Research on Vehicle and Pedestrian Detection Based on Improved RT-DETR

Jingshu LI
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, China
E-mail: lijs812@163.com

Jianguo Wang
State and Provincial Joint Engineering Lab. of Advanced Network, Monitoring and Control
Xi'an Technological University
Xi'an, 710021, China
E-mail: wjg_xit@126.com

***Abstract*—This paper proposes a vehicle and pedestrian detection model based on an improved RT-DETR to address the issues of high redundancy in feature extraction and insufficient accuracy for small targets in existing real-time detection models, especially in complicated traffic scenarios. The core of this improved model is to embed a parameter free SimAM (Simple Attention Module) attention mechanism in the backbone network. The SimAM mechanism dynamically generates three-dimensional attention weights through energy functions, effectively enhancing the expression ability of fine-grained features of pedestrians and vehicles. This improvement not only reduces redundant information in the feature extraction process, but also improves the detection accuracy of the model for small targets, enabling the model to more accurately identify and locate small targets when dealing with complex traffic scenes. The experimental results show that on the BDD100K dataset, the improved model achieved an average precision of 73.6%, which is 3.7 percentage points higher than the original RT-DETR, effectively enhancing the model's capability to detect vehicles and pedestrians in intricate environments.**

Keywords-Object Detection; RT-DETR; Attention Mechanism; Autonomous Driving

I. INTRODUCTION

Today, the technology for detecting vehicles and pedestrians stands as a key component in multiple domains. Especially in the realm of autonomous driving, precisely and swiftly recognizing vehicles and pedestrians is fundamental to guaranteeing the safety and reliability of self-driving cars. However, real-world traffic scenarios have brought many challenges to detection technology. For example, frequent occurrences of mutual occlusion between vehicles and partial occlusion of pedestrians by roadside obstacles make detection algorithms prone to missed or false detections. In addition, under low light conditions, such as night or rainstorm weather, the image clarity and contrast will be significantly reduced, which undoubtedly brings great challenges to the detection task.

The fast-paced growth of deep learning has catalyzed substantial advancements in vehicle and pedestrian detection methods. Early detection techniques were largely facilitated by manually designed characteristics and conventional machine learning algorithms like SIFT, HOG, etc. However, the effectiveness of these methods was not ideal and there were many limitations. Subsequently, deep learning based object detection methods gradually gained prominence, which are broadly classified into single-stage object detectors and two-stage object detectors. Single stage object detectors, such as YOLO series [1], SSD [2], RetinaNet [3], directly forecast the location and type of the target on the input image without the need for complex candidate region generation steps, thus having high detection accuracy. Two stage object detectors like Fast R-CNN [4], Cascade R-CNN, CBNet, etc., they begin by creating potential regions, followed by classification and bounding box regression for each one. Although the detection accuracy is high, the process tends to be slow. In these detection methods, a large number of anchor boxes are generated during the detection phase, however, for an object, only one detection box is actually needed to represent it. Therefore, it is necessary to discard overlapping detection boxes through Non Maximum Suppression (NMS) methods to guarantee that each target is identified by a single box. In addition, intricate parameter tuning is

necessary during network training to optimize detection performance.

In view of this, researchers have shifted their focus to the Transformer structure, which has shown outstanding performance in the field of natural language processing, hoping to bring new breakthroughs to computer vision with its powerful feature extraction and modeling capabilities. As a result, a series of new structures based on Transformers emerged, such as Vision Transformer (VIT) [5] and Detection Transformer (DETR). However, although the DETR series models based on Transformer adopt a non-maximum suppression (NMS) architecture, which solves the problem of slow inference speed caused by traditional object detection models relying on NMS, their high computational cost cannot meet real-time detection requirements, and they have not shown significant advantages in inference speed. This issue limits the widespread adoption of DETR series models [6] in practical applications, especially in scenarios that require high real-time performance. To solve this problem, researchers have continuously improved and optimized the model, resulting in many excellent variants such as Deformable DETR, Conditional DETR, DAB-DETR (Dynamic Anchor Boxes Are Better Questions for DETR), RT-DERT, etc. These variants have improved the model's efficiency and performance through various novelty strategies, upholding the strengths of the Transformer architecture, making it closer to the requirements of practical applications [7].

Among numerous improved DETR models, RT-DETR has received widespread attention for its excellent real-time performance and detection accuracy. However, RT-DETR still has some shortcomings in vehicle and pedestrian detection tasks. For example, RT-DETR may experience a decrease in detection accuracy due to background interference when dealing with complex scenes. In addition, the detection performance of RT-DETR needs to be improved for small and occluded targets.

In response to these issues, this article chooses RT-DETR as the baseline model for improvement. By replacing some HGBlock modules in the backbone network with HGBlock_SimAM modules, the model can focus on important information in the image earlier, thereby preventing introducing too many extraneous or duplicate features in the initial stage. This advancement raises the model's detection accuracy and also enhances its operational performance to a notable degree.

II. Related Work

A. Application and Improvement of RT-DETR in Vehicle Inspection

RT-DETR, as an efficient real-time object detection model, has demonstrated significant performance in vehicle detection tasks. However, there is still room for improvement in detection accuracy in complex traffic backgrounds, especially when dealing with small targets and background interference. To overcome these challenges, researchers have proposed various improvement strategies. For example, Azimjonov [8] proposed a vehicle detection algorithm based on improved RT-DETR, which significantly improves the model's detection ability for small targets by introducing multi-scale feature fusion and global information. In addition, Ghosh [9] suggested an approach utilizing Faster R-CNN for vehicle detection under different weather conditions, improving accuracy via the enhancement of the Region Proposal Network (RPN). These studies provide new ideas for the application of RT-DETR in vehicle detection.

B. Application and Improvement of RT-DETR in Pedestrian Detection

Spotting pedestrians is a key responsibility in computer vision, regularly used in applications like autonomous driving systems and video security. Although RT-DETR performs well in pedestrian detection, there are still some shortcomings when dealing with occlusions, complex backgrounds, and small targets. In order to improve the performance of RT-DETR in pedestrian detection, researchers have made multiple improvements. For example, Ma [10] et al. presented a fuzzy-logic enhanced pedestrian detection strategy using DETR, which significantly improved the model's detection accuracy for occluded pedestrians by introducing dynamic deformable convolution and cascaded Transformer decoders. In addition, Xing [11] et al. proposed a multispectral pedestrian detection Transformer (MS-DETR), which further enhances the detection capability of the model in complex

environments by fusing visible light and thermal imaging features.

C. Multi-scale Fusion and Small Target Enhancement of RT-DETR

In order to further improve the performance of RT-DETR in vehicle pedestrian detection, researchers have also proposed improved methods such as multi-scale feature fusion and small target enhancement. For example, Wei [12] et al. proposed an improved model RT-DETR-MSS based on RT-DETR, which significantly improves the model's detection ability for small targets by introducing a Multi Scale Fusion Module and a Small Object Enhancement Structure. In addition, the study also introduced GSConv and Slim Neck structures to optimize the network structure and improve computational efficiency. The experimental results indicate that RT-DETR-MSS performs well on CrowdHuman and WiderPerson datasets mAP@50 Increased by 1.7% and 0.8% respectively, mAP@50:95 increased by 2.2% and 1.2% respectively.

D. Real Time and Accuracy Optimization of RT-DETR

In practical applications, the real-time capabilities and precision in detection of RT-DETR are crucial. In order to meet real-time requirements, researchers have further improved the efficiency of RT-DETR by optimizing the model structure and training strategy. For example, Sadik [13] et al. proposed a deep learning framework construct using YOLOv8 and RT-DETR for the immediate recognition of vehicles and pedestrians. This study conducted experiments within complex urban environments, and the results showed that the YOLOv8 Large version performs well in identifying pedestrians, offering high precision and reliability. In addition, the study emphasizes the importance of maintaining high accuracy and reliability under different environmental conditions, such as crowded streets, changing weather, and different lighting scenarios.

In summary, RT-DETR has demonstrated significant performance in vehicle and pedestrian detection tasks, but there are still some challenges such as small object detection, background interference, and real-time performance. To overcome these challenges, researchers have proposed various improvement strategies, including multi-scale feature fusion, small target enhancement, dynamic deformable convolution, and cascaded Transformer decoders. These improvements significantly enhance the detection precision and efficiency of RT-DETR in complex scenarios, providing strong support for applications in the fields of autonomous driving and video surveillance.

III. TECHNICAL MODEL

A. RT-DETR Model

RT-DETR [14] is an efficient real-time object detection model with a well-designed architecture consisting of three key components, a backbone network, an efficient hybrid encoder, and a Transformer decoder that includes a supplementary prediction head. Specifically, the backbone network of RT-DETR is responsible for extracting multi-scale features of images. In this model, the features of the last three stages (S3, S4, S5) of the backbone network are sent to the encoder. An efficient hybrid encoder is one of the core components of RT-DETR, which transforms multi-scale features into a series of image features with rich semantic information through intra scale feature interaction (AIFI) and cross scale feature fusion (CCFM). This feature fusion method can effectively capture detailed information and overall contextual data within images, providing robust feature representation for subsequent object detection. Subsequently, RT-DETR adopts an uncertainty minimization query strategy, selecting a fixed number of features from the encoder output as the initial object query. These initial queries are then refined through iterative optimization in a decoder with auxiliary prediction heads, ultimately generating the target category and bounding box. The overall architecture of the RT-DETR model is shown in Figure 1, which clearly illustrates the various components of RT-DETR and the data flow between them.

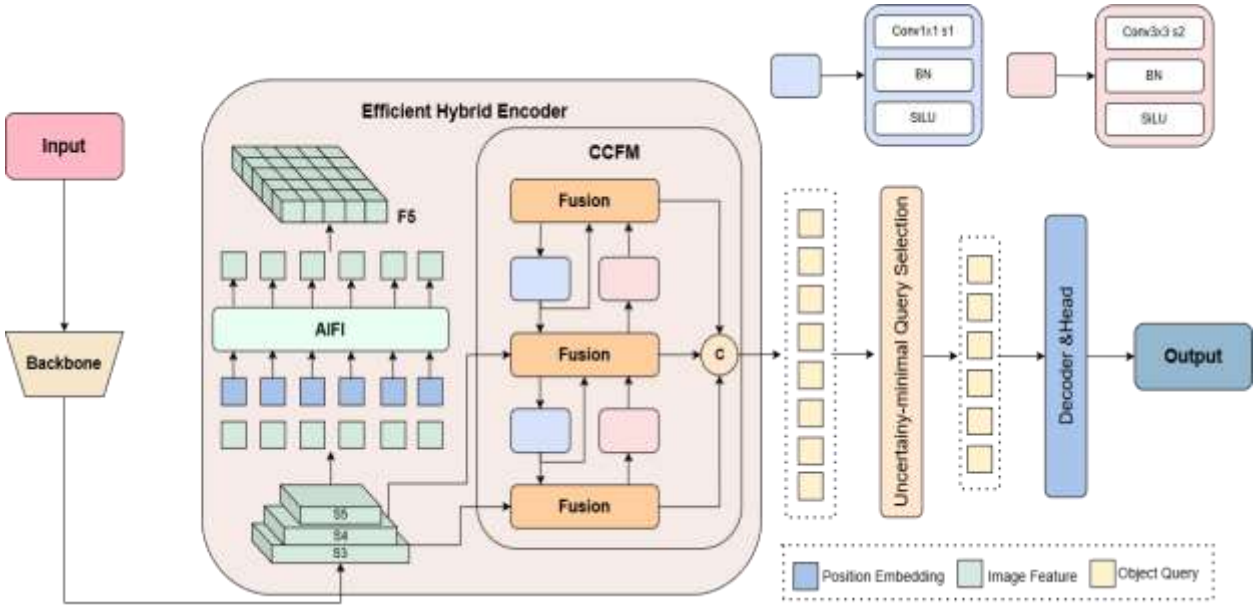

Figure 1. Network architecture of RT-DETR

B. Improved RT-DERT

In the original RT-DETR model, although it performs well in real-time, there are still some limitations. For example, when dealing with complex scenes, models may introduce too much irrelevant or redundant feature information, which not only reduces the detection accuracy of the model, but also elevates the computational workload. To overcome these potential obstacles, this paper recommends substituting some HGBlock modules in the backbone network with HGBlock_SimAM modules. By integrating SimAM attention mechanism in the HGBlock module [15], the model can focus on important information in the image earlier, thereby avoiding introducing too many irrelevant or redundant features in the initial stage. The enhancement not only boosts the model's detection precision but also enhances its operational efficiency to a certain degree. The improved RT-DETR structure is shown in Figure 2. Through experimental verification, the improved RT-DETR model has significantly improved detection accuracy in complex scenes compared to the original model.

Figure 2. Network architecture of improved RT-DETR

C. Principle of SimAM (Simple Attention Module) Attention Mechanism

SimAM is a simple and parameter free attention mechanism designed to provide an efficient and parameter free attention module for neural networks. It is based on the spatial inhibition theory

in neuroscience, using optimization of specific energy functions to measure the importance of each neuron. Specifically, SimAM is implemented through the following steps.

1) Energy function optimization

The aim in optimizing the energy function is to identify the most suitable weights for each neuron, which will represent their importance within the feature map. The energy function defined by SimAM is as follows.

$$E(t)=\sum_{i\in N}\left(\mathrm{y_i}-\hat{\mathrm{t}}\right)^2+\lambda\sum_{i\in N}\sum_{\mathrm{j}\in N}\left(\mathrm{y_i}-\mathrm{y_j}\right)^2 \quad (1)$$

Among them, t is the target neuron, N is the domain of the target neuron, y_i is the activation value of the i-th neuron, $\hat{t}$ is the predicted value of the target neuron,and λ is the regularization parameter used to balance the weights of the two terms.

2) Quick analytical solution

SimAM proposed a fast analytical solution for efficiently calculating the weights of each neuron. The analytical solution formula is as follows.

$$w_i=\frac{1}{k}\sum_{j\in N_i}s(f_i,f_j) \quad (2)$$

$$s(f_i,f_j)=-\|f_i-f_j\|_2^2 \quad (3)$$

Among them, w_i is the attention weight of the i-th neuron, k is the normalization constant, N_i is the domain of the i-th neuron, $s\left(f_i,f_j\right)$ is the similarity between the i-th neuron and the j-th neuron.

3) Attention weight calculation

Finally, SimAM calculates the attention weight of each neuron using the following formula.

$$w_i=\frac{1}{e^*} \quad (4)$$

Among them, e^* is achieved by optimizing the energy function. The greater the attention weight w_i, the more significant the role of the neuron i.

SimAM differs from traditional channel attention mechanisms and spatial attention modules in that it can directly infer the three-dimensional attention weights of feature maps without increasing the original network parameters, referencing Figure 3. In vehicle and pedestrian detection tasks, SimAM helps the model better focus on the target area, thereby improving detection accuracy.

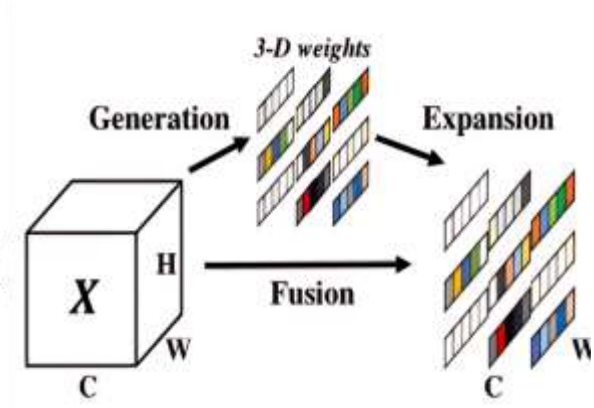

Figure 3. Full 3-D weights for attention

IV. Experiment and Analysis

A. Experimental environment

The experiments in this article were conducted on servers provided by AutoDL. The parameters configured for the experimental conditions are shown below, the system operates on Ubuntu 22.04 and is powered by 24 virtual CPU cores of the Intel (R) Xeon (R) Platinum 8255C model, running at a 2.5GHz clock speed, complemented by two RTX 3080 GPUs, each with 10GB of storage. In terms of software framework, PyTorch version 2.3.0, Python version 3.12, and CUDA version 12.1.0 are used. The explicit experimental training parameters are depicted in Table 1.

TABLE I. EXPERIMENTAL PLATFORM

Hyper-parameters	Value
Inputs	640x640
Epochs	100
Batchsize	16
Lr0	0.001
Lrf	0.0001
Momentum	0.9
Warmup-decay	0.0005
Warmup-epochs	5

B. Experimental environment

In order to evaluate the performance of improving RT-DERT detection of vehicles and pedestrians, the dataset used in this paper is BDD100K (Berkeley DeepDrive 100k). BDD100K is a large-scale and diverse dataset designed specifically for autonomous driving research, containing 100000 driving scene images covering complex scenarios such as different weather conditions (sunny/rainy/snowy), lighting conditions (day/dusk/night), and geographic regions (urban and rural roads in the United States/Asia). This dataset can be used to specifically test the performance of our detection system in scenarios with dense vehicular and pedestrian traffic. Table I presents detailed information on dataset sampling and sample partitioning in this study, covering key aspects such as dataset size, sampling method, sample size, and sample partitioning ratio.

In order to further improve the generalization ability of the model under small sample conditions, this study adopts transfer learning strategy and uses a pre trained model on the COCO2017 dataset. COCO2017 is a widely used dataset that contains rich categories and complex scenes, and its pre trained models can provide a good initialization weight. Following that, perform adjust the pre-trained model utilizing the 3000 extracted instances. Experimental results have shown that compared to training models directly from scratch on small sample datasets, this strategy of transfer learning combined with fine-tuning significantly improves the mean Average Precision (mAP@50) on the validation set during the initial training phase.

TABLE II. DATASER SAMPLING SITUATION

Content	Detailed information
Dataset size	69534 valid training samples
Sample method	Randomly select samples
Sample quantity	Sampling 3000 samples
Tag filtering	Filter other category tags
Division ratio	8:2
Training	2400 training images
Verify	600 verification images

C. Evaluation

This experiment uses Precision, Recall mAP@50 and mAP@50:95 measures to assess the model's effectiveness, with comprehensive descriptions of these criteria outlined below.

1) Precision reflects the model's precision in identifying positive samples by showing the percentage of correct positive predictions. The calculation formula is as follows.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Among them, TP is the abbreviation for true positive, which is the count of positive samples that were accurately predicted, and FP is for false positive, which is the count of positive samples that were inaccurately predicted.

2) Recall measures the proportion of samples that are actually positive and correctly predicted as positive by the model. The calculation formula is as follows.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Among them, FN, meaning false negative, is used to describe the number of negative samples that were misidentified.

3) MAP (Mean Average Precision) determines the model's average performance across all categories by averaging the AP (Average Precision) values for each one. The calculation formulas are as follows.

$$AP = \int_0^1 Precision\left(Recall\right) d\left(Recall\right) \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}_i \quad (8)$$

Among them, N is the total number of categories and is the AP value of the i-th category.

mAP@50 is the value of mAP when the IoU (Intersection over Union) threshold is 0.50. In object detection tasks, IoU is used to measure the degree of overlap between the predicted bounding box and the ground-truth bounding box. mAP@50 calculates the average of the AP (Average Precision) values of all categories at IoU=0.50.

mAP@50:95 is the average value of mAP when the IoU threshold ranges from 0.50 to 0.95. The specific calculation method is to calculate the AP values for all categories for each IoU value (from 0.50 to 0.95, usually in steps of 0.05), and then take the average of these AP values to obtain the mAP value. Finally, the mAP values corresponding to all IoU values are averaged to obtain mAP@50:95.

D. Experimental results

Extensive experiments were conducted to validate the effectiveness of adding SimAM attention mechanism in the RT-DETR backbone network. The experimental results indicate that this enhancement strategy augments the model's performance.

1) Improved detection accuracy

The changes in evaluation metrics of the RT-DETR model under different configurations are depicted in Figure 4. The left chart shows the situation of the RT-DETR model without SimAM attention mechanism. The accuracy (Rrecall) oscillates frequently in the range of 0.72 to 0.78, and although the recall remains above 0.7, there is a periodic decline of 20%. mAP@50 The final convergence is around 0.65, and the growth is weak, while the sum of strict detection ability is measured mAP@50:95 has remained stagnant below the 0.35 threshold for a long time. In contrast, after using the HGBlock_SimAM module on the right side, the curves of various indicators are smoother, including accuracy, recall, mAP50, and mAP@50:95 steadily improved during the training process, mAP@50 Breaking through 0.7, this demonstrates the positive effect of introducing SimAM attention mechanism on model performance optimization.

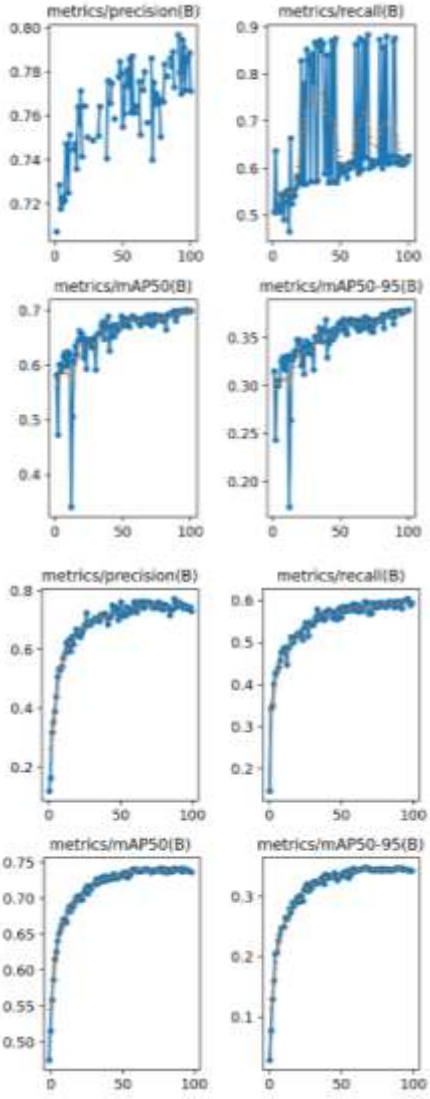

Figure 4. Comparison before and after improvement

Specifically, mAP@50 The RT-DETR has increased from 0.699 to 0.736, an increase of 3.7 percentage points. This result indicates that the SimAM attention mechanism allows the model to concentrate on crucial information in the image earlier, thereby detecting target objects more accurately. The specific experimental outcomes are depicted in Table 3.

TABLE III. EXPERIMENTAL RESULTS

RT-DETR	Without SimAM	Added SimAM
Precision	0.779	0.793
Recall	0.621	0.624
mAP@50	0.699	0.736
mAP@50:59	0.379	0.383

2) Training and reasoning efficiency

The implementation of the SimAM attention mechanism has not led to a significant decrease in the model's training and inference efficiency. As training progresses, the model's rate of convergence remains steady, and the training duration is equivalent to the original RT-DETR's. In the inference step, the advanced RT-DETR accomplishes real-time object detection, with a minor decline in inference speed relative to the original model, indicating that the SimAM mechanism improves performance while maintaining the efficiency of the model.

3) Comparison with other models

For validation of the proposed method's overall efficacy, a comparison and analysis were made with the traditional object detection algorithm YOLOv8. Experiments were executed on the BDD dataset, and findings are detailed in Table 4.

According to Table 4, the model proposed in this paper is more effective than YOLOv8. In detail, the accuracy has jumped from 0.721 with YOLOv8 to 0.793, mAP@50 has also increased from 0.692 to 0.736, and mAP@5095 has escalated from 0.372 to 0.383. The findings demonstrate that incorporating the SimAM attention mechanism has successfully enhanced the model's detection precision and its ability to perform fine-grained detections.

TABLE IV. CAMPARISON RESULTS

Model	YOLOv8	Ours
Precision	0.721	0.793
Recall	0.645	0.624
mAP@50	0.692	0.736
mAP@50:59	0.372	0.383

4) Visualization results

The actual detection results in different scenarios of the BDD dataset test set are shown in Figure 5.

Figure 5. Visualization results

The improved RT-DETR model proposed in this study demonstrates significant advantages in complex scenarios. Specifically, when dealing with partial occlusion, the model secured an 80% accuracy in identifying the locations of both vehicles and pedestrians. This indicates that the improved model can more accurately identify and locate targets when dealing with complex scenes and occlusion problems, thereby significantly improving detection performance.

V. CONCLUSIONS

This article proposes an improved model based on RT-DETR, which achieves an average accuracy value (mAP) of 73.6% on the BDD dataset by replacing some HGBlock modules in the backbone network with HGBlock_SimAM modules. This result is superior to the original RT-DETR model, fully demonstrating the effectiveness of introducing SimAM attention mechanism in the field of vehicle and pedestrian detection.

Moving forward, we intend to keep refining and enhancing the algorithm of this model. On the one hand, the plan is to further reduce the number of parameters in the model to enhance its computational speed and real-time capabilities, making it more suitable for use in resource constrained environments, such as real-time detection systems for embedded devices or autonomous vehicles. On the other hand, efforts

will be geared towards enhancing the model's detection precision by introducing more advanced feature extraction and fusion techniques to further strengthen its detection capabilities for small targets, occluded targets, and complex backgrounds. In addition, the scalability and adaptability of the model will be explored to better cope with changes in different scenarios and datasets. Through these optimization and improvement measures, the improved RT-DETR model is expected to play a greater role in the field of autonomous driving detection, providing strong technical support and reference for the development of autonomous driving technology.

REFERENCES

[1] Hidayatullah, P.; Syakrani, N.; Sholahuddin, M. R.; Gelar, T.; Tubagus, R. YOLOv8 to YOLOv11: A Comprehensive Architecture In-Depth Comparative Review.

[2] Zhang, X.; Zhang, Y.; Gao, T.; Fang, Y.; Chen, T. A Novel SSD-Based Detection Algorithm Suitable for Small Object. IEICE Trans. Inf. Syst. 2023, E106.D (5), 625–634.

[3] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. 2020, 42 (2), 318–327.

[4] Arora, N.; Kumar, Y.; Karkra, R.; Kumar, M. Automatic Vehicle Detection System in Different Environment Conditions Using Fast R-CNN. Multimed. Tools Appl. 2022, 81 (13), 18715–18735.

[5] Abd Alaziz, H. M.; Elmannai, H.; Saleh, H.; Hadjouni, M.; Anter, A. M.; Koura, A.; Kayed, M. Enhancing Fashion Classification with Vision Transformer (ViT) and Developing Recommendation Fashion Systems Using DINOVA2. Electronics 2023, 12 (20), 4263.

[6] Fahad, I. A.; Arean, A. I. H.; Ahmed, N. S.; Hasan, M. Automatic Vehicle Detection Using DETR: A Transformer-Based Approach for Navigating Treacherous Roads. arXiv February 25, 2025.

[7] Cheng Xinmiao, Zhang Xuesong, Cao Bingjie, Song Cunli Research on Improving the Small Object Detection Method of RT-DETR [J]. Computer Engineering and Applications, 1-21.

[8] Azimjonov, J.; Özmen, A. A Real-Time Vehicle Detection and a Novel Vehicle Tracking Systems for Estimating and Monitoring Traffic Flow on Highways. Adv. Eng. Inform. 2021, 50, 101393.

[9] Ghosh, R. On-Road Vehicle Detection in Varying Weather Conditions Using Faster R-CNN with Several Region Proposal Networks. Multimed. Tools Appl. 2021, 80 (17), 25985–25999.

[10] Wu, T.; Li, X.; Dong, Q. An Improved Transformer-Based Model for Urban Pedestrian Detection. Int. J. Comput. Intell. Syst. 2025, 18 (1), 68.

[11] Xing, Y.; Yang, S.; Wang, S.; Zhang, S.; Liang, G.; Zhang, X.; Zhang, Y. MS-DETR: Multispectral Pedestrian Detection Transformer with Loosely Coupled Fusion and Modality-Balanced Optimization. IEEE Trans. Intell. Transp. Syst. 2024, 25 (12), 20628–20642.

[12] Song, Y.; Qian, P.; Zhang, K.; Liu, S.; Zhai, R.; Song, R. An Improved Multi-Scale Fusion and Small Object Enhancement Method for Efficient Pedestrian Detection in Dense Scenes. Multimed. Syst. 2025, 31 (2), 151.

[13] Sadik, M. N.; Hossain, T.; Sayeed, F. Real-Time Detection and Analysis of Vehicles and Pedestrians Using Deep Learning. arXiv April 11, 2024.

[14] Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. arXiv April 3, 2024.

[15] Xu, Y.; Du, W.; Deng, L.; Zhang, Y.; Wen, W. Ship Target Detection in SAR Images Based on SimAM Attention YOLOv8. IET Commun. 2024, 18 (19), 1428–1436.