Traffic Sign Detection and Recognition Based on Deep Learning

H. Zhang, J. Zhao

***Abstract*—As the core technology in the field of intelligent transportation, traffic sign detection and recognition play a significant role in unmanned driving, high-precision map navigation, and auxiliary driving. However, due to the inevitable extreme weather and human impacts in the process of driving on the road, traffic sign detection and recognition remains a difficult problem. In this paper, a traffic sign detection and recognition algorithm based on the improved YOLOv4 is proposed to improve the original Feature Pyramid Network, so that the feature layer has more semantic information. Additionally, the coordinate attention mechanism module is added to the model, which fully considers the relationship between channels and position information. An adaptive feature fusion module is also added before the detection head to solve the problem of unclear semantic information caused by scaling in target detection. The experimental results indicate that the detection accuracy of the improved YOLOv4 algorithm is increased by 5%, and the detection performance is better than that of other target detection algorithms.**

***Index terms*— adaptive feature fusion, attention mechanism, traffic sign detection, YOLOv4**

I. Introduction

THE appearance, colour, and significance of traffic signs in different countries vary due to the different driving regulations and traffic sign design concepts. This study is based on experiments exclusively conducted on traffic signs in China.

According to China's road traffic signs and line marking criteria [1], domestic road traffic signs are generally of two types: main and auxiliary signs. Each of them has its own distinct role on the road. The main signs are set on the side of the road to inform the drivers about rules applicable in the vicinity, whereas the auxiliary signs are located below the main signs to explain the role. Domestic road traffic signs can be divided into the following categories: warnings, prohibitions, and instructions.

This study focuses on the aforementioned three types of traffic signs, which are representative and widely used as experimental objects to test the recognition of traffic signs.

A. Warning signs

Most warning signs are placed in areas where traffic accidents frequently occur to inform drivers or pedestrians that the road ahead is dangerous. The warning signs are triangular in shape with the top angle facing upwards and are yellow and black in colour, as illustrated in Fig. 1.

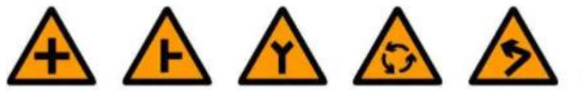


Fig. 1. Warning signs

B. Prohibition signs

Prohibition signs are set near the starting points of prohibited and restricted road sections. They also indicate the lifting of a road section to prohibit or restrict vehicles, such as no traffic or parking, except for some signs with white background, red circle, red bar, and black graphics, as shown in Fig. 2.

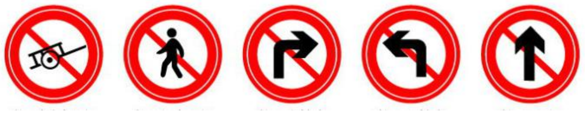


Fig. 2. Prohibition signs

C. Indication signs

Indication signs instruct vehicles and pedestrians to drive in specific directions and places. The colour of the indicator is blue and white, and its shape is round, rectangular, and square, as shown in Fig. 3.

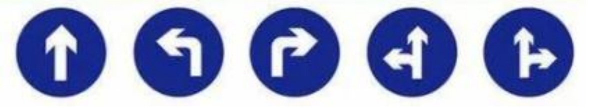


Fig. 3. Indication signs

Experts have proposed several detection and recognition algorithms in recent years, but the results indicate that they cannot fully cope with the detection and recognition tasks in various environments. Traffic sign recognition relates to many fields, such as artificial intelligence, intelligent transportation, and image processing. However, the current stage of traffic sign detection and recognition is presented with several difficulties, including complex weather conditions, shooting angles, object types, and damaged traffic signs, as illustrated in Fig. 4.

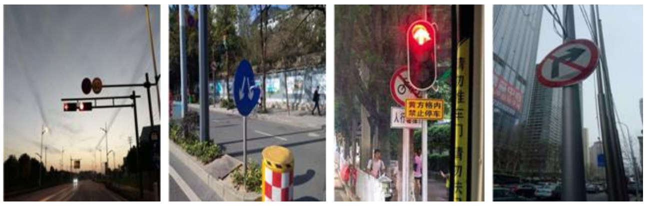


Fig. 4. Difficulties in traffic sign detection and recognition

For the model algorithm selection, we considered both speed and accuracy and chose YOLOv4 [2] as our prototype

Manuscript received August 31, 2021; revised February 14, 2022.This work was supported by the Natural Science Foundation of Liaoning Province, grant number 20180551048.

H. Zhang is a master student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan 114051, China (e-mail: 208696553@qq.com).

J. Zhao is a Professor of School of Computer Science and Software Engineering, University of Science and Technology, LiaoNing, Anshan114051, China (corresponding author, phone: 86-13998086167; e-mail: zhaoji@ustl.edu.cn).

algorithm.

In this study, the spatial pyramid pooling structure, multi-scale feature extraction, and other parts of the YOLOv4 were developed as follows:

1. The Reverse Cycle Feature Fusion (RCFF) module was added to improve the original Feature Pyramid Network (FPN) [3] traditional feature fusion module. Due to the circulation structure, feature fusion was carried out more than once to achieve the effect of repeated extraction of information.

2. The coordinate attention [4] mechanism module was added to the space pyramid structure to fully consider the relationship between channels and position information so that the model could locate and identify the target region more accurately.

3. The Adaptively Spatial Feature Fusion（ASFF）module [5] was added before the detection head to solve the inconsistent scale of different features in object detection, improve the problem of unclear semantic information caused by information transformation, and reduce the cost of model reasoning.

II. Related Studies

This section primarily reviews the development of algorithms for traffic sign detection and recognition. Traditional traffic sign detection algorithms can be divided into three categories: traffic sign detection algorithms based on colour features, shape features, and machine learning.

A. Traffic sign detection methods based on colour features

The main principle of a traffic sign detection algorithm based on colour features is to distinguish the detection image information, which is obtained through processing or the block containing traffic sign information in a specific frame in the video by considering colour as the main information feature. Ruta *et al.* [6] used a method based on colour distance transformation to reduce the negative impact of light on colour, made comparisons with simple and powerful images of spatial information, and used them to detect traffic signs.

Xin *et al.* [7] used a method based on colour standardization; the proposed site image was projected onto a colour standardization image, composed of eight different colours, to ensure a uniform influence of light on the colour during image detection and through data mining in different colour feature regions. Thus, the method could detect and identify different traffic signs using set features. Fang *et al.* [8] used primary colours (i.e. red, yellow, blue) for the extraction of image information and the rest of the white and black colours to identify the characteristics of a symbol. In 2004, Huang *et al.* [9] used the illumination of the segmentation algorithm based on the RGB colour space model; the proposed method was not affected by the weather, and the real-time performance of the algorithm was enhanced. At the same time, it avoided the traditional method of colour segmentation, i.e. the RGB colour space was converted into the HSI colour space. In 2012, Yang *et al.* [10] proposed a multi-stage extraction algorithm. In the beginning, the probability model of each colour was calculated, and the error was reduced using integral channels. Additionally, the real-time information detection improved significantly.

Based on the colour feature extraction method, the accuracy and real-time performance of traffic sign detection have improved. However, the problem of bad traffic sign detection under extreme environmental conditions has not yet been addressed, damage in perennial traffic signs is unavoidable, and the morphology of deformation defects can cause low detection accuracy.

B. Traffic sign detection methods based on shape features

Traffic signs mostly possess rectangular, triangular, and circular shapes. Hence, geometric features are also a standard method used by researchers to detect and recognize traffic signs. In 1962, Hu [11] first proposed a theory based on algebraic invariants; the proposed method considered traffic signs at the centre of the image as a grey value based on the corresponding distance to describe the characteristics of different grayscale images. However, this method of traffic sign recognition and detection led to significant computational problems; due to long calculation times and the need for high-level computer hardware, the testing accuracy with general equipment was mostly insufficient. Gavrila [12] proposed a novel method, called the Hough transform, for detection of traffic signs. The primary idea was to use the Hoffman transformation within a specific range to achieve the purpose of testing using certain geometric transformations; circular lines were used to detect circular traffic signs, and vertical lines were used to detect rectangular and triangular traffic signs. Moreover, Yu *et al.* [13] used a traffic sign detection algorithm based on the Fourier interpretation and saliency maps. The idea was mainly derived from the frequency tuning method, through which the image obtained via this method was determined by the value obtained from the depolarization method to detect the traffic sign detection range region. In 2013, Boumediene *et al*. [14] obtained certain required position nodes through programming decoding, determined the desired symmetric detection lines in the node coding, and subsequently changed the previously obtained lines into simple line segments for detection. This method can be used to detect warning traffic signs.

C. Traffic sign detection methods based on machine learning

In the development of today's computer vision, machine learning holds a dominant position because it performs well for the detection task and is widely used by many researchers [15]. For the traffic sign detection method that is based on machine learning, the general idea is that the classifier focuses on the area of information of the study. Subsequently, the trained classifier carries out feature matching on the learned data samples, following which traffic sign detection and recognition can be carried out.

In 1999, David [16] used scale invariance to carry out a feature transformation algorithm (called SIFT). In 2004, experts added some upgrades to this algorithm, which could obtain the required infographic in many key points of the scale without actually changing the scale, and the infographic contained all the information features. In 2005, Dalal *et al.* [17] proposed a new idea called the Histogram of Oriented Gradient (HOG) algorithm. The HOG has many features that remain unchanged in some fixed regions. In 2010, Creusen *et al.* [18] obtained the HOG features by calculating the RGB colour space. The information obtained via these three channels could better enhance the accuracy.

III. Traffic Sign Detection and Recognition Based on Improved YOLOv4

The aim of this section is to provide a concise description of the steps involved in our proposed method.

A. YOLOv4 Network Architecture

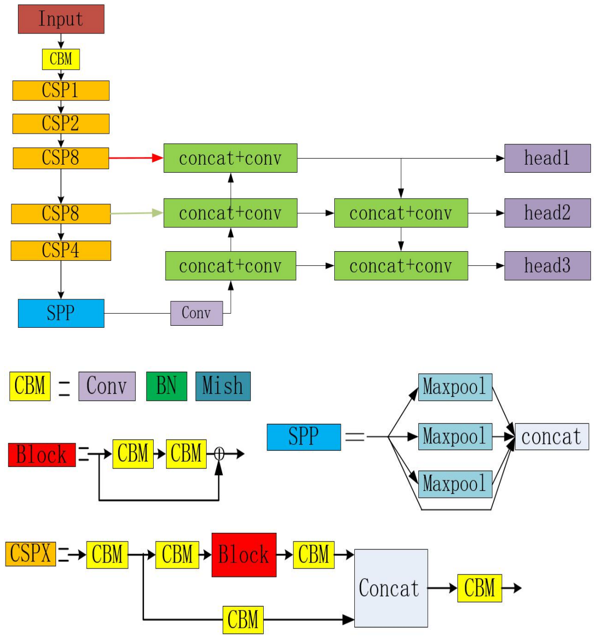


Fig. 5. YOLOv4 network architecture

As the fourth version of the YOLO series, YOLOv4 does not represent a substantial improvement compared with the previous generation. In essence, it is an improved model based on YOLOv3 [19], which is often used in various detectors to improve the performance. In comparison with the previous generations, YOLOv1 [20] and YOLOv2 [21], YOLOv4 can significantly improve the detection performance of the model under the condition of ensuring adequate speed.

The architecture of YOLOv4, as illustrated in Fig. 5, is CSPDarknet53+Spatial Pyramid Pooling (SPP)+Path Aggregation Network (PAN)+YOLO Head. A CSPDarknet53 with strong image feature extraction ability was selected as the backbone structure of the network, which was composed of 5 resblock_bodies (1, 2, 8, 8, 4 for each resblock_body); subsequently, SPP and PAN were used.

The SPP [22] structure was constructed at the last feature layer of CSPDarknet53. A total of three operations (i.e. convolution, batch normalization (BN), and leaky Rectified Linear Unit activation function) were performed. The pooling cores of four different scales were used for maximum pooling operations in image processing, and their pooling cores were 13*13, 9*9, 5*5, 1*1.

YOLOv4 adds the PAN [23] module, which is a multi-directional improvement on the mask region-based convolutional neural network [24] and takes feature fusion to an extreme degree, such as introducing FPN and Bottom-up Path Augmentation.

Bottom-up Path Augmentation considers the importance of superficial features of the network because they are generally edge features. In FPN, features are extracted in a bottom-up order and transferred from the shallow layers to the top layers. With tens or even hundreds of network layers, the information in the shallow layer is gradually lost after several layers of feature extraction; hence, we can transfer the feature map along with Bottom-up Path Augmentation to the top layer on top of the original FPN and no more than 10 layers. Overall, its structure is less than ten layers, so it can better preserve the semantic information at the shallow layer. Finally, three detection heads with different resolutions are connected for the information prediction of images with different scale sizes.

B. Improved YOLOv4 network architecture

The improved YOLOv4 algorithm, proposed in this study, is illustrated in Fig. 6. FPN was replaced by RCFF, an attention mechanism was added on the basis of the SPP spatial pyramid, and an ASFF module was added before the detection head.

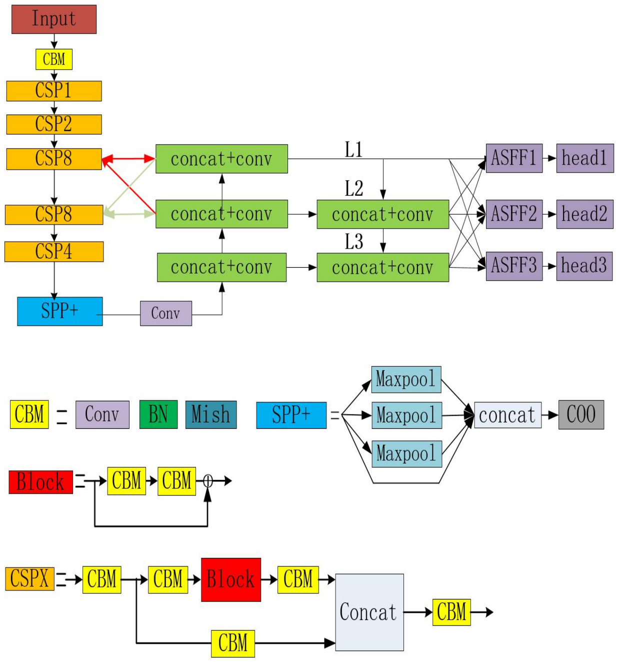


Fig. 6. Improved YOLOv4 network architecture

1) Reverse Cycle Feature Fusion

In this study, an RCFF module was proposed, which is depicted in Fig. 7. In the process of continuous exploration and development of the FPN mechanism, we discovered that humans must use visual perception to find objects of interest [25]. This perception absorbs the characteristics of the recursive feature pyramid through the feedback mechanism. After combining it with the top-down feature pyramid structure network and the feedback connection mechanism of feature fusion, the current target detection algorithm proposes the principle of "think before you act" and repeatedly collects, uses, and extracts feature information. Fig. 7 indicates that the RCFF module is built on FPN. Its main idea is to feed the semantic information of each layer in the FPN back to the bottom-up backbone network. Moreover, the top in the reverse connection of the characteristics of rich semantic information and the underlying characteristics of numerous details thoroughly mix, so that our model will be two images; in the main backbone structure, from bottom to top, and at the same time, the characteristics of FPN pyramid twice in different scale due to fusion. Therefore, considering the problem in this manner, we can maximize the detection accuracy and also facilitate the detection of small targets.

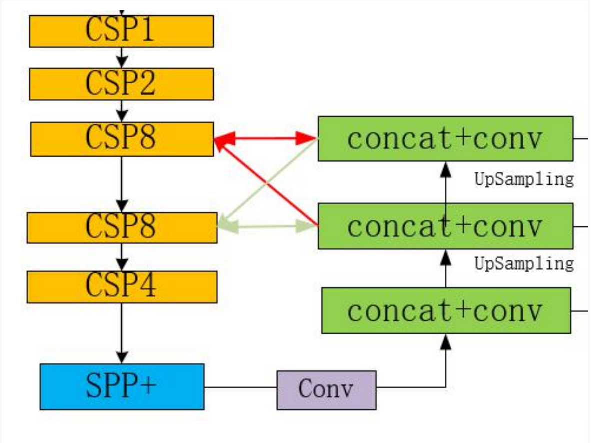


Fig. 7. Reverse cycle feature fusion

2) SPP+

To ensure that the size of the input image is suitable for recognition, the image is generally cut. However, if the above processing is carried out on the image, it is very likely to lose a lot of information features when extracting image features.

The function of the SPP module is based on the fact that when the convolutional neural network is used for image detection and recognition tasks, the target detection tasks can be well completed irrespective of the image size.

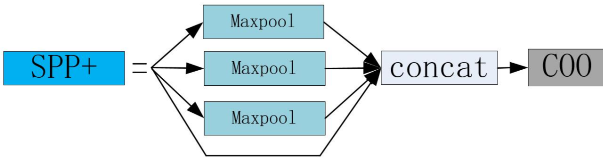


Fig. 8. SPP+

The structure was constructed at the last convolution layer of CSPDarknet53. At the last feature layer of CSPDarknet53, a total of three operations (convolution, BN, and leaky Rectified Linear unit activation function) were performed. Four different pooling cores were used for maximum pooling operations in image processing. The pool nuclei were 13*13, 9*9, 5*5, 1 *1.

On this basis, we added an efficient and lightweight coordinate attention mechanism module, as shown in Fig. 8, to avoid certain unnecessary parameters and save calculation costs when obtaining large information areas.

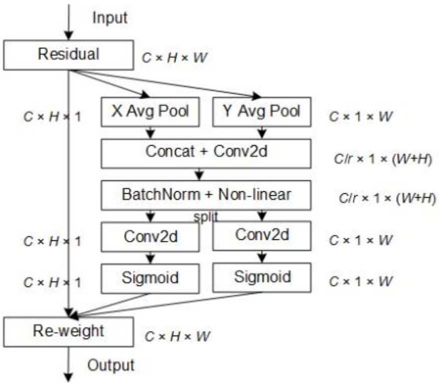


Fig. 9. Coordinate Attention mechanism

The location awareness of the two global pooling operations was carried out in two different directions. These two feature maps had the location feature information of specific directions; each feature map could capture the near and far dependence of the input feature map along a spatial direction, and the location information could be saved in the generated feature map.

As illustrated in Fig. 9, the specific operation carries out the average pooling operation in the corresponding horizontal and vertical directions. Subsequently, it carries out the stack Concat operation and 1*1 convolution method in the corresponding spatial dimension to compress the channel layers. Following this, BN and non-linear methods are used to encode the spatial information in both vertical and horizontal directions, the feature map is segmented, and the corresponding 1*1 convolution operation is performed to obtain the same number of channels in the input feature map and carry out the normalized weighting.

For the aforementioned operation, not only the number of parameters and operation cost were significantly reduced, but because of the attention of the reference module and the corresponding filter after the weights to generate some unwanted attention, (and give their attention a weighting of 0), we could discard those we did not need.

3) Adaptive Spatial Feature Fusion

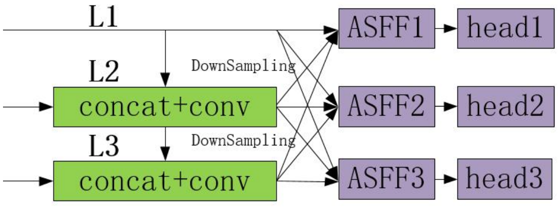


Fig. 10. YOLOv4 network architecture

The characteristic of the FPN method is to resolve the target detection when faced with different characteristics and inconsistent scales. Thus, this research used the adaptive feature fusion module, as shown in Fig. 10, as the solution for filtering the spatial information of conflict. This is a change caused by the information characteristics of scaling the semantic information in unknown problems, and it can reduce the cost of model reasoning. The main idea is described as follows: after the original PAN, we added the feature maps of three scales in the same channel and dimension and added a parameter that could carry out self-learning so that feature fusion could be adequately carried out according to the learning [26].

After considering ASFF-3 as an example —where L1, L2, and L3 represent the information obtained from the three feature layers, respectively, and are multiplied by the corresponding weight coefficients α_{ij}^{l}, β_{ij}^{l}, and γ_{ij}^{l} —,the results are added together to obtain the fused new feature ASFF-3, as presented in (1).

$$y_{ij}^{l} = \alpha_{ij}^{l} \cdot x_{ij}^{1\to l} + \beta_{ij}^{l} \cdot x_{ij}^{1\to l} + \gamma_{ij}^{l} \cdot x_{ij}^{1\to l} \qquad (1)$$

Since addition is used, the size of the output feature map and the number of channels should be the same. When obtaining the ASFF3 feature information, the size of L1 and L2 should be adjusted, the feature layer should be adjusted to the same value as the 1*1 convolution, and the size should be adjusted to the same size with upsampling.

The weight parameters α, β, and γ are obtained by performing a 1*1 convolution operation on the feature graphs of L1, L2, and L3 after adjusting their size. Note that the three parameters are defined by using the softmax function with $\lambda_{\alpha_{ij}}^{l}$, $\lambda_{\beta_{ij}}^{l}$, and $\lambda_{\gamma_{ij}}^{l}$ as the control parameters, respectively, and their sum is presented in (2).

$$\alpha_{ij}^{l} = \frac{e^{\lambda_{\alpha_{ij}}^{l}}}{e^{\lambda_{\alpha_{ij}}^{l}} + e^{\lambda_{\beta_{ij}}^{l}} + e^{\lambda_{\gamma_{ij}}^{l}}} \qquad (2)$$

The effectiveness of adding the ASFF module can be explained from the perspective of back propagation. Having added the chain rule of the FPN mechanism as an example, the gradient calculation formula for backward processing can be given by (3).

$$\frac{\partial \mathcal{L}}{\partial x_{ij}^{1}} = \frac{\partial y_{ij}^{1}}{\partial x_{ij}^{1}} \cdot \frac{\partial \mathcal{L}}{\partial y_{ij}^{1}} + \frac{\partial x_{ij}^{1\to 2}}{\partial x_{ij}^{1}} \cdot \frac{\partial y_{ij}^{2}}{\partial x_{ij}^{1\to 2}} \cdot \frac{\partial \mathcal{L}}{\partial y_{ij}^{2}} + \frac{\partial x_{ij}^{1\to 3}}{\partial x_{ij}^{1}} \cdot \frac{\partial y_{ij}^{3}}{\partial x_{ij}^{1\to 3}} \cdot \frac{\partial \mathcal{L}}{\partial y_{ij}^{3}} \qquad (3)$$

Therefore, we can set both options to a constant value by default. To better express (3), we thus assume that $\frac{\partial x_{ij}^{1\to l}}{\partial x_{ij}^{1}} = 1$, and $\frac{\partial y_{ij}^{l}}{\partial x_{ij}^{1}} = 1$, and $\frac{\partial y_{ij}^{l}}{\partial x_{ij}^{1\to l}} = 1$, as shown in (4).

$$\frac{\partial \mathcal{L}}{\partial x_{ij}^{1}} = \frac{\partial \mathcal{L}}{\partial y_{ij}^{1}} + \frac{\partial \mathcal{L}}{\partial y_{ij}^{2}} + \frac{\partial \mathcal{L}}{\partial y_{ij}^{3}} \qquad (4)$$

C. Chinese traffic sign data set

The Chinese Traffic Sign Detection Benchmark (CCTSDB): CSUST was derived from a real-time Chinese traffic sign detection algorithm based on the modified YOLOv2 training data set. There are 15,723 sample images in this data set, and there are three types of image labels in that year: warning, prohibition, and indication signs.

This data set was established based on the traffic signs in China (CCTSDB) according to the original data expansion, which covers the most frequently encountered road traffic signs. The data set itself has a very clear and concise structure; its classification is average according to the data once every 700 images enhancement processing. Among them, 0−699 are the original images, 700−1399 are the images with their lengths and widths swapped based on the original size, 1400−2099 are images with added salt-and-pepper noise, 2100−2799 are images with a certain angle of tilt, and 2800−3499 are images with adjusted brightness, as shown in Fig. 11. The rest of the images are selected from each video frame captured while driving on the road. So far, the data have been enhanced by rotation, brightness adjustment, elongation, blurring, and other data processing.

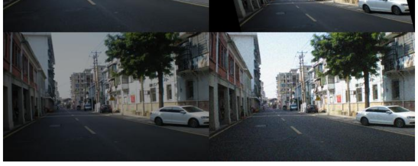


Fig. 11. Sample image of data set after data enhancement

In this experiment, a DELL T7920 graphics workstation is used, in which the hardware environment of the server uses an Intel Xeon E5 processor, three built-in mechanical hard disks with 3 TB capacity (a total of 9 TB), two NVIDIA GTX TITAN XP graphics cards with 12 GB video memory, and the memory of the server was 128 GB . The operating system for model training was Ubuntu16.04, and Python libraries such as Python2.7, CUDA, CUDNN, PyTorch, Torchvision, and Numpy were installed. All the programs used in this experiment were implemented on the server.

A. Experimental analysis

The detection traffic sign map with the improved YOLOv4 algorithm is depicted on the right side of Fig. 12. On the left side of Fig. 12, the detection regression map with the original YOLOv4 algorithm before the improvement is presented. As can be seen from the figure, in comparison with the original model, the improved algorithm could better detect the three types of traffic signs such as warning, prohibition, and indication, and at the same time, the detection accuracy was significantly improved.

The diagram on the left side of the detection did not detect a ban and directional signs, and for the accuracy of traffic signs that have been detected, the minimum score was 62%. It can be seen that the original algorithm was not satisfactory in detection tasks; however, on the right side of the figure, it is clear from the improved model test results that all the marks were detected. In addition, except for one whose accuracy was 79%, the rest had an accuracy of above 90%. Therefore, it can be concluded that the improvement in the original algorithm dramatically improved the detection and recognition of traffic signs. Even when the threshold was set to 0.7, the detection task was completed well.

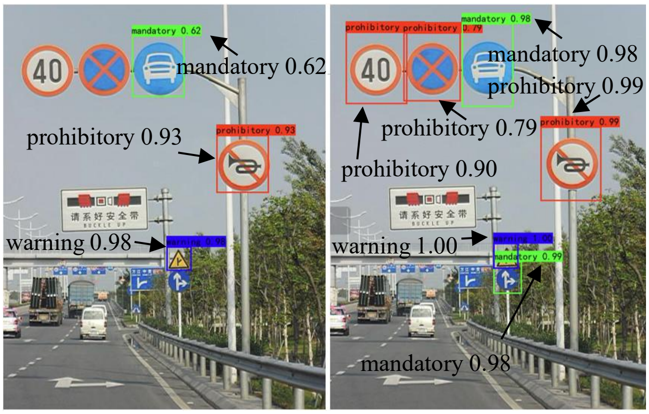


Fig. 12.Comparison of YOLOv4 before and after improvement

B. Comparison of traffic sign detection under complex conditions

As shown in Fig. 13, for the traffic sign detection and recognition under shadow conditions, the detection accuracy of the YOLOv3 algorithm was 55%, which is not a good rate. The following two algorithms were used: efficientdet-D2 and YOLOv4, and the detection accuracy was improved by more than 10 points compared with the original basis. However, there exists a major problem that the three models are missing; while the final accuracy rose, a leak problem occurred. In the case of the last one where the improved model after the

YOLOv4 algorithm was used, we could see that our model had the highest accuracy (i.e. 98%) compared to the three previous model results. However, the missed detection was detected again. Although the accuracy was only 56%, it was enough to prove that our model had been greatly improved compared with the previous three algorithms.

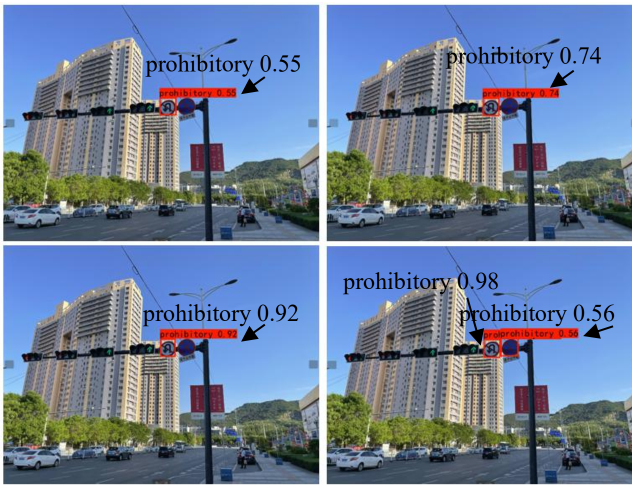


Fig. 13. Comparison of YOLOv4 before and after improvement under shadow condition

As shown in Fig. 14, YOLOv3 did not detect the broken traffic sign, while Efficientdet-D2 and YOLOv4 detected it; however, the detection accuracy was not sufficiently good. After using the improved algorithm, proposed in this study, YOLOv4 was able to detect the broken traffic sign. It can be seen that the detection accuracy reached 99% for typical traffic signs and 96% for damaged ones. In comparison with the unrecognized model, the detection accuracy was significantly improved from that of other models.

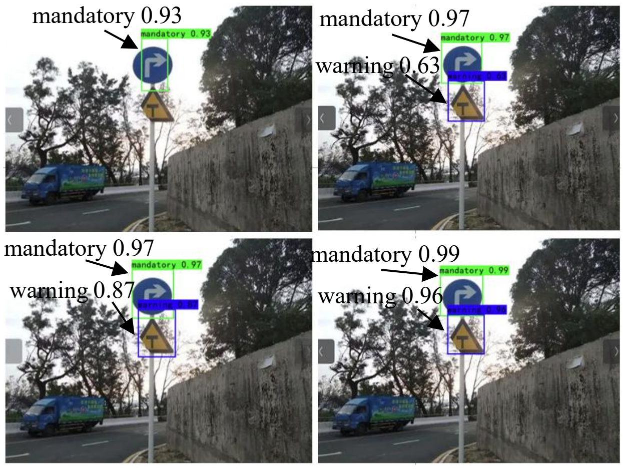


Fig. 14. Comparison of YOLOv4 under damaged condition before and after improvement

In our experiment, we used Python code for image processing, as it is easier to simulate the traffic signs in foggy weather conditions, as shown in Fig. 15. Additionally, it significantly improved the recognition difficulty of detection due to bad weather. As shown in Fig. 15, in the first image on the left side, only a ban on the traffic signs was detected; the effect was not satisfactory, and the model was not very stable in the extreme weather conditions. After this, an attempt was made to use the improved YOLOv4 algorithm to identify traffic signs; however, there was still a lack of recognition. Nevertheless, five traffic signs were detected. This represented a significant increase in the precision and presented successful detection in foggy weather conditions with the target being predicted correctly.

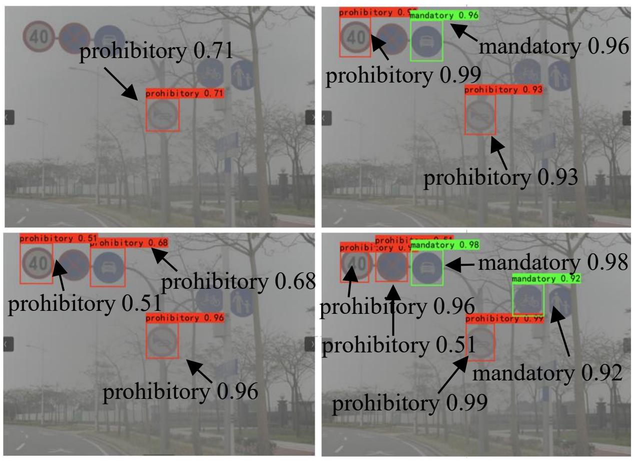


Fig. 15. Comparison of YOLOv4 before and after improvement under foggy weather conditions

To observe the differences between the algorithms on a more intuitive level, we compared not only different algorithms under different conditions but also included the current state-of-the-art target detection algorithms. For the same CCTSDB set of test data sets, it can also be seen from Table I that the accuracy of this algorithm is superior to other target detection algorithms when using this data set.

TABLE I
COMPARISON WITH CLASSICAL ALGORITHMS

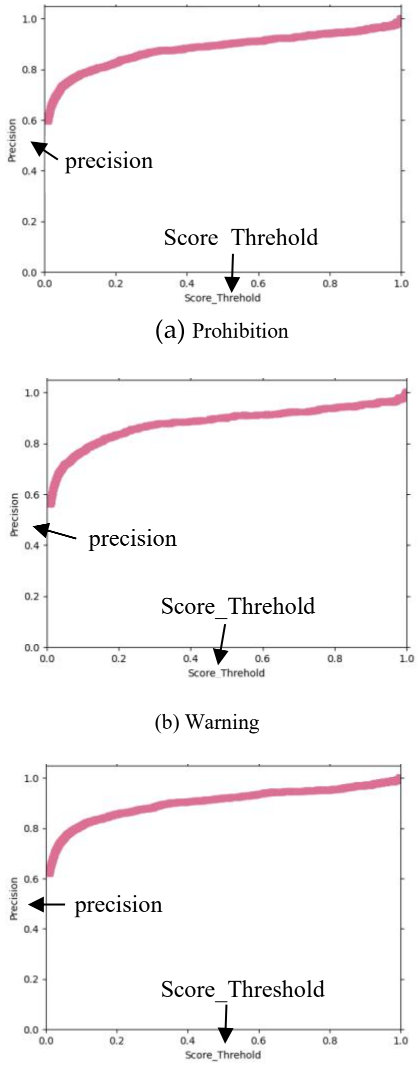
Algorithm	Backbone	AP_{30}	AP_{50}	AP_{70}
EfficientDet-D2	Efficient-B2	88.2%	83.3%	69.3%
Faster R-CNN	ResNet-50	81.4%	76.9%	65.7%
YOLOv3	Darknet-53	86.1%	82.1%	64.4%
YOLOv4	CSPDarknet-53	89.9%	84.3%	68.2%
YOLOv5	BottleNeck CSP	90.2%	85.6%	71.9%
PP-YOLO	ResNet50	90.8%	87.4%	70.4%
Improved YOLOv4	Darknet-53	93.4%	89.3%	79.3%

Precision and recall are critical indicators in measuring the performance of machine learning models, especially in the case of unbalanced data set distributions. Fig. 16 shows the precision of different threshold effects and the recall rate.

Table II describes the values of precision and recall rate under different thresholds.

TABLE II
PRECISION AND RECALL RATE VALUES

Threshold (t)	0.1	0.2	0.5	0.7
Precision	0.5811	0.8122	0.8956	0.9268
Recall	0.9011	0.8891	0.7967	0.5562



(a) Prohibition

(b) Warning

(c) Indication

Fig. 16. Comparison of accuracy for different thresholds

To verify the improvement in the SPP structure effect after the coordinate attention mechanism was added, as part of this research, an attention comparison test was performed. Two widely used attention mechanisms, Squeeze-and-Excitation (SE) and Convolutional Block Attention Module (CBAM), were added to the SPP module to compare their effects with the improved SPP+, as shown in Table III.

TABLE III
COMPARATIVE EXPERIMENT OF ATTENTION MECHANISMS

Method	AP_{30}	AP_{50}	AP_{70}
SPP	89.9%	84.3%	68.2%
SPP+SE	88.1%	84.6%	65.1%
SPP+CBAM	87.7%	83.9%	76.1%
SPP+CA	93.4%	86.5%	79.3%

To verify the difference between the improved YOLOv4 and other mainstream algorithms in terms of inference time, a comparison test of the models' inference time was performed. It was discovered that the improved YOLOv4 had a significant improvement in terms of accuracy, and the model's inference time was relatively satisfactory, as shown in Table IV.

TABLE IV
COMPARATIVE TEST OF INFERENCE TIME

Algorithm	Backbone	Size	Inference time
EfficientDet-D2	Efficient-B2	768×768	121ms
Faster R-CNN	ResNet-50	416×416	346ms
YOLOv3	Darknet-53	416×416	186ms
YOLOv4	CSPDarknet-53	416×416	115ms
YOLOv5	BottleNeckCSP	416×416	108ms
PP-YOLO	ResNet50	416×416	116ms
Improved YOLOv4	Darknet-53	416×416	129ms

C. Analysis of ablation experiment

In this paper, a total of three improvements were proposed. To better compare each proposed method and observe the resulting improvement in the model, we designed the experimental ablation module.

The general process of this module is to carry out the gradual accumulation of each novel part from scratch under the original YOLOv4 algorithm and then check the accuracy of the model at each stage along with its effect. We added improvement parts in the following order: SPP+, RCFF, and ASFF. The improved YOLOv4 network structure and ablation experimental accuracy results are shown in Table V.

TABLE V
ABLATION EXPERIMENTS

Experiment	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6
SPP+	√				√	√
ASFF		√		√		√
RCFF			√	√	√	√
mAP(%)	86.5	85.3	86.1	87.1	88.2	89.3

V. CONCLUSIONS

This study used a convolutional neural network to complete the detection and recognition tasks of warning, prohibition, and indication traffic signs in China. In addition, the CCTSDB data set was amplified to a certain extent using data enhancement, and the algorithm of the YOLOv4 model was partially improved. The experimental results indicate that our improved YOLOv4 model excelled in performance and effect.

However, there are still many areas that need to be improved and optimized to enhance the detection accuracy of the model in the future. Firstly, many experiments are based on their own data sets for learning and training as well as on their own methods for calculating the accuracy of the recognition tasks. Consequently, comparison between methods is difficult and inconsistent; hence, the creation of a unified data set for the current image recognition and detection research would be a significant achievement. Secondly, due to the uncertainty of multiple realistic scenarios, such as hardware problems, vague imaging makes the image-recognition process flawed, which is still an unaddressed problem. Lastly, during the experimental process, we found that there were still cases of false positives or missed detections. Moreover, this algorithm is currently only applicable to a single data set, so it is necessary to develop the algorithm to improve the identification and detection accuracy of traffic signs for other data sets.

REFERENCES

[1] W. Su, L. Du. *Amendment Explanation for National Standard "Specification and Test Method for Road Traffic Markings"*Transport standarization, Beijing, China, 2009.

[2] B. Alexey, W. Chien-Yao, M. L. Hong-Yuan "YOLOv4:Optimal Speed and Accuracy of Object Detection," arXiv:2004.10934, 2020.

[3] "Networks for Object Detection*,* " *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936-944.

[4] Q. Hou, D. Zhou, J. Feng. "Coordinate Attention for Efficient Mobile Network Design," *Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 13713-13722.

[5] S. Liu, D. Huang, Y. Wang. "Learning Spatial Fusion for Single-Shot Object Detection,", arXiv:1911.09516, 2019.

[6] A. Ruta, Y. Li, X. Liu. "Real-time traffic sign recognition from video by class-specific discriminative features," *Pattern Recognition*, vol. 43, pp. 416-430, 2010.

[7] X. Liu, S. Zhu, "Ken Chen.Method of Traffic Signs Segmentation Based on Color-Standardization," *Proceedings of the 2009 International Conference on Intelligent Human-Machine Systems and Cybernetics*, pp. 193-197.

[8] C. Y. Fang, C. S. Fuh, P. S. Yen, S. Cherng, S. W. Chen. "An automatic road sign recognition system based on a computational model of human recognition processing," *Computer Vision and Image Understanding,* vol. 96, pp.237-268, 2004.

[9] Z. Huang, G. Sun, Fang Li."Traffic Sign Segment Based on RGB Vision Model*" Microelectronics and Computer*, pp. 147-148, 2004.

[10] X. Liu, S. Zhu, K. Chen. "Real-Time Traffic Sign Detection via Color Probability Model and Integral Channel Features," *Proceedings of the 6th Chinese Conference CCPR 2014*, pp. 545-554.

[11] M. Hu, "Visual pattern recognition by moment invariants, " *in IRE Transactions on Information Theory*, vol. 8, pp. 179-187, 1962.

[12] D. M. Gavrila,"Traffic Sign Recognition Revisited," *Proceedings of the Pattern recognition* (1999), pp. 86-93.

[13] C. Yu, J. Hou Ch. Hou. "Traffic sign detection based on saliency map and Fourier descriptor," *Computer engineering*,pp. 28-34, 2017.

[14] M. Boumediene, C. Cudel, M. Basset, A. Ouamri. "Triangular traffic signs detection based on RSLD algorithm," *Machine Vision and Applications*, vol. 24, pp. 1721-1732, 2013.

[15] T. T. Yang, S. Y. Zhou, and A. J. Xu, "Rapid Image Detection of Tree Trunks Using a Convolutional Neural Network and Transfer Learning, "*IAENG Intl. J. Comput. Sci*., vol. 48, no. 2, pp. 257-265, 2021.

[16] D. G. Lowe. "Distinctive Image Features from Scale-Invariant Keypoints,"*International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.

[17] N. Dalal, B. Triggs,"Histograms of oriented gradients for human detection," *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*,pp. 886-893.

[18] I.M. Creusen, R.G.J. Wijnhoven, E. Herbschleb, P.H.N. de With. "Color exploitation in hog-based traffic sign detection," *Proceedings of the 2010 IEEE International Conference on Image Processing,* ,pp. 2669-2672, 2010.

[19] J. Redmon, A. Farhadi, "YOLOv3: An incremental improvement," arXiv:1804.02767, 2018.

[20] J. Redmon, S. Divvala, R. Girshick, "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016,* pp. 779-788.

[21] J. Redmon, A. Farhadi, "YOLO9000: Better, faster, stronger," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017,* pp. 7263–7271.

[22] K. He, X. Zhang, S. Ren, J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,* pp. 1904-1916.

[23] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, "Path Aggregation Network for Instance Segmentation," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition, 2018*, pp. 8759-8768.

[24] K. He, G. Gkioxari, P. Dollár, R. B. Girshick "Mask R-CNN," *Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017,* pp. 2961-2969.

[25] S. Santosa, R. A. Pramunendar, D. P. Prabowo, and Yonathan P. Santosa, "Wood Types Classification using Back-Propagation Neural Network based on Genetic Algorithm with Gray Level Co-occurrence Matrix for Features Extraction, " *IAENG Intl. J. Comput. Sci.,* vol. 46, no. 2, pp. 149-155, 2019.

[26] M. El Alaoui, and M. Ettaouil, "An Adaptive Hybrid Approach: Combining Neural Networks and Simulated Annealing to Calculate the Equilibrium Point in Max-stable Problem,"*IAENG Intl. J. Comput. Sci.,* vol. 48, no.4, pp. 893-898, 2021.