ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

A Lightweight Improved YOLOv5s Model for Vehicle Detection

Si Shen[2,*], Shimin Qu[3], Xingfeng Wang[1,3], Wei Zheng[3], Ziqi Wang[1,2]

[1] College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China

[2] Huzhou Research Institute of Zhejiang University, Huzhou, Zhejiang 313000, China

[3] Inner Mongolia Huomei Hongjun Aluminum Electric Co., Inner Mongolia 029200, China

*2022388721@stu.zjhu.edu.cn

Abstract

With the rapid development of computer vision and deep learning methods, object detection has emerged as a hot research topic in fields such as autonomous driving, robot navigation, and intelligent transportation. To address the issues of numerous parameters and slow detection speed associated with traditional vehicle target detection algorithms, this article introduces a lightweight YOLOv5s-based algorithm, which emphasizes its efficiency and accuracy in real-time scenarios. First, a lightweight GhostNet is incorporated for feature extraction. Subsequently, the streamlined MBConv layers are utilized to replace the traditional convolutional layers in the backbone, thus achieving a more efficient and less complex model. Finally, by integrating the CBAM attention mechanism into the neck network of YOLOv5s, the image feature information is fully utilized, thereby enhancing detection accuracy. Experimental results show that the average accuracy of this algorithm reaches 91.5%, with a 30.9% reduction in parameters. Clearly, the method proposed in this paper fulfills the real-time requirements for vehicle detection.

Keywords

Object Detection; YOLOv5s; Lightweight Network; GhostNet; CBAM.

1. Introduction

In the era of rapid economic development, the number of motor vehicles in China have shown a rapid growth trend, making traffic safety a major concern across various sectors. According to relevant traffic accident data from both domestic and international sources, traffic accidents caused by vision blind spots, slow reaction times, inattention, and other factors account for 87% of all incidents. If vehicles can accurately and quickly detect all targets on the road, many accidents could be prevented. Therefore, it is crucial for vehicles to conduct real-time monitoring of their surroundings and achieve efficient identification of vehicles and pedestrians.

Traditional object detection methods mainly rely on manually selected features for extraction, which performs well in simple scenes but are limited when dealing with challenges such as complex scenes, lighting changes, and object occlusions [1]. In recent years, deep learning-based object detection methods have demonstrated excellent recognition capabilities in complex scenes by automatically learning features from training data. These methods include RCNN [2], Fast R-CNN [3], Faster R-CNN [4], SSD [5], and YOLO [6-9], among others. Among them, YOLO-based deep models are favored by many researchers for their high accuracy and fast detection speed.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

Song Zhang et al. [10] proposed an improved YOLOv5 for pepper harvesting in complex field environments. In their approach, the CSPDarknet53 backbone was replaced with GhostNet, and the CBL module in the neck was substituted with Ghost convolution. This resulted in a 46.6% reduction in model size, achieving a lightweight network model with only a minor loss in accuracy. Jianwei Zhao et al. [11] developed the YOLOv5-SC3FB model to facilitate landmark detection and avoidance tasks for mobile robots. By incorporating ShuffleNetV2, they reduced the number of network parameters and computations, minimizing the model size. They also constructed the C3Faster module to enhance detection speed in dynamic scenarios and employed a simplified BiFPN structure for feature fusion. This reduced the model size to 0.72M with a parameter volume of 0.19M, doubling the detection speed and enabling deployment on mobile robot platforms. Zihao Xu et al. [12] presented the YOLOv7x-CM model tailored for complex driving scenarios. By integrating the CBAM attention mechanism and the MPDiou loss function, they enhanced the efficiency and accuracy of object detection. Results showed improvements of 5.3% and 6.8% in mean accuracy on the KITTI and BDD100K datasets, respectively, along with a 35.4% increase in FPS. Qi Zhang et al. [13] tackled the challenge of balancing positioning accuracy and computational complexity in traditional VSLAM (Visual Simultaneous Localization and Mapping) systems by adopting the lightweight GhostNet module as the backbone of the YOLOv5s object detection network and introducing an attention mechanism. This approach allowed for the continuous and accurate capture of dynamic factors, enhancing both real-time performance and accuracy. Currently, vehicle detection still faces the dilemma of balancing detection accuracy against model parameters and computation speed.

To address the issues of low efficiency in vehicle and person detection, we propose an improved strategy based on the YOLOv5s model. This strategy meets the requirements for precise positioning and real-time detection. We have made the following main contributions:

1) This paper proposes an improved GhostNet as the backbone network. Compared to the CSP-Darknet53 used in the YOLOv5s network, our model reduces the number of parameters by 30.9% without sacrificing detection accuracy.

2) MBConv is used to replace traditional convolutional layers in the neck of the model, significantly reducing the number of parameters and improving the model's speed.

3) The introduction of the CBAM attention mechanism into the neck network enhances the network's ability to focus on important features of image targets, leading to improved detection accuracy.

2. YOLOv5s

YOLOv5 is subdivided into four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. These versions share the same overall structure but differ in the depth and width of the network, resulting in an increasing trend in detection accuracy and performance. However, as the model size and complexity increase, the detection speed decreases accordingly. Considering the requirements for both detection speed and accuracy in this paper, the YOLOv5s model is selected as the fundamental framework for object detection. Figure 1 illustrates the network structure of this model, which consists of four parts: Input, Backbone, Neck, and Head.

The input of YOLOv5s incorporates mosaic data augmentation, adaptive anchor box calculation, and adaptive image scaling. The backbone network extracts various levels of feature information from the target image. This network includes key components such as Convolutional (Conv) modules, C3 modules, and Spatial Pyramid Pooling Fast (SPPF) modules (Fig. 1). The C3 module utilizes residual features from the Cross Stage Partial Network (CSPNet), splitting the original input into two branches. These branches are recombined, maintaining consistent input and output channel counts. This method reduces the number of parameters while preserving accuracy. Convolution operations halve the channel count, and branch one undergoes a Bottleneck×N operation. Finally, the channels from both branches are stacked, ensuring that the input and output channel numbers remain unchanged, thus facilitating parameter reduction without compromising accuracy. The SPPF module uses pooling layers of different sizes to downsample the image, increasing the network's receptive field and fusing

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

features from various scales of the same feature map. This addresses the issue of multi-scale target transformations.

The Neck network employs a Feature Pyramid Network (FPN) to propagate high-level semantic information features top-down and a Path Aggregation Network (PAN) to transmits localization information features bottom-up [14]. By merging these semantic and localization features, the Neck network not only captures information from different layers of the backbone network but also enhances the detection of dense targets.

The head network is the output layer of YOLOv5s, where each output layer comprises a convolutional layer and a fully connected layer. It is used to predict multi-scale and multi-class targets, directly deriving the location, category, and confidence level information of objects in the image, thus completing the output of object detection results.

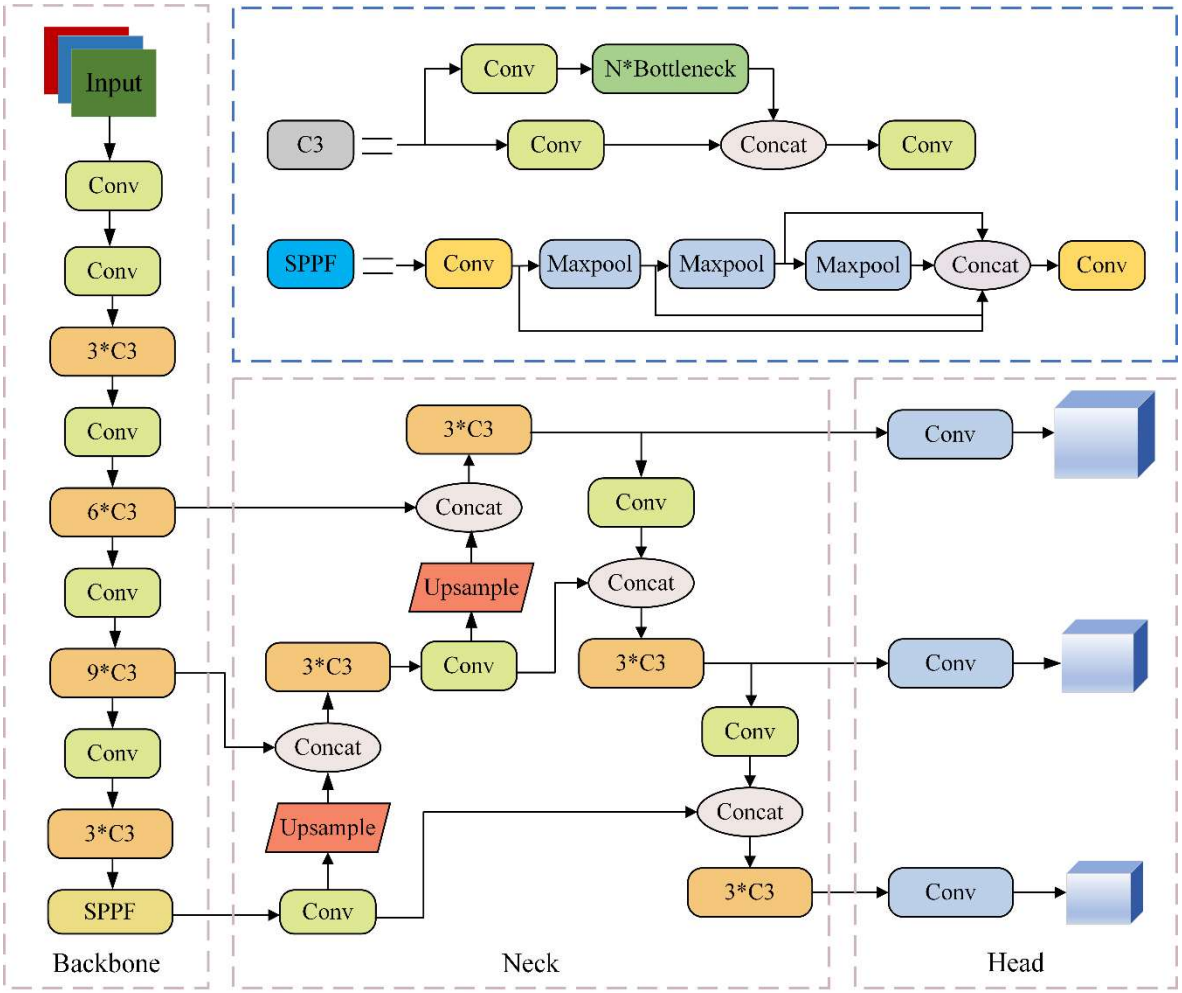

Figure 1. YOLOv5s network structure

3. Proposed Method

3.1 Backbone Network Improvements

Han, Kai et al. [15] proposed a novel network architecture called GhostNet to reduce redundancy in feature maps and achieve a lightweight network. The core idea of GhostNet is the staged convolutional computation module, known as the Ghost module, as illustrated in Figure 2. The Ghost module integrates standard convolution into two steps: the first step performs a small number of convolutional kernel operations on the input feature map to reduce the number of channels, obtaining a feature condensation. The second step utilizes simple linear operations to generate more feature maps through layer-wise convolutions. The compressed feature map is then concatenated directly

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

with the extracted feature map to obtain the complete convolutional output. Compared to ordinary convolution, the GhostConv module reduces the computational cost and the number of parameters by a factor of 1/S [16]. This algorithm effectively decreases the network's parameters and computations without altering the output feature map size, thus achieving model lightweighting.

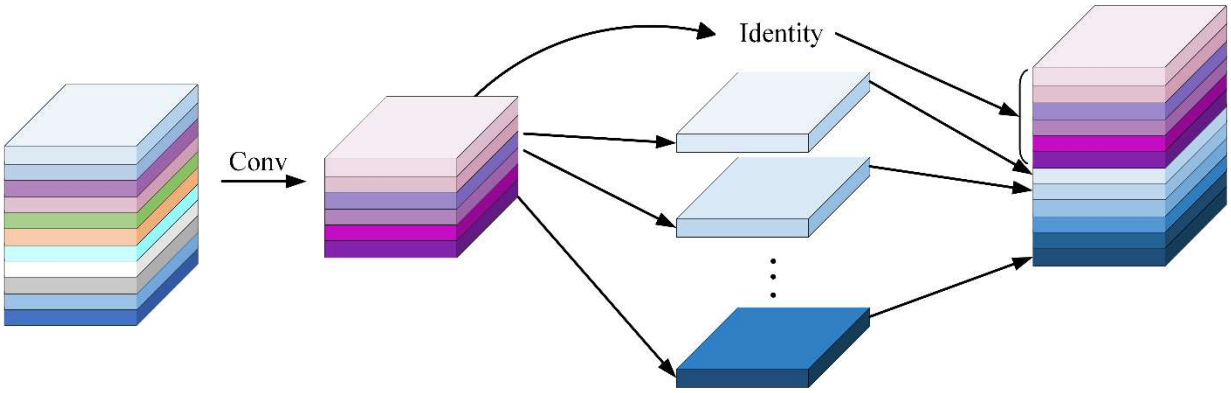

Figure 2. Module diagram of GhostConv

GhostNet also introduces the Ghost bottleneck structure, which consists of two modules, as shown in Figure 3. The first Ghost bottleneck structure stacks two Ghost modules with a stride of one in the main branch, serving as an expansion layer to increase the network's width. The second Ghost bottleneck structure connects two Ghost modules using a depthwise convolution with a stride of two in the main branch. In this paper, the Ghost bottleneck is used to replace the standard residual modules in the C3 module, forming a new C3Ghost module. By reducing most of the original 3×3 traditional convolutions in the structure, resulting in a compressed model with lower computational cost. The structure of the C3Ghost module is illustrated in Figure 3.

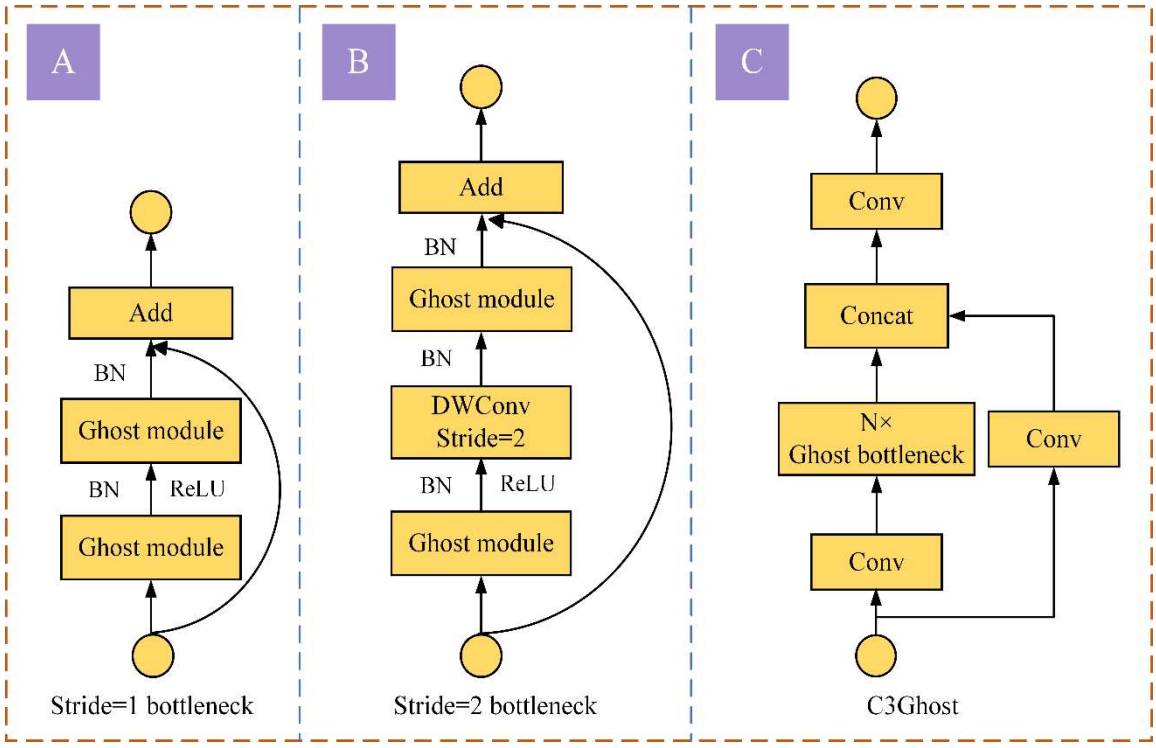

Figure 3. The GhostNet module structure diagram is built in this study. A) Ghost bottleneck module (stride =1); B) Ghost bottleneck module (stride =2); C) C3Ghost module

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

The model design in this paper aims to achieve lightweight and high-precision object detection. Taking the YOLOv5s network as the basic framework, it integrates the C3Ghost module with Ghost bottleneck and the Ghost convolution module to form a new feature extraction network. In the neck network, the MBConv module replaces regular convolutions, and the CBAM attention module is inserted before the detection head to enhance small object detection performance. The improved network structure is illustrated in Figure 4.

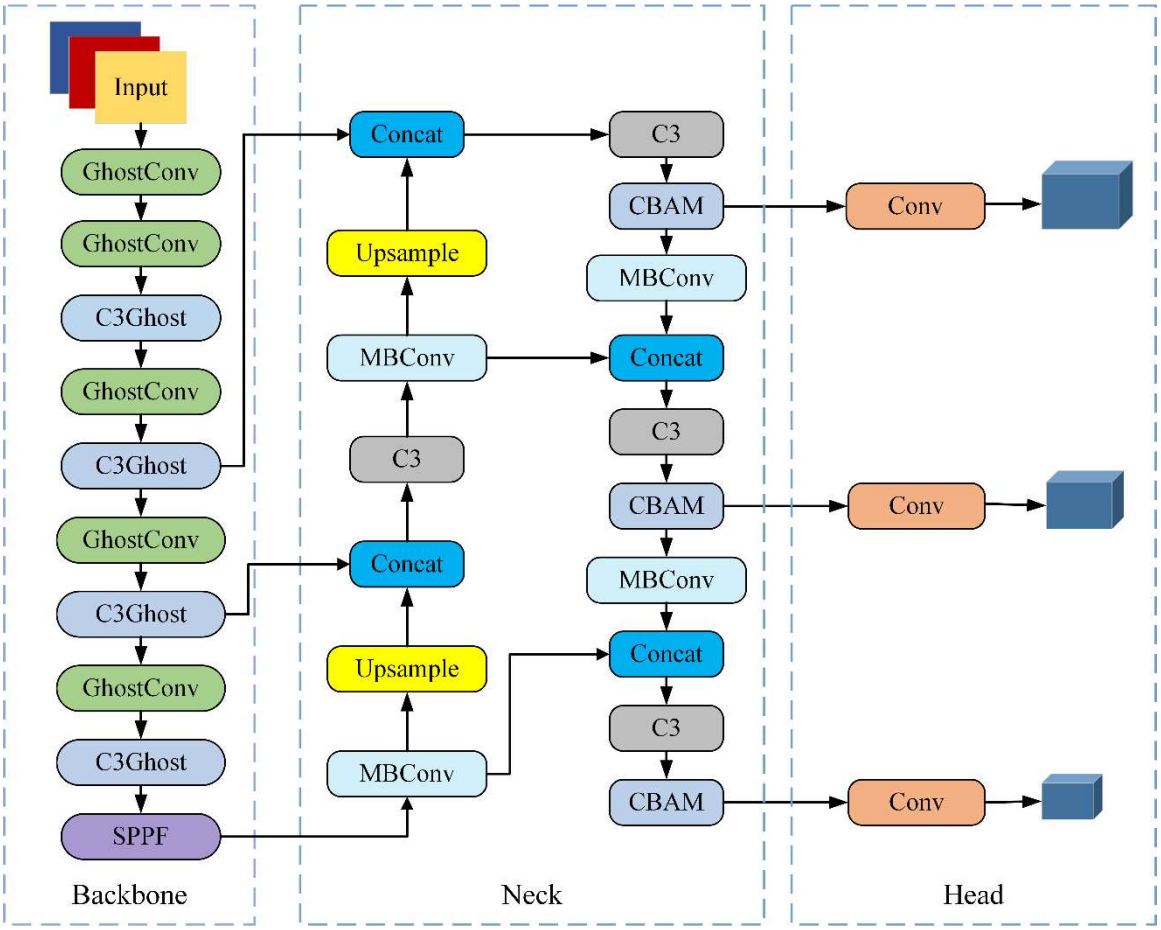

Figure 4. Improved YOLOv5s structure diagram

3.2 Neck Network Improvements

The network structure of EfficientNetV2 primarily utilizes the MBConv module and introduces a new Fused-MBConv module in the shallower layers of the network. MBConv increases the number of channels using a 1x1 convolution in the main branch, and then applies 3x3 depthwise separated convolutions in high dimensions. The computational cost of depthwise separable convolution is approximately one-ninth of a regular convolution, significantly reducing computation [17]. The SE attention mechanism is then used to optimize the feature maps, followed by a 1×1 convolution to reduce the number of channels. The Fused-MBConv module replaces the 3×3 depthwise separable convolution and the expanded 1×1 convolution in MBConv with a single regular 3×3 convolution [18]. This modification significantly improves training speed in the shallower layers of the model but is not suitable for deeper layers. The structures of depthwise separable convolution, MBConv, and Fused-MBConv are illustrated in Figure 5. Replacing regular convolutions in the neck network with MBConv can significantly reduce the number of parameters in the network, thus improving the model's speed.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

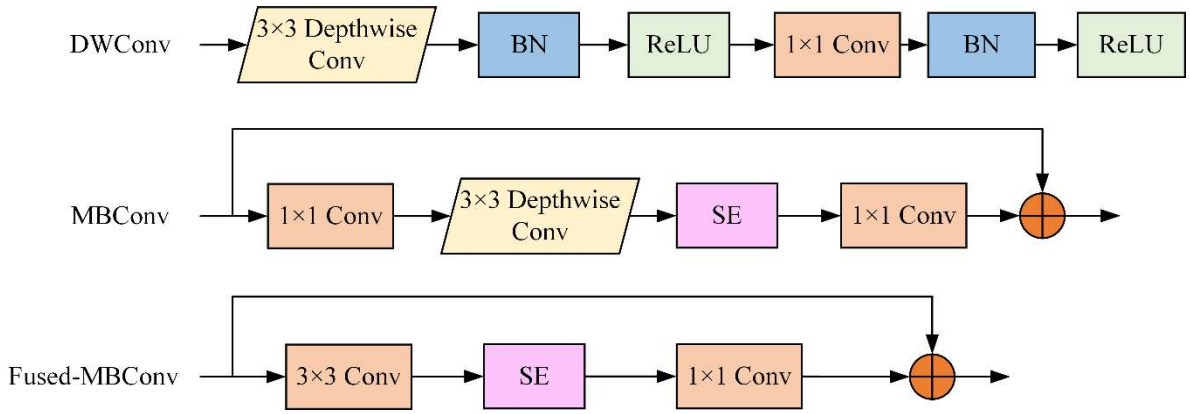

Figure 5. Architecture diagram of convolutional layer

Moreover, the attention mechanism allows the model to selectively focus on important aspects, enhancing feature extraction. The Convolutional Block Attention Module (CBAM) [19] is a simple yet effective feed-forward convolutional neural network that leverages the attention mechanism to "concatenate" the channel attention module (CAM) and the spatial attention module (SAM). The channel attention module focuses on the relationships between different channels, assigning higher weights to key information channels and suppressing irrelevant ones. The spatial attention module instructs the neural network to focus more on critical regions or positions in the feature maps. As illustrated in Figure 6, we placed the CBAM modules after the outputs of three feature maps with different scales. This placement further optimizes the feature fusion layer and the feature scales of the multi-scale detection layer, allowing the trained model to better adapt to small object detection.

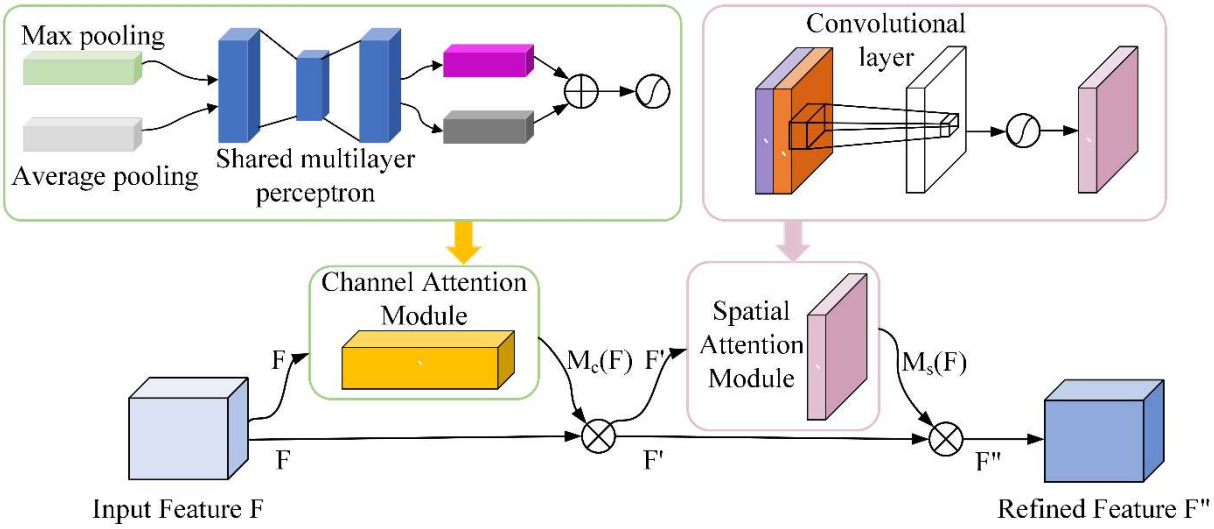

Figure 6. CBAM attention mechanism module

4. Experimental Process and Analysis

4.1 Experimental Environment and Dataset

The experimental dataset consists of 6147 publicly available images of vehicles, categorized into two classes: people and vehicles. With a ratio of 8:1:1, the dataset is divided into training, test, and validation sets. Models were trained using adaptive moment estimation (Adam), with input images of 640×640 pixels, a batch size of 24, a learning rate of 0.001, and a momentum factor of 0.9. As shown in Table 1, the experiments were conducted in the same environment.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

Table 1. Experimental environment configuration

Hardware	Configuration	Tool	Version
OS	Ubuntu 20.04	Python	3.9.0
RAM	32GB	Pytorch	2.3.0
CPU	Intel Core i9-10900KF	Pycharm	2022.3.3
GPU	RTX 3090(24G)	CUDA	11.8

4.2 Experimental Environment and Dataset

4.2.1 Evaluation Index

We use precision (P), recall (R), and mean Average Precision (mAP) as evaluation metrics for model accuracy. Additionally, we assess the model's complexity using the number of parameters, Frames Per Second (FPS), and model size. The calculation formulas for each metric are shown in equations (1) to (3):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$mAP = \frac{1}{classes} \sum_{i=1}^{classes} \int_{0}^{1} P(R)d(R) \tag{3}$$

In the formulas, TP is the number of true positive samples correctly predicted by the model, FP is the number of false positives incorrectly predicted as positive, and FN is the number of false negative samples incorrectly predicted as negative. The Mean Average Precision (mAP) is the average of all precision scores.

4.2.2 Ablation Experiment in Backbone Network

To validate the effectiveness of the improvements made to each module, ablation experiments were conducted, comparing different enhancement methods across various network layers.

The backbone network employs a relatively complex C3 network structure. Repeated use of C3 modules can generate many redundant features during the critical feature extraction process, leading to high computational costs and a large number of parameters in the entire network model. By adopting lightweight networks such as GhostNet [15], ShuffleNetV2 [20], and MobileNetV3 [21] as the foundational structures for the backbone network, we can maintain performance while reducing computational costs. In this study, we compare GhostNet with MobileNetV3 and ShuffleNetV2. The experimental results, shown in Table 2, demonstrate that GhostNet reduces the number of parameters and the model size, ultimately achieving a top accuracy of 91.5%.

Table 2. Ablation experiments of the backbone network

Model	Precision	Recall	mAP@0.5	Parameters	Model size(MB)
X+Shuffnetv2	0.855	0.882	0.889	2946249	6.3
X+MobileNetV3	0.865	0.885	0.894	3294983	7.1
X+GhostNet	0.887	0.913	0.915	4847597	10.2

[1] Note: X is the part of the improved model after removing the backbone network.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

4.2.3 Ablation Experiment in Neck Network

In the neck network of YOLOv5s, we adopt efficient convolutional modules to replace the ordinary convolutions, aiming to improve the model's inference speed. This study compares three lightweight convolutions methods: Depthwise Separable Convolution (DWConv) [22], Grouped Spatial Convolution (GSConv) [23], and Ghost Convolution (GhostConv). The experimental results are shown in Table 3. Although the GhostConv module has a higher number of parameters compared to the DWConv and GSConv modules, it achieves a higher precision of 1.1% and 0.5% respectively. The average detection accuracy is also improved by 1.1% and 0.9%.

Table 3. Ablation experiments of the neck network

Model	Precision	Recall	mAP@0.5	Parameters	Model size(MB)
X+DWConv	0.876	0.904	0.904	4184269	7.0
X+GSConv	0.882	0.908	0.906	4642797	9.8
X+GhostConv	0.887	0.913	0.915	4847597	10.2

[2] Note: X is the part of the improved model after removing the neck network.

4.2.4 Ablation Experiment in Attention Module

Attention mechanisms excel at capturing key features while ignoring secondary ones. In this study, we introduced and compared four attention mechanisms to enhance image target information: Squeeze-and-Excitation networks (SE) [24], Coordinate Attention networks (CA) [25], Efficient Multi-scale Attention(EMA) [26] and Convolutional Block Attention Module (CBAM). We placed these attention modules after the output of three different scale feature maps. The experimental results, shown in Table 4, indicate that with nearly equal model size and parameters, CBAM achieves the highest detection accuracy. It surpasses the SE attention module by 0.7%, the CA attention module by 0.9%, and the EMA attention module by 1.1%.

Table 4. Ablation experiments of the attention module

Model	Precision	Recall	mAP@0.5	Parameters	Model size(MB)
X+SE	0.873	0.915	0.908	4885927	10.3
X+CA	0.878	0.915	0.906	4878599	10.3
X+EMA	0.875	0.904	0.904	4846359	10.2
X+CBAM	0.887	0.913	0.915	4847597	10.2

[3] Note: X is the part of the improved model after removing the attention module.

4.2.5 The Comprehensive Ablation Experiments of the Improved YOLOv5s Model

To validate the influence of the improvement strategies on the original model, we conducted comprehensive ablation experiments. First, YOLOv5s introduces the lightweight GhostNet to reconstruct the feature extraction network (Model 1), reducing the model size by 3.8M. Next, the MBConv module replaces the Conv module in the feature fusion network (Model 2), While the number of parameters is reduced, the accuracy improves by 0.1%. Finally, the CBAM attention mechanism is introduced to enhance feature fusion, improving accuracy by 1%, and increasing average detection accuracy by 0.8% (as shown in Table 5). These improvements achieve an optimal balance between model size and detection accuracy. Compared to the original YOLOv5s model, our enhanced model reduces the number of parameters by 30.9%, decreases the model size by 29.2%, accelerated FPS detection speed, and achieved minor improvements in precision, recall, and mean average precision for detection.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

Table 5. Ablation experiments of the improved YOLOv5s module

Model	Precision	Recall	mAP@0.5	Parameters	Model size(MB)	FPS
YOLOv5s	0.881	0.908	0.908	7015519	14.4	83.6
Model1	0.876	0.912	0.907	5081671	10.6	94.7
Model2	0.877	0.906	0.907	4842919	10.2	103.0
Ours	0.887	0.913	0.915	4847597	10.2	103.5

4.2.6 Performance Comparison of Six Object Detection Networks

To further validate the performance of the proposed method, a comparative analysis was conducted with lightweight networks, including YOLOv3-tiny and YOLOv7-tiny. Figures 7 to 9 illustrate the comparisons of different algorithm models in terms of mean Average Precision (mAP), the variation of bounding box loss training curves, and real-time detection speed, respectively. The training results for each model are summarized in Table 6.

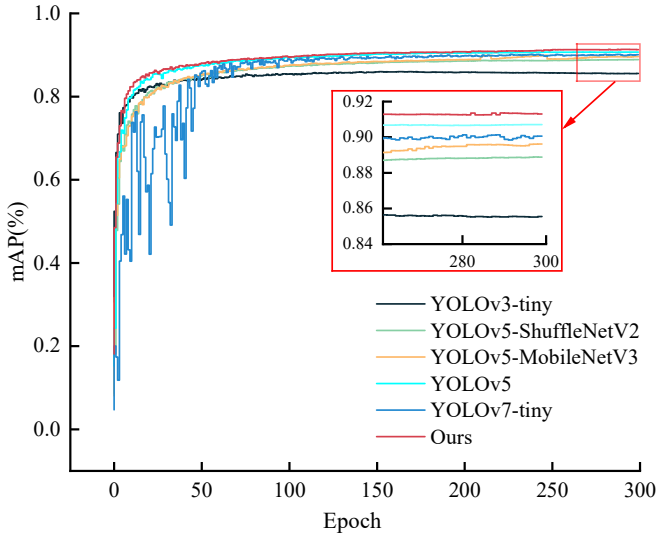

Figure 7. mAP variation curves

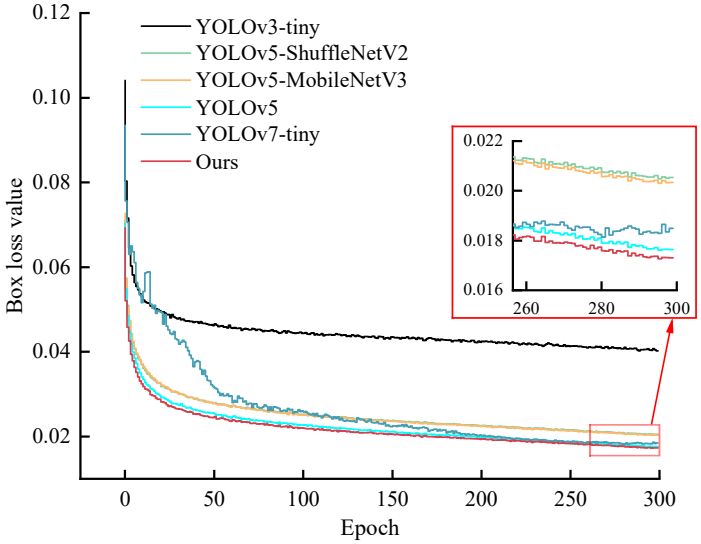

Figure 8. Boundary box loss curves

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

In Figure 7, all models show a steady increase in average detection accuracy after training for 100 epochs. After 300 epochs of training, the improved model converges the fastest and achieves the highest average detection accuracy compared to YOLOv3-tiny, YOLOv5-ShuffleNetV2, YOLOv5-MobileNetV3, YOLOv5s, and YOLOv7-tiny.

In the boundary box loss curve, the improved network shows a rapid decline within the first 75 epochs of training and stabilizes after 150 epochs. Throughout the entire training phase, the loss curve of our model consistently remains lower than that of the other models.

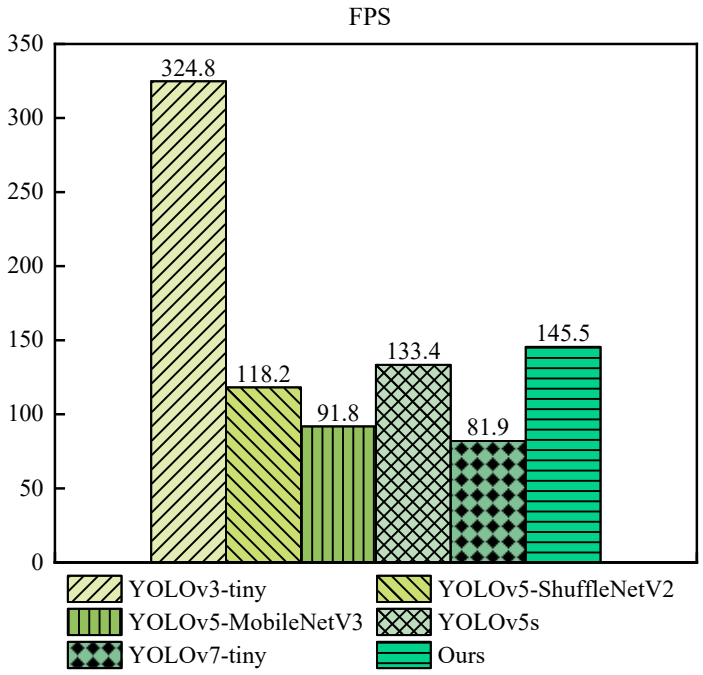

Figure 9. Real-time speed comparison detection

Figure 9 shows that while YOLOv3-tiny achieves the highest real-time detection speed at 324.8 FPS, its average detection accuracy is only 85.7% (as evident from Figure 7). Our model, with a detection speed of approximately 145.5 FPS, surpasses other advanced models in terms of both speed and accuracy. Compared to the original YOLOv5s algorithm, the proposed algorithm in this paper offers a slightly faster detection speed, demonstrating its efficiency and effectiveness.

Table 6. The experimental results between the improved YOLOv5s model and other models

Model	Precision	Recall	mAP@0.5	Parameters	Model size(MB)	FPS
YOLOv3-tiny	0.858	0.82	0.857	8669002	17.4	324.8
YOLOv5-ShuffleNetV2	0.855	0.882	0.889	2946249	6.3	118.2
YOLOv5-MobileNetV3	0.865	0.885	0.894	3294983	7.1	91.8
YOLOv5s	0.881	0.908	0.908	7015519	14.4	143.4
YOLOv7-tiny	0.872	0.901	0.902	6009343	11.7	81.9
Ours	0.887	0.913	0.915	4847597	10.2	124.5

As shown in Table 6, the improved algorithm demonstrates exceptional performance, achieving an average detection accuracy of 91.5%. This represents notable improvements of 5.8%, 2.6%, 2.1%,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

0.7%, and 1.3% respectively over YOLOv3-tiny, YOLOv5-ShuffleNetV2, YOLOv5-MobileNetV3, YOLOv5s, and YOLOv7-tiny, respectively. Its detection speed has slightly increased compared to YOLOv5s, trailing only behind YOLOv3-tiny. Regarding model size, it is 41.3%, 29.2%, and 12.8% smaller than YOLOv3-tiny, YOLOv5s, and YOLOv7-tiny, respectively. Although the model is larger than YOLOv5-ShuffleNetV2 and YOLOv5-MobileNetV3, it offers a clear advantage in accuracy. Overall, the proposed method achieves an excellent balance between lightweight design, real-time performance, and detection accuracy, resulting in the best overall performance compared to the other evaluated algorithms.

5. Conclusion

This study proposes a lightweight YOLOv5s model for vehicle detection. The GhostNet replaces the original YOLOv5s backbone, while the Neck network utilizes mobile inverted bottleneck convolutions to reduce model size and computational costs. Additionally, the CBAM attention mechanism is introduced before the detection head to enhance feature information. The Focal-EIOU loss function is adopted to optimize the loss function calculation, enabling the bounding box regression process to focus more on high-quality anchor boxes, resulting in better detection performance compared to other loss functions. The improved network achieves an average detection accuracy of 91.5%, with a 0.7% increase in mAP over YOLOv5s, a 30.9% reduction in parameter count, and a 29.2% decrease in model size. This algorithm ensures real-time detection while maintaining high accuracy.

Acknowledgments

This research was funded by Smart Aluminum Plant Technological Innovation Project of Inner Mongolia Huomei Hongjun Aluminum Electric Co. (K20231334).

References

[1] Su J, An Y, Wu J, et al. Pedestrian Detection Based on Feature Enhancement in Complex Scenes[J]. Algorithms, 2024, 17(1): 39.

[2] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.

[5] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

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

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

[8] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arxiv preprint arxiv:1804.02767, 2018.

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

[10] Zhang S, **e M. Real-time recognition and localization based on improved YOLOv5s for robot's picking clustered fruits of chilies[J]. Sensors, 2023, 23(7): 3408.

[11] Zhao J, Liu Y. Research on Road Sign Detection and Visual Depth Perception Technology for Mobile Robots[J]. Electronics, 2023, 12(14): 3202.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202409_10(9).0005

[12] Xu Z, Meng Y, Yin Z, et al. Enhancing autonomous driving through intelligent navigation: A comprehensive improvement approach[J]. Journal of King Saud University-Computer and Information Sciences, 2024: 102108.

[13] Zhang Q, Yu W, Liu W, et al. A Lightweight Visual Simultaneous Localization and Map** Method with a High Precision in Dynamic Scenes[J]. Sensors, 2023, 23(22): 9274.

[14] Wang Y. Improved Traffic Sign Recognition Algorithm for YOLOv5[J]. International Core Journal of Engineering, 2024, 10(5): 63-76.

[15] Han K, Wang Y, Tian Q, et al. Ghostnet: More features from cheap operations[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 1580-1589.

[16] Wang T, Huang D, Tang X, et al. Aerial Photo Insulator Defect Detection based on Lightweight YOLOv5s[J]. International Core Journal of Engineering, 2024, 10(2): 96-108.

[17] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[18] Tan M, Le Q. Efficientnetv2: Smaller models and faster training[C]//International conference on machine learning. PMLR, 2021: 10096-10106.

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

[20] Ma N, Zhang X, Zheng H T, et al. Shufflenet v2: Practical guidelines for efficient cnn architecture design[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 116-131.

[21] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.

[22] Chollet F. Xception: Deep learning with depthwise separable convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1251-1258.

[23] Nascimento M G, Fawcett R, Prisacariu V A. Dsconv: Efficient convolution operator[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 5148-5157.

[24] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[25] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 13713-13722.

[26] Ouyang D, He S, Zhang G, et al. Efficient multi-scale attention module with cross-spatial learning[C]// ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[27] Zhang Y F, Ren W, Zhang Z, et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing, 2022, 506: 146-157.

[28] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.

[29] Gevorgyan Z. SIoU loss: More powerful learning for bounding box regression[J]. arXiv preprint arXiv:2205.12740, 2022.

[30] Tong Z, Chen Y, Xu Z, et al. Wise-IoU: bounding box regression loss with dynamic focusing mechanism[J]. arXiv preprint arXiv:2301.10051, 2023.