance*, no. 7, pp. 27-33, 2023.
- [11] Y. Yang, "Exploration of the Engineering Template Principle of the STM32 Microcontroller Firmware Library," *Nonferrous Metals Design*, vol. 50, no. 1, pp. 93-96, 2023.
- [12] J. Wei, F. Li, Q. Zhang, and L. Xie, "Design of an IoT Temperature Control Platform System Based on STM32," *Modern Electronic Technology*, vol. 46, no. 4, pp. 52-56, 2023.
- [13] H. Zhang, Y. Zhang, S. Sun, and Z. Li, "Lost and Found System Based on ESP8266 Wi-Fi Module and MQTT," in *Proceedings of the 19th Shenyang Science and Technology Academic Annual Conference*, Shenyang, China: Shenyang Municipal Party Committee, Shenyang Municipal People's Government, Shenyang Science and Technology Association, 2022, pp. 3.
- [14] Y. Huang, "Design of an IoT Data Acquisition System Based on STM32 Microprocessor," *Information Recording Materials*, vol. 23, no. 8, pp. 185-188, 2022.
- [15] G. Jin, H. Xiong, S. Bi, and C. Fu, "Design and Development of an Embedded Core Board Based on STM32 Microprocessor," *China Modern Educational Equipment*, no. 9, pp. 18-19, 30, 2022.
- [16] Z. Pang, "Research on Defect Detection Method of Aiming Light of Barcode Scanner Based on Machine Vision," M.S. thesis, Suzhou University, Suzhou, China, 2021.

Infrared Weak and Small Target Detection Algorithm Based on Deep Learning

Lei Wang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 2531361795@qq.com

Jun Yu

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: 763757335@qq.com

Abstract—In the infrared imaging scene where the target is at a long distance and the background is cluttered, due to the interference of noise and background texture information, the infrared image is prone to problems such as low contrast between the target and the background, and feature confusion, which makes it difficult to accurately extract and detect the target. To solve this problem, firstly, the infrared image is enhanced by combining DDE and MSR algorithm to improve the contrast and detail visibility of the image. For the RT-DETR network structure, the EMA attention mechanism is introduced into the backbone to enhance the feature extraction ability of the model by extracting context information. The CAMixing convolutional attention module is introduced into CCFM, and the multi-scale convolutional self-attention mechanism is introduced to focus on local information and enhance the detection ability of small targets. The filtering rules of the prediction box are improved, combined with Shape-IoU, and the convergence speed of the loss function in the detection and the detection accuracy of small targets are improved by paying attention to the influence of the intrinsic properties of the bounding box itself on the regression. In the experiment, the infrared weak target image dataset of the National University of Defense Technology was selected, labeled and trained. Experimental results show that compared with the original DETR algorithm, the average precision of the improved algorithm (mAP) is increased by 3.2%, and it can effectively detect infrared weak and small targets in different complex backgrounds, which reflects good robustness and adaptability, and can be effectively applied to infrared weak and small target detection in complex backgrounds.

Keywords-RT-DETR; EMA; CAMixing; Shape-IoU

I. INTRODUCTION

As an important thermal measurement technology, infrared imaging technology uses

infrared detectors to receive infrared thermal radiation in different wavelengths on the surface of the scene and convert it into images. This technique offers a variety of advantages, such as passive imaging, long range, ease of concealment, and ability to work day and night. This makes infrared imaging widely used in military, security, medical, industrial testing and other fields. However, there are also some challenges faced by infrared imaging devices in practical applications. First of all, because the imaging mechanism of infrared images is different from that of visible light, the contrast between the target object and the background is usually low, which makes it difficult to identify and detect the target. Secondly, when the background interference is strong and the target signal is weak, the signal-to-noise ratio of the infrared image is usually low, resulting in the target image often showing a small target with incomplete structure [1].

In addition, the problems of noise, scattering, and radiation inhomogeneity that are prevalent in infrared imaging further increase the difficulty of effectively detecting small targets in infrared images. With the development of computer vision technology, how to accurately and quickly detect and identify small targets in complex backgrounds has become one of the hot and difficult problems in research. The purpose of this paper is to explore and study the methods and technologies to improve the detection performance of small objects in infrared imaging, in order to provide new ideas and solutions for this field.

II. RELATED WORKS

In the early stages of micro-object detection, traditional algorithms were mainly based on filters and wavelet transforms to achieve single-frame and multi-frame detection [2]. Some commonly used techniques include median filtering, high-pass filtering, wavelet transform, and threshold segmentation. Yuan Shuai et al. [3] proposed a method to separate the target from the background by comparing the difference between the target area and the inner and outer double-layer neighborhoods, so as to enhance the local contrast of bright and weak small targets and effectively suppress complex background noise. Liu Delian et al. [4] proposed the concept of stagnation point connection, which was used as a benchmark to calculate the difference between the gray scale of each pixel and the datum to determine the target position. The improvement of these single-frame and multi-frame algorithms has the advantages of relatively simple computation and low complexity, but the detection performance is limited when the target contrast feature is not obvious in complex and changeable real scenes.

With the development of deep learning, effective and non-traditional solutions have been introduced to solve complex problems in the field of computer vision. At present, deep learning networks are mainly divided into two types in object detection: two-stage object detection algorithms (represented by R-CNN series) and single-stage object detection algorithms (represented by SSD and YOLO series). Both algorithms rely on deep convolutional neural networks (CNNs) to learn high-level features of images, capture semantic information in images, and use multi-scale strategies to detect targets at different resolutions, thereby improving detection performance [5]. Li Mukai et al. [6] introduced SEblock based on the idea of calibrating features according to weights in SENet, and improved the accuracy of YOLOv7 for small target detection to 83.97%. In view of the problems existing in the infrared image itself, Jiang Zhixin et al. [7] chose to combine the histogram equalization with the MSR shown in the image preprocessing to enhance the image, and at the same time, the loss function of the Faster-CNN network was improved,

and the mAP was improved by 6.11% compared with the original network. However, these deep learning algorithms based on prior boxes still have certain difficulties in processing small targets in images, especially for small targets that are scattered and do not overlap or are occluded. The introduction of attention mechanism to improve the backbone network alleviates this problem to a certain extent, but there are still limitations of transformer decoder in feature representation. Therefore, the DETR network based on transformer architecture can effectively improve the detection ability of small targets in complex backgrounds through global modeling capabilities and encoded location information.

III. ALGORITHMS MODEL

A. RT-DETR network

RT-DETR is the first real-time object detector in the DETR series and consists of four models. Different networks use different backbones, among which rtdetr-l uses HGNetv2-l as the backbone network, which has the best performance under the same conditions with fewer parameters and computational costs. The network structure is shown in Figure 1. The RT-DETR-L network structure consists of three parts: Backbone, Neck, and Decoder. The backbone network HGNetv2 consists of four HG Stage modules, each of which is mainly composed of HG Blocks. The backbone network extracts feature maps of three scales at different levels as the input of the hybrid encoder. The hybrid encoder is composed of two modules, the AIFI encoder and the CCFM feature fusion, in which the AIFI is still a multi-head transformer in essence, and only the deepest S5 feature layer is processed, and the F5 feature layer is finally output, which is used as the input of the CCFM module together with the S3 and S4 features, and the upper feature fusion and the lower feature fusion are carried out twice, respectively. For the fusion result, the anchor frame is selected through IOU-ware Query Selection, and the top-k300 is finally selected as the input into the decoder for prediction output.

B. Image preprocessing

In view of the problems of low contrast, high noise and unclear details in infrared images, the

detection performance of the model can be improved by selecting an appropriate algorithm for image enhancement before the model is processed.

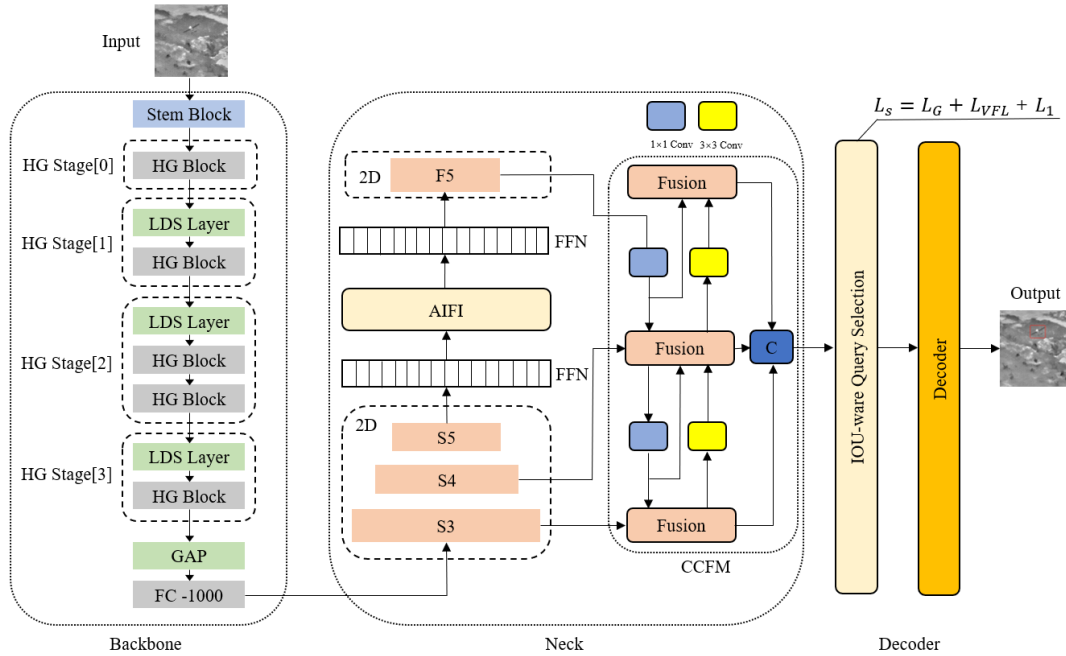


Figure 1. RT-DETR network structure

Dynamic Detail Enhancement (DDE) [8] and MSR algorithm are two commonly used image enhancement techniques in the preprocessing of infrared images. DDE separates the base layer and the detail layer of the image through the filtering algorithm, and enhances the detail layer to achieve significant enhancement of the details. However, the brightness and contrast of the base layer are limited, and it is easy to amplify the noise while amplifying the details. The multi-scale MSR algorithm improves the brightness and global contrast of the image through digital transformation and smoothing. However, the processing effect at the level of detail is limited, in complex scenes, the improvement of local contrast is not obvious. Therefore, the combination of DDE and MSR algorithm can dynamically adjust and enhance the detail and clarity of the image according to the complexity of the image content.

The detailed design process of the algorithm is as follows.

Bilateral filtering is used to decompose the original image into basic component I_B , and the output is:

$$I_B = \frac{1}{W_q} \sum_{p \in S} G_s(p) * G_r(p) * I_p \quad (1)$$

where the I_p represents the original infrared image, the G_s represents the pixel value weight, the G_r represents the spatial distance weight, and the W_q represents the sum of the weights of each pixel value in the filter window, which is used for the normalization of the weights.

The detail component is the result of subtracting the base component from the original image, The formula is shown in the following formula.

$$I_D = I_p - I_B \quad (2)$$

The basic component contains the large-scale structure and lighting information of the image, and the local details are enhanced by MSR for the basic component R_B .

$$R_B = MSR(I_B) \quad (3)$$

Finally, the detail component is recombined with the enhanced base component to obtain the final infrared image.

C. Attention mechanisms

In order to solve the problem that small and medium-sized targets are difficult to detect in infrared image object detection, the original rt-detr-l uses HGNetv2, which is essentially still a CNN network structure, and uses HGBlock to extract features through deep convolution operations. Therefore, an improved RT-DETR network model is designed, and the EMA attention mechanism module is added to the backbone network of RT-DETR, so that it can extract multi-scale feature information that is rich in global context information and differentiated features, especially small target feature information.

Figure 2 shows how to add an EMA module.

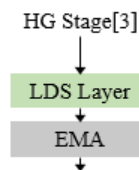


Figure 2. EMA module

At the same time, in order to suppress the interference of background and noise in the infrared image, the attention of small targets is improved. The introduction of CAMixing convolution-attention module enables CCFM to suppress the interference of irrelevant information and improve the denoising performance when multi-scale feature fusion, which is conducive to enhancing the modeling of global and local features and improving the detection rate of small targets.

Add the CAMixing module to the network, as shown in Figure 3.

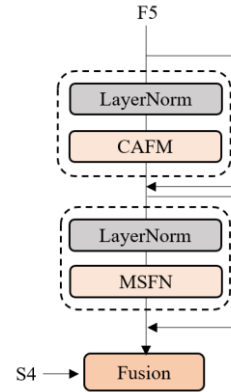


Figure 3. CAMixing module

In RT-DETR, the encoder is only used in S5, and comparative experiments are used to verify that it not only helps to significantly reduce the computational effort and improve the computational speed, but also does not cause significant damage to the performance of the model [9]. Therefore, the addition of CAMixing to the two-way downward feature fusion with S5 helps to improve the detection accuracy and reduce the amount of computation.

D. Loss Function

The original IOU-aware Query Selection uses GIoU to optimize the query selection process of the prediction box, and introduces an external rectangle to reduce the loss between the real box and the prediction box of the large target. However, for small target detection, a slight shift in the target may cause a significant change in its position information. GIoU does not directly consider the aspect ratio difference between the prediction box and the real box, and cannot fully capture the subtle changes in position information. Therefore, Shape-IoU is used instead of GIoU for small targets, and the loss is calculated by paying attention to the shape and scale of the bounding box itself, so as to optimize the detection accuracy of small targets. The regression loss calculated by Shape-IoU is shown in the following formula.

$$L_s = 1 - IOU + D^s + 0.5\Omega^s \quad (4)$$

Among them, IOU is the intersection and union ratio, S is the zoom factor, and its value is related to the size and number of targets in the dataset. and, D^S is the distance loss, which is calculated as follows:

$$ww = \frac{2 \times (w^{gt})^s}{(w^{gt})^s + (h^{gt})^s} \quad (5)$$

$$hh = \frac{2 \times (h^{gt})^s}{(w^{gt})^s + (h^{gt})^s} \quad (6)$$

$$D^S = hh \times \frac{(x_c - x_c^{gt})^s}{c^2} + ww \times \frac{(y_c - y_c^{gt})^s}{c^2} \quad (7)$$

Where w^{gt} , h^{gt} are the width and height of the GT box, x_c , y_c , x_c^{gt} , y_c^{gt} are the coordinates of the center point of the prior box and the GT box. ww and hh are calculated from the coordinates as the weight coefficients of the horizontal and vertical directions, and the loss of D^S can be adjusted by adjusting the value of s.

Ω^S is the loss of shape. It follows the formula of SIOU:

$$\Omega^S = \sum_{t=w,h} (1 - e^{-w_t})^\theta \quad (8)$$

Where w,h are the width and height of the prior frame, respectively, θ determines the size of the shape loss, and in order to avoid the shape loss accounting for a relatively heavy proportion of the overall loss and the position change of the prediction frame, the genetic algorithm takes the value of 4.

In the case of an infrared small target image, it is mainly composed of the background, and only a small part is occupied by the target. It is easier to learn the features of the background than the features of the target during training. Therefore, the ATFL loss function is used to replace the original VFL classification loss, which decouples the target from the background, and uses the

adaptive mechanism to adjust the loss weight, forcing the model to allocate more attention to the small target features. The ATFL expression is as follows:

$$\begin{cases} -(\lambda - p_t)^{-\ln(p_t)} \log(p_t) & p_t \leq 0.5 \\ -(1 - p_t)^{-\ln(p_t)} \log(p_t) & p_t > 0.5 \end{cases} \quad (9)$$

Where p_t represents the current average prediction probability value, and λ (>1) is the hyperparameter. ATFL is balanced by improving the adaptive factor in TFL $\gamma - \ln(p_t)$. The weights of the samples [10] increase the contribution of small targets while reducing the time consumption of adjusting hyperparameters.

IV. EXPERIMENTS

A. Experimental Environment

Table I shows the experimental environment in this paper, which is based on the Ubuntu 18.04 operating system, the graphics card model is RTX2080Ti, and the memory is 16GB. The experiment basically uses the parameters recommended by RT-DETR, builds the model based on Python3 and Pytorch framework, and uses the standard SGD optimizer, with batch-size set to 8 and epochs set to 100.

TABLE I. EXPERIMENTAL ENVIRONMENT

Experimental environment	Version
CPU	IntelCorei7-11800H
GPU	NVIDIA GeForce RTX2080 Ti
Language	Python3.8
Deep Learning Framework	Pytorch1.14.0
CUDA	11.8.0

B. Dataset

The infrared aircraft small target dataset [11] used in this experiment includes a total of 22 annotated data folders, and the image content is mainly based on the ground background, sky background, multiple aircraft, aircraft distance, aircraft approaching, etc. A total of 12177 infrared images were selected, with an image resolution of

256*256, a channel count of 1, and a bit depth of 24. This dataset is widely used in tasks such as object detection and target tracking.

C. Evaluation Metrics

In this experiment, Precision (P), Recall (R) and Average Precision (AP) are mainly used as network evaluation indicators, and their mathematical expressions are as follows:

$$precision = \frac{TP}{TP + FN} \quad (10)$$

Where TP stands for True Positives, FP stands for False Positives, and Precision measures the proportion of instances that the model predicts to be positive. Used to evaluate the accuracy of instances predicted to be positive samples.

Recall, also known as recall, is the proportion of the sample predicted as positive to the predicted sample, and its mathematical expression is:

$$recall = \frac{TP}{TP + FN} \quad (11)$$

The mathematical expression for the average precision (AP) is:

$$AP = \int P(R) dR \quad (12)$$

AP is used to evaluate the Precision-Recall curves of the model at different thresholds, and is the average value obtained by integrating the Precision-Recall curves. It measures the average accuracy of the model's predictions at different thresholds.

In addition, in order to evaluate the processing effect of the image preprocessing algorithm, PSNR and SSIM similarity were used to distinguish the similarity of the images before and after processing, the information entropy (Entropy) was used to reflect the retention degree of image information, and the average gradient (AG) and edge intensity (EME) were used to evaluate the clarity of the image.

D. Algorithm verification results

The objective evaluation method is used to evaluate the quality of the improved images, and the effectiveness of the improvement is verified compared with other algorithms.

The final enhancement is shown in Figure 4.

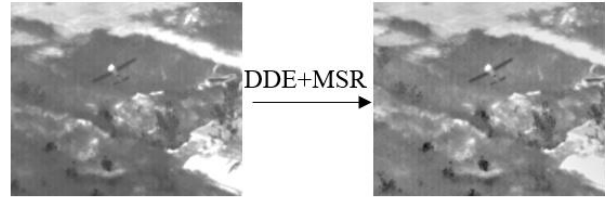


Figure 4. Image enhancement effect

Table II lists the verification results.

E. Shape-IoU parameter validation

In this paper, the Shape-IoU loss function is introduced into the network model [12]. which has a scale parameter whose size can be determined by optimizing the parameters, and the s parameter is finally set to 0.4 by comparing different parameters on the dataset. Comparative experiments are shown in Table III.

TABLE II. COMPARISON OF ALGORITHM ENHANCEMENTS

index algorithm	PSNR	SSIM	Entropy	AG	EME
Original image			6.2850	41.9135	2.6388
SSR	28.2970	0.85522	5.7150	44.2675	2.8967
MSR	28.7772	0.8676	6.2176	44.9135	2.8932
DDE	36.0989	0.9679	6.4594	44.2206	2.6857
Bilateral filtering	34.6621	0.8395	6.3155	21.7407	1.4891
DDE+MSR	28.7581	0.8436	6.2793	46.8969	2.8882

TABLE III. COMPARATIVE EXPERIMENTS OF DIFFERENT PARAMETERS OF SHAPE-IOU

s	0.1	0.2	0.3	0.4	0.5	0.6	1.0
mAP(%)	83.5	83.4	84.9	85.3	84.8	83.5	83.4

F. Network comparison experiment

In order to verify the effectiveness of each improvement point of the network, based on the RT-DETR-L network, six sets of comparative experiments were carried out on the dataset. and the environment and parameter settings were uniform. The experimental results are shown in Table IV, and "√" indicates that the corresponding method was used. It can be seen from Table 4 that after adding the EMA and CAMixing modules and improving the loss function, the algorithm achieves the best detection accuracy, and the AP value is 3.2% higher than the original RT-DETR,

which proves the effectiveness of the improved module in this paper. Figure 5 shows the AP variation curve of the improved method more intuitively. The detection fluctuates greatly due to the influence of the dataset, but the improved network detection accuracy is significantly stable and the accuracy increases. Figure 6 shows the detection results of the original network and the improved network, and the prediction score increases significantly, which verifies the effectiveness of the improved method and proves that the improved method is better for the detection of dense targets.

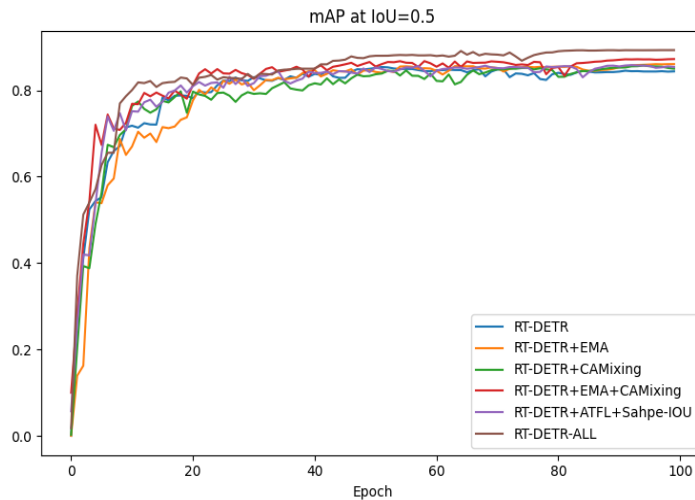


Figure 5. AP change curve

TABLE IV. COMPARES THE EXPERIMENTAL RESULTS

EMA	CMAixing	Shape-IoU	ATFL	P/%	R/%	AP/%	Param/10 ⁶
				73.2	75.2	84.6	32.81
√				72.2	76.3	85.5	33.40
	√			74.8	75.2	85.6	34.97
√	√			74.1	75.0	86.2	35.23
		√	√	75.5	76.2	85.9	32.81
√	√	√	√	75.4	77.1	87.8	35.23

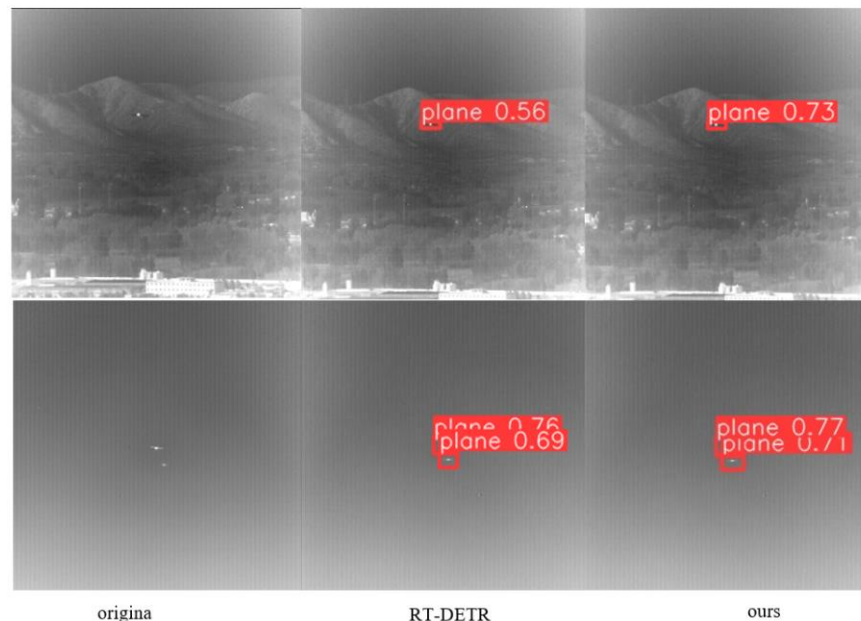


Figure 6. Improved detection results

V. CONCLUSION

In order to solve the challenge of target detection in infrared imaging scenes, this paper effectively enhances the contrast and detail visibility of infrared images by combining DDE and MSR algorithms, improves the RT-DETR network structure, introduces EMA attention mechanism and CAMixing convolutional attention module, and significantly improves the model's detection ability and overall detection accuracy of small targets. At the same time, the Shape-IoU and ATFL loss functions are combined to improve the regression ability of small targets under infrared conditions. Experimental results show that the improved algorithm is better than the original DETR algorithm in detecting infrared weak and small targets in complex backgrounds, and the mean average accuracy (mAP) is increased by 3.2%, showing good robustness and adaptability. However, considering the richer training data, especially the infrared images containing different weather, time and terrain conditions, it is necessary to further process the contrast between the target and the background to enhance the generalization ability of the model. In addition, real-time performance is very important in practical applications, and the detection speed can be improved by optimizing the model structure

and algorithm, making it better suitable for real-time infrared object detection scenarios.

REFERENCES

- [1] Guo Yujie. Research on micro target detection and recognition method based on information enhancement [D]. Guangdong Technical Normal University, 2023.
- [2] Liu Ying, Sun Haijiang, Zhao Yongxian. Research on infrared weak and small target detection method in complex background based on attention mechanism [J]. Computer Software and Computer Applications, 2023, 38(11):1455-1467.
- [3] Yuan Shuai, Yan Xiang, Zhang Yugeng, et al. Infrared and Laser Engineering, 2022, 51(4):20221071.)
- [4] Liu D, Zhang J, Dong W. Temporal profile based smallmoving target detection algorithm in infrared image sequences [J]. International Journal of Infrared and Milli-meter Waves, 2020, 28(5):373-381.
- [5] LI B Y, XIAO C, WANG L G, et al. Dense nested attention network for infrared small target detection [J]. IEEE Transactions on Image Processing, 2023, 32:1745-1758.
- [6] Li Mukai. Research on small-scale infrared pedestrian detection technology based on deep learning [D]. Shanghai:University of Chinese Academy of Sciences,Shanghai Institute of Technical Physics, Chinese Academy of Sciences, 2022.
- [7] Jiang Zhixin. Research on infrared small target detection method at sea based on deep learning [D]. Dalian: Dalian Maritime University, 2019.
- [8] ZHENG Z H, WANG P, LIU W, et al. Distance-IoU loss: faster and better learning for bounding box regression [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York: AAAI, 2020:12993-13000.

- [9] LIU X, PENG H, ZHENG N, et al. Efficientvit: Memory efficient vision transformer with cascaded attention [C]//Proceedings of group the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 14420-14430.
- [10] LI Y, HOU Q, ZHENG Z, et al. Large selective kernel network for remote sensing object detection [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 16794-16805.
- [11] HUI B W, SONG ZY, FAN HO, et al. A dataset for infrared image dim-small aircraft target detection and track-ing under ground air background [J]. Scientific Database, Chinese Academy of Science, 2020, 5(3):291-302
- [12] Zhou Mengran, Wang Ao. Object Detection Algorithm for Lightweight Remote Sensing Image Based on DETR[J/OL]. Journal of Chongqing Technology and Business University (Natural Science Edition). <https://link.cnki.net/urlid/50.1155.N.20240328.1703.004>.

3D Reconstruction of Indoor Scenes Based on 3DGS Models

Hanghua Li

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 2451734651@qq.com

Lipeng Si

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 37648537@qq.com

Abstract—With the rapid development of computer vision and artificial intelligence technologies, indoor scene reconstruction has been more and more widely used in the fields of virtual reality, augmented reality and architectural design. In this paper, we study an indoor scene reconstruction method based on the 3DGS model, which has been widely used in computer graphics and vision processing with powerful scene representation and rendering capabilities. In this study, we optimize the 3DGS model to enhance the detail preservation and realism of the reconstruction results by adjusting the opacity of the Gaussian function. We used the Replica dataset and the self-harvested dataset for model training. Through experimental validation, the peak signal-to-noise ratio as well as the structural similarity ratio of the reconstruction results of the optimized model have an improvement effect of more than 1%, which indicates that the optimized model has a significant improvement in detail retention and realism, and the reconstructed scene performs more realistically in terms of texture details and light and shadow effects.

Keywords-3DGS; Indoor Scene; 3D Reconstruction

I. INTRODUCTION

In recent years, there are more and more demands for indoor refined 3D models in smart cities, cultural relics protection, indoor navigation, virtual reality, etc. 3D reconstruction of indoor scenes has become one of the important research topics in the field of computer vision and computer graphics. Scene 3D reconstruction refers to the acquisition of image or video data of the indoor environment, the use of computer vision technology and 3D reconstruction algorithms to analyze the image content and geometric information, to infer the layout of the room, the position and size of the furniture, the geometry of

the walls and floors and other structured information, and ultimately to construct a real indoor 3D model.

Traditional scene reconstruction methods rely on camera position and pose information as well as data from depth sensors, but these methods usually suffer from limitations in real-time, accuracy, and cost. With the advent of the 3DGS algorithm, it enables high-quality real-time rendering and scene optimization by efficiently modeling the scene using Gaussian functions. The technique starts from a sparse point cloud, represents the scene as a differentiable 3D Gaussian set, and constructs an accurate and compact representation of the scene by optimizing its properties such as position, opacity, and covariance. During the rendering process, the 3DGS algorithm utilizes fast GPU sorting and tile-based rasterization to achieve efficient visibility-aware rendering with anisotropic splash support, thus achieving real-time rendering while ensuring rendering quality.

In this paper, we will use 3DGS based algorithm to reconstruct the indoor scene, which can reconstruct the complex indoor scene efficiently and reliably. And the model is trained on Replica dataset as well as self-collected dataset and further optimized. The experimental results show that the performance of the model is improved.

II. RESEARCH BACKGROUND

In recent years, 3D scene reconstruction technology has made significant progress in the fields of computer graphics and computer vision, which refers to the transformation of two-

dimensional image or video data into interactive 3D models through computer vision, 3D reconstruction, deep learning and other technical means. This process involves multiple technical fields, including computer graphics, image processing, machine learning, etc. and aims to accurately restore the physical space and generate 3D digital models with a high degree of realism.

The 3D scene reconstruction technique mainly includes several steps of data acquisition and processing, feature extraction, 3D reconstruction, and result optimisation. Among them, the dataset is mainly categorized into point cloud, mesh and voxel, and the 3D reconstruction mainly includes indoor scene reconstruction based on deep learning and 3D reconstruction by traditional methods. PointNet [1], a deep learning network that directly processes point cloud data, proposed by researchers at Stanford University, provides an effective solution for 3D reconstruction of indoor scenes. NeRF Neural Radiation Field, an emerging technique proposed by Ben Mildenhall et al [2], utilizes a neural network model to achieve fast reconstruction and rendering of indoor scenes by training on captured scene images to learn attributes such as lighting, material, and depth of the scene, and then generating realistic images of new perspectives. Neural radiation fields have become an important area of research in subsequent research efforts. The main ones include D-NeRF [3], which is able to learn dynamic deformable fields from image view sequences, NSFF [4], a scene flow field algorithm for the free synthesis of spatio-temporal views of dynamic scenes, NeRV [5], a neural reflective and visible field algorithm for view and illumination resynthesis, and GIRAFFE [6], a composable generative feature algorithm for editable scene representations. While all of the above techniques can be reconstructed for different scenes, the aforementioned neural radiation field models are mainly deep convolutional neural networks, which take much longer to train compared to traditional shader and illumination techniques, possibly several times longer than these techniques. It also requires significant computational resources to support its training and rendering process.

Recently, a 3DGS technique based on neural radiation field was proposed by Bernhard Kerbl et al [7], which has attracted much attention due to its high efficiency and real-time performance. This technique realizes efficient reconstruction and real-time rendering of the scene by utilizing the Gaussian function to represent the spatially continuous distribution of the data, which provides a new way of thinking for the reconstruction of the indoor scene. The 3DGS technique has already shown its advantages in accurate modeling and detail 3DGS technology has demonstrated its advantages in accurate modeling and detail preservation. It not only preserves the geometric information of the scene, but also retains the rich texture and lighting effects during the rendering process.

In conclusion, with the continuous development of research and technology, the advantages of 3DGS algorithm in terms of accuracy, real-time and efficiency will become more and more significant. These advantages will not only improve the quality of indoor scene reconstruction, but also bring more possibilities for interior design, virtual reality, augmented reality and many other fields. At the same time, the continuous progress of 3DGS will not only change the way we understand and utilise interior space, but also provide a broad stage for future technological applications and innovations.

III. INDOOR SCENE 3D RECONSTRUCTION

Indoor scene reconstruction using 3DGS mainly includes the following steps: data acquisition and processing, indoor scene reconstruction and model optimisation. Data acquisition is to collect data from the scene to be reconstructed by using the shooting tool, and then the collected data are preprocessed to be converted into SfM point cloud data, and the scene reconstruction results are obtained by training the 3DGS model. The process of reconstruction is shown in Fig. 1.

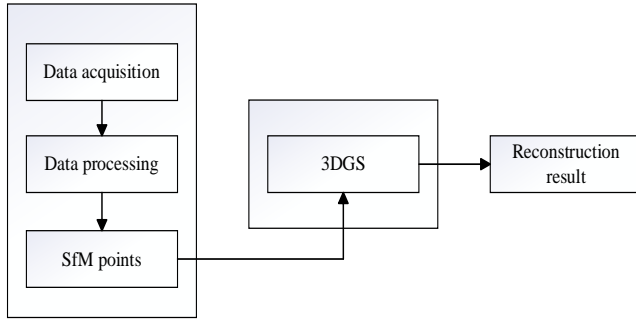


Figure 1. 3D Reconstruction Process Map.

A. Data Acquisition And Processing

Considering the scene equipment variability of the shot, after converting the shot into a continuous image, the denoising process is performed on this continuous set of images. Common image noise processing schemes mainly include mean removal filter for mild noise, Gaussian filter for Gaussian noise, and bilateral filter for filtering noise while preserving image edge information. Therefore, we introduce bilateral filtering to preprocess the data, and realize the removal of noise and smoothing of local edges on the basis of retaining regional information by comprehensively considering the spatial information of the image within the filter and the similarity of pixel gray values, some of the detail processing results are shown in Figure 2.



Figure 2. The result of adding bilateral filters for indoor scenes.

B. Model Introduction

3DGS modeling is an innovative scene reconstruction and rendering technique based on Gaussian distribution. Its core idea is to use 3D Gaussian distribution as the basic element to represent the geometric and color information in the scene, and to render this information onto a 2D plane by rasterization showing unique advantages in many application scenarios. It mainly includes the following steps:

1) Creating Gaussian Functions

A 3D Gaussian was chosen for this experiment, which can be easily projected into a 2D image, thus allowing for fast blending rendering. Starting with a set of sparse point clouds, each feature point is represented in 3D space by a Gaussian function. The Gaussian function is defined by several parameters, including position, covariance matrix, color, and transparency. Our Gaussian is defined by the full 3D covariance matrix Σ defined in the world space, centered at the mean point:

$$G(x) = e^{-\frac{1}{2}x^T \Sigma^{-1}x} \quad (1)$$

At the same time, an affine transformation is needed to project the 3D Gaussian to 2D for rendering, letting the covariance matrix in the camera coordinate system be Σ . Given a scaling matrix S and a rotation matrix R , we can find the corresponding Σ :

$$\Sigma = RSR^T S^T \quad (2)$$

2) Adaptive density control

Starting from the initial sparse point of the SfM, the initially sparse set of Gaussians is changed into a denser set of Gaussians by controlling the number and density of Gaussians per unit volume to better represent the scene. Gaussian densification is performed every 100 iterations and removes essentially transparent Gaussians, i.e., Gaussians with α less than a threshold. Our adaptive control part of the Gaussian needs to be corrected by moving the Gaussian for regions with missing geometric features and regions where the Gaussian covers a large area of the scene. For small Gaussian areas with missing geometric features, they need to be covered. For large Gaussians that are large for the Gaussian coverage of the scene need to be split into smaller Gaussians, using two new Gaussians to replace these Gaussians. In the first case, the need to increase the total volume and the number of Gaussians is detected and handled, and in the second case, the larger Gaussians are split into multiple smaller Gaussians.

3) Rasterization

The screen is first partitioned into 16×16 blocks and the Gaussians with 99% confidence intervals

that intersect the view vertebrae are retained. At the same time, each Gaussian is instantiated according to the number of tiles they cover, and then each instance is assigned a key that combines the depth of the view space and the tile ID. The Gaussians are then sorted based on those keys.

After sorting the Gaussian, the entries are sorted by recognising the first and last depth of the splat to which they were sputtered, and a list is generated for each tile. Rasterisation is then performed, and for each tile a thread block is started, each of which loads the packet containing the Gaussian into shared memory and traverses the list from front to back through the colour and opacity values based on the given pixels, thus simultaneously loading and processing the data. When we reach the target saturation level during the accumulation of pixels in the continuous traversal, the corresponding thread stops, and the processing of the whole tile terminates when the opacity of all pixels is 1.

C. Extracting 3DGS Model Optimization

Since the SfM point cloud data generated by colmap increases to millions when the dataset is large, these data take relatively long time in scene training, which affects the efficiency of 3DGS splatting. In this experiment, a pruning operation on the Gaussian volume is added in the middle, with the main purpose of removing the high-density and high-volume regions in the scene due to false positives or redundancy, in order to improve the rendering quality and reduce the amount of computation. The specific flow of this experiment is shown in Fig. 3.

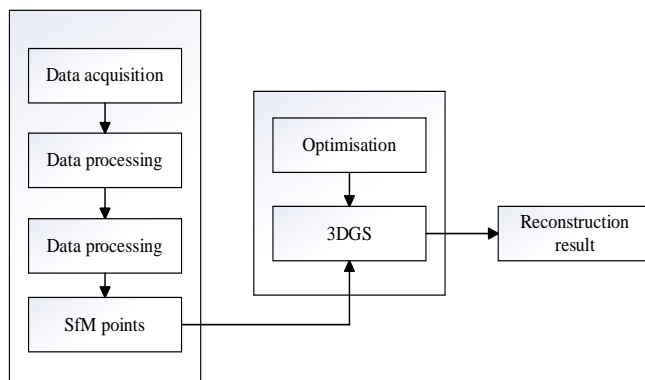


Figure 3. Optimized 3D Reconstruction Process Map.

First, a series of 3D Gaussian bodies are defined within the view body, which are rendered by

projecting them to the camera viewpoint so that their impact can be observed in the 2D image plane. For each Gaussian body, we calculate its contribution to each pixel or ray. This is done by determining whether the Gaussian body intersects a light ray. Iterating over all the pixels in the training view, the number of ‘hits’ on a pixel is calculated for each Gaussian as an initial significance score. In addition to the basic number of ‘hits’ on a pixel, we also consider the volume and opacity of the Gaussian to further refine the score. So the summary is as follows:

$$GS_j = \sum_{i=1}^{MHW} L(G(X_j), r_j) \cdot \sigma_j \cdot \gamma(\sum_j) \quad (3)$$

Where j is the Gaussian index, i is the pixel, and M , H , and W are the number of training views, image height, and image width, respectively. l is the indicator function, which determines whether or not the Gaussian function intersects a given ray. However, the use of Gaussian volume tends to exaggerate the importance of the background Gaussian distribution, leading to excessive pruning of the Gaussian distribution for complex geometric models. Therefore, we introduce a more adaptive method to measure the size of its volume.

$$\gamma(\sum) = (V_{norm})^\beta \quad (4)$$

$$V_{norm} = \min(\max(\frac{V(\sum)}{V_{max90}}, 0), 1) \quad (5)$$

The range of Gaussian volumes was limited to 0 to 1 by sorting all Gaussian volumes and normalizing the top 90% of maximum values in the Gaussian volume as a benchmark, thus avoiding overly high or underly high floating-point Gaussian values obtained directly from the original 3DGS.

IV. EXPERIMENTAL ANALYSIS

In this paper, we firstly reconstructed the indoor scene using the base 3DGS model, and partially optimised the base 3DGS model, and evaluated the reconstruction quality of this paper's model in the Replica dataset and the self-picked data scene, and quantitatively analysed the reconstruction results by two metrics: PSNR and SSIM.

This paper first uses the original model to reconstruct indoor scene scenes for the Replica dataset, which is rich in scene details with dense meshes, high dynamic range textures, semantic layers, and reflective properties, and record the training results, and then we use the optimized model in this paper to reconstruct each scene in this dataset and record the results to compare with the former results.

The results of comparing the peak signal-to-noise ratio and structural similarity obtained by training the Replica dataset after optimizing the opacity of the original 3DGS model are shown in Table I.

TABLE I. Experimental Results

Dataset	PSNR		SSIM	
	3DGS	ours	3DGS	ours
office0	38.45	39.05	0.925	0.934
office1	36.97	37.17	0.938	0.954
office2	36.39	36.68	0.945	0.962
office3	35.85	36.35	0.941	0.958
office4	36.54	36.66	0.939	0.941
room0	37.26	38.06	0.915	0.923
room1	36.28	37.75	0.929	0.935
room2	34.98	35.14	0.913	0.927

By reconstructing the Replica dataset with diversity, we observe that the optimised method performs well on image quality assessment metrics. Both the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM) show significant performance improvement in all test scenarios. This consistent improvement not only reflects the robustness of the optimised algorithm in dealing with different types of indoor environments, but also indicates its good adaptability in scene reconstruction.

The peak signal-to-noise ratio results obtained by training the Replica dataset after optimizing the opacity of the original 3DGS model are shown in Fig. 4.

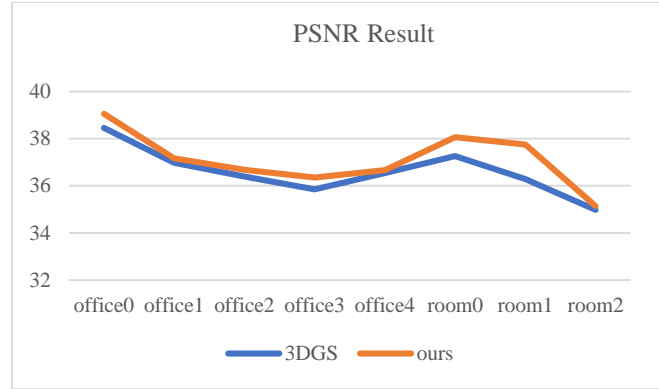


Figure 4. Performance of peak signal-to-noise ratios of 3DGS and ours models for eight different indoor scenes, respectively.

The structural similarity ratio results obtained by optimizing the opacity of the original 3DGS model after training on the Replica dataset are shown in Fig. 5.

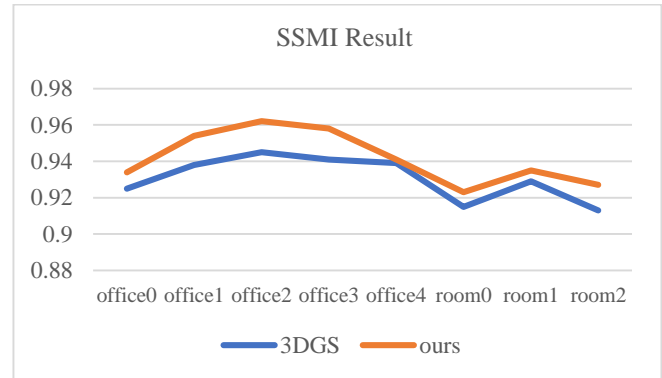


Figure 5. Representation of structural similarity between 3DGS and ours model for 8 different indoor scenes, respectively.

In multiple scenes of Replica dataset, the reconstruction results of the optimized 3DGS model show better results than the original 3DGS model in both PSNR and SSIM. With the increase of the number of Gaussian primitive, the PSNR value shows an obvious upward trend, which verifies the positive correlation between the model scale and the reconstruction quality. Meanwhile, the reconstruction results of the 3DGS model on the Replica dataset are not only highly similar to the real scene in terms of brightness and contrast, but also maintain good consistency in structural information. This indicates that the model is able to accurately capture the structural features of the scene, thus generating visually more realistic reconstruction results.

A. Indoor scene reconstruction on a self-built dataset

In order to evaluate the applicability of the model in this paper more comprehensively, a series of own environmental data were also collected in this study. It is used to verify the performance of the optimized 3DGS model in the paper in real scenarios and to compare the results, as shown in the following table II.

TABLE II. Experimental Results

Dataset	PSNR	SSIM
3DGS	30.41	0.879
ours	31.89	0.897

In the self-collected data scenes, the optimized 3DGS model also shows superior reconstruction quality to the original 3DGS model. Especially in the area with complex light and rich texture, the model can better restore the scene details, and the PSNR value is kept at a high level. the SSIM value is also kept at a high level, which better adapts to the scene changes and maintains the structural similarity of the reconstruction results.

Through this experiment, we verified the excellent reconstruction performance of the model in the Replica dataset and self-collected data scenes, and both PSNR and SSIM metrics show that the model can efficiently and accurately reconstruct the 3D scene and retain rich structural information and visual details. In the future, we can further optimize the model structure and training algorithm to improve its reconstruction efficiency and quality in large-scale and complex scenes.

V. CONCLUSIONS

In order to solve the traditional neural radiation field for indoor scene reconstruction problem, and

through the literature research is the scene of the method, the final choice of 3DGS model, selected Replica dataset and self-collected data scene reconstruction quality In the future, we can further optimise the model structure and training algorithms, in order to improve its reconstruction in large-scale, complex scenes in the efficiency and quality, for virtual reality, game development, film production and other fields to provide more possibilities.

REFERENCES

- [1] Chen C , Fragonara L Z , Tsourdos A .GAPointNet: Graph Attention based Point Neural Network for Exploiting Local Feature of Point Cloud[J].Neurocomputing, 2021, 438(7553).
- [2] Mildenhall B , Srinivasan P P , Tancik M ,et al.NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[C]//2020.
- [3] Pumarola A, Corona E, Pons-Moll G and Moreno-Noguer F. D-NeRF: Neural Radiance Fields for Dynamic Scenes [C] . IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:10313-10322.
- [4] Li Z, Niklaus S, Snavely N and Wang O. Neural Scene Flow Fields for Space-Time View Synthesis of Dynamic Scenes [C] . IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:6494-6504.
- [5] Srinivasan P, Deng B, Zhang X, Tancik M, Mildenhall B and Barron J. NeRV: Neural Reflectance and Visibility Fields for Relighting and View Synthesis [C] . IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021:7491-750.
- [6] Niemeyer M and Geiger A. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields [C] . IEEE/CVF Conference on Computer Vision and Pat- tern Recognition (CVPR), 2021:11448-11459.
- [7] Niemeyer M and Geiger A. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields [C] . IEEE/CVF Conference on Computer Vision and Pat- tern Recognition (CVPR), 2021:11448-11459.

9D Rotation Representation-SVD Fusion with Deep Learning for Unconstrained Head Pose Estimation

Jiaqi Lyu

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: lvjiaqi@st.xatu.edu.cn

Changyuan Wang

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: Cyw901@163.com

Abstract—Accurately estimating human head pose poses a significant challenge across various application domains. To address the inherent limitations of previous approaches, this research proposes an unconstrained head pose estimation strategy. The method combines deep learning with rotation matrices, utilizing nine-dimensional vectors output by the neural network, which are projected back to rotation matrices in SO (3) space through singular value decomposition. This ensures both the smoothness and uniqueness of the rotation representation. The approach demonstrates distinct advantages in handling the rotation estimation task, particularly when the rotated representation is used as the model output. It not only avoids the discontinuity and double-coverage issues associated with prior methods but also enhances the stability of the representation in high-dimensional space, thereby improving the learning process. Additionally, the geodesic loss function is incorporated to train the network. The proposed strategy surpasses previous state-of-the-art methods, as evidenced by experiments conducted on the AFLW2000 and BIWI datasets.

Keywords—Head Pose Estimation; Efficientnetv2; Rotation Matrix; Geodesic Loss

I. INTRODUCTION

In fields such as human-computer interaction [1] and augmented reality [2], head pose estimation has become a core technology driving immersive experiences and precise interactions. There are two main types of current methods: those that use landmarks and those that don't [3]. Landmark-based algorithms find important facial points in pictures and then use these points to map them to a 3D model of the head to figure out the 3D head position. Although this method is highly accurate, it is directly limited by the precision of key point localization. Occlusions and extreme rotation

angles can make key points difficult to identify, leading to deviations in their positions and affecting the accuracy of the final head pose estimation.

Advancements in deep learning have significantly improved the accuracy of head pose estimation algorithms that do not depend on landmarks, because of the utilization of deep neural networks. HopeNet [4] proposes a multi-task learning approach that discretizes continuous head pose angles into several categories. It captures the discrete distribution of head poses through classification tasks while refining continuous angle values with regression tasks, using multi-task learning to predict Euler angles. QuatNet [5] employs a dual-branch structure for classification and regression. One branch uses a recurrent neural network for Euler angle classification, while the other represents head pose regression with quaternions. HPE [6] enhances head pose estimation by using a two-stage ensemble and a top-k regression. Multiple models independently predict in the first stage, and the top k optimal predictions are integrated in the second stage. WHENet [7] uses a single-branch model but increases the number of head pose angle categories. FSA-Net [8] utilizes a dual-branch architecture and fine-grained attention mechanism to effectively merge local and global image features, resulting in more accurate Euler angle predictions. TriNet [9] uses vectors to represent head direction instead of traditional Euler angles. FDN [10] introduces a feature decoupling method that helps the model focus on head pose-related features, ignoring background noise and other

irrelevant factors. LwPosr [11] is a lightweight network that employs a two-stream, three-stage structure for fine-grained regression. This structure combines a depth-separable convolution with a transformer encoder, enabling the network to efficiently predict head pose with a low number of parameters and high accuracy.

Many of the above methods split the rotation representation into bins for classification and combine it with regression for stable prediction, a practice that has become common. However, binning the angles can result in fragmented information. Additionally, choosing the appropriate rotation representation method is crucial for optimal performance. Most current methods use Euler angles or quaternions to train networks. While effective in some scenarios, they suffer from numerical discontinuities when handling large-scale and continuous rotations, such as the gimbal lock issue with Euler angles and the double coverage problem with quaternions. Zhou et al. [12] demonstrated that any rotation representation with four or fewer dimensions is discontinuous, making it unsuitable for neural network learning.

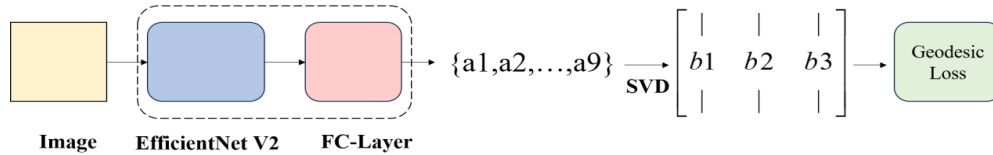


Figure 1. Overview of the proposed method

II. METHOD

A. Feature Extraction Network

Many existing neural networks use depthwise separable convolution to extract features, and although its structure possesses fewer parameters as well as smaller FLOPs compared to normal convolution, it is usually not able to fully utilize the gas pedal with the available hardware. This paper utilizes EfficientNet V2 as the feature extraction network, which is a more advanced and lightweight convolutional neural network model compared to EfficientNet. It is characterized by low number of parameters, high accuracy, and excellent training and inference speed. A notable improvement is the substitution of EfficientNet's

Geist et al. [13] summarized the characteristics of various rotation representations and their impact on gradient-based optimization methods. Building on this study, a head pose estimation technique is proposed that does not rely on landmarks but instead utilizes rotation matrices to accurately determine head pose direction. EfficientNetV2-S [14] is employed as the feature extraction network. Rather than directly predicting the rotation matrix, the neural network generates a nine-dimensional vector. This vector is subsequently transformed into a 3×3 matrix and converted into a valid rotation matrix using Singular Value Decomposition (SVD).

The network was trained using a geodesic loss function instead of the more commonly employed mean squared error (MSE) loss function. This choice was made because the geodesic loss function more effectively captures the differences in the manifold's rotations. The proposed approach is illustrated in Figure 1. The following sections provide a more detailed explanation of each component.

shallow MBCConv with the Fused-MBCConv module. The Fused-MBCConv module substitutes the expansion 1×1 convolution and depthwise 3×3 convolution in the primary branch of the original MBCConv structure with a 3×3 convolution. This solves the problem of employing depth-separable convolution in the initial layer of the network. The issue of slowdown caused by using of depth-separable convolutions in the shallow layers of the network is effectively solved, resulting in an important enhancement in training speed. Figures 2 and 3 show the structure of the MBCConv and Fused-MBCConv modules, respectively. Table 1 shows the EfficientNet V2-S structure. The method proposed in this paper is adapted by changing the final fully connected layer output to 9.

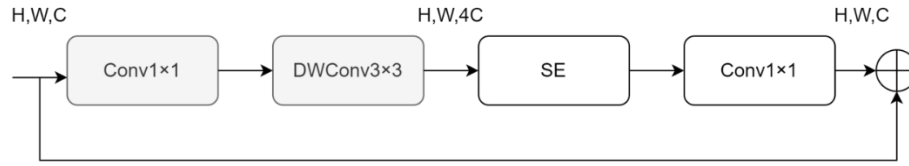


Figure 2. MBConv

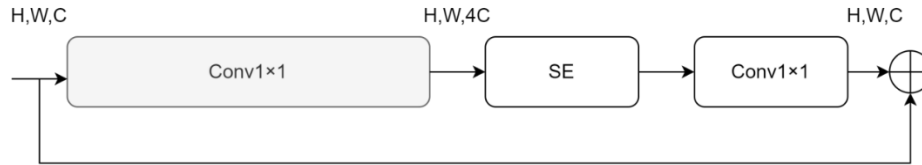


Figure 3. Fused-MBConv

TABLE I. EFFICIENTNETV2-S ARCHITECTURE

Stage	Operation	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1,3x3	1	24	2
2	Fused-MBConv4,3x3	2	48	4
3	Fused-MBConv4,3x3	2	64	4
4	MBConv4,3x3,SE0.25	2	128	6
5	MBConv6,3x3,SE0.25	1	160	9
6	MBConv6,3x3,SE0.25	2	256	15
7	Conv1x1&Pooling&FC	-	1280	1

B. R9+SVD

Choosing a suitable approach for representing rotation is vital for accurately estimating head posture. Traditionally, Euler angles have been employed. Nevertheless, this method of representing rotation is not ideal because to its susceptibility to gimbal lock. In such cases, specific sequences and angles of rotation can cause the loss of one of the three independent rotation axes. Another rotation representation is the quaternion method, which is not affected by gimbal lock but has the issue of double coverage. This means that for each rotation, there are two corresponding quaternions. While these two representations are physically equivalent, they exhibit a significant numerical discontinuity. Therefore, neural networks struggle to learn accurate poses in the presence of numerical discontinuities.

Figure 4 shows two examples of pictures with comparable visual presentation from the 300W-LP dataset. A comparison of the two pictures shows that the Euler angles and quaternions have distinct labeling values, notably the second value, yaw, in

the Euler angles. Positive and negative numbers suggest entirely opposing attitudes. A better rotation representation is the rotation matrix, which is a continuous representation, only the rotation matrix can reflect the similarity of pose appearance. In $SO(3)$ space, the rotation matrix is a 3×3 matrix that satisfies the orthogonality criteria $RR^T = I$, where R^T represents the transpose of R , and I represents the identity matrix. The R9+SVD method can work with any 3×3 matrix and turn it into a valid rotation matrix in $SO(3)$. R9 refers to using the neural network to directly output 9 vector values that describe a 3×3 rotation matrix. The SVD method was created because directly estimating 9 variables might not work because rotation matrices have certain qualities, such as being orthogonal and having a positive determinant.

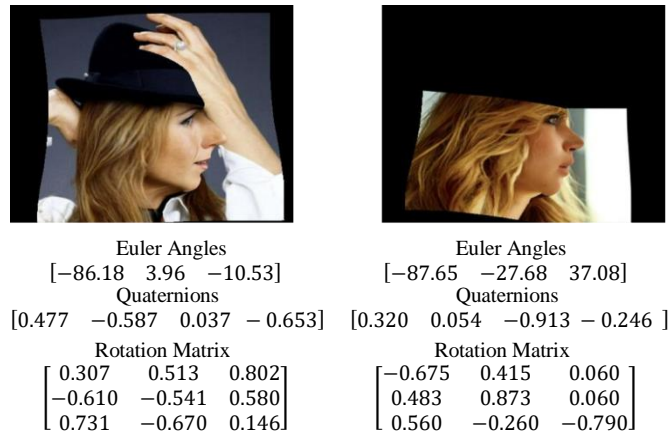


Figure 4. Image samples from 300W-LP dataset with different rotation representations

Given a 3×3 matrix, its singular value decomposition (SVD) is expressed as:

$$M = U\Sigma V^T \quad (1)$$

Here, U and V are 3×3 orthogonal matrices, and Σ is a diagonal matrix containing the singular values of matrix M . To project M onto the rotation matrix R , R must satisfy two conditions:

- 1) The column vectors of R must be of unit length and orthogonal to each other.
- 2) The determinant of R must equal 1.

Therefore, after adjusting the singular values, R is constructed as:

$$\Sigma^+ = \text{diag}(1, 1, \det(UV^T)) \quad (2)$$

Reconstruct the rotation matrix using the adjusted singular value matrix:

$$R = U\Sigma^+V^T \quad (3)$$

Hereby, $\det(UV^T)$ ensures that the determinant of R is 1, while the combination of U and V^T ensures that the column vectors of R are orthogonal. Therefore, the neural network predicts 9 parameters, which are then transformed into a 3×3 rotation matrix while sticking to the orthogonality requirement.

The advantages of this method are:

1) Smoothness: It provides a continuous and smooth representation, allowing optimization algorithms like gradient descent to converge effectively while avoiding issues such as the singularities of Euler angles or the double coverage problem of quaternions.

2) Robustness: SVD can be seen as a model architecture where the three column vectors of the matrix contribute equally to the prediction. This enhances robustness to input noise.

C. Geodesic Loss

The loss function commonly used in previous head pose estimation tasks is the L2 loss function, and the calculation method is equation (4). However, in the head pose estimation task, there

are some problems when using the L2 loss function to measure the difference between rotations. First, the L2 loss does not take into account the periodicity of the rotation angle, which makes it impossible to correctly evaluate the similarity between rotations close to 360 degrees or -360 degrees. Secondly, the L2 loss assumes that all dimensional changes are independent and linear, which is inconsistent with the geometric structure of the rotation matrix or quaternion.

$$\text{loss}(x, y) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \quad (4)$$

The geodesic loss function measures the distance between two rotation matrices along the shortest path on the manifold, known as the geodesic. The geodesic loss function is calculated based on the trace of the rotation matrices, with the formula given as:

$$d(R_1, R_2) = \cos^{-1}\left(\frac{\text{tr}(R_1 R_2^T) - 1}{2}\right) \quad (5)$$

R_1 and $R_2 \in SO(3)$, representing the predicted rotation matrix and the true rotation matrix, respectively. The trace (tr) denotes the sum of the diagonal elements of a matrix. This distance will be used as the loss function for the neural network in subsequent experiments.

III. EXPERIMENTS AND RESULTS

A. Datasets

This work trained and evaluated its method on various types of datasets. The most commonly used publicly available datasets for head pose estimation are 300W-LP [15], AFLW2000 [16], and BIWI [17].

1) The 300W-LP dataset consists of 66,225 facial pictures, which are increased to 122,450 samples using image flipping augmentation. It encompasses a diverse array of postures and comprehensive 3D annotation data. The ground truth is given as Euler angles, which were transformed into matrix representation following the method described by Hempel [18].

2) The AFLW2000 dataset includes the

initial 2,000 face photos that were chosen from the AFLW dataset. These images are accompanied by 68 key point annotations. It includes a range of face positions, including various degrees of rotation and emotions.

3) The BIWI dataset includes video sequences of 24 individuals, totaling 15,678 images. Each frame provides detailed 3D head pose and key point annotations, covering various head pose variations in real-world scenarios. The MTCNN [19] facial detection algorithm was used to extract the head region from the images.

B. Evaluation Metrics

The head pose estimate error is measured using the Mean Absolute Error (MAE) of Euler angles, which is the most widely used metric. This is represented by Equation (6).

$$MAE = \frac{1}{N} \sum_{i=1}^N (|x_g - x_p|) \quad (6)$$

N refers to the total number of face images, x_g represents the true values of the head poses, and x_p represents the predicted values of the head poses.

C. Implementation Details and Results

This work employed PyTorch to create the whole model, with EfficientNetV2-S serving as the backbone network, and trained the network for 30 epochs with the Adam optimizer. The initial learning rates for the backbone network and the final fully connected layer were set to $1e-5$ and $1e-$

4, respectively, with each learning rate halving every 10 epochs. The batch size was set at 64.

In the first experiment, the network was trained using the synthetic 300W-LP dataset and subsequently tested on two real-world datasets: AFLW2000 and BIWI. The evaluation metric used was the mean absolute error (MAE) of Euler angles, which required transforming the predicted rotation matrices into Euler angles for comparison purposes. Table 2 presents the results of the first experiment, comparing the proposed approach to other state-of-the-art landmark-free head pose estimation methods. The experimental results demonstrate that the proposed strategy outperformed the current best methods by approximately 22% and achieved the lowest error rates in pitch, yaw, and roll angles on the AFLW2000 dataset. On the BIWI dataset, the approach exceeded seven out of eight of the most advanced algorithms in terms of MAE. Figure 5 illustrates the results of the method on the AFLW2000 dataset after converting the predicted rotation matrices to Euler angles.

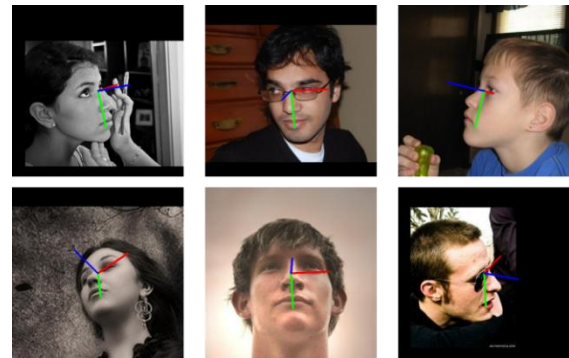


Figure 5. Example images of Euler angle visualization using rotation matrix transformation from AFLW2000 dataset

TABLE II. COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE AFLW2000 AND BIWI DATASET

Models	AFLW2000				BIWI			
	Yaw	Pitch	Roll	MAE	Yaw	Pitch	Roll	MAE
HopeNet[4]	6.40	6.53	5.39	6.11	4.54	5.15	3.37	4.36
FSA-Net[8]	4.50	6.08	4.64	5.07	4.64	5.61	3.57	4.61
HPE[6]	4.80	6.18	4.87	5.28	3.12	5.18	4.57	4.29
QuatNet[5]	3.97	5.62	3.92	4.50	2.94	5.49	4.01	4.15
WHENet[7]	5.11	6.24	4.92	5.42	3.99	4.39	3.06	3.81
TriNet[9]	4.04	5.77	4.20	4.67	4.11	4.76	3.05	3.97
FDN[10]	3.78	5.61	3.88	4.42	4.52	4.70	2.56	3.93
6DRepNet[18]	3.63	4.91	3.37	3.97	3.24	4.48	2.68	3.47
9D-EfficientNet	3.57	4.69	3.28	3.85	4.08	4.17	2.94	3.73

In the second experiment, the method outlined by FSA-Net was followed, with the BIWI dataset randomly split into training and testing sets in a 7:3 ratio. The results were compared with other networks that employed the same experimental approach. Table 3 presents the results of the second experiment. The proposed method outperforms other methods in terms of MAE, and shows superior performance in yaw and pitch, with roll being better than most. These experimental results demonstrate the robustness of the proposed method, as it achieves stable and accurate results in both Euler angle and MAE across different datasets.

TABLE III. EULER ERROR COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE 70/30 BIWI DATASET

Models	BIWI			MAE
	Yaw	Pitch	Roll	
HopeNet[4]	3.29	3.39	3.00	3.23
FSA-Net[8]	2.89	4.29	3.60	3.60
TriNet[9]	2.93	3.04	2.44	2.80
FDN[10]	3.00	3.98	2.88	3.29
MDFNet[20]	2.99	3.68	2.99	3.22
DDD-Pose[21]	3.04	2.94	2.43	2.80
6DRepNet[18]	2.69	2.92	2.36	2.66
9D-EfficientNet	2.62	2.36	2.51	2.50

To demonstrate the superiority of the geodesic loss function as a distance metric for head pose estimation, additional tests were conducted using the rotation matrix. To support this claim, the previous experiments were replicated by training the network with the L2 loss function. Table 4 presents the effectiveness of the proposed technique when trained with two distinct loss functions. Training the network with the geodesic loss function yields superior results compared to the L2 loss.

TABLE IV. COMPARISON OF THE MAE BETWEEN L2 AND GEODESIC LOSS

Loss function	AFLW2000	BIWI	70/30 BIWI
	MAE	MAE	MAE
L2 Loss	3.90	3.92	2.71
Geodesic Loss	3.85	3.73	2.50

This paper also examines the influence of different backbone networks on the results

obtained by employing geodesic loss. In order to do a comparison, this research employed the ResNet [21] network as an illustrative example. The findings shown in Table 5 demonstrate that the approach outlined in this research achieves exceptional results when used to the EfficientNet V2-S backbone network. By employing ResNet18 as the foundation network, this approach surpasses the majority of previous approaches in terms of performance on both the AFLW2000 and BIWI datasets. This illustrates that employing a suitable rotation representation greatly enhances the accuracy of head pose estimation.

TABLE V. COMPARISON OF MAE BETWEEN RESNET AND EFFICIENTNETV2 BACKBONE NETWORKS

Models	AFLW2000	BIWI	70/30 BIWI
	MAE	MAE	MAE
ResNet18	4.37	3.70	2.64
EfficientNetV2-S	3.85	3.73	2.50

In the final experiment, the THOP library was used to compare the proposed method with 6DRepNet in terms of parameter count and floating-point operations (FLOPs). As shown in Table 6, the proposed approach achieved a lower MAE while requiring fewer parameters and FLOPs.

TABLE VI. COMPARISON OF PARAMETERS AND FLOPS BETWEEN 6DREPNET AND OUR METHOD

Models	Params	FLOPs
6DRepNet	43.752M	9.844G
9D-EfficientNet	20.189M	2.901G

IV. CONCLUSIONS

In this research, provide an appearance-based, unconstrained, end-to-end head posture estimation approach. Following the assumption that rotation matrices are better suited to deep learning in 3D rotation problems, and provide a continuous 9D vector + SVD technique for head pose estimation. In addition, this research uses the geodesic loss function rather than the usual MSE to better correspond with the rotation matrix representation. Experiments show that using the EfficientNetV2 backbone network, this approach surpasses other most advanced methods on the AFLW2000 dataset

and most methods on the BIWI dataset. In further experiments, this investigated the effects of various loss functions and backbone networks on the findings, as well as comparisons of parameter count and floating-point operations. All of the experiments show that this approach is robust, reliable, and lightweight.

REFERENCES

- [1] Strazdas Dominykas, Hintz Jan, AlHamadi Ayoub. Robo-hud: interaction concept for contactless operation of industrial cobotic systems [J]. Applied Sciences, 2021, 11(12):5366-5366.
- [2] CHARISSIS V, FALAH J, LAGOO R, et al. Employing emerging technologies to develop and evaluate in-vehicle intelligent systems for driver support: infotainment ar hud case study [J]. Applied Sciences, 2021, 11(4):1397-1397.
- [3] Werner, P., Saxen, F., & Al-Hamadi, A. Landmark based head pose estimation benchmark and method. In ICIP, 2017.
- [4] Ruiz, N., Chong, E., & Rehg, J. M. Fine-grained head pose estimation without keypoints. In CVPR, 2018.
- [5] Hsu H W, Wu T Y, Wan S, et al. QuatNet: Quaternion-Based Head Pose Estimation with Multiregression Loss [J]. Multimedia, IEEE Transactions on, 2018. DOI:10.1109/TMM.2018.2866770.
- [6] Huang B, Chen R, Xu W, et al. Improving head pose estimation using two-stage ensembles with top-k regression [J]. Image and Vision Computing, 2019, 93:DOI:10.1016/j.imavis.2019.11.005.
- [7] Zhou Y, Gregson J. WHENet: Real-time Fine-Grained Estimation for Wide Range Head Pose [J]. 2020. DOI:10.48550/arXiv.2005.10353.
- [8] Yang T Y, Chen Y T, Lin Y Y, et al. FSA-Net: Learning Fine-Grained Structure Aggregation for Head Pose Estimation from a Single Image [J]. IEEE, 2020. DOI:10.1109/CVPR.2019.00118.
- [9] Cao Z, Chu Z, Liu D, et al. A Vector-based Representation to Enhance Head Pose Estimation[C]//Workshop on Applications of Computer Vision. IEEE, 2021. DOI:10.1109/WACV48630.2021.00123.
- [10] Zhang H, Wang M, Liu Y, et al. FDN: Feature Decoupling Network for Head Pose Estimation [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7):12789-12796. DOI:10.1609/aaai.v34i07.6974.
- [11] Dhingra N. LwPosr: Lightweight Efficient Fine-Grained Head Pose Estimation [J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2202.03544.
- [12] Zhou Y, Barnes C, Lu J, et al. On the Continuity of Rotation Representations in Neural Networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019. DOI:10.1109/CVPR.2019.00589.
- [13] Geist A R, Frey J, Zobro M, et al. Learning with 3D rotations, a hitchhiker's guide to SO (3) [J]. arxiv preprint arxiv:2404.11735, 2024.
- [14] Tan M, Le Q V. EfficientNetV2: Smaller Models and Faster Training [J]. 2021. DOI:10.48550/arXiv.2104.00298.
- [15] Xiangyu Zhu, Zhen Lei, Xiaoming Liu 0002, et al. Face alignment across large poses: a 3d solution. [J]. CoRR, 2015.
- [16] Zhu X, Lei Z, Yan J, et al. High-fidelity Pose and Expression Normalization for face recognition in the wild [J]. IEEE, 2015. DOI:10.1109/CVPR.2015.7298679.
- [17] Fanelli G, Matthias Dantone. Random Forests for Real Time 3D Face Analysis [J]. International Journal of Computer Vision, 2013, 101(3):437-458. DOI:10.1007/s11263-012-0549-0.
- [18] Hempel T, Abdelrahman A A, Al-Hamadi A .6D Rotation Representation for Unconstrained Head Pose Estimation [J]. arXiv e-prints, 2022. DOI:10.48550/arXiv.2202.12555.
- [19] Zhang K, Zhang Z, Li Z, et al. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks [J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503. DOI:10.1109/LSP.2016.2603342.
- [20] Liu H, Fang S, Zhang Z, et al. MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation [J]. IEEE Transactions on Multimedia, 2021, PP (99): 1-1. DOI:10.1109/TMM.2021.3081873.
- [21] Aghli N, Ribeiro E. A Data-Driven Approach to Improve 3D Head-Pose Estimation[C]//International Symposium on Visual Computing. Springer, Cham, 2021. DOI:10.1007/978-3-030-90439-5_43.
- [22] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. IEEE, 2016. DOI:10.1109/CVPR.2016.90.

Vector Storage Based Long-term Memory Research on LLM

Kun Li

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: 38190985@qq.com

Chengang Jing

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: jcg050980@163.com

Xin Jing

School of Computer Science and Engineering
Xi'an Technological University
Xi'an, China
E-mail: jingxin@xatu.edu.cn

Abstract—Current large language model (LLM) intelligences face the challenges of high inference cost and low decision quality when dealing with complex tasks, and are especially deficient in maintaining context coherence during long tasks. This research presents an innovative vector storage long-term memory mechanism model (VIMBank) to enhance the long-term context retention ability and task execution efficiency of LLM intelligences by storing and retrieving historical interaction data through a vector database. VIMBank utilizes a dynamic memory updating strategy and the Ebbinghaus forgetting curve theory to efficiently manage the memory of intelligences and reinforce critical information, forgetting unimportant data, and optimizing storage and reasoning costs. The experimental results show that VIMBank significantly improves the decision quality and efficiency of LLM intelligences in multi-tasking scenarios and reduces the computational cost. Compared with different agents, the success rate of task decision is increased by 10% to 20%, and the reasoning cost is reduced by about 23%, which provides an important theoretical basis and practical support for the future development of intelligences with long term memory and adaptive learning ability.

Keywords—Large Language Model; Long Term Memory; Vector Storage

I. INTRODUCTION

Recent revolutionary advances in large language model-based intelligences have dramatically changed our interactions with AI

systems, with LLM intelligences capable of autonomously fulfilling user commands and demonstrating impressive performance in a wide range of tasks. However, LLM intelligences still suffer from the fatal problems of high reasoning cost and low quality of decision making for complex problems. Long-term memory mechanisms aim to improve the decision quality and reduce the reasoning cost of LLM intelligences by storing external knowledge and historical interactions. For example, in conversations that require long interactions with the user, long-term memory helps the intelligent body to maintain contextual coherence, remember the user's historical behaviors and preferences, and provide more accurate or personalized answers, and it can also avoid repetitive reasoning on repeated historical tasks and obtain the reasoning results of similar tasks directly from long-term memory, which saves the arithmetic resources and improves the efficiency of task execution. Therefore, the long-term memory mechanism in LLM intelligent body systems is crucial for maintaining contextual understanding by storing information perceived from the environment and utilizing the recorded information to assist with future instruction tasks that will be performed. This mechanism is one of the important capabilities that have become indispensable for application scenarios such as

continuous dialogue systems, personalized recommendations, healthcare, and education. Therefore, in order to improve the quality of decision making and reduce the arithmetic resource consumption of LLM intelligences, it is necessary to study a more effective long-term memory mechanism.

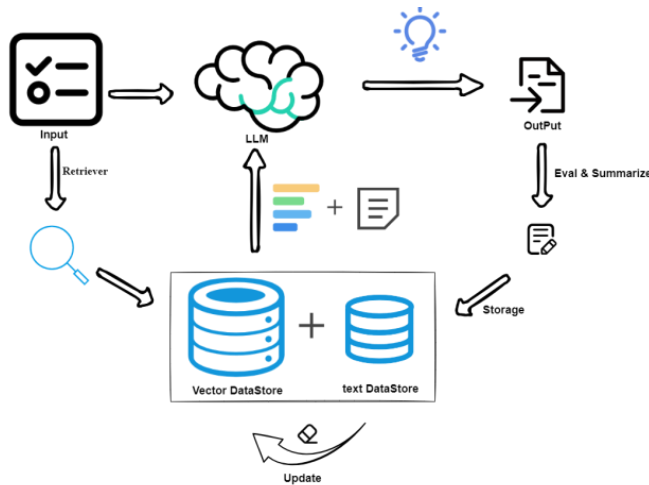


Figure 1. VIMBank Framework

Therefore, this research proposes a long-term memory mechanism model, VIMBank, which aims to improve the decision quality and decision efficiency of intelligences. As shown in Fig. 1, VIMBank uses a vector database as the underlying basic tool to support the storage of historical information, which enables LLM intelligences to effectively store different historical interaction information, such as knowledge information, dialog information, and related task information, which are sliced and vectorized before being stored as long-term memory. The LLM intelligences can effectively store different historical interaction information, such as knowledge information, dialog information and related task information, and different types of information from the interaction process of the intelligences are stored as long-term memory in the vector database after slicing and vectorization, which facilitates the subsequent retrieval and updating of related memory information. In order to realize the accurate and efficient retrieval of different types of information in the decision-making process of the intelligent body, different retrieval strategies are designed for different types of interaction information to effectively improve the recall rate of relevant

information from long-term memory in the decision-making process of the intelligent body, so as to enhance the quality of decision-making of the intelligent body's task instructions. Meanwhile, based on Ebbinghaus's forgetting curve theory, which describes the law of human brain's forgetting of new things, VIMBank further combines human's own dynamic memory mechanism for new things, and VIMBank introduces this dynamic mechanism as a long-term memory updating strategy, realizing that the LLM is able to selectively forget the long-term memories with the passage of time and strengthen its memory of the LLM can selectively forget long-term memories over time and enhance the memorization of more frequent memory information. Overall, VIMBank is a memory mechanism that improves the quality and efficiency of decision-making of intelligent body tasks on the basis of storing, retrieving and updating long-term memories.

VIMBank is generic in the sense that it is able to adapt closed-source large models such as ChatGPT, as well as open-source models such as Qwen2-7b, chatglm3-6b, and other models.

This research is based on the ALFWorld [1], HotpotQA [2] and KAgentBench [3] multi-task datasets, and tests the performance of open source large models such as chatglm3-6b and Qwen2-7b in planning and decision-making on different tasks, to understand the proficiency of the large models in understanding, planning and decision-making, and to analyze the problems existing in the large models in the reasoning and decision-making process. Furthermore, based on the large models, this research also considered the performance of the large models on various tasks after giving them the ability to decompose tasks and reflect based on intelligent agent models such as ReACT [4] and InterACT [5]. For example, the success rate of the chatglm3-6b model on the ALFWorld task set reached 62%, and after being equipped with ReACT, the task success rate increased to 70%. Although the large models themselves have certain understanding and reasoning capabilities, experiments have found that in the multi-round task process, due to the limitations of the large model context window, there is a semantic loss phenomenon caused by the overflow of the context

window, which leads to the failure of task decision-making. Therefore, the necessity of long-term memory for large language models is verified.

In order to evaluate the effectiveness of VIMBank, this research uses ChatGPT4-o to generate new task datasets similar to existing tasks based on the characteristics of the existing datasets to expand them, so as to verify the effectiveness of long-term memory in the long-trajectory decision quality and multi-round reasoning efficiency of complex tasks in large language models. The experimental results demonstrate the ability of VIMBank in long-trajectory decision quality and multi-round execution efficiency of tasks. The main contributions of this research are summarized as follows:

This research verifies that the semantic environment of large language models is missing in the task reasoning process due to the limitation of context windows.

This research proposes a new long-term memory mechanism VIMBank, which improves the decision quality and efficiency of LLM in multi-round planning reasoning of complex tasks.

This research demonstrates the versatility of the VIMBank mechanism, which can be adapted to existing open-source large models such as Qwen2-7b and chatglm3-6b, as well as current mainstream closed source large models such as ChatGPT3.

II. RELATED WORKS

In recent years, the field of large language models (LLMs) has undergone significant transformation, demonstrating its powerful capabilities in a variety of natural language processing tasks. Models including GPT-3, OPT and FLAN-T5 have achieved outstanding results in multiple fields. At the same time, the latest closed-source models such as PaLM, GPT-4, and ChatGPT continue to demonstrate wide adaptability and gradually become an auxiliary tool for many people's daily decision-making. However, the closed-source nature limits researchers and companies from in-depth study of the internal mechanisms of LLM and hinders the development of applications adapted to specific fields. Therefore, many open source LLM projects have emerged in

the community, such as LLaMa, ChatGLM, Alpaca, Vicuna, and Qwen. These models usually contain 6 billion to 14 billion parameters and have achieved remarkable results in multiple benchmark tests.

Nonetheless, these models still have some shortcomings. A significant drawback is that they lack strong long-term memory capabilities. This limitation hinders LLM's ability to maintain context over long periods of time and retrieve relevant information from past interactions. Therefore, in order to improve the decision-making quality and reasoning efficiency of LLM in complex tasks, it is particularly important to research and develop effective long-term memory mechanisms.

With the rapid progress of large language models, researchers have also conducted in-depth research on the context window limit of LLM. There are two main directions involved: one is to adjust the model to increase the context window limit, and the other is to introduce a long-term memory mechanism to enhance the ability of LLM to process long texts through retrieval enhancement. A representative application is the retrieval enhancement generation system based on LLM. For example, Wang et al. [6] proposed a self-knowledge guided retrieval enhancement, which aims to improve the reasoning and generation capabilities of the model by combining external knowledge and LLM's own knowledge, especially in the face of complex problems and task scenarios that require context understanding. Sun et al. [7] proposed a Think-on-Graph method that uses knowledge graphs to provide structured information to guide and optimize the reasoning process of LLM, thereby improving the accuracy and coherence of the generated results. However, the RAG method focuses on external knowledge and is used for tasks that require combining external knowledge to generate answers, but cannot be used to record and manage the historical information of LLM during user interaction. Long-term memory is more like an internal storage mechanism of LLM, which can store external knowledge, past conversation content, task execution history, and other information. This information can be stored and retrieved and used in subsequent interactions with the LLM, thereby helping the agent maintain contextual coherence, improve decision-making

quality, and demonstrate higher intelligence in long-term interactions. At present, some research has made preliminary progress. For example, Liu et al. [8] pointed out that when LLM processes long-term tasks, due to the limitation of the context window, it cannot effectively maintain the memory of past information, resulting in poor reasoning and decision-making. Therefore, the research team proposed the Think-in-Memory method to enhance the long-term memory ability of the model through the mechanism of recall and post-thinking. In order to ensure the consistency and contextual coherence of LLM in long-term open dialogues, Lu et al. [9] introduced a memorandum mechanism to enhance the performance of LLM in long-term dialogues. However, the above studies all have high storage and retrieval overhead and correctness problems. At the same time, memory also needs to support dynamic updates to avoid LLM referencing outdated or irrelevant task-related information, thereby affecting the accuracy of task decisions.

In general, although significant progress has been made in the field of LLM in the past two years, long-term memory support is still needed to enhance LLM when persistent interaction is required and accuracy and efficiency are guaranteed in multi-task scenarios. This research uses VIMBank as a new approach to address this challenge.

III. VIMBANK MECHANISM

This section first introduces the overall workflow of the proposed long-term memory mechanism framework. Then each stage of VIMBank is described in detail, including the storage of long-term memory, the retrieval of stored information, and the principles of memory updating.

A. Overall framework

1) *Task Definition*: The purpose of long-term memory is mainly for context retention and

decision optimization during the execution of complex tasks in LLM. For example, given a complex task query that requires N steps of reasoning to execute $Q = \{s_1, s_2, \dots, s_N\}$, where s_i denotes the first state of the task execution process state, $i \in (1, N)$; each task state is composed of a series of task context information $S_i = \{c_1, c_2, \dots, c_M\}$, where denotes the first state of a task state; each task state is composed of a series of task context information, where the first task context information of a task state, $j \in (1, M)$. At the same time, based on the long-term memory existing query Q' and historical decision responses R' , the parameter pair $p = \{(Q'_1, R'_1), (Q'_2, R'_2), \dots\}$, the new task decision-making process is optimized and excited for each round. Then through the updating principle, the task context information of the current round is updated in the long-term memory, formalized as $F(s_i | (Q'_i, R'_i)) \rightarrow (Q_i, R_i)$, where $(Q'_i, R'_i) \in P$.

2) *Overview of the framework*: Given a task query, the main goal is to enable the LLM to generate more accurate decisions based on previous knowledge and experience, while updating the information from the previous afternoon of the task in each round to form a long-term memory to ensure the contextual coherence of the new upcoming query. The proposed VIMBank enables LLM to retain useful historical information during the processing of multiple rounds of tasks. As shown in Fig. 1, VIMBank is a unified mechanism consisting of three core functional modules: (1) memory storage (2) memory retrieval and (3) memory update.

B. memory storage component

Q: Musk is a genius, how old is he now?	
External knowledge memory	Chat Memory
Elon Musk, a recognized genius and disruptive innovator, was born in 1971 and has achieved tremendous success in the fields of technology and business. As of 2023, Musk is 52 years old, but his spirit of innovation and passion for the future seem forever young, just like a young person.	User: I've been thinking recently, Elon Musk can really be called a modern genius. What do you think? Assistant: Absolutely, I completely agree. Musk's achievements in technology innovation and business are indeed extraordinary. User: He must be in his forties now, right? Considering the tremendous achievements he has made in electric vehicles, space exploration, and even artificial intelligence at this age, it's truly impressive. He is a genius.
Task Memory	
<pre>[{ "task_name": "Calculate Elon Musk's Age", "command": {"name": "time_delta", "args": {"start_time": "1971-06-28 00:00:00", "end_time": "2023-11-28 14:24:19" } }], "task_id": 2, "result": "The time difference between 1971-06-28 00:00:00 and 2023-11-28 14:24:19 is: 19146 days, 14:24:19; 52 years 5 months 16 days"]</pre>	

Figure 2. Different types of memory

The memory storage component caches task-related context information during the execution of task instructions by the LLM, which is categorized into three different types of memory banks: knowledge memory, dialog memory, and task-related memory. As shown in Figure 2, the knowledge memory mainly captures and stores task-related data and documents that need to be not trained or outdated by external search engines or LLM pre-training; the dialog memory records the query and response pairs in each round of dialogs between the LLM and the user to ensure the context coherence and consistency of the subsequent dialogs, and the task memory records the task decision-making process of the LLM. After each task instruction is given to the LLM, the LLM performs strategic planning through task planning, tool selection, and observations obtained from the tool, and task-related information is systematically stored in memory for effective utilization in subsequent dialogues. In order to improve the validity of historical information during the process of storing LLM interaction information in the memory bank, an evaluation strategy is introduced to ensure that all historical experiences memorized by the LLM are valid, and at the same time, Prompt is pre-set to activate the ability of the intelligent body to recognize positive and negative sample

experiences, to ensure that the information memorized by the intelligent body is correct as much as possible. Specifically, when memorizing conversational experiences, positive samples refer to real multi-round conversations in which each round is logically interrelated. On the contrary, negative samples refer to pseudo-multi-round conversations, in which there is no logical correlation between the conversation history and the latest task query, and the intelligent body does not need to refer to previous conversational experiences when answering the query. When constructing knowledge-based experiences, knowledge experiences need to synthesize possible conflicting, irrelevant and relevant information in the knowledge. Therefore, retrieved knowledge experiences are defined into three categories: relevant, irrelevant and conflicting information. Relevant knowledge refers to experiences from which the answer to a query can be found directly. For example, if a query is made about someone's age and the retrieved knowledge experience contains personal information about the person and his/her age, this part of the knowledge is relevant knowledge. Irrelevant knowledge experiences are those that are related to the task at hand but do not provide a direct answer. Conflicting type of knowledge refers to the existence of two

contradictory historical experiences about the same task information in the knowledge experience base, for example, the retrieved knowledge contains two different ages of the same person, which the intelligent should be able to distinguish. The memory of task types is similar to external knowledge memory, which is also likely to have three types of memory: relevant, irrelevant, and conflicting, and the LLM's ability to utilize complex historical experience information to better adapt to different tasks is enhanced by this memory recognition method.

For each type of memorized information text, the text is divided into fixed-length segments. These segments are converted into vector representations on the basis of a vector database for subsequent efficient vector-based retrieval using FAISS. The specific formalization is:

First a piece of text is sliced and divided into n segments $T = \{S_1, S_2, \dots, S_N\}$ each of which is of length n . Each segment is of length n . For each text fragment, this research uses CoBERT [10] as the encoder model to pre-code it into a vector representation, and its vector representation can be expressed in Equation 1:

$$v_i = E(S_i) = \{v_i^0, v_i^1, \dots, v_i^m\}, v_i \in \mathbb{R}^d \quad (1)$$

Where E is the function used to convert text segment into vectors, and v_i is the first i vector representation of the text fragment, $i \in (1, m)$. Store all the fragment vectors in a shared vector database for subsequent retrieval.

The vector database as the basis of the whole VIMBank is mapped to the vectorized representation of each text using vector indexing, which is able to capture the semantic information of the text by mapping the text to a high-dimensional vector space, but the vector indexing approach is not able to support fast keyword matching, and when the LLM recalls the long-term memory information, on the one hand, the semantic information is very important, and on the other hand, the precision against the keyword matching to the historical memory information should also be captured. As a result, VIMBank needs to use Elasticsearch technology to map text segments and

store them as sparse vectors, create an index for each keyword, and record all documents containing the keyword to realize fast processing and precise matching of large amounts of text data.

By mixing the two approaches, vector databases map a large number of text segments as dense vectors, and Elasticsearch stores text maps as sparse vectors, and in the subsequent retrieval phase, vector indexes are good at semantic understanding, while Elasticsearch-based backward indexes are good at fast and accurate keyword matching, while utilizing the advantages of both to obtain more comprehensive and accurate retrieval results.

C. Memory Retrieval Component

The quality of the relevance of information retrieved from long-term memory can have a critical impact on LLM reasoning and decision-making in the process of LLM executing reasoning and decision-making, so the retrieval mechanism is a crucial stage in long-term memory.

Based on text memory vectorization and keyword storage, VIMBank adopts FAISS [11] retrieval method for densely embedded vectors and BM25 retrieval model for sparse vectors [12] for these two storage methods respectively. The core idea of FAISS is to generalize, partition, or quantize high-dimensional vectors so as to reduce the search space and improve the retrieval speed, which can be achieved through a combination of various indexing techniques such as inverted files, product quantization, HNSW, and GPU acceleration to achieve efficient retrieval on large-scale datasets. And keyword storage uses the Elasticsearch search engine to represent these text segments as sparse vectors through inverted indexing, of which the BM25 retrieval model is the most commonly used inverted indexing-based retrieval model, which not only considers word frequency and inverse document frequency, but also introduces factors such as lexical item saturation, which better balances the impact of high-frequency words. VIMBank improves the information retrieval speed by integrating the vector retrieval and the keyword retrieval results to improve the overall effectiveness of information retrieval. This process involves merging and de-duplicating the results from the two retrieval strategies. With these two retrieval

techniques, LLM is able to achieve more comprehensive and accurate information recall.

On the other hand, in order to ensure the recall of information related to LLM retrieval and the current task, VIMBank designs different retrieval strategies based on three different types of memory information, namely external knowledge, dialog, and task. Each memory type has its own adapted retrieval mechanism during each round of interaction in LLM.

For knowledge experience, this research wants to increase the number of recall results to access more relevant information. In a vector database, documents and instructions are represented as high-dimensional vectors. Vector databases enable semantic similarity search by storing and retrieving these high-level vectors. The vector similarity of different texts can be defined by approximate nearest search method using cosine similarity distance function, which is implemented as:

$$\text{dist}(q, d) = \frac{q \cdot d}{\|q\| \|d\|} \quad (2)$$

In Equation 2, q is the query vector, d is the document vector, $\|q\|$ is the parameter of vector q and $\|d\|$ is the parameter of vector d . Suppose that $q = [q_1, q_2, \dots, q_n]$ and $d = [d_1, d_2, \dots, d_n]$, then the distance function of cosine similarity is specified as:

$$\text{dist}(q, d) = \frac{\sum_{i=1}^n q_i d_i}{\sqrt{\sum_{i=1}^n q_i^2} \cdot \sqrt{\sum_{i=1}^n d_i^2}} \quad (3)$$

Cosine similarity can measure whether two vectors are in the same direction or not without considering the size of the two vectors, this method focuses more on the relative importance of the words and can maximize the retrieval of relevant document information for similar words. In contrast, knowledge-experience ES retrieval can increase the

number of returned results by tuning the retrieval parameters.

For conversational experience, ensure recall of conversation rounds that are relevant to the query context. The main performance is semantically identical and logically coherent. Therefore, for conversational memory retrieval focuses more on pre-temporal weighting and content semantic similarity. ES retrieval introduces temporal weighting, while semantics can be realized directly based on vectorized retrieval, and the quality of retrieval recall for conversational information needs to consider both kinds of retrieval results comprehensively:

$$S_{call} = \omega \cdot \alpha + (1 - \omega) \cdot s \quad (4)$$

In Equation 4, α and s are ES scores and similarity scores, ω is the time weights, assuming that the current time is T and the timestamp of the dialog round is t , and the timestamp of the conversation round as:

$$\omega = e^{-\lambda(T-t)} \quad (5)$$

For task memory, the focus is on recalling recent task-relevant information with increased temporal weighting. The retrieval strategy is similar to conversational memorization. For example, for the query "What year was Xi'an University of Technology founded?", the information retrieved from the memory bank with different memory types is shown in Figure 3. The background color marked in gray indicates the part of the retrieval process related to the query, while the text marked in red indicates the information related to the text and the answer to the query. As the dialog progresses, the session memory and task memory are continuously updated to ensure that the LLM can effectively handle dynamically changing dialog situations and task requirements. This design not only enhances the flexibility of the LLM, but also improves the efficiency and accuracy of its application in practical tasks.

Q: How many years has Xi'an Technological University been established?	
External knowledge memory	Chat Memory
<p>The school was established in 1955 as a military industrial support project under the country's "First Five-Year Plan" and one of the 156 key construction projects. In 1965, it was renamed Xi'an Institute of Technology, becoming the only undergraduate institution in the ordnance industry in Northwest China. For a long time, the school was managed by the Ministry of Ordnance Industry, and in 1999, it was transferred to the former China North Industries Group Corporation...</p>	<p>Me: Xi'an Technological University truly is a university with a long history. Assistant: Yes, Xi'an Technological University indeed has a long history. It has strong academic strengths in engineering and technology. Me: Do you know how many years it has been since Xi'an Technological University was founded? Assistant: Xi'an Technological University was founded in 1955, so it has been nearly 70 years since its establishment.</p>
Task Memory	
<pre>[{ "task_name": "Check the founding year of Xi'an Technological University.", "command": { "name": "hybrid_search", "args": { "entity": "Xi'an Technological University.", "question": "What is the founding year of Xi'an Technological University?" } }, "task_id": 1, "result": "title: Xi'an Technological University/navigation: A well-known institution in the fields of engineering and technology.\nbody: Xi'an Technological University was established in 1955 and is located in Xi'an, Shaanxi Province, China. The university is a comprehensive institution with a focus on engineering and technology, particularly renowned for its contributions to the defense industry and engineering education.\nurl: https://example.edu/entity/xian_industrial_university/info" }]</pre>	

Figure 3. Retrieval memory

D. Memory Update Component

Through the memory storage and retrieval mechanism, LLM can break through the limitations of the context window and obtain contextually coherent or relevant task information from long-term memory, and the LLM memory capacity is greatly enhanced. However, as time goes by, there may be a large amount of redundant information in the LLM long-term memory, resulting in too much data in the long-term memory, which will greatly affect the retrieval efficiency of the entire VIMBank mechanism, and may even cause LLM decision errors due to outdated and inaccurate information. Therefore, in order to ensure the correctness and efficiency of information in long-term memory, a memory update mechanism needs to be designed.

To address the above two problems, this research was inspired by the forgetting curve theory proposed by Ebbinghaus, and designed the human brain's law of forgetting new things in the updating strategy of LLM long-term memory. According to the forgetting curve theory, clearing those secondary memory segments that occurred long ago and have not been frequently recalled can avoid the LLM memory module from occupying too much memory, and also avoid the influence of old and outdated information on the LLM decision-making. The LLM long-term memory updating strategy is mainly guided by the following

principles: (1) memory forgetting; (2) speed of forgetting; and (3) review effect. Memory forgetting aims to simulate the decline of human memory over time, forgetting speed aims to reflect the rate of forgetting of different frequency and importance of information over time, and the review effect aims to express that the steepness of the forgetting curve can be effectively slowed down when the learning content is regularly reviewed or memorized over and over again to enhance the durability of memory.

The Ebbinghaus forgetting curve can be described by an exponential decay model:

$$R(t) = R_0 \cdot e^{-\frac{t}{\lambda}} \quad (6)$$

In Equation 6, $R(t)$ denotes the memory retention rate at time t , that is, the proportion of information retained, t denotes the time that has elapsed since the information was learned, e is approximately equal to 2.71828. λ denotes the parameter of forgetting rate, which controls the speed of memory decline and is affected by factors such as learning depth and number of repetitions. In order to simplify the process of memory updating. The λ modeled as a discrete value and initialized to 1 at the time of the first memory. when a memory segment is retrieved by the LLM decision-making process will increase the forgetting rate parameter of the segment pair by 1, and will reset to 0, which

makes the segment less likely to be forgotten in the future and thus retained in memory for a longer period of time.

In summary, VIMBank builds a more comprehensive LLM long-term memory mechanism through the entirety of these key components. This mechanism improves the decision quality of LLM while reducing the reasoning cost, providing new possibilities for LLM applications.

IV. EXPERIMENTS

A. Experimental Environment

The experimental environment is shown in Table 1.

TABLE I. EXPERIMENTAL ENVIRONMENT

Experimental Environment	Version
CPU	Intel Core i9-10900K
GPU	NVIDIA Tesla V100 PCIe 32G
Language	Python 3.9
Framework	LangChain

B. Task Set

In order to verify the effectiveness of the long-term memory mechanism of VIMBank, this research conducted experiments on three different task sets, ALFWorld, HotpotQA, and KAgentBench.

The ALFWorld dataset combines task execution in virtual environments and natural language

instruction comprehension, and contains about 8,000 test tasks covering a wide range of scenarios and complex tasks. HotpotQA is a dataset for studying and evaluating complex question-answering tasks, especially multi-step reasoning tasks, and contains 7405 test samples and is commonly used as a benchmarking dataset for testing models on complex problems. KAgentBench is a benchmarking tool for evaluating the capabilities of knowledge-based intelligences, which contains multiple tasks that cover different knowledge-based reasoning capabilities, decision-making capabilities, as well as natural language understanding and generation capabilities. The above three textual task sets are shown in Figure 4:

Examples from different task sets	
ALFWorld	You are in the middle of a room. Looking quickly around you, you see a towelholder 1, a toilet 1, a bathtubbasin 1, a drawer 4, a handtowelholder 2, a drawer 6, a drawer 1, a countertop 1, a sinkbasin, a drawer 2, a drawer 3, a toiletpaperhanger 1, a drawer 5, a handtowelholder 1, a towelholder 2, a sinkbasin 2, and a garbagecan 1. Your task is to: put a clean cloth in bathtubbasin.
HotpotQA	Paragraph A: LostAlone were a British rock band ... consisted of Steven Battelle, Alan Williamson, and Mark Gibson... Paragraph B: Guster is an American alternative rock band ... Founding members Adam Gardner, Ryan Miller, and Brian Rosenworcel began... Q: Did LostAlone and Guster have the same number of members? (yes)
KAgentBench	My friend recommended a few movies to me with the IMDb IDs tt0001702, tt0001856, tt0001856. I'd like to know their rankings and ratings in the most popular movie series, as well as their basic information.

Figure 4. Examples from different task sets

C. Benchmark Evaluation

In order to assess the effectiveness of VIMBank, the evaluation follows the principle of "Unity of knowledge and action", which integrates planning decisions and corresponding actions at each step.

$$S_{plan} = \frac{1}{M} \sum_{j=1}^M \max EM(T_{n,i}, T'_{n,j}) \cdot \Gamma(T_{h,i}, T'_{h,j}) (1 \leq i \leq N) \quad (7)$$

$$S_{action} = \frac{1}{M} \sum_{j=1}^M \max EM(T_{n,i}, T'_{n,j}) \cdot \sum_{k=1}^{K_i} EM(a_{k,i}, a'_{k,j}) \cdot \Gamma(v_{k,i}, v'_{k,j}) \quad (8)$$

In Equation 7 and Equation 8, M is the number of complex tasks decomposed into subtasks by the LLM and N is the number of real decisions made

by the LLM. $T_{n,i}$ is the number of i true decision, and $T'_{n,j}$ is the prediction result of the j -th subtask. Γ is the ROUGE-L evaluation metric function to

measure the similarity between the decision results and the true results. Specifically, ROUGE-L is based on the principle of the longest common subsequence, and measures the recall and precision of the decision results through F1-Score comprehensively.

$$R = \frac{LCS(T_{n,i}, T'_{n,j})}{|T_{n,i}|} \quad (9)$$

$$P = \frac{LCS(T_{n,i}, T'_{n,j})}{|T'_{n,j}|} \quad (10)$$

$$F1 = \frac{2 \times R \times P}{R + P} \quad (11)$$

The calculation based on the longest common subsequence can well capture the order information between sequences and is suitable for measuring the similarity between the generated sequences and the desired results in the decision-making process.

EM (Exact Match) is another evaluation metric to assess the exact match between the generated text and the reference text. When evaluated using EM, it tends to 1 if the generated text (R) and the

reference text (G) are basically the same, and tends to 0 if most of the text is different from the reference text, which can be expressed as:

$$EM(R, G) = \begin{cases} 1, & \text{if } R \approx G \\ 0, & \text{if } R \neq G \end{cases} \quad (12)$$

In order to comprehensively assess the performance of LLM in decision making and execution, a penalty factor is introduced $p \in \{0, 1\}$, the comprehensive assessment formula is as follows:

$$S_{total} = (1 - p) \cdot (0.7 \cdot S_{plan} + 0.3 \cdot S_{action}) \quad (13)$$

D. Results

In order to evaluate the effectiveness of VIMBank in improving the decision quality and reducing the reasoning cost on multi-tasks, this research replicates the performance of some intelligences on task datasets based on some open source LLMs and compares VIMBank with the replication results. The performance of various LLMs on task reasoning decisions is shown in Table 2.

TABLE II. EXPERIMENTAL RESULTS ON VARIOUS DATASETS

DataSet	Model	NoAgent	ReAct	InterAct	VIMBank
ALFWorld	Qwen2-7b	48.8	54.7	60.1	72.3
	ChatGLM3	46.2	49.2	55.8	64.9
HotpotQA	Qwen2-7b	51.6	57.3	63.4	76.3
	ChatGLM3	45.9	51.8	59.7	71.5
KAgentBench	Qwen2-7b	34.2	48.5	52.6	58.7
	ChatGLM3	32.6	44.7	46.3	54.2

The experimental results show that compared to ReAct and InterAct, which incorporate short-term memory, the LLM with the introduction of VIMBank's long-term memory mechanism improves the quality of decision making on different tasks, which is still a gap compared to the original paper that uses a closed-source LLM such as ChatGPT, which may be due to the overall poorer capability of the open-source models compared to ChatGPT (Qwen2-7b and ChatGLM) are less capable overall. In order to analyze the

effectiveness of long-term memory in enhancing LLM decision-making, experiment observed the ALFWorld task environment when LLM performs the pick 2 task during which LLM needs to find two identical items, such as "find two books and put them in bookshelf. "Since only short-term memory can forget the previous position due to the limitation of the context window, which leads to task failure, especially in the case of performing the same task in the same environment, it is more likely to cause inefficiency in task execution, and the

advantage of long-term memory mechanism in the similar task environment comes to the forefront. token test on different LLMs, with NoAgent as the benchmark, this research recorded the token consumption of LLMs with the process of multiple rounds of tasks, which effectively proved the effectiveness of VIMBank in reducing the cost of LLMs' reasoning for multiple rounds of tasks, as shown in Table 3.

TABLE III. REASONING COST OF ALFWORLD ENVIRONMENT

	200	600	1000
NoAgent	63.2K	164.7K	334.7K
VIMBank	56.8K	142.6K	258.3K

In addition, this research observes the reasoning trajectories of different intelligences in the ALFWorld environment as shown in Fig. 3, which is analyzed to show that the introduction of the long-term memory mechanism can guide the new reasoning steps through previous experiences, thus avoiding repeated reasoning and improving the retrieval efficiency to a large extent.

V. CONCLUSIONS

In this paper, this research presents a new long-term memory mechanism, VIMBank, to enhance LLM context-awareness without model fine-tuning, to improve decision quality, and to reduce the inference cost of similar tasks with the help of long-term memory. This research evaluates this mechanism in a variety of different task environments, and it significantly improves in decision accuracy compared to some baseline models, by more than 10% compared to the InterAct model built on the ReAct model, and by more than 20% compared to the NoAgent. Meanwhile, in terms of LLM reasoning cost, NoAgent reasoning cost basically grows in a linear trend, while our VIMBank mechanism reduces the

reasoning cost by nearly 25% as the task is executed, through the accumulation of long-term memories and specific retrieval and update mechanisms. This highlights the great potential of VIMBank in LLM-driven AI systems.

REFERENCES

- [1] SHRIDHAR M, YUAN X, CÔTÉ M A, et al. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning[J]. 2021. arXiv:2010.03768.
- [2] YANG Z, QI P, ZHANG S, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering[C]Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium. 2018:2369-2380.
- [3] PAN H, ZHAI Z, YUAN H, et al. KwaiAgents: generalized information-seeking agent system with Large Language Models[J]. 2023. arXiv:2312.04889.
- [4] YAO S, ZHAO J, YU D, et al. ReAct: Synergizing Reasoning and Acting in Language Models [J]. 2023. arXiv:2210.03629.
- [5] CHEN P L, CHANG C S. InterAct: Exploring the Potentials of ChatGPT as a Cooperative Agent [J]. 2023. arXiv:2308.01552.
- [6] WANG Y, LI P, SUN M, et al. Self-Knowledge Guided Retrieval Augmentation for Large Language Models [J]. 2023. arXiv:2310.05002.
- [7] SUN J, XU C, TANG L, et al. Think-On-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph [J]. 2023. arXiv:2307.07697.
- [8] SHEN Y. Think-in-Memory: Recalling and Post-Thinking Enable LLMs with Long-Term Memory [J]. 2023. arXiv:2311.08719.
- [9] LU J, AN S, LIN M, et al. MemoChat: Tuning LLMs to Use Memos for Consistent Long-Range Open-Domain Conversation [J]. 2023. arXiv:2308.08239.
- [10] KHATTAB O, ZAHARIA M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT[J]. SIGIR, 2020:39-48.
- [11] JOHNSON J, DOUZE M, JEGOU H. Billion-scale similarity search with GPUs [J]. IEEE Transactions on Big Data, 2021: 535-547.
- [12] AKLOUCHE B, BOUNHAS I, SLIMANI Y. BM25 Beyond Query-Document Similarity[M/OL]/Lecture Notes in Computer Science. String Processing and Information Retrieval. 2019: 65-79.

Improved Pedestrian Vehicle Detection for Small Objects Based on Attention Mechanism

Yanpeng Hao

Xi'an University of Technology
School of Computer Science and Engineering
Xi'an, China
E-mail: hnrnk@163.com

Chaoyang Geng

Xi'an University of Technology
School of Computer Science and Engineering
Xi'an, China
E-mail: 541211200@qq.com

Abstract—This study aims to solve the low detection accuracy and susceptibility to false detection and omission in pedestrian and vehicle detection by proposing an improved YOLOv5s algorithm. Firstly, a small target detection module is added to better acquire and determine the information of pedestrians from long-range vehicles. Secondly, the multi-scale channel attention CBAM attention module is added, and the dual attention mechanism is not only flexible and convenient, but also improves the computational efficiency. Finally, the MPDIoU loss function based on minimum point distance is introduced to replace the original GIoU loss function, and this change not only enhances the regression accuracy of the model. At the same time, the convergence speed of the model is accelerated. KITTI data set was used for experiments, and the experimental results showed that the average accuracy of the model trained by the improved YOLOv5s algorithm on the data set reached 84.9%, which was 3.7% higher than that of the original YOLOv5s algorithm. It is verified that the model is suitable for high accuracy of pedestrian and vehicle recognition in complex environments, and has high value for promotion.

Keywords—Deep Learning; Small Target Detection; CBAM; MPDIoU; Vehicle Pedestrian Detection

I. INTRODUCTION

Along with the continuous progress of computer technology, the field of computer vision is rapidly rising, and the detection of pedestrians and vehicles is becoming more and more critical in many real-life scenarios [1]. The detection of pedestrians and vehicles is a basic task in the field of computer vision, which has a wide range of applications in the industries of automatic driving, intelligent transportation, video surveillance and human flow

analysis [2]. Along with the increase of urban population and the rapid prosperity of finance, the need for pedestrian detection and vehicle detection is increasing day by day. Pedestrian and vehicle detection is a key aid in reducing traffic accidents and alerting drivers. Target detection plays a key role in traffic management, surveillance security and the development of smart cities. Despite the maturity of existing technologies, missed and false detections still occur in the field of vehicle pedestrian detection, especially due to the relatively small size of the pedestrians and their frequent occlusion. Therefore, it is urgent and important to overcome this challenge by continuously optimising the performance of small target detection.

The YOLO algorithm proposed by Redmon et al. is a regression-centred algorithm, which converts the target detection problem into a regression problem by partitioning the target into a grid, an innovation that enables target detection. On the basis of YOLO, Redmon team further developed YOLOv2 version [3]. YOLOv3, which adopts Darknet-19 as a solid foundation and incorporates the residual module, significantly improving the performance of feature extraction [21]. YOLOv3, using Darknet-19 as a solid foundation and incorporating the residual module, significantly improves the performance of feature extraction. The logistic regression improvement layer is used to improve the accuracy of multi-label classification. In order to comprehensively acquire the multi-scale features of the target, the feature

pyramid (FPN) concept is added, which fuses the semantic depth of the high level of the image with the detailed visual information of the low level, which in turn improves the comprehensiveness of the target detection. While Bochkovskiy et al. inherited and developed YOLOv3, the innovative YOLOv4 is the refinement and enhancement of the previous algorithm. YOLOv4 algorithm combines the Cross-Stage Partial Network (CSPNet) with Darknet-53 as the backbone, which enhances its deep feature extraction function, and introduces the SPP (Spatial Pyramid Pooling) to deal with different scales of physical information, and the PANet.

Researchers Zhiyong Ju et. al [6]. innovatively designed Vit-YOLOv4 , a model that incorporates the Transformer architecture and deeply separable convolutional techniques to improve detection accuracy. Zihao Jia et al [22]. added a lightweight sub-pixel convolutional layer in front of the detection layer of the YOLOv5 model, which significantly improved the inference speed of the model despite sacrificing some of the detection accuracy in the lightweighting process. Although the above studies have partially contributed to the field of pedestrian detection, however, the challenge still exists when dealing with distant and dense scenes, and the problem of false and missed detections still needs to be overcome [7].

In this paper, YOLOv5s is chosen as the base model, and the model is lightweighted by improving the convolutional block, introducing a small target detection layer and a multi-scale channel attention mechanism CBAM. MPDIoU loss function is used to optimize the model training, and experiments are carried out on the constructed vehicle and pedestrian data sets. The experimental data prove that the improved YOLOv5s algorithm has improved the accuracy and practicality.

II. INTRODUCTION TO THE YOLOv5 ALGORITHM

YOLOv5 provides a series of models of different sizes, which can be classified into five structures: n, s, m, l, and x, based on the network depth and model size of the network model [8]. The depth of these five models ranges from shallow to deep, and the speed of detection ranges from fast to

slow, so users can choose the appropriate model according to the application scenario. In this paper, for vehicle pedestrian detection, we need to ensure the real-time detection, so we choose YOLOv5s, which has a better balance between accuracy and speed, as the base model [9]. Figure 1 below shows the algorithm flow of YOLOv5s.

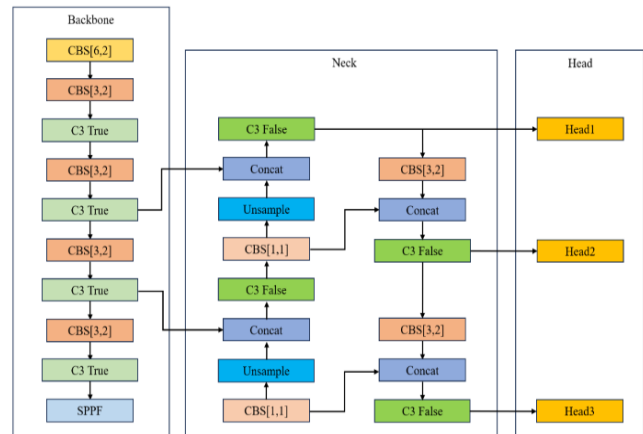


Figure 1. Flowchart of YOLOv5s algorithm

As a lightweight model of the YOLOv5 family, the YOLOv5s is designed to significantly reduce the need for computing resources while maintaining high detection accuracy, making it particularly suitable for resource-constrained environments. Its design is exquisite, thanks to the powerful function of CSPDarknet53, it can efficiently extract features and traverse the multi-level information of the input image [10]. YOLOv5s uses a multi-scale prediction strategy to process three feature maps with different resolutions, ensuring excellent detection performance regardless of the size of the target. More innovative is that it introduces an adaptive anchor frame mechanism to dynamically generate a frame that matches the target size, improving the detection accuracy [11]. In addition, the model incorporates training optimization methods such as data enhancement and more refined learning rate adjustment, and the integration of these strategies significantly improves the overall performance of YOLOv5s in various detection tasks.

YOLOv5s is designed with a hierarchical architecture, including an input preprocessing module; a powerful Backbone network, through which the core features are extracted; a feature

fusion neck network, which acts as a connecting point for feature fusion and effectively integrates the features at different levels; and finally an output prediction Head, which is responsible for generating the prediction results [12].

The core feature extraction network is a key component of the main network, and the feature extraction module is carefully designed, which combines the C3 structure module, the CBS convolution block and the SPPF spatial pyramid pooling module. The CBS module is the basic architecture, which employs traditional convolutional layers and selects the SiLU nonlinear activation function to enhance the expressive capability [13]. The C3 module introduces residual connectivity, which significantly improves the deep learning effect of the model; while the SPPF module is upgraded from the SPP module, which uses the stacked connection of small-sized pooling kernels of the same size to construct receptive fields of different sizes, which are used to obtain more detailed size of object information

The basic idea of the neck link network is to organically integrate the characteristic pyramid network and the route network, which requires effective integration of multidimensional characteristics in order to improve the accuracy and reliability of vehicle and pedestrian detection. Feature pyramid network structure transmits deep semantic content to shallow features in a top-down manner, while the path aggregation network structure transmits shallow location content to deep features. This mutually complementary structural design achieves multi-scale feature fusion by integrating semantic and location information in the feature map, through which the feature map incorporates rich semantic and precise location information, thus improving the detection accuracy and depth of feature maps at different scales [14].

The head prediction network performs classification and regression operations on small, medium and large targets respectively to achieve the prediction of target category and location. YOLOv5s is a popular target detection algorithm, which has stable detection effect on many datasets [15]. However, there is still room for improvement in detection accuracy when facing small target detection and target detection in complex situations

[16]. In this study, it is intended to make improvement and optimization so that YOLOv5s can better fulfill the detection task in various missions, so as to improve the detection accuracy when facing small targets and complex situations [17].

By integrating Mosaic data enhancement technology, YOLOv5s enhances the diversity of training data, thus significantly enhancing the generalization performance of the model. The adaptive anchor frame strategy optimizes the detection process and automatically adjusts to the characteristics of the training data, thereby improving the detection accuracy. At the same time, the model can flexibly adapt to the size change of the target to ensure that the target can be accurately and stably detected at different scales. Within the system, a smart combination of Mish focusing and activation functions optimizes model performance. The use of non-maximum suppression techniques can effectively eliminate redetection and ensure the accuracy of the search results. This design structure enables YOLOv5s to achieve a high level of performance and efficiency [18].

III. IMPROVEMENT OF YOLOv5s

A. Small Target Detection

The Head of the YOLOv5 network uses three detection modules at different scales. In real-world scenarios, distant pedestrian and vehicle targets often appear relatively tiny in videos and pictures, and their visible visual size is usually extremely limited, reflecting the extremely small image size. Since the features of smaller targets are not easy to identify, the situation of missed detection and false detection often occurs in the detection process. Facing the problem of missed detection and misdetection due to small target recognition, we adopt the YOLOv5 model and introduce an additional high-resolution detection layer, i.e., Level P1 160 by 160 is designed to improve the accuracy of detection of small objects. The key mission of P1 is to introduce high-resolution feature mapping, which aims to enhance the ability to accurately capture and locate remote pedestrian vehicles in detail. Pedestrians at long distances tend to present smaller sizes in images and videos, and traditional algorithms are prone to miss these

targets, leading to missed detections. The addition of the P1 layer enables the network to better handle the detection of small target pedestrians, better capture and localise the detailed information of long-distance pedestrians, and reduce the occurrence of misdetection and missed detection.

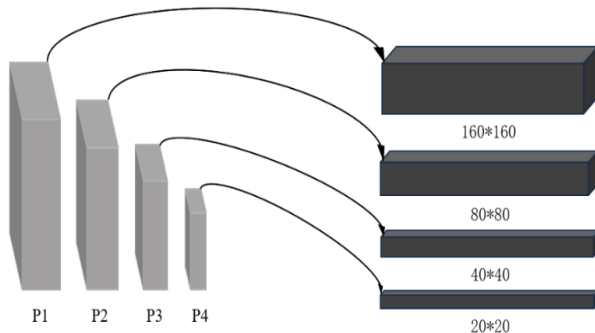


Figure 2. Multi-scale detection structure

This study proposes to introduce a strategy to achieve efficient detection of complex targets by embedding an additional fourth module at the head of the detection layer without extending the classical neck network architecture. This strategy avoids the information loss associated with increasing the depth of the network, while maintaining the savings in computational resources [19].

B. Hybrid Attention Mechanism CBAM

Inspired by human cognitive systems, the attention mechanism enables models to better understand the associations between different locations, and it can highlight important details, providing significant support for deep learning models [20]. Notably, the same attention design can be adapted to handle different data patterns and can be easily integrated into large network models. In addition to this, multiple complementary attention mechanisms can be integrated into the same network.

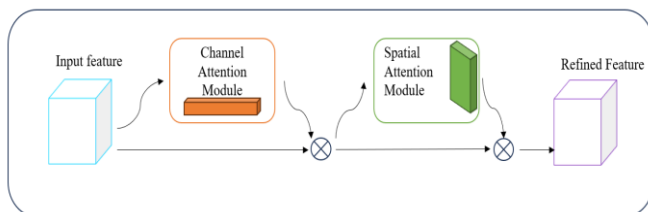


Figure 3. Structure of CBAM module

CBAM (Convolutional Block Attention Module) is an effective and relatively simple attentional mechanism that can be easily integrated into most well-known CNN frameworks, and the structure of the CBAM module is shown in Figure 3. CBAM can perform adaptive function cleaning by having specific function cards flow from two parallel directions of channel and space, respectively, and then multiply function cards to perform adaptive function cleaning. Integration of CBAM into different models has resulted in significant performance improvements on various classification and detection datasets [21]. In vehicle detection, since a large detection area usually contains multiple complex information, using CBAM allows focusing on a part of the area, which helps the network model to resist the ponderous information and focus its core attention on the beneficial objects.

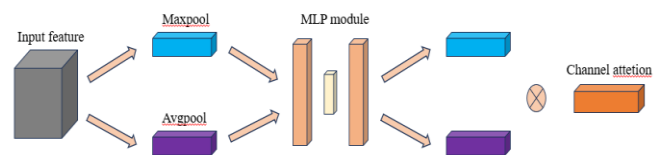


Figure 4. Channel Module

As can be seen from Fig. 4 above, in Channel Attention, the spatial information of the feature maps is integrated, and subsequently, the feature maps after fusing the features are put into it. The maximum maxpool level and the average avgpool level produce two different spatial descriptor contexts, respectively, to obtain a two-element map $1 \times 1 \times c$ [22]. Both cards are connected to an artificial neural network, MLP (Multilayer Perceptron), to increase the number of feature channels of the original feature maps, where the MLP has one hidden layer, the number of neurons in the first stage c/r (R-reduction rate) [23]. The number of neurons in the second level is c . This design controls the model complexity to some extent and avoids overfitting. complexity, avoids overfitting, and efficiently extracts higher-order representations of the input features.

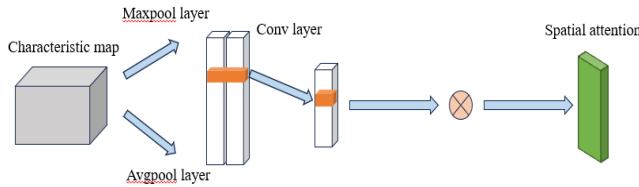


Figure 5. Space module

As shown in Figure 5, in Spatial Attention, maximum pooling on the channel dimension is performed on the feature map to find a maximum value at all positions in the feature map, which in turn produces a completely new feature map. At the same time, this study used the average pooling technique on the channel dimension to produce a new feature map containing aggregated information by averaging the channels over the feature maps at all positions. This processing can better preserve the feature information. The topic map is constrained by a maximum average pool based on channel orientation, and two output maps are superimposed on the channel to format the mapping of elements containing two pools of information.

C. Loss Function (Loss Function) Improvement

In deep learning and machine learning, the loss function plays a non-trivial role in judging the degree of difference between the model's predictions and the true labels. In the vast majority of tasks, choosing the appropriate loss function can directly affect the convergence speed and final performance of the model. By optimising the loss function, the optimal solutions of the model parameters are clearly identified, significantly improving the accuracy of the predictions. During the training phase, model variables such as neural network weights are uninterruptedly adjusted to reduce the loss function values.

In earlier versions of the YOLO series, a loss function dependent on the mean square error was constructed using the position of the centroid of the predicted frame with respect to the real frame and dimensional information (width and height).

In the case of mean square error as a loss function, the centroid coordinates and edge length information are treated as independent variables, however, in essence they are not independent and are related to each other. Therefore, the mean

square error is not a good metric to represent the intersection and integration ratio (IOU) relationship between the bounding boxes [24], as in equation (1).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

The intersection and concatenation schematic of the IoU prediction frame and the real frame is shown in Figure 6.



Figure 6. Schematic diagram of intersection and concatenation of IoU prediction and real frames.

In YOLOv5, the GIoU_Loss loss function is applied instead of the traditional IoU. The GIoU takes into account the relative positional relationship between the predicted box and the real box, which not only focuses on the overlapping part, but also measures the difference in the outer join matrix between the two. The optimisation of the cost function is enhanced by adding a penalty term which calculates the difference between the concatenation and closure of the two boxes, the smaller the difference the smaller the penalty term. The GIoU loss function is calculated as in equation (2).

$$GIoU = IoU \cdot \frac{|C / A \cup B|}{|C|} \tag{2}$$

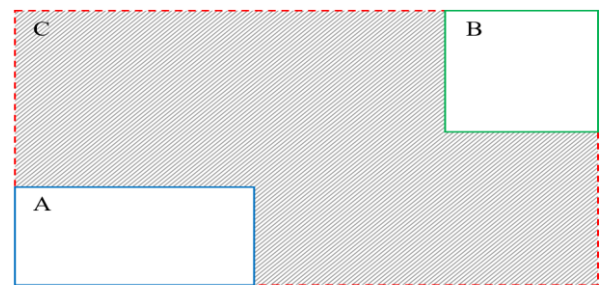


Figure 7. GIoU penalty content for minimising the area of the shaded region

Figure 7 shows the schematic diagram of the GIoU principle, where C is the area of the smallest outer rectangle, A is Ground Truth, and B is Bounding Box.

If there is an inclusion relationship between two bounding boxes, then the GIoU becomes IoU, which cannot accurately express the relative positional relationship between them. The GIoU is also heavily dependent on the IoU, and when the two boxes intersect, horizontal and vertical errors are large, convergence is slow, and do not accurately reflect the size of the overlap between the two squares [25].

The MPDIoU loss function was proposed by MA S L et al. proposed in July 2023, achieved good performance when training the target detection model on the PASCAL VOC dataset [26]. The loss function calculates similarity by calculating the minimum deviation between the desired image and the actual image, that is, taking into account the distance of the overlap area, the center distance, and the width. Through optimisation, the model can converge faster and get more accurate prediction results [27]. In order to enhance the training effect, accelerate the convergence speed of the model and improve the regression accuracy, choose MPDIoU as a new loss function. The MPDIoU calculation principle shown in figure 8 simplifies the calculation process by minimizing the distance between the upper left and lower right corners of the prediction and control cells.

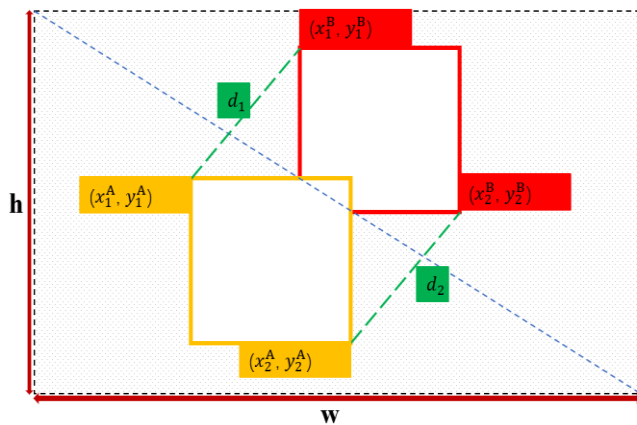


Figure 8. Computational schematic of MPDIoU

The main calculations are shown in Eqs. (3) to (5). d_1 and d_2 signify the distances that separate the

upper left corner from the lower right corner of the predicted frame B and its corresponding real frame A. The (x_1^A, y_1^A) and (x_2^A, y_2^A) are the coordinates of the upper-left and lower-right corners of the real frame, and the boundary of the prediction frame is described by two points: the upper-left corner (x_1^B, y_1^B) coordinates and the lower right corner (x_2^B, y_2^B) coordinates, adopting L_{MPDIoU} as an evaluation metric, which measures the overlap between the predicted box and the real box. The width and height of the input image are denoted as w and h , respectively.

$$d_1^2 = (x_1^B - x_1^A)^2 + (y_1^B - y_1^A)^2 \quad (3)$$

$$d_2^2 = (x_2^B - x_2^A)^2 + (y_2^B - y_2^A)^2 \quad (4)$$

$$L_{MPDIoU} = 1 - \left(\frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \right) \quad (5)$$

IV. EXPERIMENTAL CONFIGURATION AND ANALYSIS OF RESULTS

A. Experimental environment

The operating system trial environment is Windows 11 (64-bit), running on 32GB of RAM, the graphics driver is RTX2070 SUPER, and it is powered by a 12th generation Intel processor i5-12400F.

B. Data sets and evaluation indicators

This study relies on the KITTI dataset, which was created jointly by the Karlsruhe Institute of Technology, Germany, and the Toyota Technological Institute, Chicago, USA. The dataset scenarios are collected from a variety of complex road environments such as rural, motorway and urban from different perspectives and time periods [28]. Especially in vehicle and pedestrian detection tasks are considered important benchmarks for measuring the performance of the technology. 9200 images were selected from the preprocessed dataset for the experiments, dividing the training, validation and test sets in a ratio of 7:1:2.

When evaluating the performance of a model, it is common to rely on two core metrics, precision (P)

and recall (R), which together reveal the accuracy and completeness of the model. (6) to (8) are the calculation formulas used in this study.

$$P = \frac{T_p}{T_p + F_p} \quad (6)$$

$$R = \frac{T_p}{T_p + F_N} \quad (7)$$

$$mAP = \frac{\sum_{i=1}^N P_{A_i}}{N} \quad (8)$$

T_p represents the number of instances that are accurately identified as positive; F_p represents the

number of instances that are actually negative but are judged to be positive; F_N represents those instances that were actually positive but were misidentified as negative; N is the number of categories of samples in the data set.

C. Impact of different attention mechanisms on the model

In order to investigate the effect of multiple attention mechanisms on the model performance, the C3 module was replaced with ECA, SE, CCA, SA-Net, MS-CAM, CBAM. The results of the experiments with six attention mechanisms are shown in Tables 1 and 2.

TABLE I. EFFECT OF INTERNAL PARAMETERS ON THE MODEL

Batch Size	Average accuracy	accuracy	recall rate	confidence level
	Mean average precision	Precision P/%	Recall	(math.)
	mAP percent		R/%	Confidence/%
10	87.4	92.0	96.1	86.0
13	88.3	93.4	96.0	84.0
16	88.7	93.7	95.0	84.0
18	89.4	94.3	95.9	82.0
20	90.3	95.2	95.7	86.0

TABLE II. INCORPORATION OF MULTIPLE ATTENTION MECHANISMS

Attention Mechanism	mAP%	P/%	R/%
+SE	85.3	91.8	95.0
+ECA	86.6	92.3	94.1
+CCA	86.4	92.0	94.6
+SA-Net	87.8	92.5	94.7
+MS-CAM	87.5	92.7	95.2
+CBAM	88.5	93.2	95.8

The loss function is improved by adding a small target detection layer and introducing an attention mechanism. From the experimental data in Table 2, it can be seen that after the introduction of the

CBAM attention module, the precision and recall of the model show an improvement, which effectively improves the performance metrics compared to other attention mechanisms.

D. Experimental result

The model of this article has been compared with other universal algorithms on Kitty recyclable data sets, which count as shown in table 3.

Compared with the benchmark model YOLOv5s, mAP@0.5 increases by 4.0%. Although the volume of the improved algorithm increases by 1.7MB, and the detection speed is reduced, but it still meets the real-time requirement.

TABLE III. COMPARISON OF DIFFERENT ALGORITHMS

Model	Volume	mAP@0.5%
YOLOv5s	14.0	86.9
YOLOv8s	22.4	86.5
YOLOv6s	37.4	83.0
YOLOv4	245.9	79.5
SSD	100.3	71.0
YOLOv5l	93.7	89.7
ours	15.7	90.9

E. Ablation experiments

In the study, some cutting-edge technologies were added to the YOLOv5s model, and three improvement points were proposed, namely, improvement of small target detection, denoted by X in Table 4, addition of the multi-scale channel attention mechanism CBAM, denoted by C in Table 4, and improvement of the MPDIoU loss function, denoted by L in Table 4, and denoted by ticks for whether it is added or not, resulting in an advanced detector, naming it XCL-YOLO. the improvement points are added to the YOLOv5s model to carry out the experiments, respectively.

According to the data in Table 4, the accuracy of each optimisation point for the detection of small targets has been improved to a different degree, and the improvement is significant. After the introduction of the small target detection module, there is an improvement of 3.1% on Person class and 1.4% on mAP. After adding the multi-scale channel attention mechanism CBAM, the improvement is 2.6% on Person class and 1.6% on mAP. Finally, improving the loss function from GIoU to MPDIoU improved 3.3% on Person class and 2.1% on mAP. A comparison of the algorithm before and after improvement is shown in Figure 9.

TABLE IV. ABLATION EXPERIMENTS

Methods	X	C	L	Person/%	Car/%	mAP%
YOLOv5s				79.2	94.3	86.9
X-YOLOv5s	√			82.3	94.4	88.3
C-YOLOv5s		√		81.8	95.2	88.5
L-YOLOv5s			√	82.5	95.5	89.0
XC-YOLOv5s	√	√		83.1	96.2	89.6
XL-YOLOv5s		√	√	83.4	96.6	90.0
CL-YOLOv5s	√		√	83.9	96.3	90.1
XCL-YOLOv5s	√	√	√	84.7	97.0	90.9



Figure 9. Comparison between before and after algorithm improvement

V. CONCLUSIONS

In this paper, an improved XCL-YOLO model algorithm is proposed, especially for the occlusion and small object detection problems in complex environments, and focuses on solving the object and missing detection problems often faced by existing models. The improved algorithm adds a small target detection layer, increases the attention mechanism, replaces the loss function, and adds the multi-scale channel attention mechanism CBAM. These changes adapt to the relationships and elements at different levels and on different channels, thus enhancing its ability to recognise targets and improve tracking accuracy. Recognition rate of vehicles and pedestrians in complex background is improved, and the approximation degree between boundary boxes can be accurately considered. The model can converge better in the training process, and the average accuracy mAP is increased by 1.4%, 1.6% and 2.1% respectively.

The new XCL-YOLO has been trained and tested on the KITTI dataset with an improved overall detection accuracy of 4.0% compared to the original model. The detection speed decreases slightly, but it still meets the detection requirements. This improved algorithm is more suitable for vehicle and pedestrian detection tasks. There will be overfitting risk in the training process, and future research can start from how to reduce the overfitting phenomenon and improve the generalization ability of the algorithm on various data sets. In addition, with the advancement of related technologies, we expect that such

algorithms can be more widely used in the field of vision.

REFERENCES

- [1] XU Yanwei, LI Jun, DONG Yuanfang, et al. A review of YOLO series target detection algorithms[J/OL]. Computer Science and Exploration, 1-19[2024-07-30].
- [2] LI Yunpeng, HOU Lingyan, WANG Chao. Motion target detection in autonomous driving based on YOLOv3 [J]. Computer Engineering and Design, 2019, 40(04):1139-1144.
- [3] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 6517-6525
- [4] Lv Wo-Feng, Lu Hua-Cai. Research on traffic sign recognition technology based on YOLOv5 algorithm [J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(10):137-144.
- [5] ZHAO H, FENG Y B. Research on traffic sign detection based on CGS-Ghost YOLO [J]. Computer Engineering, 2023, 49(12):194-204.
- [6] JU Zhiyong, LI Yuming, XUE Yongjie, et al. Pedestrian detection algorithm based on improved YOLOv4 model [J]. Control Engineering, 2023, 30(10):1912-1926.
- [7] SHAO Yanhua, ZHANG Duo, CHU Hongyu, et al. A review of YOLO target detection based on deep learning [J]. Journal of Electronics and Information, 2022, 44(10):3697-3708.
- [8] ZHOU Miaosen, TANG Quanwu, SHI Sweet, et al. Crack detection algorithm for railway track surface based on improved YOLOv5s [J]. Liquid Crystal and Display, 2023, 38(05):666-679.
- [9] 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.
- [10] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
- [11] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. the

- pas- cal visual object classes (VOC) challenge. *int. j. Comput. Vis.*, 88(2):303-338, 2010.
- [12] Sung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal loss for dense object detection," in *Proc. ICCV*, 2017, pp. 2980-2988.
- [13] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection [J]. *arXiv preprint arXiv:2004.10934*, 2020.
- [14] ZHANG Y F, REN W Q, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. *Neurocomputing*, 2022, 506(9):146-157.
- [15] LIU Hui, LIU Xinman, LIU Dadong. Optimisation of YOLOv5 algorithm for complex road target detection [J]. *Computer Engineering and Applications*, 2023, 59(18):207-217.
- [16] YANG Feng, DING Zhitong, XING Mengmeng, et al. A review of improved target detection algorithms for deep learning [J]. *Computer Engineering and Applications*, 2023, 59(11):1-15.
- [17] Heng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.
- [18] Yang G, Feng W, Jin J, et al. Face mask recognition system with YOLOV5 based on image recognition[C]//2020 IEEE 6th International Conference on Computer and Communications (ICCC). IEEE, 2020: 1398-1404.
- [19] LI Xiang, HE Miao, LUO Haibo. An improved YOLOv3 algorithm for occluded pedestrian detection [J]. *Journal of Optics*, 2022, 42(14):160-169.
- [20] ZHANG Y F, REN W Q, ZHANG Z, et al. Focal and efficient IOU loss for accurate bounding box regression [J]. *Neurocomputing*, 2022, 506(9):146-157.
- [21] CHEN Jianzhu, WANG Yue, ZHU Xiaofei, et al. Wildlife video target detection method by fusing multi-feature maps [J]. *Computer Engineering and Application*, 2020, 56(07):221-227.
- [22] JIA Zihao, WANG Wenqing, Liu Guangcan. Improved Light-weight Traffic Sign Detection Algorithm of YOLOv5 [J]. *Journal of Data Acquisition and Processing*, 2023, 38(6):1434-1444.
- [23] MA S L, XU Y M. MPDIoU: A Loss for efficient and accurate bounding box regression[J]. *arXiv preprint arXiv:2307.07662*, 2023.
- [24] XU X, ZHANG X, ZHANG T. Lite-YOLOv5: A lightweight deep learning detector for on-board ship detection in large-scene sentinel-1 sar images [J]. *Remote Sensing*, 2022, 14(4):1018
- [25] ZHAO Lulu, WANG Xueying, ZHANG Yi, et al. Research on vehicle target detection technology based on YOLOv5s fusion SENet [J]. *Journal of Graphics*, 2022, 43(05):776-782.
- [26] LI R, WU Y. Improved YOLOv5 wheat ear detection algorithm based on attention mechanism [J]. *Electronics*, 2022, 11(11):1673
- [27] Sun Z, Li P, Meng Q, et al. An improved YOLOv5 method to detect tailings ponds from high-resolution remote sensing images [J]. *Remote Sensing*, 2023, 15(7): 1796.
- [28] LIU Jiaye, WANG Chao, SHENG Long. Research on pedestrian detection method based on YOLOv5 [J]. *Computer and Information Technology*, 2024, 32(01):37-41.