Research on Road-Sign Detection Algorithms Based on Depth Network

Huaixu Gao, Ying Tian

Abstract—**Unmanned system research has attracted increasing academic and corporate attention in recent years. Traffic-sign and road detection systems play an important role in unmanned systems. However, current algorithms perform poorly in reading traffic signs from a long distance and cannot satisfy the real-time requirements of accurate and rapid detection. Therefore, Yolov5s−Swin model is proposed in this paper. First, the Swin Transformer module and convolution module are fused, the ResUnit module in the Yolov5s backbone network is improved, the Crswin module is proposed, and the PAFPN in the network is modified to strengthen the ability of the model to capture local feature information and improve network detection accuracy. The Swin Transformer module uses the moved-window attention function to segment the image into different windows of a certain size but cannot establish a connection between neighboring windows. Consequently, we designed a method for hyperbolic window attention. This approach updates the window partition method such that the window block for attention calculation at each pixel changes, which can improve the receptive field, increase the information extraction ability of the target, and solve the information loss caused by decreased resolution during image training. Experimental results show that the proposed improved Yolov5s-Swin model exhibits significantly better detection accuracy than existing models.**

Index Terms— **Road sign detection, Yolov5s, Transformer, Swin-Transformer**

I. Introduction

Road-sign detection is highly important in automatic management systems for vehicle traffic. With the development of artificial intelligence technology, road identification can provide effective data support for driving assistance systems and be used to study unmanned driving technology in the field of road traffic. Therefore, in-depth research on road identification and detection systems can not only improve road safety but also play a decisive role in the development of unmanned driving technology. Object-detection technology is an important technical component for traffic-sign recognition systems [1-3]. Traditional object-detection methods are mainly based on image matching of color and shape features. Commonly used methods are the RGB (Red-Green-Blue) and HIS (Hue-Saturation-Intensity) color spaces [4], the Hough transform [5], and HSV (Hue-Saturation-Value) space and edge information [6]. A support vector machine (SVM) can classify the characteristics of a detected image. By dividing specific colors to create significance mapping, candidate areas can be detected quickly and efficiently. However, this approach is highly susceptible to illumination and cannot meet detection requirements under severe weather conditions.

Shape-based detection methods use the particularity of the shape of traffic signs to detect their edges; however, when the traffic-sign image is blurred or blocked, the accuracy of the algorithm decreases. Therefore, the detection results are easily affected by the environment, and when a traffic sign is damaged, blocked, or faded, the detection results become uncertain, resulting in missed detection [7-8]. Traditional algorithms have weak generalization abilities and a poor ability to deal with large amounts of data; therefore, they cannot meet the requirements of complex traffic scenarios.

With the rise of artificial intelligence technology, deep learning-based object-detection methods have been proposed. Among the first methods developed were object-detection algorithms based on candidate regions: Fast R-CNN and Faster R-CNN [9]. These algorithms mainly train the regional proposal network (RPN) and the target area detection network to achieve detection. Although the effect of detection and positioning is good, the real-time performance cannot meet practical application requirements.

Object-detection algorithms based on regression have also been proposed, including Retinanet [10], SSD [11-13], and the Yolo series [14-17]. These networks directly regress and classify input images, which greatly reduces the repeated operations in the detection process; therefore, the detection speed is significantly improved. However, most networks still cannot meet actual demand because of their complex structure and heavy model weight. Of these, the algorithms in the Yolo series have low hardware requirements, are fast, and are more widely used on small equipment platforms. Moreover, they outperform other regression-based detection algorithms with high accuracy and strong real-time performance. Therefore, Yolov5s, which has a small model size and low computing resource requirements, was selected as the improved basic model for this study.

To address the difficulties and demands for accurate road traffic-sign detection, this paper proposes a Yolov5s-Swin traffic-sign object-detection model based on the improved Yolov5s algorithm. The idea of the Swin Transformer is improved into a backbone network and feature-fusion network. This model creates a road traffic-sign object-detection network model that can give consideration to both application ability and accuracy. Moreover, the

Manuscript received, August 24,2022; revised January 18,2023.

This work was funded by the foundation of Liaoning Educational committee under the Grant No. LJKZ0310.

Huaixu Gao is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China (e-mail: 243257530@qq.com).

Ying Tian, the corresponding author, is a Professor at the School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, China. (phone: +8613898015263; fax: 0412-5929818; e-mail: astianying@126.com).

original algorithm has greatly mitigated the problem of incomplete feature extraction for low-level targets.

II. Related Work

Researchers have addressed unmanned driving systems and made great progress. Road-sign detection is an indispensable aspect of unmanned driving. To accurately detect dynamic objects such as people and cars, deep learning algorithms are preferred. The Autoware platform incorporates a variety of convolutional neural network (CNN) models, including Faster R-CNN and Yolov3.

Many researchers have also integrated these models into their devices. These algorithms perform differently in different scenarios.

Redmon has developed a new framework called Yolo. It achieved 57.9% mAP on VOC 2012, which is higher than that of R-CNN (53.3%), and it processed images at 45 frames per second (fps). Yolo stands for the "one-phase approach", which abandons networks that generate zone suggestions. The end-to-end algorithm achieves feature extraction and classification only through a CNN. Yolo is both fast and accurate. There are many versions in the Yolo series. Apart from Yolo, Yolo9000, Yolov3, Yolov4, and Yolov5 have all made some optimizations based on Yolo.

Among these Yolo versions, the classic Yolov3 algorithm uses Darknet-53 as the basic feature extraction network and a large number of residual structures to enhance model training. In addition, a feature pyramid is introduced into the neck of the network model to facilitate the interoperation of the features at the bottom and top layers, improving the recognition accuracy of the network by combining them [18].

In the Yolov3 model, developers abandoned the pooling layer and retained the convolution layer. By adjusting the convolution step size, the size of the output feature graph was controlled, and the shallow feature was integrated with the deep feature such that the shallow feature also had rich semantic information. In the training process, the Yolov3 model predicts three feature images of different scales for each input image [19-21].

The improved Yolov5 joined the adaptive model Yolov3 anchor box computing, which permits adaptive image zooming and is both faster and more accurate than Yolov3.

Because road-sign images are captured in a wide field of view, objects in the distance are too small to be detected.

For this reason, Yolov5 is more appropriate than the other versions for self-driving vehicle applications. The Yolov5 network architecture is a one-stage object-detection network proposed by Ultralytics LLC. Based on the Yolov4 network, Yolov5 integrates many excellent improvements. The network deployment is more flexible, and has higher detection accuracy and speed, and is better suited for real-time object-detection applications.

The network structure of Yolov5s is divided into three parts: the Backbone (backbone network), the Neck (multi-scale feature fusion network), and the Head (predictive classifier). Yolov5 controls the size of the model by introducing depth and width factors. The method includes four models, from small to large: v5s, v5m, v5l, and v5x. However, because it is intended for unmanned systems such as car navigation, to meet the practical application conditions, Yolov5, which has the smallest model size and computing resource requirements, was selected as the basic improved model, as shown in Fig. 1.

The CBL module consists of the Conv+BN+SiLu activation function, as shown in Fig. 1, Module 1. The Resunit module refers to the structure of Residual Networks. CBL is a sub-module of the residual module, as shown in Module 2 in Fig. 1. The CSP1_X module is composed of the CBL module and the ResUnit module and is similar to the structure of CSPNet, as shown in Module 3 in Fig. 1. The CSP2_1 module is mainly based on the structure of CSPNet, which is composed of convolution layers and some CBL modules, as shown in Module 4 in Fig. 1. The Focus module is shown as Module 5 in Fig. 1. First, the Focus structure concatenates multiple slices of results and then sends them to the CBL module. The SPP module adopts 5×5, 9×9, and 13×13 maximum pooling modes to fuse multi-scale features, as shown in Module 6 in Fig. 1.

III. Improved Model

In computer vision technology, convolutional structures still dominate. However, inspired by the success of natural language processing (NLP), a number of studies have attempted to combine CNN-like structures with attention mechanisms, and some have even replaced convolution.

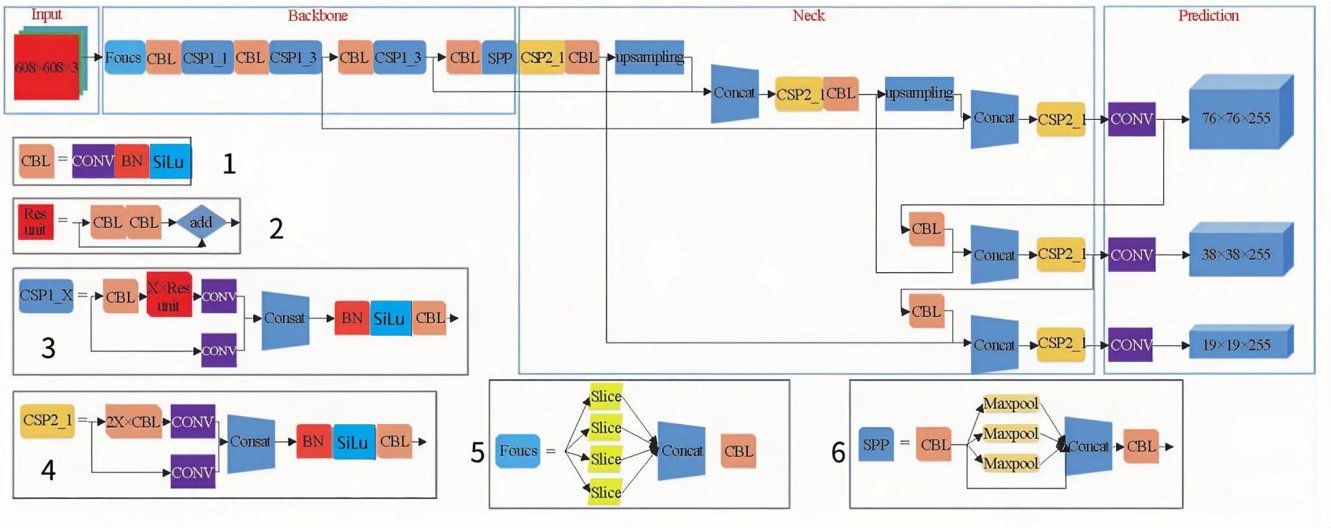

Fig. 1. Yolov5s mode

A. Transformer

The Transformer is a machine translation method proposed by Vaswani et al. in 2017 [22]. On the basis of self-attention architecture, the Transformer has become the preferred model in NLP. The main approach is to pretrain on large textual corpora, followed by fine-tuning on smaller mission-specific datasets. As models and datasets evolve, there is a lot of room for performance improvement. Although this direct substitution of convolution is theoretically possible, it has not been effectively extended to modern hardware accelerators because of the use of specialized attention modes. Therefore, the classic ResNet-like architecture is still best for large-scale image recognition. ResNet has been combined with the Transformer to apply the standard Transformer directly to images. For this purpose, images are segmented into patches, and a linear embedded sequence of these image blocks is provided to the Transformer as input. Image blocks are treated as words in NLP, and the image classification model is trained in a supervised way.

CNNs have significant advantages in extracting low-level features and visual structures, but they have limitations in modeling large-scale dependencies of low-level features. Unlike CNNs, the Transformer is capable of focusing on global information modeling. The previous standard Transformer architecture and its adaptation of image classification performed global self-attention, calculating the relationship between all tokens. Global computation leads to quadratic complexity in terms of the number of tokens, making it unsuitable for many visual problems that require a large set of tokens to make dense predictions or represent high-resolution images; global self-attention computation is generally unaffordable for large hardware.

Vision Transformer (ViT) [23] is a model proposed by Google in 2020 to directly apply the Transformer in image classification, and many subsequent works have proposed improvements based on ViT. The idea behind ViT is simple: An image is directly divided into patches of fixed size, and patch embedding is then obtained by linear transformation, which is similar to NLP words and word embedding. As the input of the Transformer is a sequence of token embeddings, the patch embeddings of the image can be sent to the Transformer for feature extraction and classification.

1) Self-attention in non-overlapped windows

Compared with the global computation of ViT, local in-window computation of self-attention is proposed for efficient modeling. These windows split the image evenly in a non-overlapping manner. Assuming that each window contains M×M patches, the computational complexity of the global MSA module (1) and W-MSA (2) for windows based on H×W patch images is as follows:

$$\Omega(\mathrm{MSA}) = 4\mathrm{hwC}^2 + 2(\mathrm{hw})^2\mathrm{C} \tag{1}$$

$$\Omega(\mathrm{W-MSA}) = 4\mathrm{hwC}^2 + 2\mathrm{M}^2\mathrm{hwC} \tag{2}$$

Window self-attention is extensible. Although the computational complexity is effectively reduced when only self-attention is performed within the window, there is no information exchange between different windows, thus limiting its modeling ability.

2) Shifted Window-based Self-Attention

The new Swin Transformer [24] model is an improved model based on Microsoft's recently proposed Transformer. It not only has the ability of the Transformer to focus on global information modeling but also realizes cross-window connections by moving windows; therefore, the model can focus on relevant information from other adjacent windows. Cross-window feature interaction expands the receptive field to a certain extent, thus achieving higher efficiency.

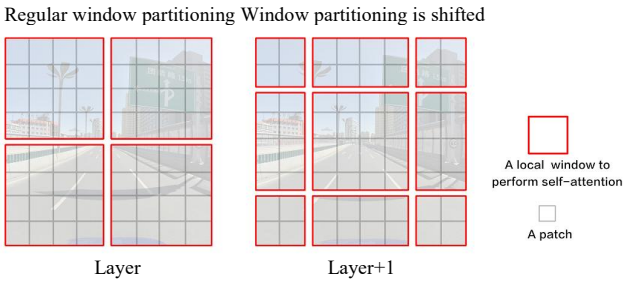

Fig. 2. Swin Transformer diagram of network module

The Swin Transformer model adopts local windows (the red box as shown in Fig. 2), which is the same as the convolution kernel idea of CNN, focusing on local areas. The difference is that the CNN's sliding windows overlap, whereas the Swin Transformer's windows do not overlap, moving the feature graph horizontally and vertically, thereby obtaining window multi-head self-attention (W-MSA) and shifted window multi-head self-attention (SW-MSA) in an alternating fashion. In the design proposed in this paper, after the Moseck enhancement of the input image, the feature data are sent to the Swin Transformer module, making full use of the high-resolution spatial information of CNN features and the global semantic information of Swin Transformer coding. The network structure of the Swin Transformer network module is shown in Fig. 2.

The window attention method proposed previously divides the image into different windows according to a certain size. The Transformer's attention is calculated only within the window at each iteration. Thus, assuming window attention alone does not send the receptive field of each pixel moving, we designed a method for hyperbolic window attention that updates the method of dividing windows. The window block that calculates the attention for each pixel changes. Thus, it enhances the receptive field. The first module uses conventional window division to divide an 8×8 feature graph into 2×2 according to the window size M = 4. The next module then shifts the window settings across M/2 pixels horizontally and vertically, producing 3×3 non-convergent windows. The moving window partition connects neighboring non-convergent windows in the upper layer, increasing the receptive field.

The Swin Transformer block is shown in Fig. 3. The Swin Transformer module can be viewed as two consecutive Transformers connected in series; the difference is that the first Transformer uses W-MSA for self-attention. The second Transformer uses SW-MSA and wires MLP. The LayerNorm (LN) layer is applied before each MSA module and before each MLP, and residual connections are applied after each module. For the hyperbolic structure, Swin Transformer blocks are calculated as follows:

$$\hat{\mathbf{z}}^{\mathrm{L}} = \mathrm{W-MSA}\left(\mathrm{LN}\left(\mathbf{z}^{\mathrm{L-1}}\right)\right) + \mathbf{z}^{\mathrm{L-1}} \tag{3}$$

$$\mathbf{z}^{\mathrm{L}} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{\mathrm{L}})) + \hat{\mathbf{z}}^{\mathrm{L}} \tag{4}$$

W-MSA and SW-MSA represent bullish self-attention using neat divisions and wires, respectively; $\hat{\mathbf{z}}^{\mathrm{L}}$ (3) and $\mathbf{z}^{\mathrm{L}}$ (4) respectively, represent the output characteristics after W-MSA and MLP.

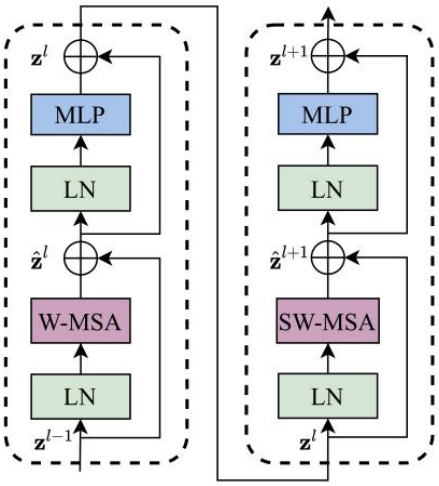

Fig. 3. Swin Transformer block

$$\hat{\mathbf{z}}^{\mathrm{L+1}} = \mathrm{SW-MSA}\left(\mathrm{LN}\left(\mathbf{z}^{\mathrm{L}}\right)\right) + \mathbf{z}^{\mathrm{L}} \tag{5}$$

$$\mathbf{z}^{\mathrm{L+1}} = \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{z}}^{\mathrm{L+1}})) + \hat{\mathbf{z}}^{\mathrm{L+1}} \tag{6}$$

$\hat{\mathbf{z}}^{\mathrm{L+1}}$ (5) and $\mathbf{z}^{\mathrm{L+1}}$ (6), respectively, represent the output characteristics of SW-MSA and MLP for block L+1. This paper proposes the Crswin module, as shown in Fig. 5, which improves the backbone network and feature fusion part of Yolov5s. CSP1_1 in Yolov5s was replaced by ResUnit, as shown in Fig. 4, with a Swin Transformer, and the second CBL in CSP2_X was replaced with a Swin Transformer; experimental comparisons were made with previous models.

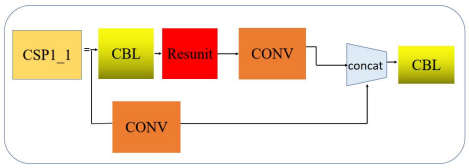

Fig. 4. CSP1-1 module in the original model Yolov5s

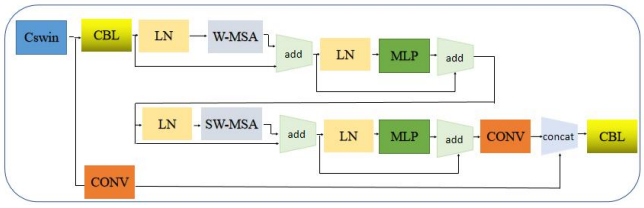

Fig. 5. Crswin module

The number of Swin Transformers used in the proposed approach gradually decreases with the deepening of the network, and the perception range of each patch expands. This design facilitates the hierarchical construction of the Swin Transformer and adapts to the multi-scale nature of visual tasks. The receptive field is enlarged, and the global features are connected with each other.

IV. Experiment and Analysis

A. Introduction to dataset

The Chinese traffic-sign dataset TT100K [25] was used in this study. TT100K has a large base and rich semantic information. It can be obtained by using high-definition cameras to acquire real street views and can restore the first driving angle, a single motor vehicle lane, or a complex urban background. However, the number of datasets is not evenly distributed. According to the statistics of instances in the TT100K dataset based on categories, it is found that among 151 categories, only 45 categories have more than 50 instances, and nearly half of them are single-digit instances, which results in seriously unbalanced data distribution and is not conducive to network training.

The TT100K dataset was screened, and only 45 categories with more than 50 instances were reserved as datasets. Some examples of the dataset are shown in Fig. 6. A total of 9,170 images were randomly divided into training and test sets in an 8:2 ratio.

Fig. 6. TT100K dataset example

B. Evaluation index

Evaluation of a model is conducive to further optimization of the model and guarantees the quality of the model. The model was evaluated in terms of mAP value and recall rate. The model evaluation index used in this study was consistent with the method provided by Zhu, the publisher of the TT100K dataset. In the experiment, the IOU threshold of the prediction frame and the real frame of the target was set to 0.5, and when the IOU value was greater than 0.5, the position of the target was correctly predicted. Then, the precision and recall of the prediction results were calculated to measure the target classification ability and object-detection ability of the model. In addition, through the confidence threshold, the precision and recall curve (P-R curve) of the model was drawn, which intuitively shows the detection effect of the

model. When calculating the precision and recall indices of the model, first the detection results were divided into the following four categories according to the true labels: true examples, true counterexamples, false positive examples, and false counterexamples.

$$\text{precision}=\frac{TP}{TP+FP} \tag{7}$$

Precision (7), also known as precision rate, can be obtained by calculating the proportion of correctly predicted samples and all predicted samples in the detection results, i.e., the proportion of correctly detected samples in the total detected samples, which can reflect the classification ability of the model for targets.

$$\text{recall}=\frac{TP}{TP+FN} \tag{8}$$

Recall (8), also known as recall rate, is obtained by calculating the ratio of the predicted correct samples to all real samples in the detection results, i.e., the ratio of the correctly detected samples to the real samples, which can reflect the model's detection ability to the target. The two indexes, precision and recall, influence each other. When the set IOU threshold and object confidence threshold are higher, the calculated precision value is higher, and the recall value is lower. Therefore, to comprehensively compare network performance, the P-R curves of the models before and after modification were compared.

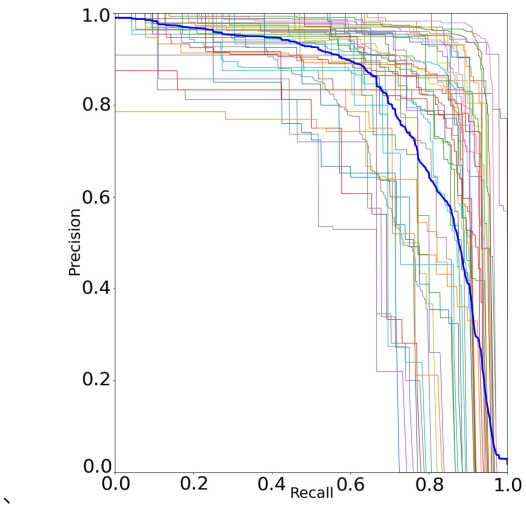

Fig. 7. Yolov5s P-R plot

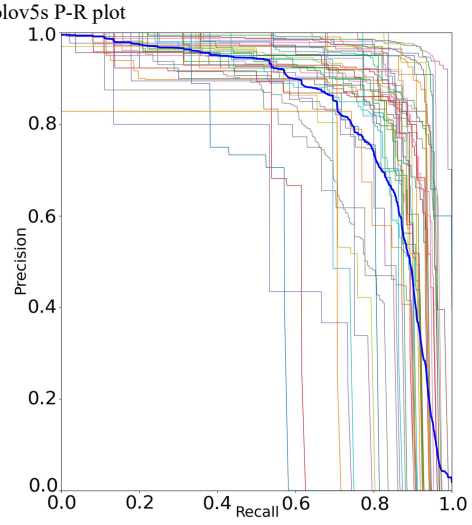

Fig. 8. Yolov5s-Swin P-R plot

The P-R curve takes precision as the abscissa and recall as the ordinate, as shown in Figs. 7 and 8. A larger area enclosed under the curve indicates better performance of the model. It can be seen that the P-R curve of Yolov5s-Swin has a larger area and better model effect. The blue line is the average of all classes, mAP%0.5. According to Table I, this represents an increase of 4.6 percentage points for the p11 class. The p26 class also grew by five percent. Similarly, classes ip and il60 grew by 1.0 and 1.5 percent, respectively.

TABLE I
COMPARISON OF P-R CURVES

Class	Yolov5s	Yolov5s-swin
p11 No honking	0.833	0.879
ip Pedestrian crossing	0.856	0.866
il60 The speed limit of 60	0.950	0.965
p26 Ban trucks	0.824	0.874
w59 Look out for inbound traffic on the right	0.753	0.842

When IOU is set to 0.5, the AP of all pictures in each category is calculated, and all categories are then averaged to calculate mAP. The graph of mAP (9) is shown in Fig. 9.

$$\text{mAP}=\frac{1}{N}\sum_{i=1}^{N}AP \tag{9}$$

In addition, the parameter size of the model before and after modification was compared, and it increased by only 0.16 MB. Among the different layers of the model, the convolutional layer and the fully connected layer handle most of the network's computational load and are the direct factors affecting the size of the network model. Reducing the computational load of the model and improving the detection speed and efficiency are important factors that cannot be ignored in the object-detection task. In terms of detection speed and efficiency, the modified model still outperforms the traditional Yolov3 and Faster R-CNN models.

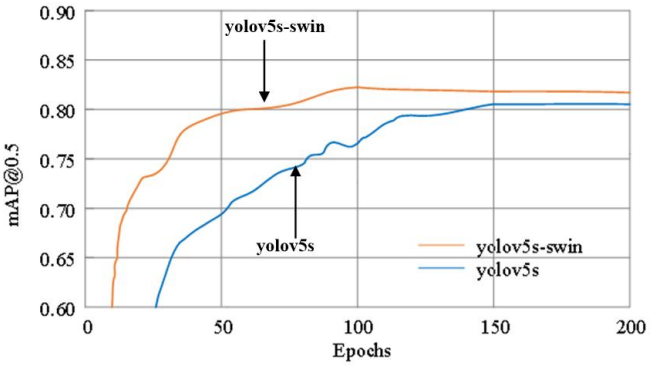

Fig. 9. mAP contrast in Yolov5s and Yolov5-Swin

C. Experimental configuration

Because of the large amount of data in the TT100K dataset, the improved Yolov5s-Swin model was trained on a graphics processing unit (GPU). See Table II for the detailed experimental configuration.

TABLE II
EXPERIMENTAL CONFIGURATION

Environment	Configuration
Operation platform	Centos7
GPU	v100
CUDA	v10. 1
Opencv	4.5
Number of threads	16
Batch size	20
Initial learning rate	0.001
Momentum	0.937

D. Experimental results and analysis

Fig. 10 shows the effects of the Yolov5s and Yolov5s-Swin models, respectively. Compared to the above figures. The p11 and pl5 classes represent "No honking" and "the speed limit is 5 kilometers an hour", respectively. They were increased by 0.7% in the Yolov5s-swin model and 1.7% in Yolov5s-swin. The po class represents the cars are forbidden to turn right, which increased 1.7% by Yolov5s. The p26 class represents A ban on truck traffic, which increased 0.6% by Yolov5s.

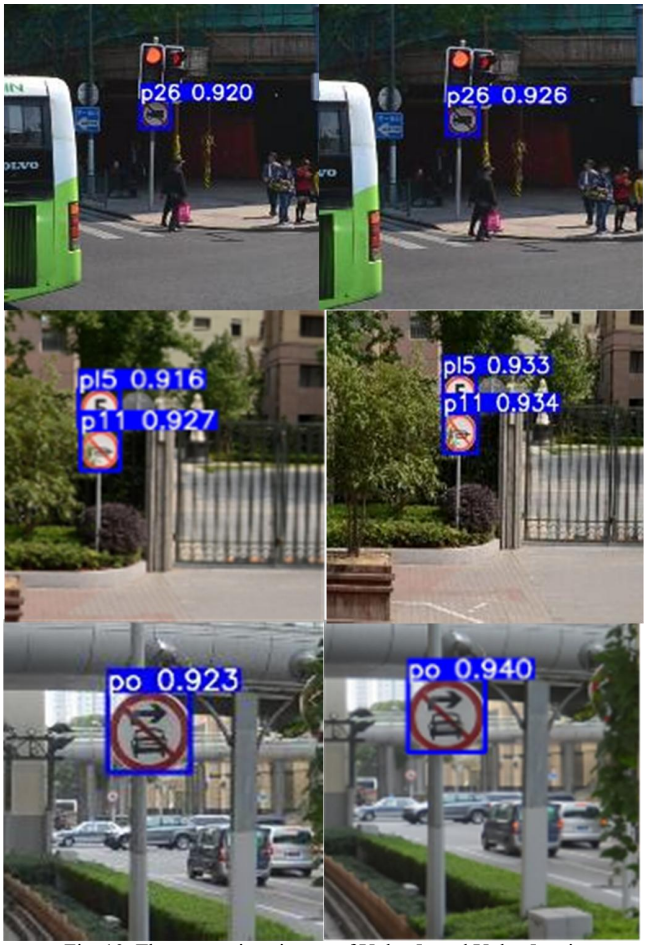

Fig. 10. The comparison image of Yolov5s and Yolov5-swin

In Yolov5s, the backbone network and PAFPN have been modified to explore whether the Swin Transformer module has different effects on each location of the network.

In Table III, backbone is mainly responsible for feature extraction, while PAFPN is responsible for feature fusion. Epochs are defined as the number of iterations, and the batch size is defined as the number of samples used to train the network.

TABLE III
BLATION STUDY

Backbone	PAFPN	Epochs	Bach size	mAP
		200	20	80.2%
√		200	20	80.9%
	√	200	20	81.4%
√	√	200	20	82.6%

To verify the influence of the improved method proposed in this paper, we compared the traffic-sign recognition performances of both the traditional Faster R-CNN model and the classic Retinanet model. A vast improvement has been observed in the accuracy of road signs under various light visibility conditions following the model's modification. Furthermore, the improved position of the Swin Transformer module was compared (see Table III). The comparison indicates that modifying the Crswin module in both backbone and PAFPN improved the model's accuracy and recall rate.

TABLE IV
COMPARISON OF STATE-OF-THE-ART MODELS

Model	Input size	mAP	Recall
Faster R-CNN	1000*600	78.8%	76.8%
Yolov3	640*640	70.9%	66.3%
Retinanet	1000*600	79.3%	75.2%
Yolov5s	640*640	80.2%	75.5%
Yolov5s-swin	640*640	82.6%	79.3%

Table IV shows the performance comparison of different models for traffic-sign recognition. The detection accuracy (mAP) of the improved Yolov5s-Swin network proposed in this paper was 3.8%, 11.7%, and 2.4% higher than that of the Faster R-CNN, Yolov3, and Yolov5s networks, respectively. However, the reasoning speed of the improved network decreased compared with Yolov5s. The original model can reach 162 fps, whereas the improved network was stable at 118 fps on average. Its computation and reasoning speed are still better than those of Faster R-CNN and Yolov3.

V. CONCLUSION

In this study, the Crswin module based on Transformer principle is improved in the basic network model Yolov5s. Aiming at the problems of complex road scenes, numerous and dense small targets, and insufficient integration of network features of the original network, it has achieved success in extracting global features and increasing receptive field. Experiments show that this model can extract features more fully and fuse features more effectively, which improves the detection performance. The improved Yolov5s-swin algorithm has excellent detection accuracy and a real-time detection effect and can be effectively applied to the detection of road traffic signs.

REFERENCES

[1] Rehman Y, Amanullah H, Shirazi M A, et al. “Small Traffic Sign Detection in Big Images: Searching Needle in a Hay.” *IEEE Access*, 2022, 10: pp. 18667-18680.

[2] Wu Y, Li Z, Chen Y, et al. “Real-time Traffic Sign Detection and Classification towards Real Traffic Scene. Multimedia Tools and Applications.” 2020, 79(25): pp. 18201-18219.

[3] Liu Z, Qi M, Shen C, et al. “Cascade Saccade Machine Learning Network with Hierarchical classes for Traffic Sign Detection. Sustainable Cities and Society.” 2021, 67: 102700.

[4] Kamiyama M, Taguchi A. “HSI Color Space with same Gamut of RGB Color Space. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences.” 2017, 100(1): pp. 341-344.

[5] Duda R O, Hart P E. “Use of the Hough Transformation to Detect Lines and Curves in Pictures.” *Communications of the ACM*, 1972, 15(1): pp. 11-15.

[6] Shaik K B, Ganesan P, Kalist V, et al. “Comparative Study of Skin Color Detection and Segmentation in HSV and YCbCr Color Space.” Procedia Computer Science.” 2015, 57: pp. 41-48.

[7] Bouti A, Mahraz M A, Riffi J, et al. “A Robust System for Road Sign Detection and Classification using LeNet Architecture based on Convolutional Neural Network.” *Soft Computing*, 2020, 24(9): pp. 6721-6733.

[8] Yu J, Ye X, Tu Q. “Traffic Sign Detection and Recognition in Multiimages Using a Fusion Model with YOLO and VGG Network.” *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[9] Ren S, He K, Girshick R, et al. “Faster r-cnn: Towards Real-Time Object Detection with Region Proposal Networks.” *Advances in neural information processing systems*, 2015, 28.

[10] Lin T Y, Goyal P, Girshick R, et al. “Focal loss for Dense Object Detection.” *Proceedings of the IEEE international conference on computer vision*. 2017: pp. 2980-2988.

[11] Liu W, Anguelov D, Erhan D, et al. “Ssd: Single Shot Multibox Detector.” *European conference on computer vision*. Springer, Cham, 2016: pp. 21-37.

[12] Fu C Y, Liu W, Ranga A, et al. “DSSD: Deconvolutional Single Shot Detector.” arXiv:1701.06659, 2017.

[13] Li Z, Zhou F. “FSSD: Feature Fusion Single Shot Multibox Detector.” arXiv:1712.00960, 2017.

[14] Redmon J, Divvala S, Girshick R, et al. “You only look once: Unified, Real-Time Object Detection.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: pp. 779-788.

[15] Redmon J, Farhadi A. “YOLO9000: Better, Faster, Stronger.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017: 7263-7271.

[16] Redmon J, Farhadi A. “Yolov3: An Incremental Improvement.” arXiv preprint arXiv:1804.02767, 2018.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. “Yolov4: Optimal Speed and Accuracy of Object Detection.” arXiv preprint arXiv:2004.10934, 2020.

[18] Wan J, Ding W, Zhu H, et al. “An Efficient Small Traffic Sign Detection Method based on YOLOv3.” *Journal of Signal Processing Systems*, 2021, 93(8): pp. 899-911.

[19] Du Luyao et al. “Improved Detection Method for Traffic Signs in Real Scenes applied in Intelligent and Connected Vehicles.” *IET Intelligent Transport Systems*, 2020, 14(12): pp. 1555-1564.

[20] Wang Jiadong et al. “An Object Detection Model for Paint Surface Detection Based on Improved YOLOv3.” *Machines*, 2022, 10(4) : pp. 261-261.

[21] Shao Yanhua et al. “AIR-YOLOv3: Aerial Infrared Pedestrian Detection via an Improved YOLOv3 with Network Pruning.” *Applied Sciences*, 2022, 12(7): pp. 3627-3627.

[22] Vaswani A, Shazeer N, Parmar N, et al. “Attention is All You Need.” *Advances in neural information processing systems*, 2017, 30.

[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” arXiv preprint arXiv:2010.11929, 2020.vit.

[24] Liu Z, Lin Y, Cao Y, et al. “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.” *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021: pp. 10012-10022.

[25] Zhu Z, Liang D, Zhang S, et al. “Traffic-sign Detection and Classification in the Wild.” *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: pp. 2110-2118.