Small Target Detection Algorithm for Traffic Signs Based on Improved RT-DETR

Nuanling Liang, Weisheng Liu

Abstract—**To tackle the issues of low detection accuracy for small traffic signs in Advanced Driver-Assistance Systems (ADAS), we introduce an enhanced model RT-DETR_ASL to make ADAS more accurate and responsive. Firstly, we lighten and optimize the backbone network by substituting the Basic Block with an inverted residual block, thereby reducing the parameter count and enhancing computational speed. Secondly, we integrate a multi-scale deformable attention mechanism into the AIFI feature extraction network, augmenting the recognition and learning capabilities for small targets, which ultimately sharpens the precision of positioning and recognition. Additionally, to bolster the model's performance in detecting small, poorly defined traffic signs, we incorporate the S2 small-target detection layer to refine and strengthen the network's capabilities. During validation, when setting the GIoU (Generalized Intersection over Union) threshold at 0.7, the RT-DETR_ASL model demonstrated a 4.1% increase in mAP50 (mean Average Precision) over the baseline model. Upon further optimizing the loss function, the mAP value soared by an additional 4.51%, surpassing four other mainstream detection methods. Our experiments confirm that the RT-DETR_ASL model significantly enhances the detection accuracy of small traffic signs while maintaining real-time performance, contributing meaningfully to the advancement of autonomous driving assistance systems. It is hoped that the results of this research can make a valuable contribution to the further development of autonomous driving technology.**

Index Terms—**RT-DETR_ASL, ASPPDAT, Traffic sign, Small target detection, ADAS**

I. Introduction

Traffic sign detection provides ADAS with important traffic information such as speed limits, bans, warnings, etc. by recognizing traffic signs on the road in real-time, thus providing ADAS with early warning information and helping to reduce traffic violations and traffic accidents. With advanced computer vision technology and deep learning algorithms, traffic signs on the road can be accurately recognized and understood, which allows vehicles to autonomously control their driving, thus enabling safer and smarter ADAS.

Manuscript received June 13, 2024; revised November 11, 2024.

This work was supported by the Special Fund for Scientific Research Construction of University of Science and Technology Liaoning, China.

Nuanling Liang is a postgraduate student at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, China (e-mail: liangnuannuan@163.com).

Weisheng Liu is a professor of the College of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, CO 114051, China (corresponding author to provide fax: 0412-5929809; e-mail: succman@163.com).

However, in actual traffic environment, traffic signs are widely distributed on roads, intersections, schools, tunnels, bridges, urban areas, and motorways. Traffic sign detection can be affected by various factors, such as the target detection area being too small and the variety of traffic signs, which may interfere with the detection effect of traffic signs and increase the difficulty of detection. At the same time, due to the wide variety and relatively small size of traffic signs, this also increases the difficulty of detection. Therefore, it is necessary to continuously optimize and improve the traffic sign detection algorithm to improve its detection accuracy and robustness in various complex environments.

Preliminary work on traffic sign detection has focused on traditional techniques based on image processing and machine learning algorithms. Deep learning techniques based on Convolutional Neural Networks (CNN) have made significant breakthroughs in the field of target detection for complex road scenes, which gradually surpass the limitations of traditional manually designed feature methods as well as show stronger robustness and generalization capabilities. Traditional manually designed feature methods often have limitations when facing complex and changing road scenes, and are difficult to cope with various challenges. Deep learning-based techniques can automatically learn deeper semantic features during the training process, thus exhibiting better performance, robustness and generalization ability than manually designed features. Such methods can learn deeper semantic features during the training process and has better performance than manually designed features, robustness and generalization. Traditional manually designed feature methods often have limitations when facing complex and changing road scenes, and are difficult to cope with various challenges. Deep learning-based techniques can automatically learn deeper semantic features during the training process, thus exhibiting better performance, robustness and generalization ability than manually designed features. Such methods can learn deeper semantic features during training and outperform manually designed features in terms of performance, robustness, and generalization.

Target detection algorithms based on CNN can be summarized into two categories according to the detection stage: two-stage target detection algorithms and one-stage target detection algorithms. Initial two-stage series of algorithms such as Faster R-CNN [1], generate object proposals are first generated, and then each region is classified and bounded by bounding box regression. In contrast, one-stage target detection algorithms, such as the YOLO series of algorithms [2-4], predict object categories and border positions directly on the image. In order to improve the detection performance of small target objects, Zhang et al. [5] propose an improved traffic sign detection

algorithm based on the YOLOv8s algorithm. The proposed integration of deformable convolution into the auxiliary backbone and the application of the coordinate attention mechanism after the SPPF layer. However, to eliminate redundant anchor boxes, the YOLO series requires non-maximum suppression (NMS) post-processing in the final prediction. It not only affects the processing speed but also reduces the accuracy of the model.

One way to solve the NMS post-processing problem is to use the Transformer-based target detector DETR (DEtection TRansformer) series of algorithms [6]. This type of algorithm cancels the NMS step, simplifies the detection process, and has received widespread attention from the academic community. For example, Xia et al. [7] proposed DSRA-DETR, which improves multi-scale detection performance, reduces feature noise, and enhances object detection capabilities at different scales by integrating deep space pyramid pooling (DSPP) and feature relabeling (FRAM) modules to improve Traffic sign detection accuracy and robustness have been improved on GTSDB [8] and CCTSDB datasets. The advantage of the DETR [9] algorithm is that it simplifies the process of target detection, but it also has shortcomings, including slow convergence and limited spatial resolution. Deformable DETR introduces a deformable attention module to improve DETR's accuracy when detecting small targets in single-scale images by focusing on the surrounding modules of the reference points. N Gray et al. [10] used Deformable DETR to increase the mAP value of the LISA dataset to 71.6%. However, Deformable DETR, which processes multi-scale images, must process a large amount of token data, significantly affecting data accuracy and target detection time. To tackle this issue, Lv et al. [11] proposed RT-DETR(Real Time DEtection TRansformer), which features an efficient hybrid encoder designed to process a large amount of token data and simplify the execution of the encoding layer, thereby improving operating efficiency. As the first real-time end-to-end DETR detector, it not only simplifies the execution of the encoder layer but also improves operating efficiency. However, the accuracy and speed when it comes to the recognition of small targets are relatively low.

In order to solve the shortcomings in small target recognition accuracy of existing research, this study constructed the RT-DETR_ASL model. The model integrates the ASPPDAT attention mechanism and the improved CCFM network structure and uses the FasterNet Block module to replace the BasicBlock module and replaces the loss function. These improvements significantly improve the accuracy of target recognition.

II. RT-DETR

The RT-DETR algorithm represents one of the most advanced target detection methods within the Transformer-based DETR (DEtection TRansformer) series. It uses deep learning networks to detect and locate targets. Its most significant advantage is its rapid processing speed and excellent real-time performance. Compared with the early DETR algorithm, RT-DETR has achieved significant improvements in accuracy and processing speed, partly due to its use of an innovative hybrid encoder. The encoder effectively processes large amounts of token data through decoupling strategies and cross-scale feature fusion techniques, thereby enhancing the algorithm's overall performance.

In the RT-DETR framework, the entire network consists of the following key components: a backbone network, an Efficient Hybrid encoder, and a decoder. As the backbone network, we chose the popular ResNet architecture, explicitly using the ResNet18 version; the Efficient Hybrid encoder consists of two parts, namely the AIFI (Attention-based Intra-scale Feature Interaction) module and the CCFM (Cross-Scale Feature Fusion Module), the combination of the two can make the network more capable of capturing the global context information of the image, helping to identify targets more accurately, thereby reducing false detections and missed detections. In order to further optimize the target detection performance, the ASPPDAT (Atrous Spatial Pyramid Pooling with Deformable Attention Transformer) attention mechanism was introduced into the AIFI module in the study, and the structure of the CCFM module was improved. The Block from FasterNet was implemented to replace the traditional BasicBlock, with the aim of improving the accuracy and speed of target recognition. In addition, although the official version of RT-DETR uses IoU as the loss function, the small target detection accuracy is relatively low because IoU is sensitive to scale changes and other issues. This study improves upon this by utilizing the Inner-GIoU loss function to optimize the algorithm's performance further.

III. IMPROVED RT-DETR ALGORITHM

A. Improve The AIFI Part Of RT-DETR

In the traffic sign detection task, due to the significant size difference of traffic signs, in order to improve the recognition accuracy of small-sized targets, this study introduced ASPPDAT technology. The traditional AIFI module only obtains feature and position information from the S5 layer through the Multi-Head Self-Attention mechanism [12], but this mechanism often ignores small-sized targets and has limitations in flexibly adapting to the target scale. Therefore, this article replaces the multi-head self-attention mechanism in the AIFI module with the more advanced ASPPDAT structure. This improvement not only helps the model identify small targets, but also promotes better convergence of the model during the training process, further improving detection accuracy. The ASPPDAT structure combines Atrous Spatial Pyramid Pooling (ASPP) [13] and Deformable Attention Transformer (DAT) [14] technologies to allow key information to be extracted from multi-scale feature maps. ASPP can increase the receptive field to detect objects of different positions and sizes, while the DAT attention mechanism reduction allows the position and shape of the convolution kernel to change according to different areas of the input image, thereby better adapting to the different locations of the target in the image. Regular shapes and postures can improve target detection accuracy and improve robustness. ASPPDAT technology can use multi-channel feature maps at different scales to improve detection accuracy. At the same time, through deformable convolution, the convolution kernel can adaptively adjust its shape and size during the convolution process, so that the network can detect different sizes and positions. Traffic signs are detected. By introducing deformable convolutional layers to better

adapt to traffic signs of different sizes. This method can more effectively extract features related to traffic signs, which not only improves the accuracy of target detection but also enhances the robustness of the system. In summary, the improvement scheme proposed in this article, by combining ASPP and DAT technology and using deformable convolution layers, allows the detection model to more accurately locate traffic signs of different sizes and shapes, as shown in Figure 1.

The ASPPDAT structure described in this article is a channel attention module. This module combines the ASPP (Atrous Spatial Pyramid Pooling) module with the channel attention mechanism and uses a normalization strategy to integrate ASPP into the attention mechanism DAT. The ASPP module captures context information of various sizes by utilizing convolution kernels with different dilation rates to provide multi-scale contextual information, thereby enhancing target recognition accuracy. Meanwhile, the DAT (Deformable Attention Transformer) attention mechanism builds upon the traditional self-attention framework of the Transformer and incorporates features of deformable convolution. This integration addresses the inadequacies of the traditional attention mechanism when handling targets with irregular shapes and demonstrates greater efficacy in detecting small targets. The primary purpose of DAT is to be more suitable for small target detection by introducing the flexibility of deformable convolution [15]. In the fusion process of building the ASPPDAT module, this article uses the three convolutional layers through the 3×3 Conv average pooling and then the concat operation, and inputs these features into the deformable convolution attention θ offsetDAT. θ offsetDAT is a deformable convolutional attention θ offsetDAT.

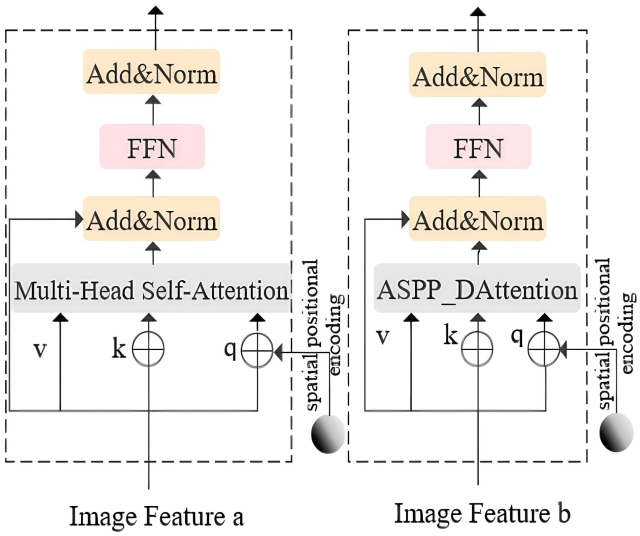

Fig. 1. AIFI structure in picture a. Addition of ASPPDAT structure diagram in picture b.

The convolutional attention module captures through 5×5 depth convolution and then uses the GELU activation function for nonlinear transformation. Then, the 1×1 Conv convolution layer is used to generate a two-dimensional offset. This design can dynamically adjust the model's focus on key feature areas. After such operations, the output representation of the attention head can be summarized in the following form: By introducing the attention mechanism in the spatial and channel dimensions, the ASPPDAT module can capture richer hierarchical information and dynamic context information of image features, enhancing the model's ability to Recognition and detection capabilities of size targets, especially small targets. The output of the attention head is expressed as follows, where xW_q is the result of the feature map being linearly projected to the query mark, xW_k, and represents the deformed key embedding and value embedding respectively, indexing the table with the relative displacement in two directions to obtain the relative position offset table, and R represents the relative position Offset. Connect the features of each head together and obtain the final output $z^{(m)}$ through projection.

$$z^{(m)} = \sigma\left(\frac{\left(xW_q\right)^{(m)}\left(\left(\tilde{x}W_k\right)^{(m)}\right)^T}{\sqrt{d}} + \varphi(\hat{B};R)\right)\tilde{v}^{(m)} \quad (1)$$

The ASPPDAT structure proposed in this article enhances the model's ability to cope with complex scenarios by combining ASPP and DAT technology. Specifically, the ASPP component uses dilated convolutions with different expansion rates to fill the spatial information in the feature representation, allowing the model to perceive image content at multiple scales, while the DAT component empowers the model by learning deformable offsets. Adaptability to local deformation. This composite structure first samples the input feature map and then uses the average pooling method to improve the model's ability to identify small and dense defects in complex backgrounds. The ASPPDAT architecture is shown in Figure 2 below.

B. CCFM Network Optimization

For small targets in traffic sign, it is necessary to obtain details that occupy only a very small area from small distant objects or images, which usually occupy fewer pixels in the image. The features extracted by the shallow network [16] are closer to the input, fewer features are lost, and the features are more similar to the original input image, thus resulting in relatively less loss of information. In contrast, deep networks [17] have low resolution and the data undergoes multiple rounds of deep processing, so the information representation of small targets is relatively weak and is more likely to be lost during processing. Therefore, in order to effectively capture the detailed information of these small targets, this paper proposes a method to integrate the S2 layer shallow network into the structure through downsampling and Conv1×1 to improve model accuracy. Subsequently, CCFM is used for multi-scale fusion [18], the S2 layer information is fused with that from CCFM, including both the underlying feature map and the shallow feature map, to provide more spatial detail information which is helpful for the detection of small objects. This optimization not only improves the accuracy of small target detection in traffic signs but also better integrates multi-scale information. The specific improved structure is shown in Figure 3.

C. FasterNet's Block Module To Improve BasicBlock

In response to the demand for real-time performance and lightweight automatic driving assistance systems, the traditional backbone ResNet, as a deep convolutional neural network, has a complex structure and numerous parameters, resulting in high model calculation costs.

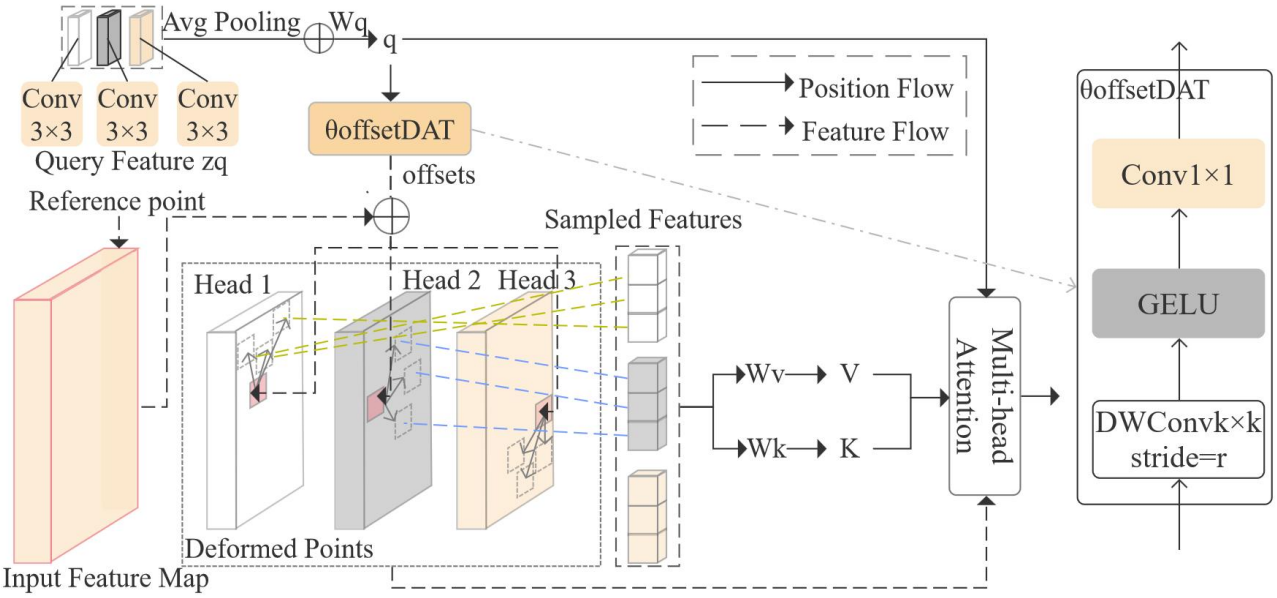

Fig. 2. ASSPDAT illustration of the structure of a module

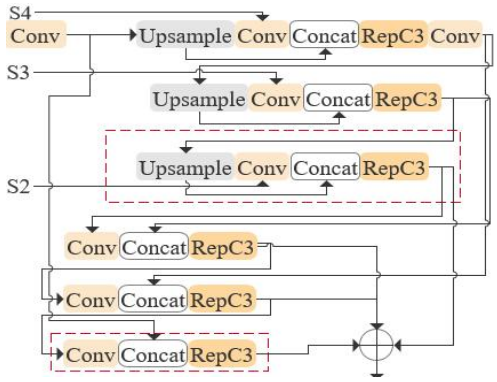

Fig. 3. Improved CCFM network structure.

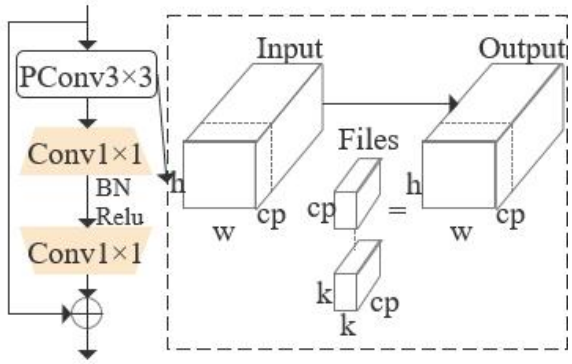

Fig. 4. Block diagram.

In autonomous driving assistance systems, this complexity may lead to a decrease in processing speed, making it difficult to meet real-time requirements. The FasterNet [19] block is an inverted residual block consisting of one PConv layer and two Conv 1×1 point convolutions, where the middle layer has an extended number of channels and a Shorcut is placed to reuse the input features. Using the structure of FasterNet's Basic module Inverted Residual Block [20], it effectively solves the redundancy caused by the fully connected channel structure of standard convolution and achieves an effective method of lightweighting the model. This replacement can help us reduce model size and computational cost while maintaining model performance. The improvement is shown in Figure 4.

D. Loss Function Improvement InnerGIoU Experimental Environment

Inner-GIoU [21] is an innovative loss function used to improve the performance of target detection algorithms. It is an extension and optimization of GIoU [22] (Generalized Intersection over Union). Inner-GIoU loss inherits certain characteristics of GIoU loss while introducing its unique features. By introducing Inner-GIoU, the degree of overlap between the predicted box and the real box is more accurately evaluated in the traffic sign detection task, and the performance of the model is improved. Inner-GIoU not only considers the intersection and union of the predicted box and the real box but also introduces the calculation of the internal area. This enables Inner-GIoU to more accurately measure the degree of overlap between the predicted box and the real box, especially in traffic scenarios when the predicted box partially overlaps or is close to the real box. The formula is shown below. Among them, A and B^{gt} represent the predicted box and the ground truth box respectively. b and b^{gt} represent the center points of the predicted box A and the real box B^{gt} respectively. w^{gt} and hgt represent the width and height of the real box respectively, and w and h represent the width and height of the anchor box. The variable-ratio is the scaling factor, which is in the value range of [0.5, 1.5].

$$IoU = \frac{|A \cap B^{gt}|}{|A \cup B^{gt}|}$$

$$L_{GIoU} = 1 - IoU + \frac{|A_c - (A \cap B^{gt})|}{|A_c|}$$

$$inter = (\min(b_r^{gt}, b_r) - \max(b_l^{gt}, b_l)) * (\min(b_b^{gt}, b_b) - \max(b_t^{gt}, b_t)) \quad (2)$$

$$union = (w^{gt} * h^{gt}) * (ratio)^2 + (w * h) * (ratio)^2 - inter(6)$$

$$IoU^{inner} = \frac{inter}{union}$$

$$L_{Inner-GIoU} = L_{GIoU} + IoU - IoU^{inner}$$

E. Model Reconstruction

In order to improve the performance of the RT-DETR r18 network in detecting smaller or dense objects, we proposed an improved RT-DETR_ASL algorithm and optimized the network structure. Compared with the original RT-DETR

algorithm, the improved RT-DETR_ASL algorithm improves the resolution and feature expression capabilities of the network and achieves better target positioning. Specific optimizations include using rectangular frames of different colors to mark the network architecture. Through the above optimization and improvements, RT-DETR_ASL significantly outperforms the original RT-DETR algorithm in detecting small targets within traffic signs, achieving more accurate target detection. The improvement is shown in Figure 8.

IV. Experiment and Result Analysis

A. Experimental Environment

The development system of this experiment is: Linux18.0, using the Pytorch1.8.1 framework, and the graphics card is GPU NVIDIA GeForce RTX 3090. The CPU is Intel(R) Core(TM) i7-13700KF@3.4GHz, with an initial learning rate of 0.01. The model converges at 100 iterations, and TT100K is the experimental data set used.

B. Experimental Data

The TT100K [23] image comes from Tencent Street View panorama, which was shot by 6 high-pixel wide-angle SLR cameras in different cities in China, with different light and weather conditions. The resolution of the original Street View panorama is 8192×2048, then the panorama is cut into four parts, and the image size in the dataset is 2048×2048. In TT100K, a total of 201 different classes appear. Among the 201 classes, 84 classes have less than 10 instances. This part of the data is of little significance and has not been added to the training; 62 classes have 10-75 instances; and only 45 classes have more than 100 instances. After re-dividing the TT100K data set, the training set is 6793 images, the verification set is 1949 images, and the test set is 996 images, for a total of 9738 images. According to the definition of small targets in COCO, 32×32 pixels and below are small targets. Small targets account for 94% of the TT100K data set, which is a small target data set.

C. Model Evaluation Indicators

In order to comprehensively and objectively evaluate the performance of the improved RT-DETR_ASL model proposed in this article, indicators such as Precision, Recall, Mean Average Precision (mAP) and F1-score are used to measure it. The specific formula is as follows. where TP is true-positive detection, FP is false-positive detection, P(R) is the precision-recall curve, i is the detection category of this paper's experiment, and c is the number of 45 categories in this paper's experiment.

$$\begin{aligned} AP &= \int_0^1 P(R)dR \\ Precision &= \frac{TP}{TP+FP} \\ mAP &= \sum_{i=1}^{c} AP_i \\ F_{1(Score)} &= 2\frac{Precsion \times Recall}{Precsion + Recall} \end{aligned} \tag{3}$$

D. Experimental Results And Analysis

Figure 5 shows the P-R curve of the original RT-DETR network, while Figure 6 displays the P-R curve of the RT-DETR_ASL network. Figure 7 shows the F1 curve The size of the area enclosed by the P-R curve and the two coordinate axes is the AP value of the corresponding classification. Comparative analysis shows that the improved P-R curve covers a larger area, and the curves are all located at the top, indicating that the improved model has good detection results. In order to further verify the advantages of this algorithm, the comparison of mAP changes during the training process of the RT-DETR_ASL network and the original RT-DETR network is shown in Figure 9 below. The improved network begins to gradually converge around 100 rounds, and the improved network can achieve better results in a limited time.

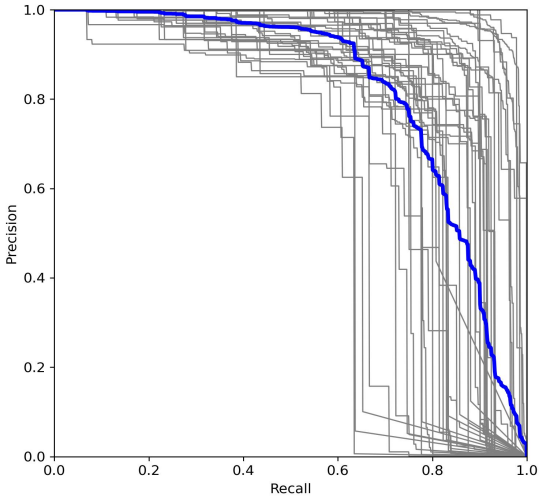

Fig. 5. Original PR curve.

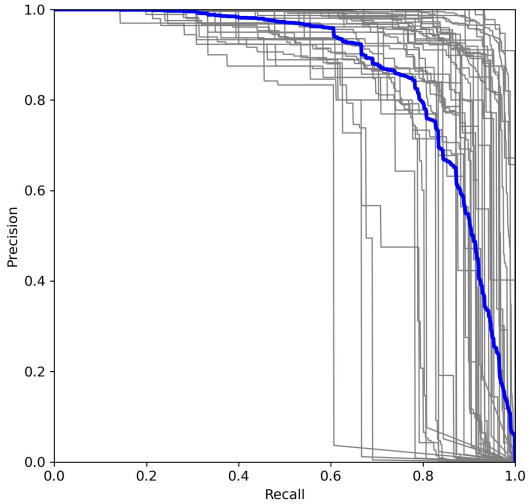

Fig. 6. Improved PR curve.

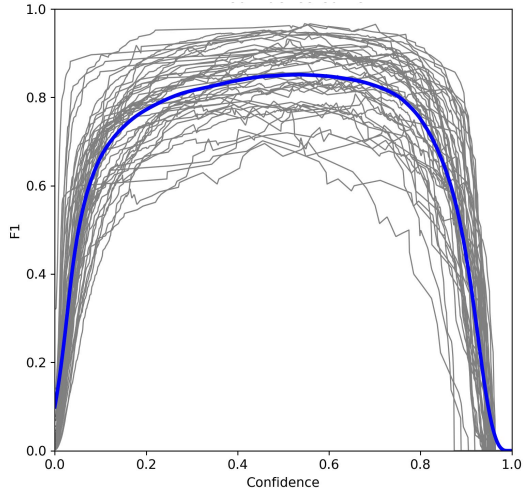

Fig. 7. F1 curve.

1) To verify the effectiveness of the Block, ASPPDAT, S2, and Inner-GIoU modules on detection accuracy, this article uses RT-DETR as the benchmark model and combines these innovative modules with the original module to conduct comparative experiments. The experimental parameters are default and the resolution is 640×640. We conducted tests and designed several different sets of experiments to analyze the impact of different improvements on network performance. Each set of experiments used the same training parameters, which √ represents the use of the corresponding improvement strategy in the model's prior work. As shown in Table 1.

2) The improved algorithm is compared with mainstream target detection algorithms and RT-DETR with improved strategies. The comparison results are shown in Table 2. The analysis shows that the mAP of the improved RT-DETR algorithm is higher than other algorithms, which is 4.51% higher than the original model is 28.45%, 13.67%, and 10.29% higher than YOLOV5, YOLOV8, and Faster-RCNN respectively. It has better detection accuracy than other mainstream target detection network models. Compared with the original RT-DETR algorithm, the accuracy and recall rates are improved. Overall, the improved RT-DETR detection performance is better than other algorithms.

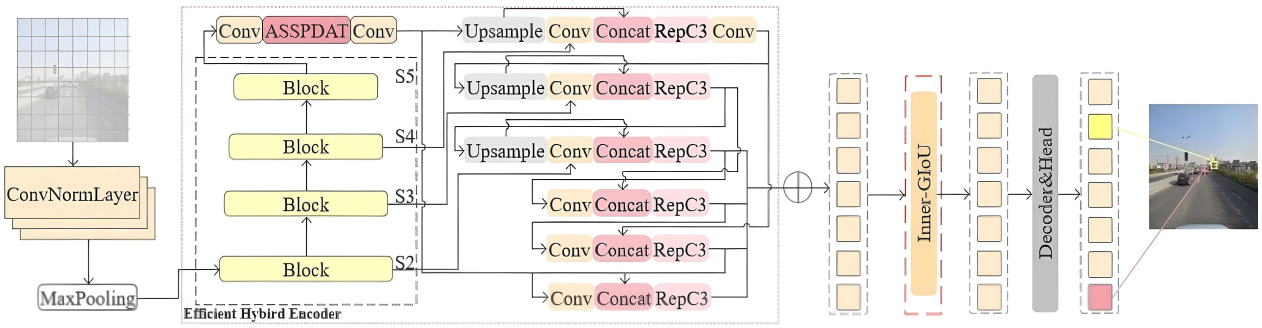

Fig. 8. RT-DETR_ASL diagram.

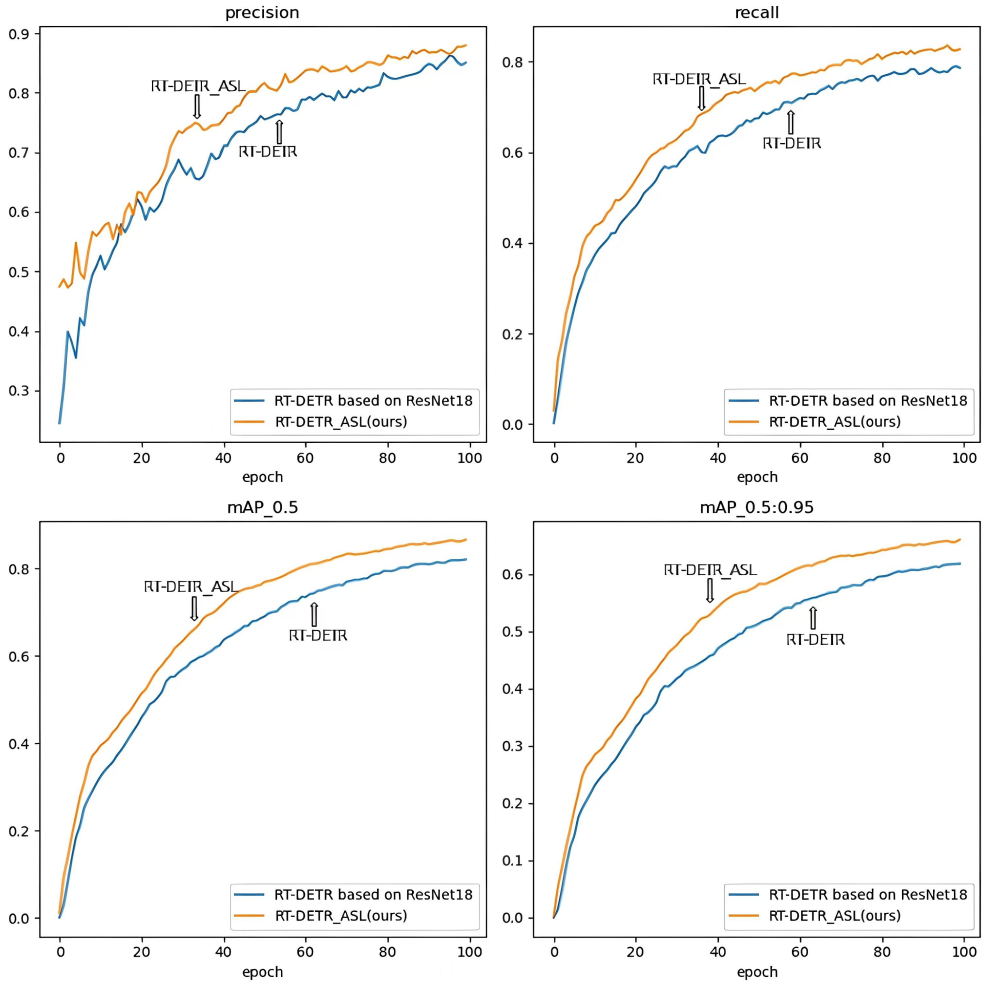

Fig. 9. Comparison of Precision, Recall, and Average Precision.

TABLE 1
ABLATION EXPERIMENT RESULT

Id	Block	ASPPDAT	S2	Inner-GIoU	mAP/%	Precious/%	Recall/%
1	-	-	-	-	0.8201	0.8502	0.7867
2	√	-	-	-	0.8316	0.8571	0.8043
3	-	√	-	-	0.8417	0.8898	0.7950
4	-	-	√	-	0.8565	0.8769	0.8192
5	-	-	-	√	0.8456	0.8621	0.8178
6	√	√	√	√	0.8652	0.8796	0.8275

TABLE 2
COMPARISON OF MAINSTREAM TARGET DETECTION MODELS

Model	Precision/%	Recall/%	mAP/%
RT-DETR	0.8502	0.7867	0.8201
Faster-RCNN	0.7980	0.7117	0.7623
YOLOV5	0.59407	0.54473	0.5807
YOLOV8	0.7461	0.6522	0.7285
Ours	0.8796	0.8275	0.8652

Fig. 11. RT-DETR_ASL diagram.

E. *Experimental Results*

Figures 10 and 11 are some visualizations of the detection results. Figure 10 is the RT-DETR detection chart, and Figure 11 is the improved RT-DETR detection chart. It can be seen from the picture that the detection effect has been improved, and the detection confidence of each target has been improved. In the detection of traffic signs, RT-DETR missed detection, while the improved RT-DETR_ASL accurately detected the logo. For some signs, the original model has false detections, but the improved model can improve this situation, and the confidence score has also improved, indicating that the improved model has strong generalization ability. Before the improvement, traffic signs were missed and misdetected. After the improvement, traffic signs can be accurately detected and it has a strong ability to deal with complex scenarios.

Fig. 10. RT-DETR_ASL diagram.

V. CONCLUSION

This study proposes the RT-DETR_ASL algorithm aims to overcome the challenges faced by DETR specifically in small target traffic sign detection. By introducing the Basic module based on the RT-DETR framework, we maintain the size of the feature map, thereby enhancing the recognition ability of to recognize small target traffic signs. In addition, the ASPPDAT module we built helps decrease the model size and enhance the inference speed. The optimized network structure also adds the S2 layer, which significantly improves the processing speed through the application of Basic blocks. Compared with four classic detectors (RT-DETR, YOLOv5, YOLOv8, and Faster-RCNN), our method shows higher accuracy and mAP value in the field of small target detection, demonstrating excellent detection results. The field of small object detection in traffic signs is full of challenges and complexities. Our study proposes a new idea and approach aimed at overcoming these difficulties. With the continuous iteration and improvement of technology, the detection efficiency and accuracy of small target traffic signs are expected to be significantly improved, thereby greatly enhancing the functions of ADAS (Advanced Driver Assistance System) and improving driving safety. The results of this study can provide valuable references for researchers in the same field and inspire new research inspiration to promote the development of this field.

REFERENCES

[1] L. Zhang, L. Lin, X. Liang, and K. He, "Is faster R-CNN doing well for pedestrian detection?" *Computer Vision-ECCV 2016: 14th European Conference*, October 11-14, Amsterdam, Netherlands, vol.14, no.2, pp443-457, 2016.

[2] C. WANG, A. BOCHKOVSKIY, and H. LIAO, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),* pp7464-7475, 2023.

[3] C. WANG, I. YEH, and H. LIAO, "YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information," *ArXiv*, vol. abs/2402.13616, 2024.

[4] J. TERVEN and D. CORDOVA-ESPARZA, "A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond," *ArXiv*, vol. abs/2304.00501, 2023.

[5] X. ZHANG and Z. ZHANG, "Traffic sign detection algorithm based on improved YOLOv7," *International Conference on Image, Signal Processing, and Pattern Recognition (ISPP 2023)*, vol.12707, pp1258-1266, 2023.

[6] Q. ZHOU, C. YU, and Z. WANG, "D 2 Q-DETR: Decoupling and Dynamic Queries for Oriented Object Detection with Transformers," *2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), pp1-5, 2023.

[7] J. XIA, M. LI, and W. LIU, "DSRA-DETR: An Improved DETR for Multiscale Traffic Sign Detection," *Sustainability*, vol.15, no.14, pp10862, 2023.

[8] J. STALLKAMP, M. SCHLIPSING, and J. SALMEN, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural networks*, vol.32, pp323-332, 2012.

[9] CARION N, MASSA F, and SYNNAEVE G, "End-to-end object detection with transformers, " *2020 European Conference on Computer Vision*, Springer, pp.213-229, 2020.

[10] N. GRAY, M. MORAES, and J. BIAN, "GLARE: A dataset for traffic sign detection in sun glare," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[11] Y. ZHAO, W. LV, and S. XU, "Detrs beat yolos on real-time object detection," *arXiv*, vol. abs/2304.08069, 2023.

[12] J.-B. CORDONNIER, A. LOUKAS, and M. JAGGI, "Multi-head attention: Collaborate instead of concatenate," *arXiv*, 200616362, 2020.

[13] L. C. CHEN, G. PAPANDREOU, and I. KOKKINOS, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.40, no.4, pp834-848, 2017.

[14] Z. XIA, X. PAN, and S. SONG, "Vision transformer with deformable attention," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp4794-4803, 2022.

[15] J. DAI, H. QI, and Y. XIONG, "Deformable convolutional networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp764-773, 2017.

[16] A. BIETTI, J. BRUNA, and C. SANFORD, "Learning single-index models with shallow neural networks," *Advances in Neural Information Processing Systems*, vol.35, pp9768-9783, 2022.

[17] Y. GUO, Y. LU, and R. LIU, "Lightweight deep network-enabled real-time low-visibility enhancement for promoting vessel detection in maritime video surveillance," *The Journal of Navigation*, vol.75, no.1, pp230-250, 2022.

[18] H. CHU, W. WANG, and L. DENG, "Tiny-Crack-Net: A multiscale feature fusion network with attention mechanisms for segmentation of tiny cracks," *Computer-Aided Civil and Infrastructure Engineering*, vol.37, no.14, pp1914-1931*,* 2022.

[19] J. CHEN, S.-H. KAO, and H. HE, "Run, Don't walk: Chasing higher FLOPS for faster neural networks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp12021-12031, 2023.

[20] M. SANDLER, A. HOWARD, and M. ZHU, "Mobilenetv2: Inverted residuals and linear bottlenecks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp4510-4520, 2018.

[21] H. ZHANG, C. XU, and S. ZHANG, "Inner-iou: more effective intersection over union loss with auxiliary bounding box," *arXiv,* 2311.02877, 2023.

[22] H. REZATOFIGHI, N. TSOI, and J. GWAK, "Generalized intersection over union: A metric and a loss for bounding box regression," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp658-666, 2019.

[23] Z. ZHU, D. LIANG, and S. ZHANG, "Traffic-sign detection and classification in the wild," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp2110-2118, 2016.