Research on Vehicle Detection Algorithms Based on YOLOv5s

Lifeng Zhang, Ying Tian

Abstract—**To tackle the challenges associated with low recognition rates and demanding computational requirements in vehicle detection, we suggest a MAGS-YOLOv5s based vehicle detection model that builds on an upgraded version of YOLOv5s. Initial efforts are focused on applying data augmentation, using mosaic augmentation and integrating mix-up augmentation to enhance the model's generalisation and robustness, thus increasing sample diversity. Secondly, the loss function for predicting bounding boxes is enhanced by including an α coefficient in addition to CIoU, facilitating adjustment to varying data scales and hastening the model's convergence speed. Moreover, Ghost Shuffle Convolution (GSC) is employed to substitute the convolutional layers in the neck of the initial network model. This replacement decreases the computational load and network parameters, whilst enhancing the feature extraction capability and robustness of the network model. Our proposed SGC block, which is built on the GC block, has been integrated to the C3 module to specifically improve the feature extraction capability of the model for occluded object detection. Experimental results indicate that the enhanced network delivers a boost of 2.9% and 5.3% in mAP (mean average precision) on the Pascal VOC and MS COCO vehicle datasets, respectively, compared to the original model. The improved network displays enhanced multi-scale representation and generalisation capabilities.**

Index Terms—**Vehicle Dection, YOLOv5s, Global Context Block, Data enhancement**

I. Introduction

Vehicle detection [1-2] is a challenging and promising research field with significant practical application value. The technologies and problems in this area play an important role in the development of computer vision and artificial intelligence. Vehicle detection has wide applications in various scenarios, such as intelligent transport systems and vehicle tracking. The growing demand for autonomous driving technology further emphasizes the importance of vehicle detection research. With the gradual development of autonomous vehicles, accurate and reliable vehicle detection technology will be crucial to ensure their safe and efficient operation. Therefore, the field of vehicle detection continues to be of great research value.

Traditional vehicles rely on feature-based detection, primarily using algorithms such as Haar [3], HOG [4], SIFT [5] for feature extraction followed by classification. Traditional detection algorithms use a set of predefined features to identify objects in images. However, in complex scenes, these features fail to recognise objects effectively due to factors such as occlusion, clutter and lighting, resulting in poor robustness. In addition, these algorithms are highly sensitive to variations in the parameters used to define the features, which reduces their robustness and ability to learn from new data, making it difficult to optimise them for different scenarios. They also require a high level of domain knowledge from operators with limited upper bounds.

With the increase in datasets and the growing demand for object detection, deep learning-based object detection algorithms have developed rapidly. Deep learning models can learn feature representations automatically through back propagation, which significantly reduces algorithm design effort. Two-stage detection methods, represented by R-CNN [6] and Faster R-CNN [7], extract image features, generate candidate regions, and then perform classification and regression. Although they achieve high detection accuracy, they cannot meet the speed requirements of vehicle object detection. On the other hand, single-stage detection methods, represented by SSD and YOLO [8-11], perform detection directly on the image, which accelerates the detection speed at the cost of some loss of accuracy.

Li Songjiang [12] introduced the channel attention mechanism ECA-Net into the backbone network to improve feature aggregation and suppress irrelevant features. In addition, the detailed information of shallow-layer features output from the backbone network is fused with the semantic information of deep-layer features. Finally, ordinary convolutions in the network feature fusion module are replaced by depth-separable convolutions.

Currently, single-stage detection models still suffer from drawbacks such as low localisation accuracy, poor performance in detecting small objects, and high training difficulty. They also have high hardware and computational requirements, which limit their application in real-time demanding scenarios such as autonomous driving and traffic control. Lightweight object detection aims to ensure detection accuracy while minimising model size and computational complexity, enabling real-time inference and deployment on embedded devices or resource-constrained systems. Consequently, many researchers have conducted studies in this area. Jin Zhi et al [13] proposed a vehicle detection algorithm called WB-YOLOv5, which is suitable for dense scenes. They designed feature selection and feature sparsity modules based on the input data structure of the ConvLSTM model, which achieved feature recalibration. By replacing the original gating structure with 1x1 convolutions, a lightweight WBConvLSTM was constructed and introduced into the neck section of the original model.Xiong

Manuscript received June 09, 2023; revised September 21, 2023.

Lifeng Zhang is a postgraduate student of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (e-mail: 3252689175@qq.com)

Ying Tian is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China. (corresponding author to provide phone: +8613898015263; e-mail: astianying@126.com)

Liyan, Tu Suocheng, Huang Xiaohui et al [14] presented a lightweight network based on MobileVit for vehicle detection. They used YOLOv4 as the base model and replaced the backbone feature extraction network to make the model lighter. In addition, they applied the GridMask image enhancement method and multi-scale techniques in the pre-processing stage to further improve the performance.

The detection accuracy can be impacted by various factors, such as occlusion and scale variations, in practical scenarios like autonomous driving and traffic detection. Therefore, this paper proposes an improved vehicle object detection method based on YOLOv5s. It uses the mixed data augmentation technique to improve the generalization ability of the dataset. Alpha-IoU is used to improve regression accuracy as a substitute for CIOU. GSConv is used to replace ordinary convolutions in the neck section, reducing the number of parameters while maintaining or even improving the accuracy of the model. Finally, the SGC module is used to improve the accuracy of small object detection.

II. Related Work

YOLOv5, as one of the models in the YOLO series, has a well-developed network structure and algorithmic concepts. It also offers greater flexibility and speed compared to other versions of YOLO and algorithms, providing significant advantages when used on embedded devices and similar platforms.

The YOLOv5 algorithm currently consists of four main versions: s, m, l and x. Of these, YOLOv5s is the smallest model in terms of size and offers extremely fast detection speed with minimal compromise on accuracy. It is particularly suitable for real-time demanding scenarios such as autonomous driving. Therefore, YOLOv5s is chosen as the base model for improvements.

YOLOv5s version 6.0 has several notable improvements over previous versions. The detection image size for YOLOv5s is 640×640. The whole network can be divided into three parts: feature extraction, feature fusion and detection heads. Mosaic augmentation is applied to the dataset, where images are randomly scaled, cropped and arranged to create different patterns. This increases the variability of the dataset.

The feature extraction part uses an improved version of CSPDarknet [15], which strengthens feature propagation between layers, reduces the number of parameters, and improves network performance and stability. Finally, the Spatial Pyramid Pooling Fusion (SPPF) module is applied, which extracts features at different receptive fields through three layers of maximum pooling.

For feature fusion, multi-scale fusion is performed using the Feature Pyramid Network (FPN) in a top-down manner and the Feature Aggregation Network (FAN) in a bottom-up manner. This allows features to be extracted at different scales, thereby improving detection performance.

The detection layer receives the fused new feature layers from the neck layer and divides them into three universal scales: 80×80, 40×40 and 20×20 for detection.

III. Improved Model

Fig. 1 depicts the MAGS-YOLOv5s network model, which features significant improvements over YOLOv5s in object recognition. Notably, the model incorporates data augmentation, the enhanced alpha-CIoU metric, GSConv, and the SGC block, all of which augment its superior performance. It overcomes the limitation of inadequate training data, accelerates the model's convergence speed, enhances its robustness, optimizes computational efficiency and accuracy, and improves the model's ability to handle occlusion.

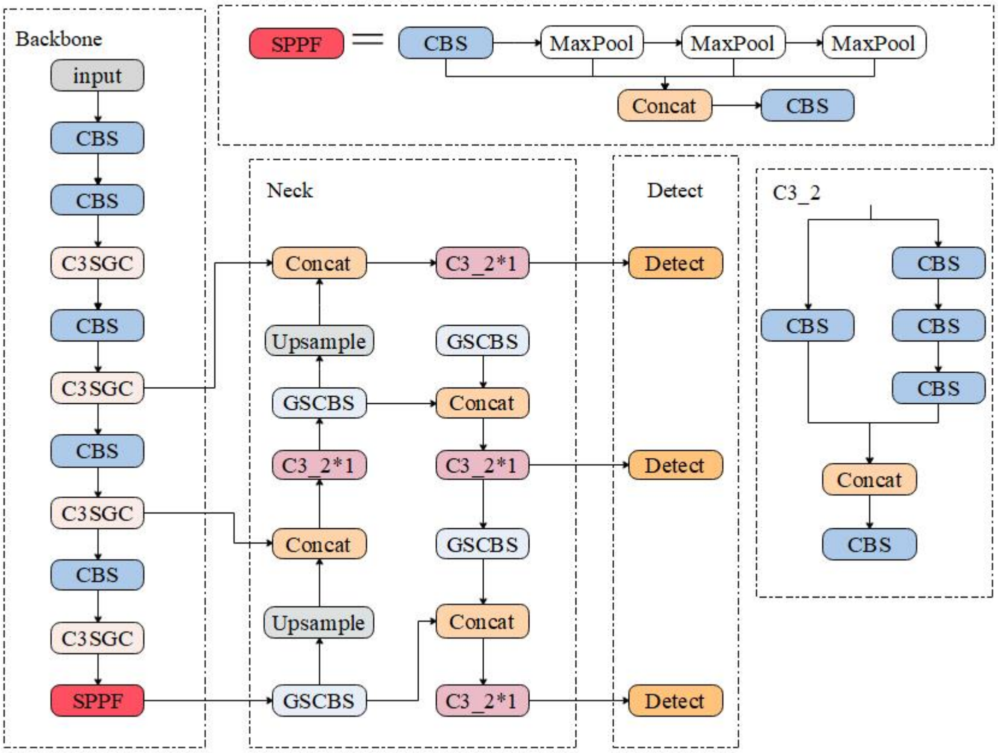


Fig. 1. MAGS-YOLOv5s model

A. Data Augmentation

In the YOLOv5s object detection model, the default data augmentation technique is mosaic augmentation. This technique involves randomly selecting four images from the dataset and then combining them through random scaling, cropping, flipping, and colour space transformations. The corresponding labels are also transformed before training. This approach enriches the dataset and reduces the risk of overfitting. In this paper, we also use mix-up augmentation in addition to mosaic augmentation. This involves selecting two images from the augmented dataset and combining their pixel values through weighted averaging, resulting in the creation of new blended images.

The images are proportionally blended, with each image contributing approximately 50% transparency, and then merged while the classification results are proportionally distributed. Expanding the number of training datasets can also increase the diversity of training samples, avoiding overfitting and improving model performance. It is particularly effective in insufficient or unbalanced training data. The mosaic-augmented and mix-up-augmented images are shown in Fig. 2 respectively.

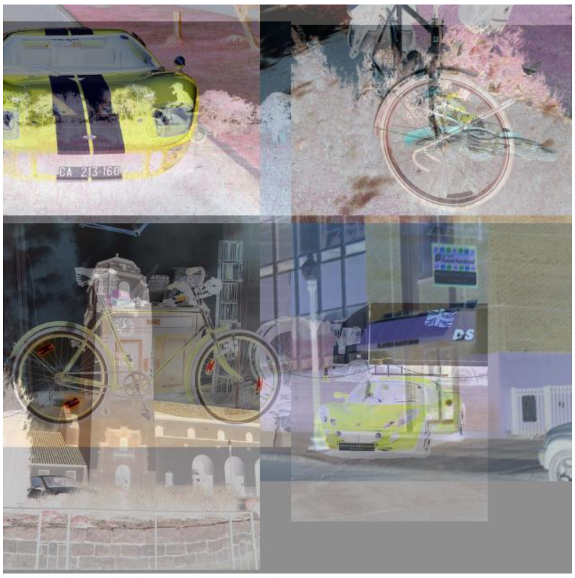


Fig. 2. Mosaic and mixup augmentation.

B. Loss Function

The loss function in the YOLO series consists of three components: confidence loss (Lconf), classification loss (Lcls) and bounding box loss (Lloc). In the YOLOv5s model, the bounding box loss uses the CIoU loss. Compared to the original IoU, CIoU addresses the problem of the gradient being 0 when the predicted box and the ground truth box do not overlap. It also takes into account the distance and aspect ratio between the two boxes, thereby improving the stability of the target box regression. Figure 3 illustrates the principle of CIoU, and the specific calculation formula is as follows:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2}{c^2} + \beta v \tag{1}$$

$$\nu = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{2}$$

$$\beta = \frac{v}{\left(1\text{-}IoU\right) + v} \tag{3}$$

IoU represents the intersection over union between the predicted box and the ground truth box, ρ represents the Euclidean distance between the centres of the two boxes, c is the diagonal length of the minimum bounding rectangle of the two boxes, ν represents the similarity of the aspect ratios between the two boxes, and β is the influence factor of ν. The meanings of the corresponding parameters are illustrated in Fig. 3.

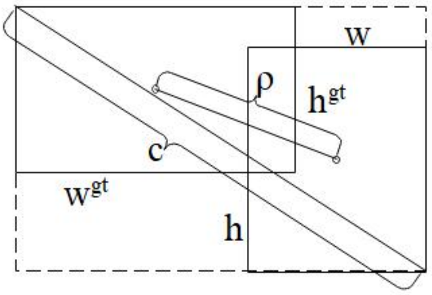


Fig. 3. CIoU

In the above figure, the green box represents the ground truth box, while the blue box represents the predicted box. (w^{gt}, h^{gt}) and (w, h) denote the width and height of the ground truth box and the predicted box, respectively.

To improve the generalisation ability of the model, the alpha-IoU [16] is introduced to improve the CIoU. According to (4), the alpha parameter is added to CIoU, making it alpha-CIoU. The adjustable parameter alpha allows the model to adapt to different datasets while avoiding cross-interference between predicted and ground truth boxes. This approach better handles overlapping objects and speeds up convergence, thereby improving the model's recognition performance. The computational process is shown in (4).

$$L_{\alpha\text{-}CIoU} = 1 - IoU^{\alpha} + \frac{\rho^{2\alpha}}{c^{2\alpha}} + (\beta v)^{\alpha} \tag{4}$$

C. Ghost Shuffle Convolution

Traditional convolution modules can extract rich feature information, but at the cost of increased computational complexity and time consumption. Depthwise Separable Convolution (DSC) [17] divides the convolution operation into two steps to reduce the computational burden: depthwise convolution and pointwise convolution. Depthwise convolution performs convolution on each input channel separately, while pointwise convolution combines the results of depthwise convolution across channels, thereby reducing training time. However, the characteristics of depthwise separable convolution can lead to loss of information when capturing small features that span multiple channels, resulting in reduced accuracy. In addition, due to the simplicity of depthwise separable convolution, more layers are required to improve model performance, which increases the network load. Therefore, in this paper, GSConv [18] is used to replace the standard convolution in the neck part of the original network. The structure of GSConv is shown in Fig. 4.

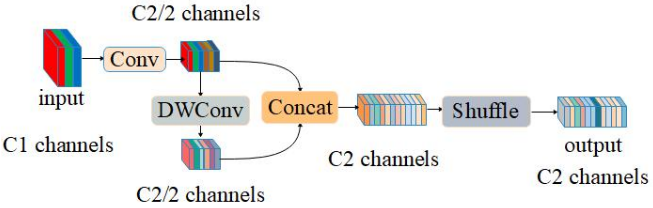


Fig. 4. GSConv model.

GSConv compresses the spatial dimension of feature maps, increasing the number of channels while preserving hidden connections between channels. This allows the model to retain a high level of semantic information, which supports accurate object recognition. In addition, the use of shuffle reduces the computational cost of convolution operations, making them more efficient and faster. However, if used at all stages of the model, the network layers would become deeper, significantly increasing inference time. When these feature maps reach the neck part, their channel dimensions reach their maximum, while the width and height dimensions reach their minimum. Therefore, transformation is no longer necessary and GSConv has the characteristic of having less redundant repetitive information. No compression is required, and the effectiveness of the attention modules is improved. Therefore, GSConv is only used in the neck layer.

D. C3SGC

Long-range dependencies help to improve performance in several visual tasks. Models typically achieve long-range dependencies by stacking convolutional layers. However, this approach can lead to low network efficiency, difficulties in information propagation, and optimisation challenges. To address these issues, this paper proposes the SGC module, which is an improvement based on the SE block and the Global Context (GC) block. The SGC module is added to the C3 module of the backbone network, resulting in C3SGC.

The non-local network uses a self-attention mechanism to obtain long-range dependencies for the network with only one additional layer. By simplifying the non-local module and the SE module [19], the GC block is designed. The GC block aggregates features from all positions to form contextual features and uses a feature transformation module to capture interchannel dependencies. Finally, a fusion module is used to fuse the global context into all positions [20]. The corresponding module structure diagram is shown in Fig. 4.

The input information is convolved to obtain a feature layer with a channel size of 1. The obtained feature layer is then passed through a softmax function to form an attention map. This attention map is multiplied element-wise with the original input. Next, the SE module is applied, which removes the global average pooling. It redistributes the weights of each channel across two fully connected layers, which are then multiplied with the original feature layer, adjusting the importance of each channel. This process produces a global information relationship vector. Finally, two convolutional layers are used to reduce the number of parameters and further extract information. Compared to previous models, the improved SGC block not only has fewer parameters and computational requirements, and captures global and channel dimensional information, resulting in improved accuracy.

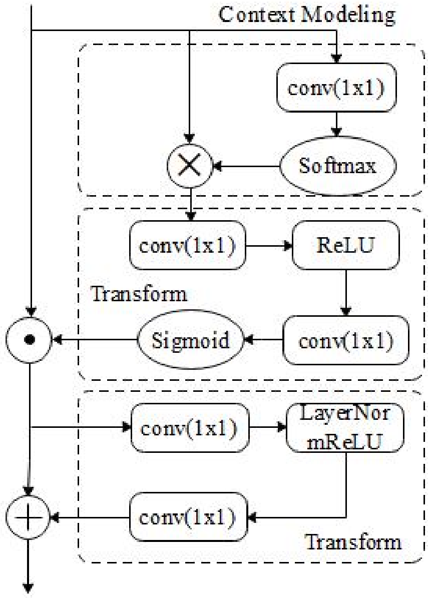


Fig. 5. SGC block.

In this paper, the SGC block is incorporated into the C3 module, which takes the feature maps generated by the convolutional layers as input. It improves the ability to capture global contextual information while maintaining the spatial resolution of the feature maps. This enables the model to improve the accuracy and robustness of object detection, especially in complex scenarios such as occluded or small objects. The structure of the C3SGC model is shown in Fig. 6.

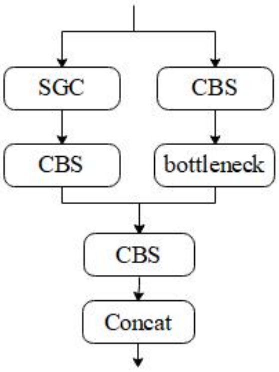


Fig. 6. C3SGC model.

IV. Experiment And Analysis

A. Datasets

The performance of the model was evaluated using the Pascal VOC07+12 and Microsoft COCO datasets. For the Pascal VOC dataset, a total of 3,074 images belonging to five vehicle categories (car, bus, train, bicycle, motorbike) were extracted from the training and validation sets of Pascal VOC 2007 and 2012. The training and validation sets were randomly split in an 8:2 ratio. Additionally, an equal number of images as the validation set were randomly selected from the Pascal VOC 2007 test set to create the test set.

Regarding the Microsoft COCO 2017 dataset, which includes six vehicle categories (bicycle, car, motorcycle, train, bus, truck), 7,000 images were randomly sampled from the training and validation sets. These images were further divided into training, validation, and test sets in an 8:1:1 ratio.

B. Evaluation Index

In the field of computer vision, there are several evaluation indicators or metrics commonly used to assess the performance of computer vision systems. Here are some of the commonly used evaluation indicators in object detection:

Precision quantifies the proportion of true positive predictions out of all positive predictions made by the model.Precision is calculated using the following formula.

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

Recall quantifies the proportion of true positive predictions out of all actual positive instances in the dataset. Recall is calculated using the following formula.

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Mean Average Precision (mAP): mAP extends the average precision (AP) by averaging AP values across multiple object classes or targets. It is commonly used in multi-class object detection tasks, where it provides an overall measure of detection performance across different target categories.

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{7}$$

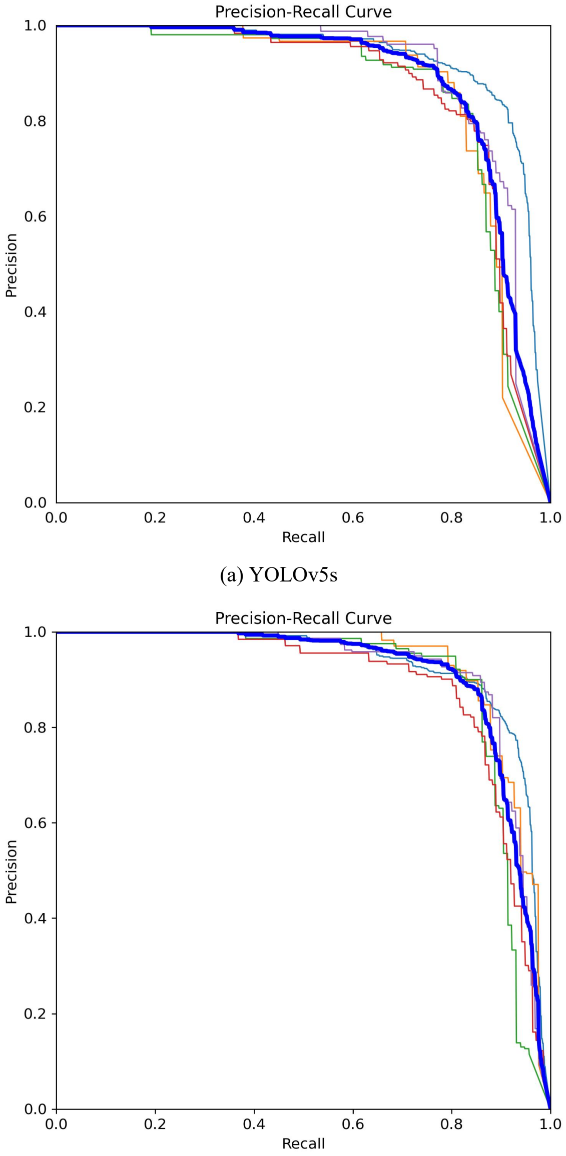
C. Experimental Configuration

In this experiment, CentOS 8.2 served as the chosen operating system, while the graphics card utilized was the Tesla T4, equipped with a remarkable 15110MiB of memory. Python version 3.7.0 was employed, and the training process revolved around the powerful PyTorch 1.8.1 deep learning framework. To maintain uniformity, all training images were resized to a consistent dimension of 640×640.

For the baseline model, the decision was made to employ YOLOv5s with a batch size of 16 and 8 threads to facilitate efficient parallel processing. The learning rate was dynamically adjusted throughout training using the cosine annealing strategy, optimizing the learning process. A comprehensive training regime of 200 epochs was conducted, allowing the model to learn and adapt extensively. As a starting point, pre-trained weights obtained from training on the COCO dataset were utilized, specifically opting for the YOLOv5s.pt weights for initializing the model.

D. Experimental Results

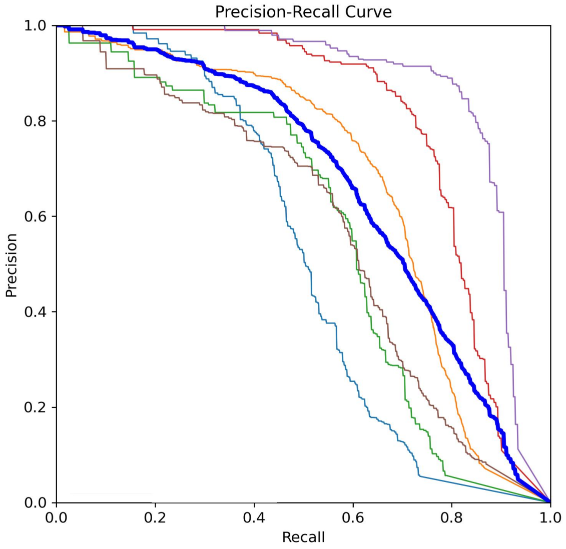
Fig. 7 and Fig. 8 show the results of the PR curves for YOLOv5s and MAGS-YOLOv5s when tested on the PASCAL VOC and MS COCO datasets. The blue line corresponds to the mAP curve at an IoU of 0.5, with the exact numerical values shown in Table I and Table II. Looking at both the figures and the tables, it is clear that the improved model shows accelerated convergence and notable progress in the recognition of the majority of categories. In particular, there is a significant improvement in the recognition of larger object categories such as buses, providing compelling evidence of the effectiveness of the proposed improvements.



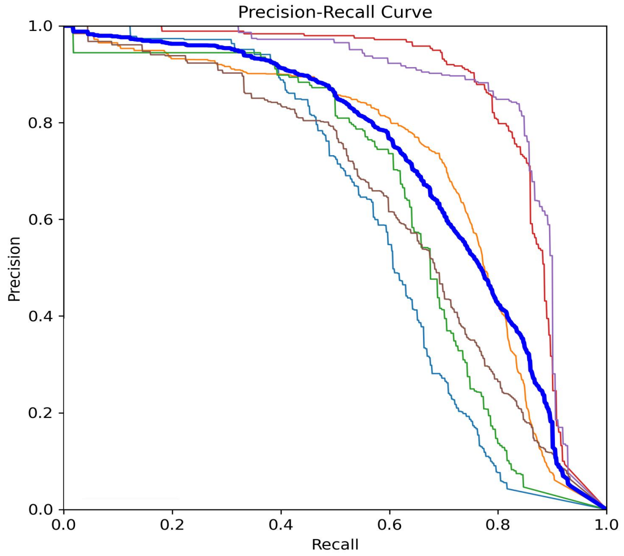
(a) YOLOv5s

(b) MAGS-YOLOv5s

Fig. 7. P-R curve on PASCAL VOC vehicle dataset



(a) YOLOv5s



(b) MAGS-YOLOv5s

Fig. 8. P-R curve on MS COCO vehicle dataset

TABLE I
THE P-R CURVES RESULTS ON PASCAL VOC VEHICLE DATASET

Class	YOLOv5s	MAGS-YOLOv5s
Car	0.926	0.934
Bus	0.865	0.933
Train	0.859	0.894
Bicycle	0.865	0.872
Motorbike	0.894	0.922

TABLE II
THE P-R CURVES RESULTS ON MS COCO VEHICLE DATASET

Class	YOLOv5s	MAGS-YOLOv5s
Bicycle	0.510	0.598
Car	0.670	0.710
Motorcycle	0.559	0.637
Bus	0.791	0.850
Train	0.869	0.852
Truck	0.568	0.637

E. Compare of Test Results

Fig. 9 and Fig. 10 present a comparison of the test results obtained from the Pascal VOC and MS COCO vehicle test sets, respectively. In each dataset, two images were selected for the purpose of comparison: the left side showcases the test result of YOLOv5s, while the right side displays the test result after applying model improvements.

By analyzing Fig. 9 and Fig. 10, it becomes apparent that the enhanced network model has yielded improvements across multiple metrics. Notably, it exhibits enhanced accuracy and recall rates, allowing it to detect smaller targets that went unnoticed by the original network. Furthermore, the model demonstrates improved detection capabilities for occluded targets, leading to a more effective detection performance overall. These enhancements have also enabled the model to better adapt to traffic driving scenarios and other relevant aspects.The detection results from various datasets strongly support the claim that the improved network model exhibits substantial progress and possesses a degree of universality. By consistently showcasing superior performance across different datasets, the model reinforces its credibility and generalizability.

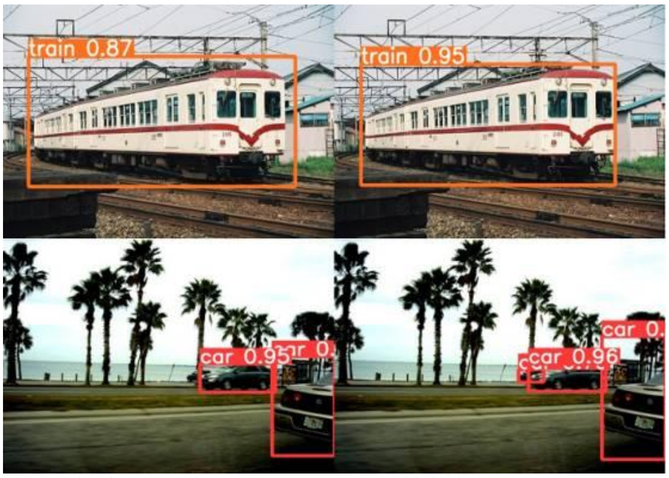


Fig. 9. The detect results on PASCAL VOC vehicle datasets

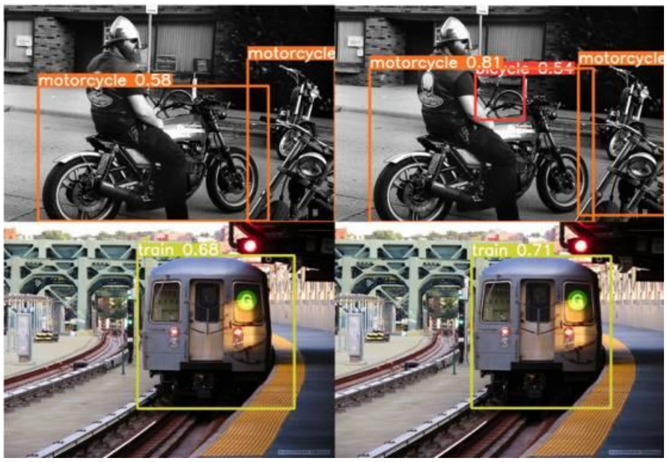


Fig. 10. The detect results on MS COCO vehicle datasets

F. Ablation Experiment

In order to validate the improvements to the YOLOv5s model proposed in this paper, a series of ablation experiments were carried out on two different datasets. The experimental results for the Pascal VOC and MS COCO vehicle datasets are presented in Table III respectively. Analysis of these tables shows that the refined network model showed significant improvements over the original model, resulting in an increase in mAP of 2.9% and 5.3% for the Pascal VOC and MS COCO datasets respectively. These results indicate that each modification resulted in a consistent increase in the mAP values of the network, highlighting the effectiveness of the proposed improvements.

TABLE III
ABLATION EXPERIMENTS RESULTS ON TWO VEHICLE DATESET

mixup	α-CIoU	GSConv	C3SGC	mAP on PASCAL VOC	mAP on MS COCO
×	×	×	×	88.2%	66.1%
√	×	×	×	89.8%	69.4%
√	√	×	×	90.7%	69.8%
√	√	√	×	91.0%	70.1%
√	√	√	√	91.1%	71.4%

G. Compare with Other Models

Table IV puts forward a performance comparison of the improved MAGS-YOLOv5s model with other models. Examination of the table shows that the improved model outperforms its counterparts. In the Pascal VOC dataset, MAGS-YOLOv5s shows an increase of 2.9%, 6.3% and 2.3% over YOLOv5s, YOLOv4-tiny and YOLOv5l respectively. Similarly, in the MS COCO dataset, MAGS-YOLOv5s achieves improvements of 5.3%, 8% and 3.3% over the same models. Notably, this model exhibits reduced computational complexity and lower equipment requirements compared to YOLOv5s. Furthermore, despite the improved model's increase in mAP, the inference time only increases slightly. Overall, the improved model has advantages in both accuracy and speed over alternative models.

TABLE IV
COMPARING WITH OTHER MODELS

Model Name	FLOPs (G)	mAP(%) Pascal VOC	mAP(%) MS COCO	Inference time(ms)
YOLOv5s	15.8	88.2	66.1	6.3
YOLOv4-tiny	13.3	84.8	63.4	5.8
YOLOv5l	107.7	88.8	68.1	25.3
MAGS-YOLOv5s	15.2	91.1	71.4	7.7

V. CONCLUSION

This section introduces the MAGS-YOLOv5s algorithm, a significant enhancement for vehicle detection based on the YOLOv5s model.

First, An extensive enhancement process is carried out on the dataset using the mixup technique. And the introduction of the α parameter further improves the adaptability of the model. This improvement is remarkable because it doesn't affect the model's computational capabilities. Another key improvement involves introducing GSConv to replace a previous component, which effectively reduces both the model's parameters and its computational complexity. This reduction significantly streamlines the model, making it more efficient and easier to use. To assess the performance of the MAGS-YOLOv5s model, extensive evaluations are performed using the widely accepted Mean Average Precision (mAP) metric. These evaluations include rigorous experimentation with two benchmark datasets, Pascal VOC and MS COCO, specifically designed for vehicle detection tasks. Furthermore, we conduct ablation experiments are performed to validate the effectiveness of the proposed enhancements, confirming their positive impact on the model's performance. The comparative analysis with other existing models further demonstrates the superiority of the proposed MAGS-YOLOv5s model. Especially, the improved model exhibits extraordinary generalization ability and can consistently perform well in different scenarios.

In summary, the MAGS-YOLOv5s algorithm represents a significant advancement in vehicle detection by harnessing the power of the YOLOv5s model. Dataset enhancement, adaptive parameter introduction, model simplification, and context awareness have greatly improved the algorithm's accuracy, performance, and universality.

REFERENCES

[1] Zehang Sun, G. Bebis and R. Miller, "On-road vehicle detection: a review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 5, pp. 694-711, May 2006.

[2] Binbin Sun, Wentao Li, Huibin Liu, Jinghao Yan, Song Gao, and Penghang Feng, "Obstacle Detection of Intelligent Vehicle Based on Fusion of Lidar and Machine Vision." *Engineering Letters*, vol. 29, no.2, 2021, pp. 722-730.

[3] S. Sivaraman and M. M. Trivedi, "Looking at Vehicles on the Road: A Survey of Vision-Based Vehicle Detection, Tracking, and Behavior Analysis," *IEEE Trans. Intelligent Transportation Systems*, vol. 14, no. 4, pp. 1773-1795, Dec. 2013.

[4] S. Tuermer, F. Kurz, P. Reinartz and U. Stilla, "Airborne Vehicle Detection in Dense Urban Areas Using HoG Features and Disparity Maps," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 6, pp. 2327-2337, Dec. 2013.

[5] Z. Yunzhou, S. Pengfei, L. Jifan and M. Lei, "Real-time vehicle detection in highway based on improved Adaboost and image segmentation," *IEEE International Conference on Cyber Technology in Automation*, Control, and Intelligent Systems, Shenyang, China, 2015, pp. 2006-2011.

[6] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580-587.

[7] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[8] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 779-788.

[9] Redmon, Joseph, and Ali Farhadi, "YOLO9000: better, faster, stronger," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.* 2017, pp. 6517-6525.

[10] Redmon, Joseph, and Ali Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv: 1804.02767*(2018).

[11] Bochkovskiy A, Wang C Y and Liao H Y M, "YOLOv4: optimal speed and accuracy of object detection," *arXiv preprint arXiv*:2004.10934(2020).

[12] Songjiang Li, Lanlan Geng and Peng Wang, "Vehicle Target Detection Based on Improved Yolov4." *Computer Engineering*. Vol. 49, no.04, pp. 272-280, 2023.

[13] Zhi Jin, Qian Zhang and Xiying Li, "Dense Road Vehicle Detection Based on Lightweight ConvLSTM." *Computer Engineering and Applications.* Vlo. 59, no.08, pp. 89-96, 2023.

[14] Liyan Xiong, Suocheng Tu, Xiaohui Huang, Junying Yu, et al. "Vehicle detection method based on MobileVit lightweight network." *Application Research of Computers*. Vol. 39, no.08, pp. 2545-2549, 2022.

[15] C. -Y. Wang, A. Bochkovskiy and H. -Y. M. Liao, "Scaled-YOLOv4: Scaling Cross Stage Partial Network," *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 13024-13033.

[16] Jiabo He, Sarah Erfani, Xingjun Ma, James Bailey, Ying Chi and Xian-Sheng Hua, "α-IoU: A family of power intersection over union losses for bounding box regression". Proceedings of the Conference and Workshop on Neural Information Processing Systems, 2021, pp. 1-10.

[17] Laurent Sifre and Stéphane Mallat, "Rigid-Motion Scattering for Texture Classification," *arXiv preprint arXiv*::1403.1687,(2014).

[18] Hulin Li, Jun Li, Hanbing Wei, Zheng Liu, Zhenfei Zhan and Qiliang Ren, "Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles," *arXiv preprint arXiv*:2206.02424(2022).

[19] Hu jie, Li Shen, and Gang Sun, "Squeeze-and-excitation networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132-7141.

[20] Y. Cao, J. Xu, S. Lin, F. Wei and H. Hu, "GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond," *IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1971-1980.