International Journal of Computer Science and Information Technology
ISSN: 3005-9690 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 2, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n2.05
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Traffic Sign Recognition Algorithm Based on YOLOv5

Zhichen Li, Hua Huo

College of Information Engineering, Henan University of Science and Technology, Kaiyuan Avenue 263, Luoyang, 471023, China

ABSTRACT

With the continuous development of driving assistance system and automatic driving technology in recent years, the requirements for traffic sign recognition technology are becoming higher and higher. Although the current mainstream target detection technology has been guaranteed in real-time and accuracy, due to the complexity of traffic signs and the fact that most traffic signs in the actual scene are small and dense, the identification accuracy of small target traffic signs is low. Based on the above situation, an improved YOLOv5 traffic sign recognition algorithm is proposed. We add CBAM attention mechanism to the model to improve the feature extraction ability of the target object. The structure based on the feature pyramid is improved to strengthen the features of small targets. The upsampling algorithm is replaced to save computing power. The overall detection and recognition ability of the model is improved. The loss function is improved to DIoU loss function, which enhances the robustness of the model.

KEYWORDS

YOLOv5; CBAM; Convolutional neural network; Traffic sign recognition

1. INTRODUCTION

In recent years, with the development of convolutional neural networks, the field of image recognition has also been rapidly improved, and the detection speed and accuracy of traffic sign recognition have become higher and higher. At the same time, models with high accuracy are detected with more layers, while models with fast detection speed are less accurate. As a result, it is difficult to ensure the accuracy of traffic sign detection, but also ensure the real-time detection[1].

At present, traffic sign recognition technology has been widely used in the field of computer vision. Traffic sign recognition refers to the process of automatically identifying traffic signs on the road through computer vision algorithms and models, and classifying or extracting relevant information. The following is an introduction to the current traffic sign recognition technology:

(1) Image processing and feature extraction: The first step of traffic sign recognition is the preprocessing and feature extraction of the input image. Commonly used image processing techniques include image enhancement, edge detection, image segmentation, etc., to extract the key features of traffic signs.

(2) Feature representation and description: The extracted image features need to be properly represented and described to facilitate subsequent classification or recognition. Common feature representation methods include local feature descriptors (such as SIFT, SURF, ORB, etc.) and deep learning features (such as CNN, ResNet, etc.).

(3) Classification and recognition algorithms: The algorithms commonly used in traffic sign recognition include traditional machine learning methods and deep learning methods. Traditional machine learning methods include support vector machines (SVM), decision trees, random forests, etc., while deep learning methods use deep neural network structures for end-to-end feature learning and classification.

(4) Data sets and training: Traffic sign recognition usually requires large scale labeled data sets for model training and evaluation. These datasets contain images of various types of traffic signs, as well as corresponding label information. Common Traffic Sign data sets include German Traffic Sign Recognition Benchmark (GTSRB) and LISA Traffic Sign Dataset[2].

(5) Real-time and robust: traffic sign recognition needs to be real-time and robust in practical applications. Real-time performance requires that the algorithm can quickly complete the recognition in a short time to meet the needs of real-time traffic monitoring and driver assistance systems. Robustness requires that the algorithm can maintain accuracy and stability under various complex scenarios, such as lighting changes and weather conditions.

(6) Development of deep learning: In recent years, deep learning technology has made remarkable progress in traffic sign recognition. By using the deep convolutional neural network (CNN) structure, more accurate classification and detection of traffic signs can be achieved. For example, CNN-based models can automatically learn the feature representation and classification patterns of traffic signs through end-to-end training.

(7) End-to-end system: Traffic sign recognition is often used as part of a complete end-to-end system. These systems may include multiple components such as image acquisition equipment, image transmission and preprocessing, sign recognition and classification, and result output.

In general, the development of traffic sign recognition technology has made great improvements in areas such as intelligent transportation systems and driver assistance systems. Through continuous research and innovation, traffic sign recognition technology will further improve accuracy, speed and robustness to meet the growing demand for traffic safety.

In the traditional traffic sign recognition algorithm, the main research focus is the extraction and classification of features. The features of traffic signs are extracted through the segmentation of color space and the shape, edge and other features of traffic signs, and then the classifier completes the feature classification to realize the recognition of traffic signs. Now, the object detection algorithm based on deep learning has gradually replaced the traditional object detection algorithm. For example, "two-stage" target detection algorithms include Faster-RCNN, Faster-RCNN, and "one-stage" target detection algorithms include YOLO algorithm and SDD algorithm. Ellahyani et al.[3] set HOG and HIS color features and adopted SVM to improve the recognition accuracy. Qian et al.[4] used the network generated after the cascade of Fast R-CNN and two CNN networks to identify traffic signs, and obtained an accuracy rate of 90.2%. However, the network has high requirements for computer hardware, and it cannot recognize traffic signs in real time. Guo Jifeng et al. applied DSConv to the feature extraction network of YOLOv4-tiny, reducing the calculation amount of the model. Khan et al.[5] proposed a new intelligent traffic sign recognition system with lighting preprocessing capability to address the problem of dark areas in traffic signs in complex environments, and enhanced the light in the dark areas of images through brightness enhancement technology. Good results are obtained on GTSDB data sets. Based on YOLOv5, Wang et al.[6] proposed a new feature fusion network AF-FPN for traffic sign detection, which significantly improved the detection accuracy of small target signs, but decreased the real-time detection speed.

In the field of traffic sign recognition, YOLO(You only look once) series of deep learning algorithms have attracted a lot of attention with their excellent recognition performance and proud detection speed. Since Redmon et al first proposed the YOLOv1 model structure in 2016, a large number of researchers have been trying to optimize it, and the YOLO model has also been continuously

optimized and improved, and the YOLO series has been constantly updated until the YOLOv5 model launched in 2020. It surpasses all the previous performance of object recognition model at one stroke, and is very suitable for application in actual scenarios [7-9]. Nowadays, many researchers and scholars have applied YOLO related models to traffic sign recognition detection and have achieved very good detection results. For traffic sign recognition, there are two main features: accuracy and real-time. In today's YOLO model, the detection speed is already very fast, and it also has a good effect on real-time performance. Therefore, we still need to further optimize the accuracy to achieve the effect required by the actual application scenario.

In this paper, the YOLOv5 algorithm with fast detection speed, small training model and high precision is adopted. And the training test is carried out on TT100K data set.

2. ORGANIZATION OF THE TEXT

2.1. YOLOv5 model introduction

According to the depth and width of the network structure, YOLOv5 can be subdivided into four specific models: YOLOv5m, YOLOv5s, YOLOv5l and YOLOv5x. For the consideration of detection speed and real-time performance, we choose YOLOv5s with the smallest model as the basic network.

YOLOv5 is mainly composed of input, Backbone, Neck and Prediction. First, the Backbone part is a convolutional neural network that aggregates and forms image features on different image fine-grained. Second, the Neck part is a series of network layers that mix and combine image features and pass the image features to the prediction layer. The Head part is to predict the image features, generate bounding boxes and predict categories.

In the backbone network part, there are 6 layers of network structure to extract image features. Two kinds of CSP structures are designed in YOLOv5 model, which are used to increase the capability of feature extraction of the network, and are applied to the backbone network and the detection module respectively. In addition, the slice structure is added before entering the backbone network to ensure that the feature extraction ability is more adequate in the process of downsampling, but the storage of more complete downsampling information.

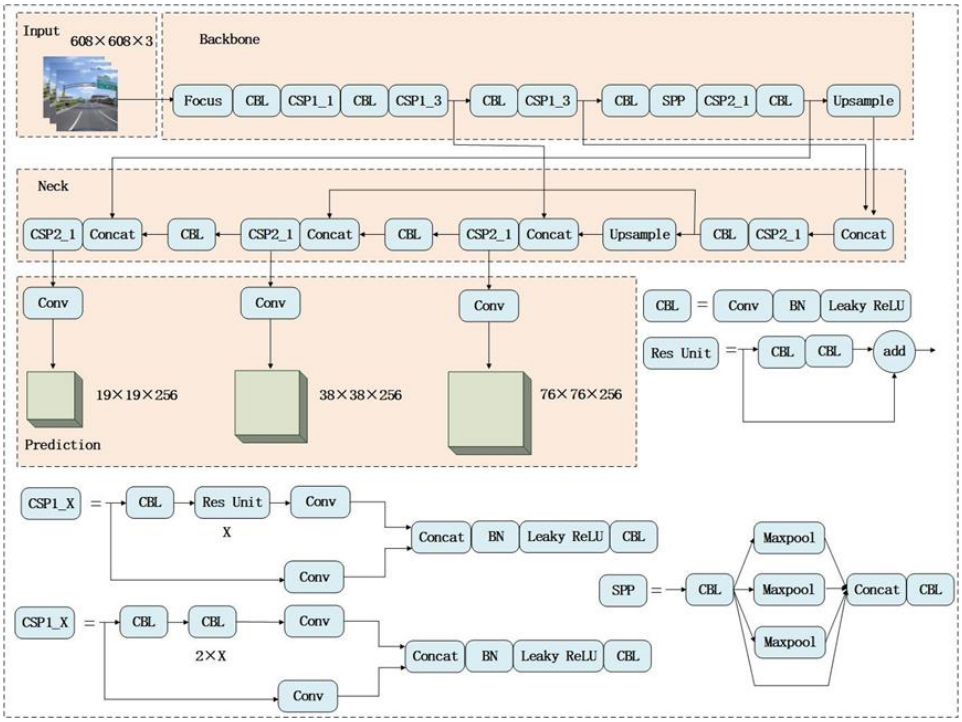

Fig. 1 YOLOv5 network structure

In the detection module, the YOLOv5 model applies the multi-scale feature fusion strategy to integrate deep and shallow layers of semantic information by using Spatial Pyramid Pooling Network (SPPNet) and path aggregation network (PANet). The feature extraction capability of the network model is enhanced by adding CSP2 structure.

For the prediction module, the model predicts and classifies the feature maps of different sizes according to different height and width prior boxes, divides them into three different target features, and then generates loss functions according to different results to feed back to the network for optimization.

According to the different depth and width, the YOLOv5 model can be divided into YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. For the consideration of real-time recognition and model size, YOLOv5s is finally adopted as the recognition network.

Figure 1 shows the network structure of YOLOv5. The network is divided into four parts, namely Input, Backbone, Neck and Prediction. In the Backbone part, the network mainly realizes the output of three different scale feature maps by multiple convolution and Concat operations on the input images. In the Neck part, after adding the perceptual adversarial network (PAN)[10] to the feature pyramid netword (FPN)[11], On the other hand, the information between different network backbones and different detection layers can be better fused. At the same time, CSP1_X and CSP2_X cross stage parital (CSPNet)[12] are applied to the Backbone and Head of the network respectively, which greatly improves the capability of network feature fusion.

2.2. Feature extraction based on lightweight attention module

In 2018, Sanghyun et al. proposed a light-weight CBAM (Convolutional Block Attention Module)[13], which can pay Attention in both channel and spatial dimensions. The CBAM module is added to the classical models such as ResNet and MobilwNet for comparative analysis and visualization. The experiment finds that CBAM pays more attention to the identification of target objects, which also makes CBAM more explanatory.

CBAM consists of two separate submodules, the Channel Attention Module (CAM) and the spatial attention module (SAM), for channel and spatial Attention respectively. This can not only save computing power, but also ensure that it can be integrated into the existing network architecture as a plug and play module.

2.2.1. Channel Attetion Module(CAM)

The specific process of CAM is: The input feature map F(H×W×C) is obtained by global max pooling(global maximum pooling) and global average pooling(full average pooling) based on width and height respectively. Two 1×1×C feature maps are obtained. They are then fed into a two-layer neural network (MLP), the first layer is C/r(r is the decrement rate), the activation function is Relu, and the second layer is the number of neurons C, which is shared by the two layers of neural networks. Then, the MLP output features are added and combined based on element-wise, and the final channel attention feature is generated through sigmoid activation. Finally, an element-wise multiplication operation is performed with the input feature graph F to generate the input features required by the Spatial attention module.

2.3. Feature Pyramid and Path Aggregation Network (FPN+PAN)

In the Neck part, YOLOv5 mainly adopts the PANet structure. PANet extracts the feature hierarchy on the FPN (feature pyramid network). The top information flow in FPN needs to be transmitted layer by layer through the Backbone. Due to the relatively large number of layers, the computation is relatively heavy.

PANet introduces a Bottom-up path on top of FPN. After the Top-down feature fusion, the bottom-up feature fusion is carried out, so that the position information at the Bottom can also be transmitted to the deep layer, thus enhancing the positioning ability at multiple scales.

2.3.1. FPN

FPN(Feature Pyramid Network) constructs a top-down hierarchical structure with lateral connections to construct high-level semantic features at various scales. FPN can be used as a general-purpose feature extractor and brings significant performance improvements on multiple tasks. FPN is applied to Faster RCNN to achieve the best single model performance on COCO. In addition, the inference speed of FPN can reach 5FPS on the GPU, so it is a method with high detection performance and inference speed that can reach practical use.

The goal of FPN is to use the hierarchical semantic features of convolutional networks to build a feature pyramid. FPN consists of two parts: the first part is the bottom-up process, and the second part is the fusion process of the top-down and lateral connections.

Bottom-up process: The bottom-up process is no different from a normal CNN. The modern CNN network is generally divided into different stages according to the size of the feature map, and the scale ratio difference of the feature map between each stage is 2. In FPN, each stage corresponds to a level of the feature pyramid, and the last feature of each stage is selected as the feature of the corresponding level in the corresponding FPN. Take ResNet as an example, the last residual block layer feature of conv2, conv3, conv4, conv5 layers is selected as the feature of FPN, denoted as (C2, C3, C4, C5). The steps of these feature layers relative to the original image are 4, 8, 16 and 32, respectively.

Top-down processes and lateral connections: Top-down processes amplify a small feature map at the top level (e.g. 20) to the same size as the feature map at the previous stage (e.g. 40) by up-sampling. The advantage of this is that it takes advantage of both the strong semantic features at the top (for classification) and the high-resolution information at the bottom (for localization). The upsampling method can be implemented with the nearest neighbor difference. In order to combine the high level semantic features with the low level accurate localization capability, we propose a lateral connection structure similar to the residual network. In the lateral connection, the features of the previous layer with the same resolution as that of the current layer after up-sampling are fused by the method of addition (here, in order to correct the number of channels, the current layer is first convolved by 1x1).

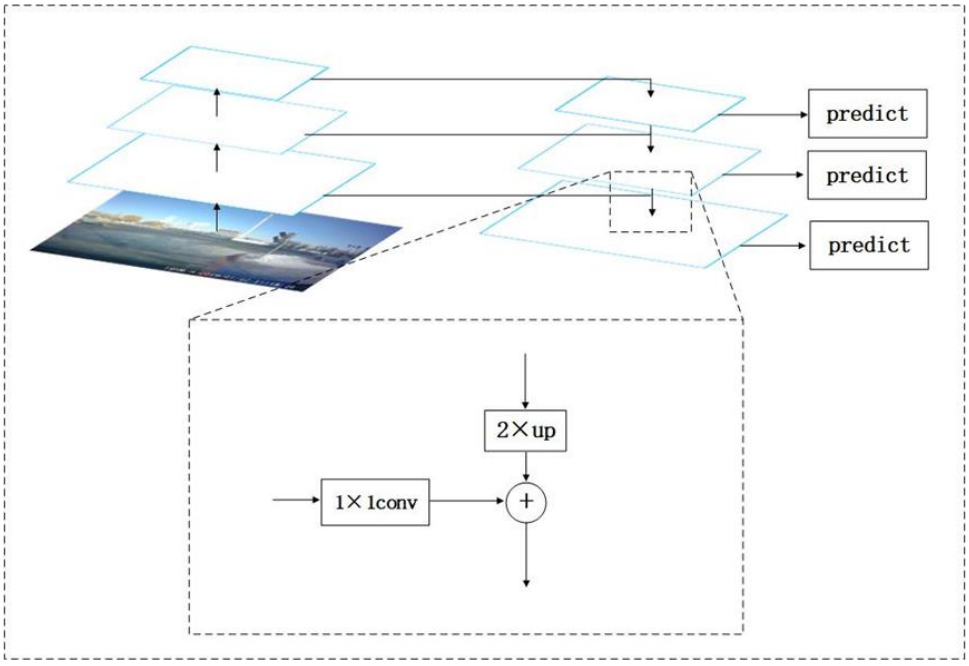

Fig. 2 Structure diagram of FPN

2.3.2. PANet

The biggest contribution of PANet (Path Aggregation Network) is to propose a bidirectional convergence backbone network from the top down and bottom up, while adding a "short-cut" between the lowest and highest layers to shorten the path between layers. PANet also proposed two modules: adaptive feature pooling and full connection fusion. Adaptive feature pooling can be used to aggregate features between different layers to ensure the integrity and diversity of features, and more accurate mask prediction can be obtained through full connection fusion.

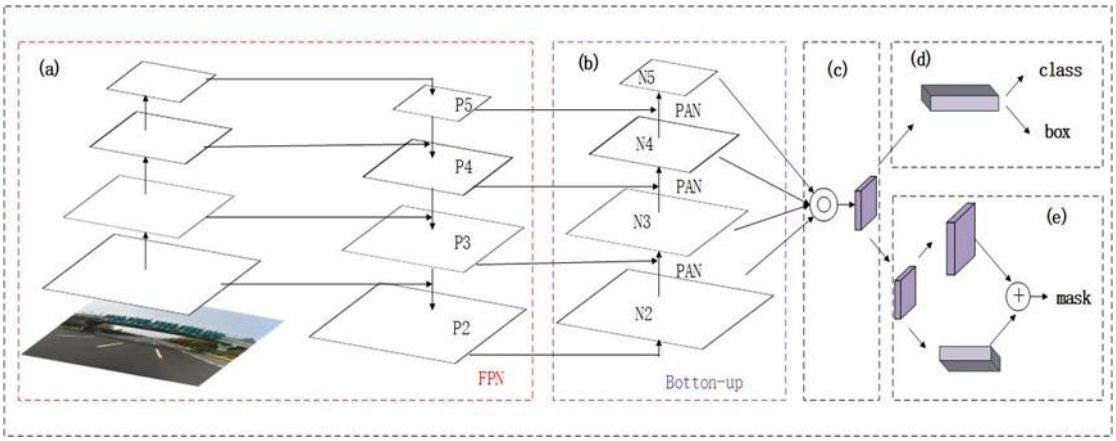

Fig. 3 Overall structure of PANet

The network structure of PANet is shown in Figure 1, which consists of five core modules. Among them, (a) is an FPN, (b) is a bottom-up feature fusion layer added by PAN, (c) is an adaptive feature pooling layer, (d) is a bounding box prediction head of PANet, and (e) is a fully connected fusion layer for predicting masks.

3. YOLOV5 NETWORK MODEL IMPROVED

Among the multiple model versions of YOLOv5 that have been released, YOLOv5s is famous for its shallow network depth, few model layers, fast detection speed and other characteristics, and is often used in real scene projects, especially the detection of road traffic signs and complex traffic scenes. Next, the YOLOv5 detection model network is improved from the aspects of feature extraction and loss function of the backbone network.

3.1. Improvement of YOLOv5 algorithm based on the addition of attention module CBAM

In traffic sign recognition, it is a key point to select important and non-important information in the process of feature extraction. Feature extraction is an important part of traffic sign recognition and a process to reduce the dimension of complex original data. In the process of feature extraction, it is more necessary to pay attention to some key target representation information, improve the attention to target location information and remove some unimportant information.

YOLOv5 algorithm based on CBAM attention module, CBAM is composed of channel attention module (CAM) and spatial attention module (SAM). It enables the network to better understand the feature information of the target object and its position in the original picture through weight redistribution.

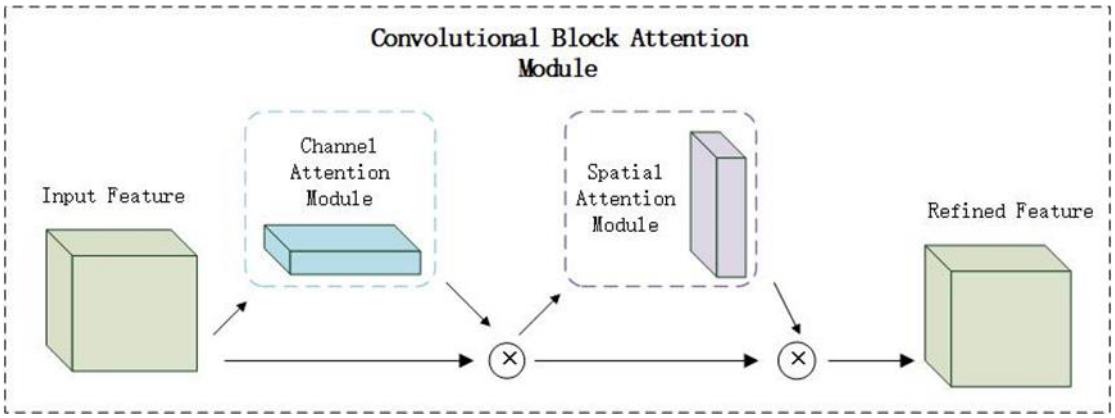

Fig. 4 CBAM structure

In order to eliminate the influence of channel dimension and space dimension on input feature maps, two one-dimensional feature maps are obtained by average pooling and max pooling in CAM and SAM modules. Convolution operations and nonlinear processing then provide channel $M_c(F) \in R^{c\times1\times1}$ and spatial attention $M_s(F) \in R^{H\times W}$ to their processing. Finally, the channel weights will be reassigned according to the obtained channel attention and spatial attention to obtain the final feature map.

The formula for calculating channel attention $M_c(F) \in R^{c\times1\times1}$ is as follows:

$$M_c(F) = \sigma((W_1(W_0(F^c_{avg}))) + W_1(W_0(F^c_{\max}))) \tag{1}$$

In the above formula: σ is the sigmoid activation function; $W_0 \in R^{c/r\times w}$ and $W_1 \in R^{c\times c/r}$ is the weight of the shared neural network.

The calculation formula of spatial attention $M_s(F) \in R^{H\times W}$ is:

$$M_s(F) = \sigma(f^{7\times7}([F^s_{avg}; F^s_{\max}])) \tag{2}$$

In formula (2) : $f^{7\times7}$ is a convolution kernel of size 7×7.

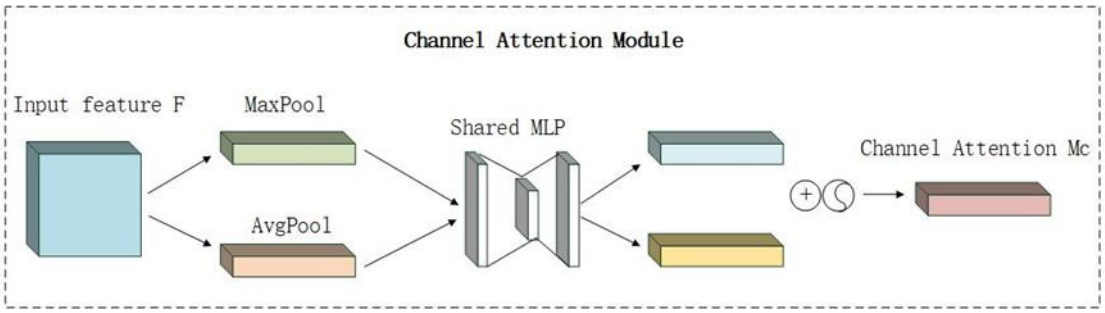

Fig. 5 CAM structure

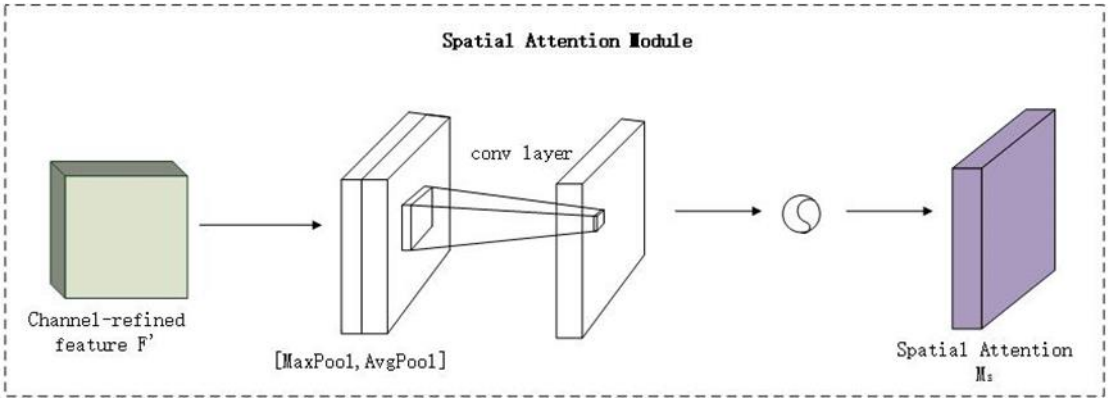

Fig. 6 SAM structure

Finally, we get the YOLOv5 network model embedded in CBAM, and its structure is as follows:

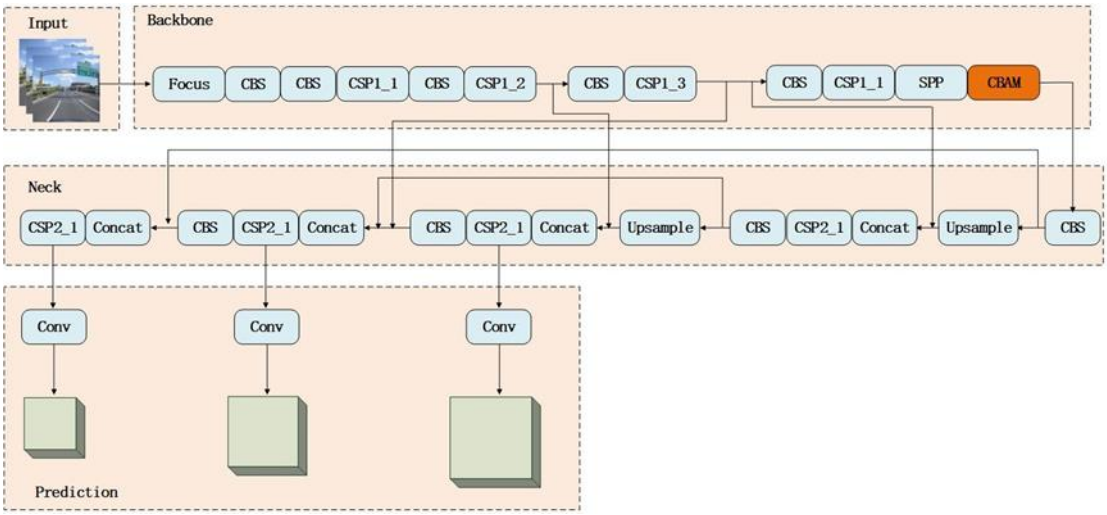

Fig. 7 YOLOv5 structure embedded in CBAM

3.2. Improvement of feature pyramid structure

(1) improvement of breed

Since the outer contour of the traffic sign is almost square, the pixel size occupied by the sign is measured by the result of downward rounding of the mean of the width and length of the sign. Statistics show that the average length and width of traffic signs are mainly concentrated in the interval [2,64], and the recognition task is mainly small targets. In FPN and PAN structures, the upper layer of the pyramid has a wide field of view, the single feature value covers a large pixel area, the features of the large target are significant, and the small target is submerged by background and noise. The original feature pyramid structure is not suitable for traffic sign recognition tasks. By reducing the structure of FPN and PAN (P/N represents N times downsampling), the network depth is reduced to solve the problem of small target features being submerged. Limit the maximum downsampling multiples. The input of the Detect layer is the result of 4, 8, and 16 times downsampling of the original input features to strengthen the low-level small target features.

(2) The upsampling algorithm is improved

PAN needs to up-sample the high-level features before transmitting them downward. The original YOLOv5 model adopts the nearest neighbor interpolation method for up-sampling, which has minimal computational overhead but low algorithm accuracy. It is intended to improve the bilinear interpolation method for up-sampling to reduce the interference caused by outliers on feature

transmission. Both the nearest neighbor interpolation method and the double-line interpolation method are based on the backward mapping principle, mapping from the target feature to the source feature. The difference is that the nearest neighbor interpolation method takes only one reference point in the source feature, while the bilinear interpolation method calculates four reference points at the same time. If the feature size is n×n after upsampling, then the time complexity of the nearest neighbor interpolation method is O(n2), and that of the bilinear interpolation method is O(4n2), and the extra computation cost is acceptable relative to the improvement of accuracy.

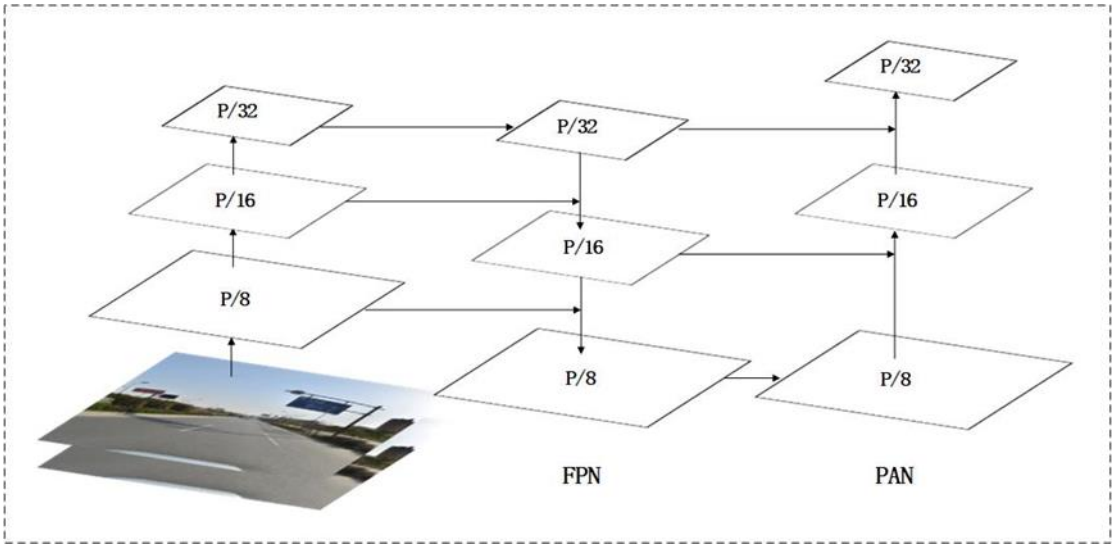

Fig. 8 Original FPN and PAN structure

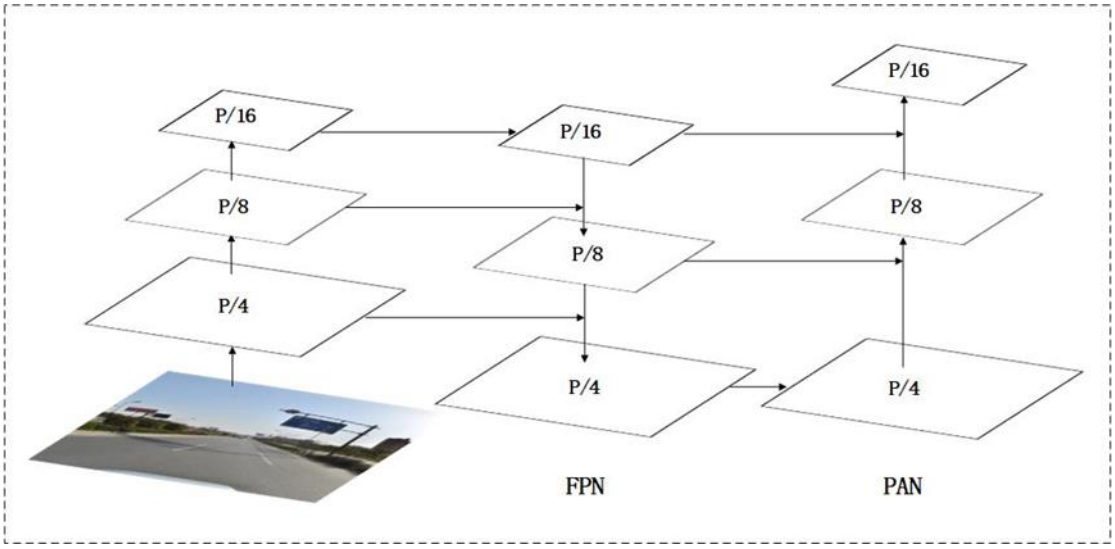

Fig. 9 Improved FPN and PAN structures

3.3. Improvement of loss function

As an operational function to measure the degree of difference between the predicted value and the real value of the model, the smaller the value, the better the robustness of the model. Therefore, it is particularly important to select the appropriate loss function in the training process.

Because of its scale invariance, symmetry, identity, non-negative and other advantages, and the output loss value is between 0-1, IoU can well express the detection effect of prediction box and real box, but there will be situations in the actual detection that the IoU loss function cannot be used.

(1) As shown in the figure: when the prediction box and the real box are completely disjoint, the output value of the IoU is 0, and the output of the loss function is 1, but it is obvious from the figure that the prediction box and the real box are close, and the output of the loss function should be smaller.

(2) As shown in the figure: when the intersection ratio between the prediction box and the real box is the same, the calculated loss function value is the same, but it cannot judge which prediction box is more accurate.

YOLOv5 network adopts GIoU Loss as the regression loss function of bounding box. Based on the characteristics of IoU, GIoU introduces the minimum external frame to solve the case that the above IoU loss function cannot be used, paying attention not only to the coincidence area, but also to other non-coincidence areas. The calculation formula is as follows:

$$\ell GIoU = 1 - IoU + \frac{A^c - B \cap B^{gt}}{B \cup B^{gt}} = 1 - \frac{B \cap B^{gt}}{B \cup B^{gt}} + \frac{A^c - B \cup B^{gt}}{A^c} \tag{3-1}$$

B and Bgt are prediction boxes and real boxes respectively.

Fig. 10 The prediction box does not intersect the real box at all

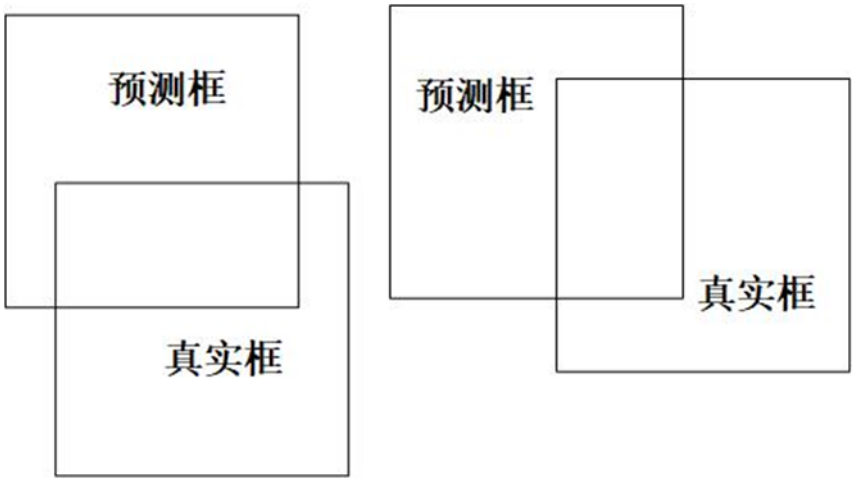

Fig. 11 The intersection ratio between prediction box and real box is the same

When the real box and the target box are included, the loss values of GIoU and IoU are equal, and the relative position relationship between the two boxes cannot be reflected. Meanwhile, in the training process, GIoU Loss chose to expand the size of the prediction box to increase the intersection with the real box, so as to maximize the overlapping area between the two boxes, so the network convergence slowed down.

To solve the above problems, DIoU Loss is proposed to be used as the network regression loss function, and the formula is as follows:

$$\ell DIoU = 1 - IoU + \frac{\rho^2(b - b^{gt})}{c^2} = 1 - \frac{B \cap B^{gt}}{B \cap B^{gt}} + \frac{\rho^2(b - b^{gt})}{c^2} \tag{3-2}$$

b and bgt are the center points of the prediction box and the real box respectively, and c is the diagonal distance of the smallest external rectangle C, representing the Euclidean distance of the two center points b and bgt.

4. EXPERIMENTAL VERIFICATION AND RESULT ANALYSIS

4.1. Improvement of loss function

In order to adapt to our country's traffic road environment better. In this paper, TT-100K data set is used for training. The TT-100K dataset, compiled and published by a joint lab of Tsinghua and Tencent, provides 100,000 images containing 30,000 traffic signs, and the images are derived from Tencent Street View panoramas taken by six high-pixel wide-angle SLR cameras in various cities in China, with different lighting conditions and weather conditions. The resolution of the original Street View panorama was 8192×2048, then the panorama was cropped into four parts, and the final data set was 8048×2048. The categories of traffic signs contained in the TT-100K dataset are relatively comprehensive, and 221 different categories appear in the whole dataset.

In order to evaluate the performance of the network model more objectively, Precision, Recall and mean average precision mAP(mean average precision MAP) are used as evaluation indicators. Precision refers to the proportion of the predicted targets predicted by the model that is correct. Recall is the percentage of all real targets that the model predicts correctly; The average accuracy is the average of all categories of prediction accuracy. The specific calculation formula is as follows:

$$\Pr ecision = \frac{TP}{TP + FP} \tag{3-3}$$

$$\mathrm{Re}\, call = \frac{TP}{TP + FN} \tag{3-4}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{3-5}$$

TP represents the number of samples correctly predicted by the model, FP represents the number of positive samples incorrectly predicted by the model, and FN represents the number of negative samples incorrectly predicted by the model. AP is obtained by dividing the sum of all accuracy rates for a class in the set by the number of images containing targets for that class.

4.2. Experimental environment and parameter Settings

The experimental environment of this paper is Windows 10 operating system and PyTorch deep learning framework. GTX 1060 with 6GB video memory is used to train the model. Some of the hyperparameters set for the network during the training process are shown in Table 1.

Table 1 Partial hyperparameter Settings

Training parameter	parameter values
momentum	0.937
weight_decay	0.0005
Initial learning rate (lr0)	0.01
Cyclic learning rate (lrf)	0.1
batch-size	16
epochs	300

The initial learning rate set in this article is 0.01, and at 100 epochs and 200 epochs the learning rate drops to 0.001 and 0,0001, respectively. Batch is set to 16 and training stops after 300 epochs.

4.3. Analysis of experimental results

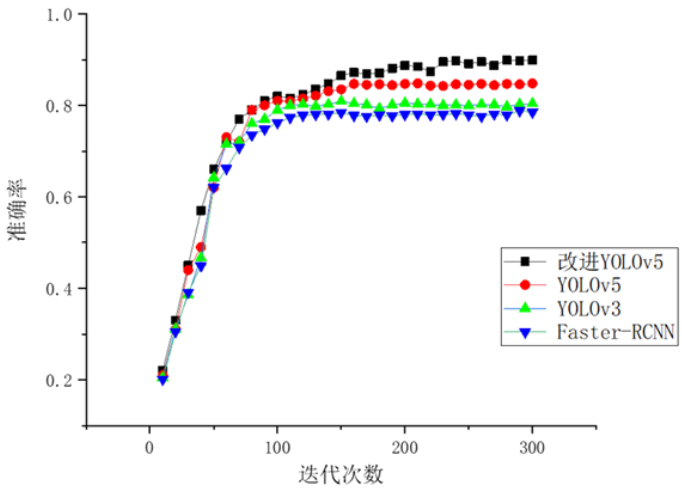

Fig. 12 Comparison of accuracy of each model

As can be seen from Table 2, the average recognition rate of the improved YOLOv5 algorithm adopted in this paper reached 88.87%, an increase of 4.09% compared with the YOLOv5 algorithm, indicating that the improved YOLOv5 algorithm has a more accurate recognition accuracy than the traffic sign.

Table 2 Comparative experimental results

model	mAP	Model size
SSD	75.38	100.23
Faster R-CNN	78.43	108.00
YOLOv3	80.42	218.84
YOLOv5	84.78	27.00
SE-YOLOv5	86.14	17.35
Improved YOLOv5 algorithm	88.87	30.17

In order to verify the effect of the above improved methods, this paper conducted ablation experiments on TT100K dataset. The results showed in Table 2 that adding CBAM attention mechanism to the model, improving feature pyramid and changing upsampling algorithm can all improve the network accuracy, and the mAP of the three can increase by 1.78%, 1.8% and 0.51%, respectively. The improved algorithm combined with the three has better results in precision rate, recall rate and mPA.

Table 3 Results of ablation experiment data

project	mAP	Model size
YOLOv5+CBAM	86.56	28.10
YOLOv5+ Improved pyramid	86.58	27.00
YOLOv5+ Upsampling algorithm	85.29	28.56
YOLOv5+DIOU	84.89	27.10
YOLOv5+CBAM+ Improved pyramid	87.93	28.21
YOLOv5+CBAM+ Upsampling algorithm	86.87	29.71
YOLOv5+CBAM+DIOU	86.49	28.10
YOLOv5+ Upsampling algorithm + Improved pyramid	87.42	28.56
YOLOv5+ Improved pyramid +DIOU	87.23	27.10
YOLOv5+ Upsampling algorithm +DIOU	86.13	28.61
YOLOv5+CBAM+ Upsampling algorithm + Improved pyramid	88.66	30.10
YOLOv5+CBAM+ Upsampling algorithm + Improved pyramid +DIOU	88.87	30.17

As can be seen from the above table, the accuracy of the network model is greatly improved compared with the original model when the parameters are unchanged. The main reason is that the improved up-sampling algorithm reduces the interference of outliers on feature sampling, the perceived feature information is more complete, and the feature extraction is more clear and accurate. The CBAM dual-channel attention mechanism improves the feature attention degree, redistributes the channel weight, saves computing power and improves the target object's attention through dual attention of channel attention and spatial attention. By reducing the network depth, the improvement of feature pyramid greatly reduces the probability of small targets being submerged by noise, improves the detection rate of small targets, and thus improves the recognition accuracy. The DIoU loss function speeds up the convergence speed and improves the robustness of the model by adjusting the overlap position between the predicted frame and the real frame.

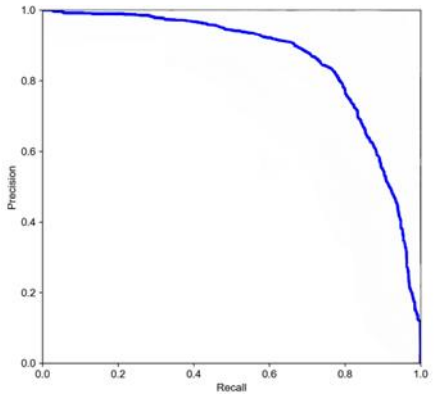

Fig. 13 Recall rate curve

The Recall curve is a graph used to evaluate the performance of a classification model, which shows the relationship between the recall and the corresponding precision of the model under different thresholds. In the binary classification task, the recall rate refers to the number of samples correctly predicted to be in the positive category as a percentage of the actual positive category sample, while the accuracy rate refers to the number of samples correctly predicted to be in the positive category as a percentage of all samples predicted to be in the positive category. In a recall curve, the horizontal axis is usually the recall rate and the vertical axis is usually the accuracy rate, with each point corresponding to a specific threshold. The shape and trend of the curve can help us understand how the model performs at different recall and accuracy rates. In general, we want the model to have a high recall rate (that is, a higher recognition rate for samples in the positive category) and a high accuracy rate (that is, the model has as few misjudgments as possible for the negative category), so the larger the upward trend of the recall curve and the larger the area under the curve, the better.

As can be seen from the recall rate curve, the curve trend of the improved Yolov5 model in this paper is relatively smooth, and the accuracy index is also relatively high with the high recall rate. It can be concluded that the experimental model has a relatively good improvement in recognition accuracy.

Next, it mainly shows the training effect of the improved model on the data set. The superiority of the improved model's performance can be seen from the data curve, thus demonstrating the validity of our most realistic and objective experimental data.

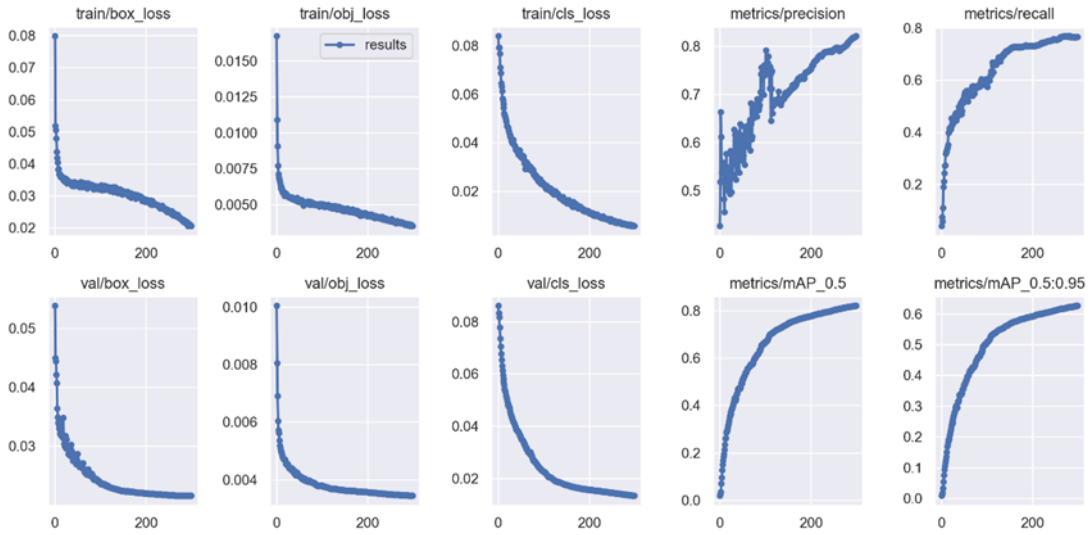

Fig. 14 Improved model training results

To sum up, the improved YOLOv5 model in this paper plays a significant role in strengthening target recognition objects, reducing noise interference to small target recognition, and enhancing model robustness, which improves the accuracy of traffic signs and can meet the accurate recognition of road traffic signs under complex road conditions in real scenes, demonstrating the effectiveness of the improved method in this chapter. Finally, the experimental effect of our improved model is shown in the figure below:

Fig. 15 Traffic sign recognition effect

5. CONCLUSION

This paper focuses on the current traffic sign recognition problem of dynamic road conditions and long-distance small targets in the road clutter scene. With the increasing number of road vehicles in our country, the road conditions are becoming more and more complicated, and the corresponding traffic signs are more and more numerous. Based on the YOLOv5 network model, CBAM dual-channel attention mechanism is added to enhance the attention to the target, while reducing the network depth of the feature pyramid in the model, improving the recognition of small targets, and solving the problem that the features of small targets are submerged by noise interference in the recognition process. Bilinear interpolation method is used for up-sampling, which weakens the disturbance caused by outliers in the process of feature transmission and improves the transmission accuracy. Finally, DIOU is used as the loss function of the model, which enhances the robustness of the model and makes the identification process more stable. The effectiveness of the improved model is proved by comparing with other models on the open traffic sign data set.

REFERENCES

[1] YURTSEVER E, LAMBERT J, CARBALLO A, et al. A survey of autonomous driving: Common practices andemerging technologies[J]. IEEE Access, 2020, 2(8): 58443-58469.

[2] JIANMING Z, MANTING H, XIAOKANG J,et al.A real-time Chinese traffic sign detection algorithm based on modified YOLOv2 [J]. Algorithms, 2017,10(4):127.

[3] ELLAHYANI A, ANSARI M AND JAAFARI I. Traffic sign detection and recognition based on random forests[J]. Applied Soft Computing, 2016, 3(46): 805-815.

[4] QIAN R, ZHANG B, YUE Y, et al.Robust Chinese traffic sign detection and recognition with deep convolutional neural network [C]. International Conference on Natural Computation,2015:791-796.

[5] KHAN J A, YEO D, SHIN H. New dark area sensitive tone mapping for deep learning based traffic sign recognition[J]. Sensors, 2018, 18(11): 3736.

[6] WANG J, CHEN Y, GAO M, et al. Improved YOLOv5 network for real-time multi-scale traffic sign detection[J]. Computer Science, 2021,arXiv:2112.08782.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[8] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[9] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[10] Li H , Xiong P , An J , et al. Pyramid Attention Network for Semantic Segmentation[J]. 2018.

[11] Lin T Y , P Dollár, Girshick R , et al. Feature Pyramid Networks for Object Detection[J]. arXiv e-prints, 2016.

[12] Wang C Y , Liao H , Wu Y H , et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE, 2020.

[13] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.