Journal of Industrial Intelligence
https://www.acadlore.com/journals/JII

An Advanced YOLOv5s Approach for Vehicle Detection Integrating Swin Transformer and SimAM in Dense Traffic Surveillance

Yi Zhang[1], Zheng Sun[2*]

[1] School of Environment, Education and Development, The University of Manchester, M13 9PL Manchester, UK
[2] School of Management and Engineering, Capital University of Economics and Business, 100070 Beijing, China

* Correspondence: Zheng Sun (32021210103@cueb.edu.cn)

Received: 02-08-2024 **Revised:** 03-12-2024 **Accepted:** 03-20-2024

Citation: Y. Zhang and Z. Sun, "An advanced YOLOv5s approach for vehicle detection integrating Swin Transformer and SimAM in dense traffic surveillance," *J. Ind Intell.*, vol. 2, no. 1, pp. 31–41, 2024. https://doi.org/10.56578/jii020103.

Abstract: In the realm of high-definition surveillance for dense traffic environments, the accurate detection and classification of vehicles remain paramount challenges, often hindered by missed detections and inaccuracies in vehicle type identification. Addressing these issues, an enhanced version of the You Only Look Once version v5s (YOLOv5s) algorithm is presented, wherein the conventional network structure is optimally modified through the partial integration of the Swin Transformer V2. This innovative approach leverages the convolutional neural networks' (CNNs) proficiency in local feature extraction alongside the Swin Transformer V2's capability in global representation capture, thereby creating a symbiotic system for improved vehicle detection. Furthermore, the introduction of the Similarity-based Attention Module (SimAM) within the CNN framework plays a pivotal role, dynamically refocusing the feature map to accentuate local features critical for accurate detection. An empirical evaluation of this augmented YOLOv5s algorithm demonstrates a significant uplift in performance metrics, evidencing an average detection precision (mAP@0.5:0.95) of 65.7%. Specifically, in the domain of vehicle category identification, a notable increase in the true positive rate by 4.48% is observed, alongside a reduction in the false negative rate by 4.11%. The culmination of these enhancements through the integration of Swin Transformer and SimAM within the YOLOv5s framework marks a substantial advancement in the precision of vehicle type recognition and reduction of target miss detection in densely populated traffic flows. The methodology's success underscores the efficacy of this integrated approach in overcoming the prevalent limitations of existing vehicle detection algorithms under complex surveillance scenarios.

Keywords: Improved YOLOv5s algorithm; Target detection; Deep learning; Swin Transformer; Similarity-based Attention Module (SimAM)

1 Introduction

With social development and the acceleration of urbanization, the scale of China's traffic system continues to expand, and its layout is constantly optimized, which puts forward higher requirements for the safety, efficiency, and intelligence of road traffic operations. The traffic system has gradually developed towards intelligent traffic systems, among which roadside object detection technology plays a crucial role. On the "highway" of information technology development, the advancement speed of artificial intelligence has made it an indispensable part. In intelligent traffic systems, video vehicle detection, as one of the foundational technologies, has evolved from simple video recording systems to semi-autonomous systems capable of detecting real-time vehicles accurately, providing effective data support for the implementation of traffic control measures [1]. Recently, Chen and Li [2] proposed an effective vehicle detection method that utilizes deep learning for the study of vehicle detection algorithms, offering new insights for algorithm improvement.

By setting up supporting sensing facilities on the roadside, relevant departments can timely collect and process current road traffic flow information for a more rational allocation of traffic resources. However, target detection in roadside scenarios still faces many challenges, such as the complexity and variability of the road environment and the severity of vehicle occlusion during road congestion, while overcoming limitations in computational resources [3, 4]. These challenges have driven the application of deep learning in target detection. For example, Wang and Jia [5]

optimized vehicle detection algorithms using deep learning techniques, constructing a vehicle detection model that is both real-time and accurate. Therefore, it is a main task for researchers to improve the detection accuracy of difficult samples, such as small targets and occluded targets, while ensuring the model's lightweight. This task holds high research value and provides more reliable support for the implementation of traffic management measures.

Common target detection algorithms can be divided into two types: those based on traditional methods and those based on deep learning. Among them, target detection algorithms based on traditional methods mainly rely on manual extraction of image features. Ravichandran et al. [6] utilized SAR image data to extract target features to complete vehicle identification. Dong et al. [7] applied Haar-like and Histogram of Oriented Gradient (HOG) features to extract vehicle features. Although traditional methods basically meet identification needs, they require a large amount of data information and preparation work during the training stage. In addition, traditional classification networks also heavily depend on predefined features and classifiers, making deep learning methods gradually gain wider application in the field of vehicle target detection.

Vehicle target detection based on deep learning does not require complex data preprocessing. The YOLO algorithm, proposed by Redmon et al. [8], is a milestone in object detection using deep learning, distinguishing itself from classical and traditional object detection methods that rely on manually designed feature extractors with slow algorithm speeds and low accuracy. The algorithm has a wide range of applications, finds detection targets through bounding boxes, and achieves high accuracy with less training resources. With the rapid development of deep learning technology, several excellent vehicle target detection algorithms have emerged, such as YOLOv5, YOLOv4, and YOLOv3 [9]. Among them, YOLOv5 is particularly popular for its easy setup, fast training speed, and user-friendly nature, making it one of the most widely applied detection networks today. For instance, Zhang et al. [3] studied a vehicle millimeter-wave radar target detection method based on deep learning, utilizing the YOLOv5 algorithm to better address the issues found in traditional Constant False Alarm Rate (CFAR) algorithms. Liu [10] replaced the backbone network in YOLOv5 with the basic structure of the MobileNet V3 network to compress the model and speed up detection. Li et al. [11] enhanced the network's feature extraction capability by adding a Convolutional Block Attention Module (CBAM) to the backbone network of YOLOv5 and replacing the neck part of the network with a bi-directional feature pyramid network (BiFPN) structure to improve the algorithm's ability to detect small targets.

The problem of traffic vehicle target detection presents the following difficulties:

• Due to the camera angle, the size of the targets in the video images varies, easily missing smaller-sized targets.

• There are mutual occlusion problems, especially in traffic jam situations, where the occlusion problem is more severe.

• High-speed moving vehicles may appear blurred and have afterimages in high-definition video surveillance, affecting recognition accuracy [12].

These factors lead to missed or false detections. The main means to improve algorithm speed and accuracy are to combine different application scenarios to select suitable YOLOv5 model weights, optimize parameter settings and replace the backbone network based on the characteristics of the targets to be detected. By taking into account factors, such as the volume of the target detection model, detection accuracy, and detection speed under high-definition video surveillance, this study selects the YOLOv5s as the basis for related research and improvement work, with the main contributions as follows:

• Swin Transformer V2 [13] is used to replace part of the original network structure, which optimizes network structure layers to improve the global searching ability of the target detection algorithm, thereby facilitating the early discovery of more targets under traffic road conditions.

• The SimAM [14] is introduced, enhancing the local discriminative feature extraction ability of the CNN network branch.

• More common improvement measures are combined, such as introducing a small target detection layer and modifying and replacing the loss function, etc. Target detection and recognition situations are compared to select the optimal improvement measures.

• The improved YOLOv5s algorithm is validated under actual road detection conditions.

In summary, this study, based on the YOLOv5s algorithm, ultimately selects the improved algorithm combining Swin Transformer V2 and SimAM, improving the target detection capability and the accuracy of target type recognition.

2 Related Work

2.1 Replacing Part of the Network Structure with Swin Transformer V2

The Swin Transformer was introduced in 2021 [15] as a backbone network for computer vision. Its name derives from the concept of a hierarchical vision Transformer using shifted windows. This design utilizes moving windows for each attention mechanism computation within the windows themselves, allowing the computational complexity to grow linearly with the image size when processing higher-resolution data rather than quadratically. Furthermore,

the use of moving windows not only brings about greater detection efficiency but also facilitates interaction between adjacent windows. Its design incorporates patch merging, which merges adjacent patches into one for feature value computation, similar to the pooling layer operations covered by the CNN in the original YOLOv5 backbone network. This enlarges the network's receptive field and enhances its multi-scale perception ability. Following this, Swin Transformer V2 [13], with 3 billion parameters, has become the largest dense vision model to date.

The structure of the Swin Transformer is shown in Figure 1. Swin Transformer V2 modifies the order of layer norm in each transformer encoder block to facilitate the normalization of outputs. It also replaces inner product similarity with cosine similarity. This compensates for the lack of long-distance modeling YOLOv5s's CNN and its inability to effectively capture global information, thereby achieving better target feature extraction effects. As shown in Figure 2, the STv2 module, or Swin Transformer V2, replaces some of the convolutional blocks in the original YOLOv5s.

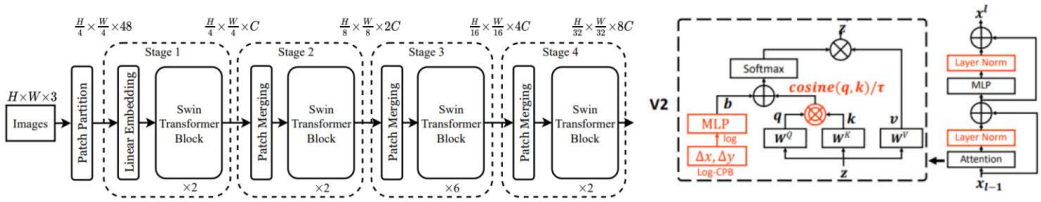

Figure 1. Structure of the Swin Transformer V2

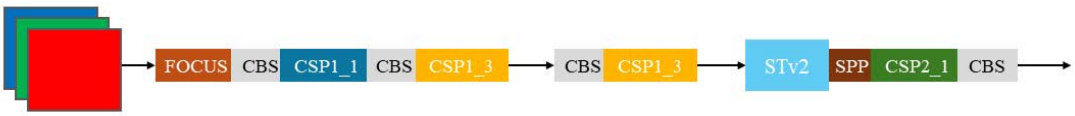

Figure 2. Module replaced with Swin Transformer V2

This study incorporates several key modules associated with Swin Transformer V2 into the training of YOLOv5s, leveraging the advanced features of Swin Transformer V2. By integrating the Swin Transformer block, window attention mechanism, and Multilayer Perceptron (MLP) module, the model's ability to capture both global and local information when processing high-resolution images is enhanced. The integration of these modules boosts the network's self-attention computation, enabling it to capture distant pixel relationships more accurately while maintaining computational efficiency, especially in densely populated vehicle scenarios. Additionally, introducing the C3STR module into YOLOv5s, with its cross-stage and cross-shape triple receptive field, further optimizes the model's detection capability for vehicles of varying sizes and under occlusion conditions. These improvements enable the model to achieve higher accuracy in identifying and classifying various vehicle targets in complex traffic environments.

2.2 SimAM

SimAM [14] is a new type of attention module in CNNs, characterized by its simplicity and efficiency. Unlike Bottleneck Attention Module (BAM) and CBAM, which focus on channel and spatial attention in parallel and serial manners, respectively [16], SimAM infers 3D attention weights (full 3D weights) for a layer's feature map without adding any parameters to the original network, effectively coordinating channel and spatial attention. SimAM can better process both types of attention and significantly improve the model's detection performance in complex traffic scenes. The energy function of SimAM is shown in Eq. (1):

$$e_i\left(w_i, b_i, y, x_i\right) = \left(y_i - \hat{t}\right)^2 + \frac{1}{M-1}\sum_{i=1}^{M-1}\left(y_0 - \hat{x}\right)^2 \tag{1}$$

where, w_i and b_i represent the weight and bias, respectively, $\hat{t}$ and x_i belong to the target neuron and other neurons in the channel, M is the number of neurons in that channel, i is the spatial index. The greater the difference between a neuron t and its surrounding neurons, the higher its importance, indicating lower energy. As shown in Figure 3, SimAM is introduced into the backbone network of YOLOv5s.

SimAM includes a rapid closed-form solution energy function that can be implemented in less than ten lines of code. One of the advantages of SimAM is that most of the operations are based on solutions to the defined

energy function, which avoids extensive structural adjustments. The module proposed by SimAM is flexible and effective, enhancing the feature extraction capability of many CNNs. It is an efficient attention module that is easy to implement and can significantly improve performance across various visual tasks. The SimAM is incorporated into the training of YOLOv5 by modifying the detection head of the backbone network in the yaml configuration file and adding the SimAM at appropriate layers. This enhances the model's performance in identifying individual vehicles while improving the accuracy and reliability of recognizing occluded or distant small-sized vehicles.

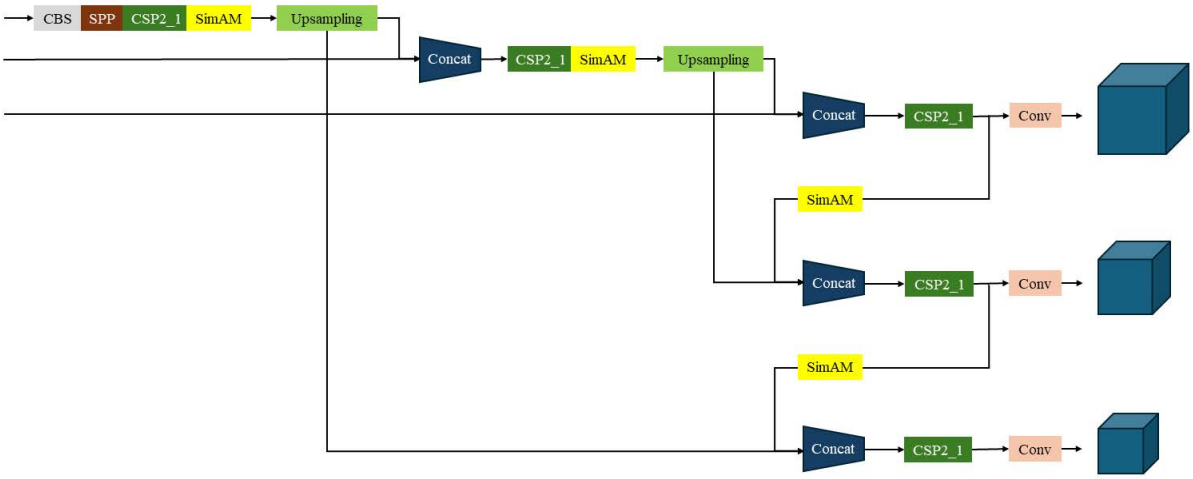

Figure 3. Introduction of SimAM

2.3 Small Object Detection Layer

The demand for faster and more accurate detectors for small objects in target detection tasks is increasing. Although the human eye can almost instantly capture contextual information even from a distance, due to limitations in image resolution and computing resources, machine detection of small objects that occupy only a few pixels in an image remains a challenging task [17]. Pretraining results show that the target detection algorithm only recognizes targets when vehicles are relatively close, and this phenomenon is more intuitive in the case of unclear input images. In the COCO dataset used in this study, objects with a pixel count less than 32×32 are considered small objects. To enhance the model's accuracy for small-sized targets, a detection head specifically for small targets is added to the original network of YOLOv5s. Based on CNN, the small object detection layer retains the original detection structure and expands with an additional convolutional layer to obtain a 160×160 feature layer. This approach increases the computational load to some extent but provides higher detection accuracy. The network structure improved by this method is shown in Figure 4.

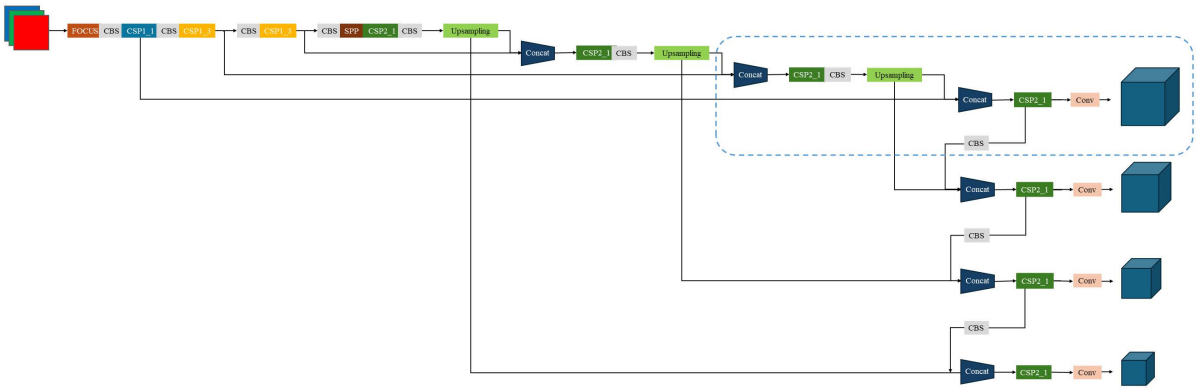

Figure 4. Small object detection head

2.4 EIoU Loss Function

In YOLOv5s, the head (head layer) uses CIoU, which is a bounding box loss function, to measure the consistency of the aspect ratio between the prediction and the target box, as shown in Eq. (2):

$$\begin{aligned} L_{CIoU} &= IoU - \rho^2\left(b, b^{gt}\right) = av \\ a &= \frac{v}{1 - IoU + v} \\ v &= \frac{2}{\pi}\left(\arctan\frac{w^{ht}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \end{aligned} \tag{2}$$

where, $\rho^2\left(b, b^{gt}\right)$ represents the Euclidean distance between the center points of the prediction and target boxes, w^{ht}, h^{gt}, w, and h are the width and height of the prediction and target boxes, respectively. The formula reflects the difference in aspect ratio rather than the actual difference in width, height, and their confidence levels, which could hinder the model's effective optimization of similarity. Addressing this issue, EIoU is proposed on the basis of CIoU by separating the aspect ratio into width and height, as shown in Eq. (3):

$$\begin{aligned} L_{EIOU} &= L_{EIOU} + L_{DIS} + L_{ASP} \\ &= 1 - IoU + \frac{\rho^2\left(v, v^{gt}\right)}{c^2} + \frac{\rho^2 + \left(w, w^{gt}\right)}{c_z^2} + \frac{\rho^2 + \left(m, m^{gt}\right)}{c_m^2} \end{aligned} \tag{3}$$

where, c_z^2, c_m^2, and ρ are the height and width of the two smallest boxes, and v and v^{gt} are the Euclidean distances between them. In other words, the original "CIoU = IoU + center point loss + aspect ratio loss" is changed to "EIoU = IoU + center point loss + width loss + height loss". During the experiments, an attempt was made to replace CIoU with EIoU.

3 YOLOv5 Target Detection Algorithm

The YOLO algorithm is a one-stage target detection algorithm, meaning it uses a single CNN model to achieve end-to-end target detection. The input image of a one-stage algorithm passes through a single network, and the generated results include both detection locations and category information. YOLO utilizes convolutional networks to extract features and fully connected layers to obtain prediction values, dividing the input image into grids, with each grid detecting targets, predicting bounding boxes and their confidence scores. Finally, non-maximum suppression is used for network prediction [18]. The most widely used YOLO algorithm is YOLOv5 [19], which has reached a high level of detection accuracy on general object detection datasets, such as MS COCO [20] and VOC [21] to date. However, vehicle-type target detection under complex traffic flow conditions requires targeted design and training. Through corresponding optimization, a detection model more suited to the target detection task is ultimately obtained.

YOLOv5, in addition to the basic model, has four different sizes of weight models for different layers of the neural network feature extraction layer, namely, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Although many studies suggest enhancing the representation ability of the target detection model in terms of depth and width to increase detection accuracy, this affects the complexity of the model. Since detection tasks require real-time performance, such models are not suitable for autonomous driving systems [22]. As the smallest model, YOLOv5s has the fastest computational speed. According to the given data, it can achieve a computational inference speed of 2 ms per image on the COCO2017 dataset while having the fewest model parameters and the fewest floating-point operations (FLOPs). Although this affects model performance, mAP@0.5 still reaches 56.8%. The YOLOv5s model is highly performant and convenient for deployment, consuming fewer resources [23]. The YOLOv5s algorithm network structure is divided into the head, neck, and backbone networks. The backbone network, consisting of Focus, Center and Scale Prediction (CSP), CBL, and SPP modules, is used for extracting key features from the input image. The SPP module and Path Aggregation Network (PA-NET) are used to integrate feature output from the backbone to the head network. The head network is responsible for the final detection steps, constructing the bounding box positions and recognition types into a neural network to form the final output vector [24, 25]. The network structure of YOLOv5s is shown in Figure 5.

In the above figure, *Conv* represents the convolutional layer, *BN* stands for batch normalization layer, *SiLU* is the activation function, *MaxPool* is the max pooling layer, *Concat* refers to the concatenation layer, and *Resunit* is the residual unit [26]. During the prediction process, YOLOv5 takes into account the distance information of the center points of the bounding frames, using the CIoU loss function [27] to calculate the ratio of the intersection and the union of two bounding boxes.

Beyond the basic model, YOLOv5 provides four different sizes of pretrained weight models for different layers of the neural network feature extraction layer, namely, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. Table 1 compares the performance of YOLO models.

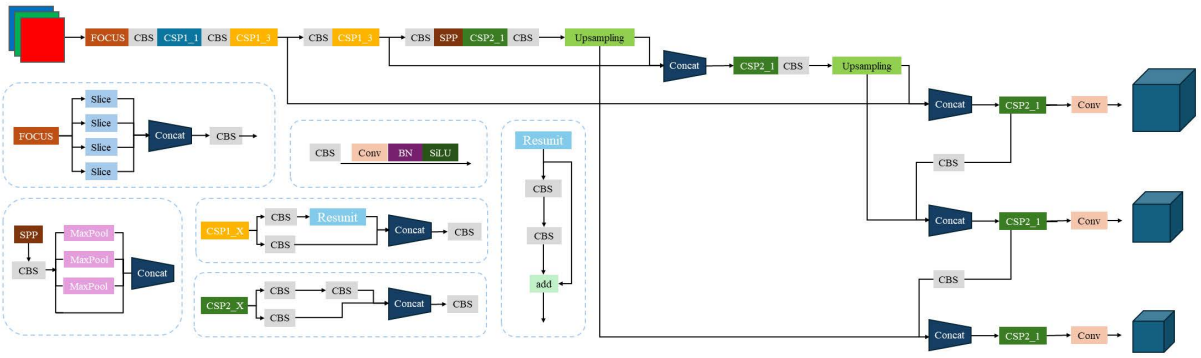

Figure 5. YOLOv5 network structure

Table 1. YOLO pre-trained model

Model	Size (pixels)	mAPval 50-95	mAPval 50	Speed CPU b1 (ms)	Speed V100 b1 (ms)	Speed V100 b32 (ms)	Params (M)	FLOPs @640 (B)
YOLOv5n	640	28.0	45.7	45	6.3	0.6	1.9	4.5
YOLOv5s	640	37.4	56.8	98	6.4	0.9	7.2	16.5
YOLOv5m	640	45.4	64.1	224	8.2	1.7	21.2	49.0
YOLOv5I	640	49.0	67.3	430	10.1	2.7	46.5	109.1
YOLOv5x	640	50.7	68.9	766	12.1	4.8	86.7	205.7

Among them, the YOLOv5s model performs very well and is convenient for deployment, consuming fewer resources. Taking everything into consideration, it is believed in this study that YOLOv5s is the best choice for the application environment of traffic flow volume statistics. The overall detection process of YOLOv5s can be summarized as follows: First, the image is scaled and padded to meet the input requirements of the YOLOv5 network. Then, the input image is processed with the YOLOv5 algorithm to obtain a series of bounding box location information and type prediction results. Next, the non-maximum suppression module is used to process the output bounding boxes, eliminating redundant duplicate detection results. Finally, the remaining bounding boxes are output as the result of vehicle detection.

To verify the effectiveness of the YOLOv5s model in complex traffic environments, an analysis predicated on polarity pretraining of the YOLOv5s model was conducted. The initial phase of training culminated in a duration of 1232 seconds, accompanied by an F1-score of 0.84. The pretraining results are shown in Figure 6.

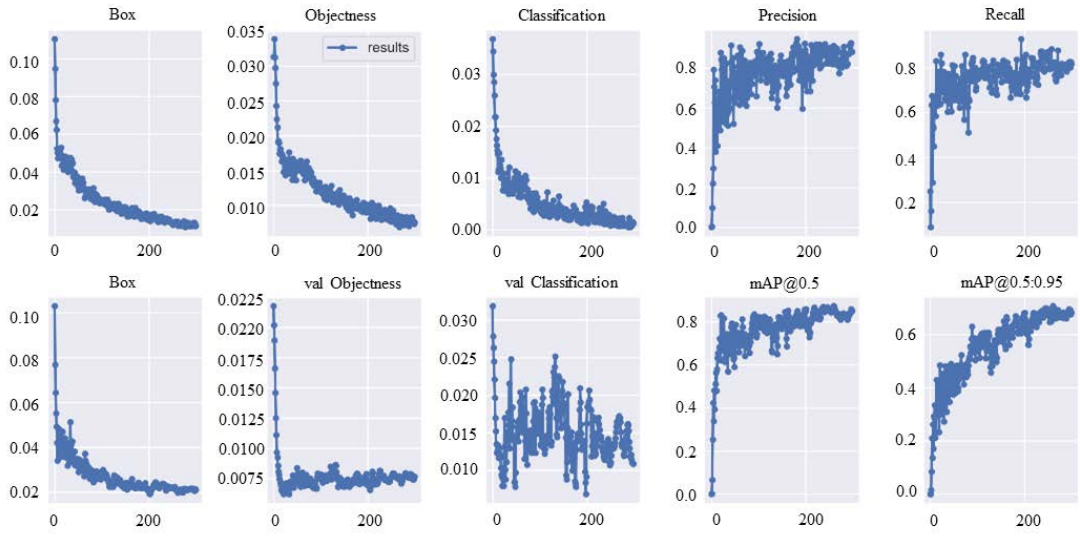

Figure 6. Pre-trained results

In the above figure, *Box* indicates the mean of the GIoU loss function. The smaller the box, the more accurate it is. *Objectness* is the mean of the object detection loss. The smaller, the more accurate the object detection. *Classification* is the mean of the classification loss. The smaller, the more accurate the classification. It can be concluded that during the training process, the means of each metric decrease steadily, and the convergence speed is ideal. Notably, the validation set exhibits significant fluctuations in the classification results, which are speculated to be caused by the small number of images in the validation set and the large difference in the number of different categories of vehicles. Subsequent expansion of the validation set and rationalization of the number of vehicle categories were conducted for repeated pretraining, and the validation results were generally satisfactory. A partial view of the validation conditions is shown in Figure 7.

Figure 7. Pre-trained validation results

4 Experiments

4.1 Experimental Environment and Traffic Flow Dataset

The experimental environment includes the Windows 11 operating system, an Intel-Core i7-8700K CPU, an NVIDIA GeForce RTX 2080 GPU, 16GB of RAM, and Pycharm IDE with Python version 3.7.11.

Traffic flow in China is characterized by high density and complex information. This study selects three of the most representative types of large motor vehicles on roads, namely, cars, buses, and trucks, to construct the initial dataset. The format of the dataset is based on the COCO dataset. The images in the dataset mainly come from VOC2007 and COCO2017. Since the shooting locations of these two official datasets are mostly in Europe and America, photos of Chinese traffic flows were added and annotated, aiming to adapt the algorithm to the application environment of this study while retaining the diversity of the original COCO dataset. This expands the coverage of the dataset and enhances its practicality in local traffic scenarios, providing support for algorithm training and optimization [28].

4.2 Experimental Results and Analysis

After implementing the aforementioned four improvements to the YOLOv5s algorithm, training was conducted on the same traffic flow dataset for each modification, with each change trained 10 times, and the average value of the experiment's metrics was taken. Table 2 shows the training results.

Preliminary analysis of the experimental results showed that, after making modifications with Swin Transformer V2 to replace part of the neural network and the SimAM, the overall average accuracy improved by about 5%, and it could even enhance the recognition capability by up to 9% under ideal conditions. The improvement in detection

capability with the addition of the small object detection layer to YOLOv5 was not outstanding, with a quite limited enhancement in detection performance. Figures 8–10 show part of the testing process for different models.

Table 2. Comparison of experimental results

Model	F1-Score	mAP@0.5	mAP@0.5:0.95	Best mAP@0.5
YOLOv5s	0.79	0.791	0.608	0.794
YOLOv5s+Swin Transformer V2	0.86	0.854	0.685	0.882
YOLOv5s+small object detection layer	0.78	0.790	0.633	0.802
YOLOv5s+SimAM	0.83	0.838	0.680	0.859
YOLOv5s+EIoU	0.78	0.785	0.645	0.793

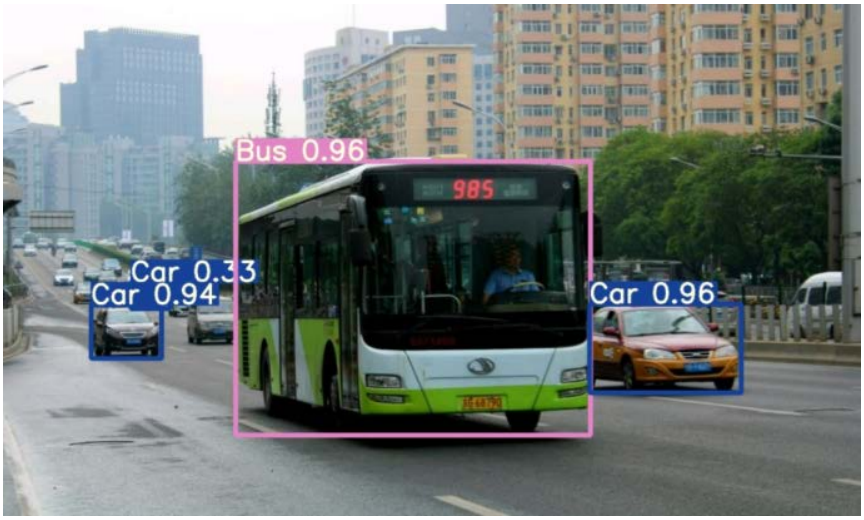

Figure 8. Model test A

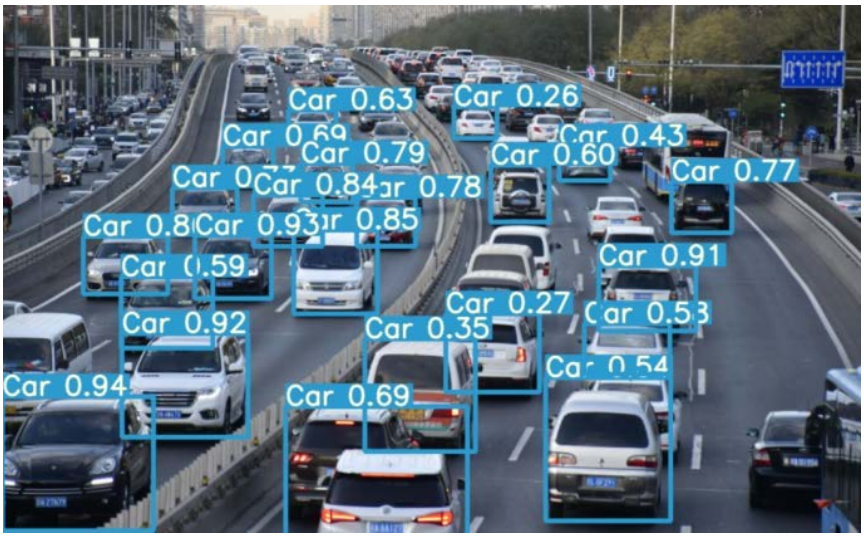

Figure 9. Model test B

The study considers optimizing the small object detection capability of YOLOv5 by combining the small object detection layer with different backbone neural networks. Similarly, the replacement of the original CIoU with EIoU on the traffic flow dataset had a limited improvement effect and even impacted the detection performance, leading to the preliminary conjecture that EIoU might not be entirely suitable for the dataset studied in this study.

To better understand the impact of various improvement modules in YOLOv5 on training and recognition effects, especially for video detection environments, various improvement modules were combined and experimented with step by step for comparison. Table 3 shows the comparison results.

Figure 10. Model test C

Table 3. Comparison of experimental results

Model	FPS	F1-Score	mAP@0.5	mAP@0.5:0.95
Swin Transformer V2+small object detection layer	55.8	0.82	0.811	0.653
SimAM+small object detection layer	64.2	0.79	0.819	0.638
Swin Transformer V2+ SimAM+small object detection layer	57.4	**0.83**	0.816	0.621
Swin Transformer V2+SimAM	**64.7**	0.82	**0.827**	**0.657**
Swin Transformer V2+EIoU	58.9	0.78	0.795	0.616
SimAM+EIoU	60.3	0.76	0.769	0.601

YOLOv5s, when combined with Swin Transformer V2+SimAM or SimAM with the small object detection layer, could produce quite good detection effects. In addition, the detection frame rate under high-definition video was ideal, basically meeting the requirements. After weighing the pros and cons, the YOLOv5s improvement method combining Swin Transformer V2+SimAM was chosen for actual application testing. Table 4 shows the comparison of algorithm evaluation metrics before and after improvement.

Table 4. Comparison of algorithm evaluation metrics before and after improvement

Model	Precision (%)	Recall (%)	Positive Detection Rate (%)	Miss Rate (%)
Initial YOLOv5s algorithm	84.72	74.21	82.91	13.53
Improved YOLOv5s algorithm	86.19	77.56	87.39	9.42

From Table 4, it can be seen that compared to the initial YOLOv5s algorithm, the improved algorithm's positive detection rate increases by 4.48%, and the miss rate decreases by 4.11%, indicating an enhancement over the initial algorithm and stronger recognition capability of the improved YOLOv5s algorithm for traffic flow detection. Figure 11 shows some of the test recognition situations.

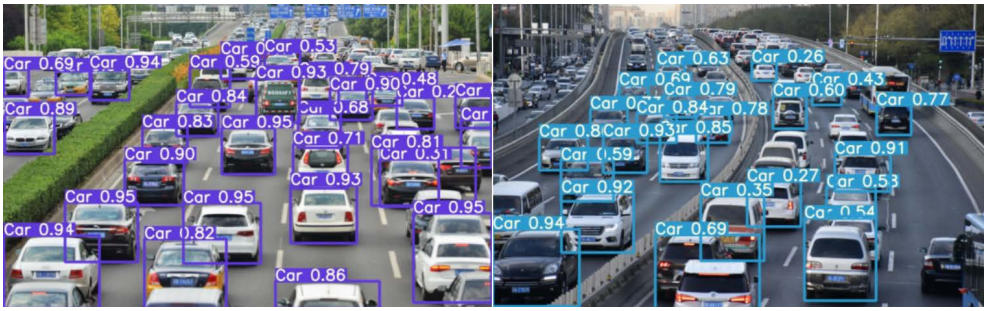

Figure 11. Partial recognition situation

4.3 Applying and Analyzing Video Stream Detection Using Different Improved Models

Based on the results of the ablation study in the previous chapter, combined with DeepSort, enhanced with Swin Transformer V2+SimAM, and modifications for small target detection layers, YOLOv5 was applied to the task of traffic flow counting. The application videos were taken in daylight with normal lighting, featuring a large number of small passenger cars and a few buses. There were instances where the model incorrectly categorized vehicle types. This study addresses these errors by marking incorrect data in the corresponding positions of multiple statistical categories as "statistical result-error quantity." When considering the accuracy of the statistics, the quantity of the corresponding category is also the result of subtraction. Table 5 shows the application results.

Table 5. Comparison of application scenario results

Model	Number of Cars	Number of Buses	Statistical Accuracy
Original video	26	3	-
Initial YOLOv5	23-1	1	79.31%
Swin Transformer V2+SimAM	24-1	2	86.21%
Swin Transformer V2+ SimAM+small target detection layer	25-1	2	89.66%

The statistical results indicate that YOLOv5, with parts of its CNN backbone network replaced by Swin Transformer V2 and combined with SimAM, can significantly enhance the accuracy of target recognition in traffic flow counting application scenarios. Furthermore, adding a small target detection layer on this basis further improves the algorithm's ability to recognize distant traffic flows.

5 Conclusions

Addressing the challenge of detecting numerous small-sized targets with mutual occlusion in traffic flow environments, which often leads to suboptimal detection performance, this study embarks on an exploration of an improved YOLOv5s algorithm. By modifying the network structure of the initial target detection algorithm and comparing several common algorithm modifications, an improved method of YOLOv5s combining Swin Transformer V2 and introducing SimAM was proposed. The experimental results proved that its detection performance was superior to the initial YOLOv5s model. The improved method proposed in this study achieved mAP@0.5 and mAP@0.5:0.95 of 0.827 and 0.657, respectively, with a model volume not significantly different from the initial model. In practical tests, both the positive detection rate and miss rate showed corresponding improvements, offering certain advantages over the initial model and being more suitable for traffic flow detection tasks. In future research, the continuous frame miss rate of the improved algorithm will be further explored, aiming to achieve higher accuracy and fewer omissions while satisfying traffic flow video monitoring requirements.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] S. Ganapathy and D. Ajmera, "An intelligent video surveillance system for detecting the vehicles on road using refined YOLOV4," *Comput. Electr. Eng.*, vol. 113, p. 109036, 2024. https://doi.org/10.1016/j.compeleceng.2023.109036

[2] Y. Chen and Z. Li, "An effective approach of vehicle detection using deep learning," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–9, 2022. https://doi.org/10.1155/2022/2019257

[3] X. Zhang, F. Liang, and X. Chen, "A target detection method of automotive millimeter wave radar based on deep learning," in *2023 8th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, China*, 2023, pp. 175–179. https://doi.org/10.1109/ccisp59915.2023.10355769

[4] Y. Li, "A review of research on deep learning-based target detection technology for automated vehicle driving systems," *Highlights Sci. Eng. Technol.*, vol. 27, pp. 19–24, 2022. https://doi.org/10.54097/hset.v27i.3716

[5] N. Wang and Y. Jia, "Research on vehicle object detection based on deep learning," in *2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL), Zhuhai, China*, 2023, pp. 412–415. https://doi.org/10.1109/cvidl58838.2023.10166204

[6] B. Ravichandran, A. Gandhe, R. Smith, and R. Mehra, "Robust automatic target recognition using learning classifier systems," *Inf. Fusion*, vol. 8, no. 3, pp. 252–265, 2007. https://doi.org/10.1016/j.inffus.2006.03.001

[7] T. Dong, T. Ruan, J. Wu, and et al., "Research on traffic video vehicle recognition method combining Haar-like and HOG features," *J. Zhejiang Univ. Technol.*, vol. 43, no. 5, pp. 503–507, 2015.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016. https://doi.org/10.1109/cvpr.2016.91

[9] K. Zhang, C. Wang, X. Yu, A. Zheng, M. Gao, Z. Pan, G. Chen, and Z. Shen, "Research on mine vehicle tracking and detection technology based on YOLOv5," *Syst. Sci. Control Eng.*, vol. 10, no. 1, pp. 347–366, 2022. https://doi.org/10.1080/21642583.2022.2057370

[10] Z. Liu, "Improved road target tracking algorithm based on YOLOv5 and DeepSort," *Pract. Automobile Technol.*, vol. 47, no. 22, p. 5, 2022.

[11] X. Li, J. Tong, Z. Chen, Y. Bao, and J. Ni, "Small object detection based on improved YOLOv5," *Comput. Syst. Appl.*, vol. 31, no. 12, pp. 242–250, 2022.

[12] T. N. Doan and M. T. Truong, "Real-time vehicle detection and counting based on YOLO and DeepSORT," in *2020 12th International Conference on Knowledge and Systems Engineering (KSE), Can Tho, Vietnam*, 2020. https://doi.org/10.1109/kse50997.2020.9287483

[13] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo, "Swin Transformer V2: Scaling up capacity and resolution," *arXiv*, p. arXiv:2111.09883, 2021. https://doi.org/10.48550/ARXIV.2111.09883

[14] L. Yang, R. Y. Zhang, L. Li, and X. Xie, "Simam: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning, PMLR 139, Virtual*, 2021, pp. 11 863–11 874.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV), Virtual*, 2021, pp. 10 012–10 022. https://doi.org/10.1109/iccv48922.2021.00986

[16] X. Hu and G. Wang, "Research on safety helmet detection algorithm based on improved YOLOv5," *Comput. Era*, vol. 372, no. 6, pp. 76–81, 2023. https://doi.org/10.16644/j.cnki.cn33-1094/tp.2023.06.016

[17] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "Yolo-Z: Improving small object detection in YOLOv5 for autonomous vehicles," *arXiv*, p. arXiv:2112.11798, 2021. https://doi.org/10.48550/ARXIV.2112.11798

[18] Y. Shao, D. Zhang, H. Chu, X. Zhang, and Y. Rao, "A survey on deep learning-based YOLO object detection," *J. Electron. Inf. Technol.*, vol. 44, no. 10, pp. 3697–3708, 2022. https://doi.org/10.11999/JEIT210790

[19] A. Malta, M. Mendes, and T. Farinha, "Augmented reality maintenance assistant using YOLOv5," *Appl. Sci.*, vol. 11, p. 4758, 2021. https://doi.org/10.3390/app11114758

[20] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV), Zurich, Switzerland*, 2014, pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal Visual Object Classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2009. https://doi.org/10.1007/s11263-009-0275-4

[22] B. Mahaur and K. Mishra, "Small-object detection based on YOLOv5 in autonomous driving systems," *Pattern Recognit. Lett.*, vol. 168, pp. 115–122, 2023. https://doi.org/10.1016/j.patrec.2023.03.009

[23] S. S. Park, V. T. Tran, and D. E. Lee, "Application of various YOLO models for computer vision-based real-time pothole detection," *Appl. Sci.*, vol. 11, no. 23, p. 11229, 2021. https://doi.org/10.3390/app112311229

[24] R. Cai, N. Cheng, Z. Peng, and et al., "Research on lightweight infrared weak small vehicle target detection algorithm based on deep learning," *Infrared Laser Eng.*, vol. 51, no. 12, pp. 357–367, 2022.

[25] X. Chen, C. Wang, and Y. Wu, "Real-time vehicle target detection method of improved YOLOv5s algorithm," *J. Harbin Univ. Sci. Technol.*

[26] F. Chollet, "Xception: Deep Learning with depthwise separable convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA*, 2017, pp. 1251–1258. https://doi.org/10.1109/cvpr.2017.195

[27] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *arXiv*, p. arXiv:1911.08287, 2019. https://doi.org/10.48550/ARXIV.1911.08287

[28] M. Durve, S. Orsini, A. Tiribocchi, A. Montessori, J. Tucny, M. Lauricella, A. Camposeo, D. Pisignano, and S. Succi, "Benchmarking YOLOv5 and YOLOv7 models with DeepSORT for droplet tracking applications," *Eur. Phys. J. E*, vol. 46, no. 5, 2023. https://doi.org/10.1140/epje/s10189-023-00290-x