

MA-YOLOv8: Road Vehicle Target Detection Based on MBConv Module and C2f_AKConv Structure Improved YOLOv8

Ting Zhang

*Artificial Intelligence, Anhui University, Hefei, China
zhang_tinghf@163.com*

Abstract. Rapid, accurate and robust detection of road vehicles is a core and fundamental task in fields such as intelligent transportation systems, autonomous driving and traffic management. In response to this demand, this paper proposes an improved YOLOv8 model, which optimizes feature extraction and scale adaptability by introducing the MBConv module and the C2f_AKConv structure. The improved model has achieved a significant improvement in overall detection performance compared to the original model: the average accuracy (mAP50-95) has been greatly enhanced from 0.352 to 0.466. The recall rate (R) increased from 0.421 to 0.539, significantly reducing the rate of missed detections. The precision rate (P) has increased from 0.632 to 0.735, effectively reducing false detections. In terms of specific categories, the improvements are particularly prominent: the recall rate of the Car category has jumped from 0.252 to 0.328, and the mAP50 has increased from 0.392 to 0.474, enhancing the ability to recognize occlusions and small targets. The mAP50-95 of the Truck category has been upgraded from 0.183 to 0.264, improving the detection of vehicles with complex shapes. The Ambulance category maintained a mAP50 above 0.9 while MAP50-95 was raised to 0.799, achieving more accurate bounding box localization. The comprehensive results show that the proposed improvement strategy effectively enhances the feature extraction ability, scale adaptability and detection robustness for weak texture targets of the model. This research provides a more effective solution for real-time target detection in complex traffic scenarios.

Keywords: Road vehicle inspection, YOLOv8, MBConv

1. Introduction

In fields such as intelligent transportation systems, autonomous driving technology, traffic management and monitoring, it is a crucial fundamental task to conduct rapid, accurate and robust detection of various types of vehicles on the road, such as cars, trucks, buses, motorcycles, bicycles, etc. [1] With the acceleration of urbanization and the sharp increase in the number of motor vehicles, road scenarios have become increasingly complex, and traffic congestion, accident risks and management difficulties have also risen accordingly. The traditional detection methods based on manual observation or simple sensors are inefficient, have limited coverage and are vulnerable to

environmental interference, making it difficult to meet the modern transportation requirements for real-time performance, all-weather operation and high precision [2]. The development of computer vision technology, especially the breakthroughs in deep learning in the field of object detection, has provided a powerful tool for the perception of automated and intelligent road traffic vehicles, enabling them to identify and locate various vehicle targets in real time from complex road images or video streams. It provides core perception capability support for subsequent traffic flow analysis, behavior understanding, assisted driving, and even fully autonomous driving decision-making [3].

The emergence of the YOLO (You Only Look Once) series of algorithms, especially its first-generation version, has brought about a revolution in the field of object detection and had a profound impact on the detection of road vehicles [3]. The core idea is to treat object detection as a single regression problem, directly predicting bounding boxes and category probabilities across the entire image, achieving true "end-to-end" detection. Compared with traditional two-stage or early complex one-stage detectors, the greatest advantage of YOLO lies in its astonishing speed. This real-time performance perfectly meets the strict requirements for low latency in application scenarios such as road monitoring and vehicle perception, making it possible to deploy high-performance vehicle detectors on embedded devices and ordinary computing units [4]. Meanwhile, while maintaining a relatively fast speed, YOLO can also achieve a rather good detection accuracy, balancing the key contradiction between real-time performance and accuracy. This has greatly promoted the practical application process of intelligent transportation and autonomous driving perception modules, and has become one of the widely adopted and continuously optimized benchmark frameworks in this field [5].

As the latest outstanding representative of the YOLO family, YOLOv8, while inheriting the real-time performance advantages of its predecessors, has elevated the performance of road vehicle detection to a new level through a series of ingenious architectural improvements and innovative training strategies [6]. It has optimized the model structure, adopted more advanced training techniques, and provided models of different scales from lightweight to high-precision to meet diverse deployment requirements. YOLOv8 significantly enhances detection accuracy and robustness in complex road scenarios while still maintaining excellent inference speed [7]. Its more concise and efficient architecture design also lowers the threshold for use and deployment. This paper improves and optimizes the YOLOv8 model based on the MBConv module and the C2f_AKConv structure for target detection of road vehicles.

2. Data set sources and data analysis

The dataset used in this paper is a private one, which contains images of multiple categories and corresponding object detection labels, totaling 250 images. Among them, there are 50 images of ambulances, 30 images of cars, 90 images of private cars, 42 images of motorcycles, and 38 images of trucks. It can be used to develop object detection algorithms and test the detection accuracy of models. The distribution of the dataset is shown in Table 1.

Table 1. Selected data sets

Class	Images
All	250
Ambulance	50
Bus	30
Car	90
Motorcycle	42
Truck	38

Some datasets are selected for display, as shown in Figure 1.

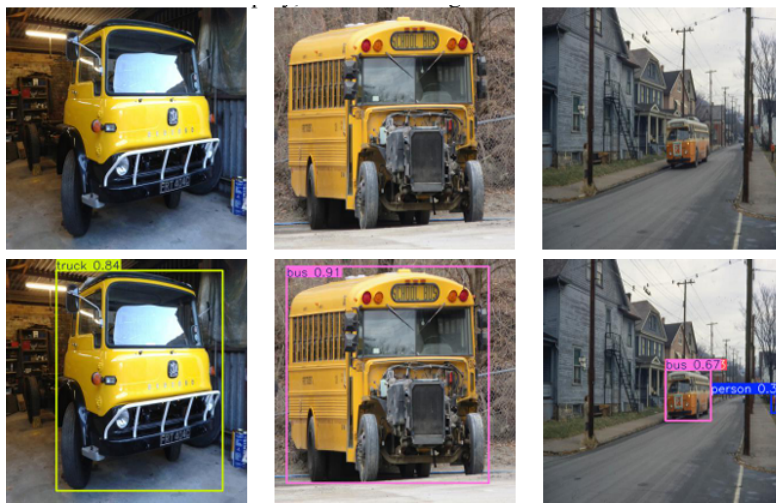


Figure 1. Partial dataset display

3. Method

3.1. MBCA module

The MBConv module is a key component of the lightweight CNN architecture. It achieves effective feature transformation at a relatively low computational cost through depth-separable convolution and channel dilation/compression operations. However, the standard MBConv has limitations in modeling long-distance spatial dependencies and precise location information, which is crucial for the detection of vehicles with variable sizes, mutual occlusion, or at edge positions in road scenarios. The Coordinate Attention (CA) mechanism can effectively make up for this deficiency: it innovatively decomposes the channel attention into two parallel 1D global pooling operations along spatial height and width, generating a pair of direction-aware feature maps [8]. Subsequently, these feature maps are concatenated, convoluted, and an attention weight map containing precise spatial position information is generated through the Sigmoid function. This weight map is ultimately applied to the input features, enabling the model to selectively enhance the response strength at specific spatial positions in the feature channels that contain important targets or key parts of the targets. The CA mechanism is embedded after the deep convolutional layer and before the point convolutional layer of MBConv to form the MBCA module. This enables MBConv to perform efficient feature transformation while obtaining the collaborative perception ability of the target

spatial position and cross-channel relationship conferred by CA. The network structure of the MBCA module is shown in Figure 2.

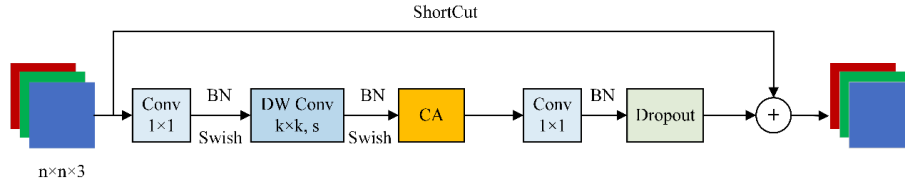


Figure 2. The network structure of the MBCA module

The two standard C2f modules at the front end of the backbone network were chosen to be replaced with MBCA modules because the shallow layer of the network is mainly responsible for processing the low-level features of the original input. In road inspection tasks, these low-level features are particularly crucial for subsequent identification of vehicle contours, differentiation of vehicle types, and location of small-sized targets. Introducing MBCA at this stage enables the model to focus on the spatial regions and channel information highly related to vehicles in the image at the initial stage of feature extraction, significantly enhancing the discriminability of shallow features. This improvement can suppress background noise interference at the source, enhance the target signal, and provide a higher-quality and more position-sensitive feature basis for subsequent deep feature abstraction and detection head prediction, thereby ultimately improving the overall detection accuracy and robustness of the model for vehicles in various complex road environments. At the same time, due to the high computational efficiency and reasonable embedding position of the CA mechanism itself, The influence on the reasoning speed of the model is relatively controllable.

3.2. C2f_AKConv structure

By using arbitrary kernel convolution (AKConv) to replace the second standard convolution of the Bottleneck structure in the C2f module, the C2f_AKConv module is constructed and applied to the front end of the neck network, aiming to endow the model with the ability to dynamically adapt to targets of different scales and shapes. Significantly enhance the geometric feature modeling and information extraction effect of multi-scale vehicles in complex road scenarios. The network structure of the C2f_AKConv structure is shown in Figure 3.

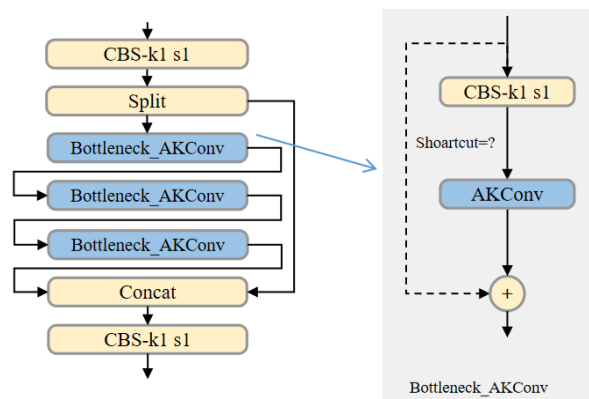


Figure 3. The network structure of the C2f_AKConv structure

The first two C2f modules of the neck network were chosen to be replaced with the C2f_AKConv module because the neck is mainly responsible for fusing the feature maps from different levels of the backbone network to construct a feature pyramid containing rich multi-scale information for the final detection and prediction [9]. The front end of the neck processes precisely the mid-to-high-level features that have been initially fused. These features not only contain strong semantic information but also retain certain spatial details. At this critical stage, the introduction of AKConv, with its dynamic deformable sampling capability, is particularly important: on the one hand, it can effectively model the multi-scale targets that are commonly present in road scenes. By adaptively adjusting the size and shape of the receptive field of the convolutional kernel, it ensures that its key features can be effectively aggregated regardless of how the target scale changes. On the other hand, for vehicles with non-rectangular contours, severe obstructions or from special perspectives, AKConv can more accurately focus on the visible and effective parts of the target, extract more discriminative local features, and suppress the interference of the background or obstructions. This refined modeling of geometric information in the feature fusion stage directly enhances the positioning accuracy and classification confidence of the bounding boxes generated by the subsequent detection heads, thereby improving the overall robustness and generalization ability of the model in complex and dynamic road environments.

3.3. Model framework

The main framework of MA-YOLOv8 is shown in Figure 4. The optimization of YOLOv8 in this paper mainly includes two core improvements. Firstly, an MBCA module integrating coordinate attention (CA) was introduced at the front end of the Backbone network. This improvement is based on the MBConv structure of MobileNetV2. The CA mechanism significantly enhances the model's perception ability of the target spatial distribution by decomposing the channel attention into position information encoding along the horizontal and vertical directions and modeling the feature spatial position relationship in the channel dimension. Specifically, CA first generates direction-sensitive feature maps through coordinate information embedding, then enhances position-aware features through coordinate attention transformation, and finally fuses position weights into the original features. After upgrading MBConv to the MBCA module and replacing the first two C2f modules of the backbone network, the network can accurately capture the geometric position information of the target in the shallow feature extraction stage, effectively improving the positioning ability of small targets and edge features [10]. Secondly, a C2f_AKConv module based on deformable convolution was designed at the front end of the Neck network. This module replaced the Bottleneck structure in the standard C2f with AKConv. AKConv dynamically learns the offsets of the convolution kernels, enabling the convolution sampling points to break through the limitations of the fixed grid and adaptively match the complex deformations and scale changes of the target. The core lies in constructing a parametric offset prediction mechanism to generate an independent spatial offset vector for each sampling point, thereby achieving flexible sampling at any spatial position. The improved C2f_AKConv module was applied to the first two C2f structures of the neck network, significantly enhancing the deformation modeling ability for non-rigid targets during the multi-scale feature fusion process and optimizing the spatial adaptability of high-dimensional features. The two improvements respectively approach from the dimensions of spatial attention enhancement and convolution kernel deformation adaptation, jointly enhancing the robustness of the model's target representation in complex scenarios.

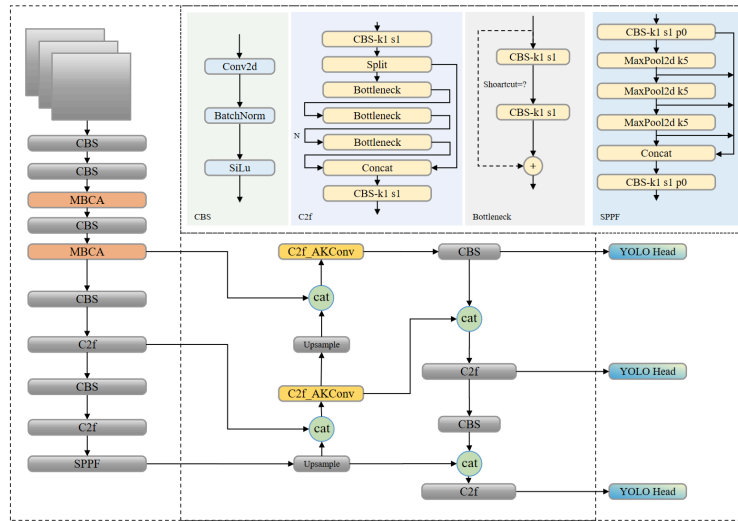


Figure 4. The main framework of MA-YOLOv8

4. Experiments and results

The experiment was run in a hardware environment equipped with NVIDIA RTX 3090 GPU (24GB video memory), Intel Xeon Silver 4210R CPU and 64GB memory. The software used was Ubuntu 20.04, CUDA 11.8 and PyTorch 2.0.1. The training uses an input with a resolution of 640×640 , with a batch size of 16, and conducts 100 rounds of optimization. The initial learning rate of the AdamW optimizer is set to 0.01, and the weight decay is 0.0005.

First, output the F1 confidence curve of the improved YOLOv8 model, as shown in Figure 5.

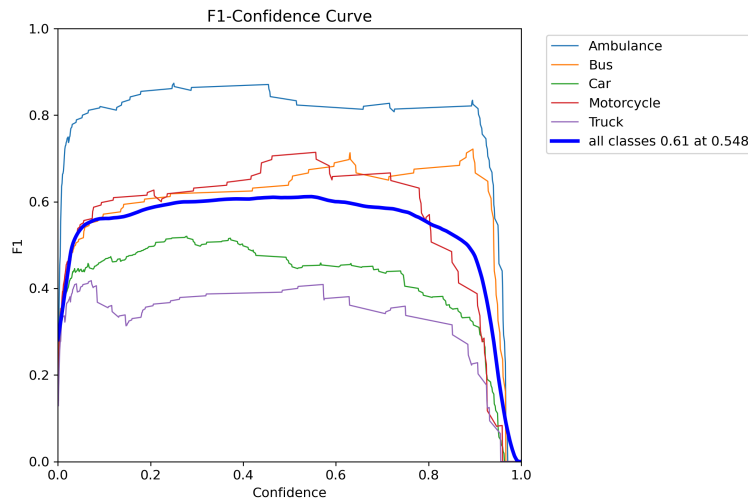


Figure 5. The F1 confidence curve of the improved YOLOv8 model

Output the model evaluation results of the YOLOv8 model, as shown in Table 2. The evaluation results of the improved YOLOv8 model are output as shown in Table 3.

Table 2. The model evaluation results of the YOLOv8 model

Class	Images	Instances	P	R	mAP50	mAP50-95
All	250	454	0.632	0.421	0.487	0.352
Ambulance	50	64	0.825	0.672	0.842	0.701
Bus	30	46	0.592	0.533	0.602	0.482
Car	90	238	0.621	0.252	0.392	0.271
Motorcycle	42	46	0.681	0.522	0.512	0.298
Truck	38	60	0.541	0.217	0.291	0.183

Table 3. The evaluation results of the improved YOLOv8 model

Class	Images	Instances	P	R	mAP50	mAP50-95
All	250	454	0.735	0.539	0.608	0.466
Ambulance	50	64	0.891	0.765	0.911	0.799
Bus	30	46	0.674	0.652	0.689	0.571
Car	90	238	0.706	0.328	0.474	0.338
Motorcycle	42	46	0.782	0.652	0.597	0.359
Truck	38	60	0.623	0.3	0.371	0.264

The improved YOLOv8 model significantly outperforms the original model in overall detection performance. From the perspective of global indicators, the mAP50-95 of the improved model has increased from 0.352 to 0.466, and the recall rate (R) has risen from 0.421 to 0.539, indicating a significant reduction in missed detections. The precision rate (P) increased from 0.632 to 0.735, indicating a decrease in the false detection rate. Specifically in terms of categories, the improvement of the model for Cars is the most prominent: the recall rate (R) jumps from 0.252 to 0.328, and the mAP50 rises from 0.392 to 0.474, demonstrating an enhanced detection capability for occlusions and small targets. The improvement of Truck is equally significant. mAP50-95 has risen from 0.183 to 0.264, resolving the issue of insufficient recognition of complex-shaped vehicles in the original model. In addition, the mAP50 of Ambulance remained at a high level of 0.9+, but the improved MAP50-95 increased to 0.799, indicating that the bounding box localization is more accurate. The comprehensive results show that the improved strategy effectively enhances the feature extraction and scale adaptability of the model, especially significantly optimizing the robustness of detection for weakly textured targets.

5. Conclusion

This paper, aiming at the core demand of road vehicle detection in the fields of intelligent transportation, autonomous driving and traffic management, proposes a YOLOv8 model based on the MBConv module and the C2f_AKConv structure improvement. By introducing a more efficient feature extraction mechanism and enhanced adaptability to spatial geometric features, the improved model demonstrates a significant advantage over the original model in overall detection performance. The global evaluation indicators have fully verified its effectiveness: the average accuracy (mAP50-95) has significantly increased from 0.352 to 0.466, and the Recall rate has significantly improved from 0.421 to 0.539, indicating that the model's ability to recognize targets

has been significantly enhanced, and the phenomenon of missed detections has been effectively suppressed. Meanwhile, the Precision has increased from 0.632 to 0.735, reflecting that the model's discriminative ability is more accurate and the false detection rate has significantly decreased. The leaps in these global indicators jointly confirm the substantial optimization of the overall performance of the model.

The advantages of the improved model have been further highlighted by focusing on specific categories of transportation. For the numerous "Car" category that often faces occlusion or scale challenges, its recall rate has jumped from 0.252 to 0.328, and mAP50 has increased from 0.392 to 0.474, strongly demonstrating that the model has a stronger recognition ability for small targets and vehicles under occlusion. For the relatively complex "Truck" class, its mAP50-95 has been upgraded from 0.183 to 0.264, significantly improving the problem of insufficient recognition of such targets by the original model. Even for the "Ambulance" class, which already has a relatively high detection accuracy, after improvement, its mAP50-95 has been raised to 0.799, while maintaining a high mAP50 level of above 0.9. This indicates that while the model maintains a high detection rate, the positioning accuracy of the bounding box is also more precise. Overall, the improved strategy integrating MBConv and C2f_AKConv significantly enhances the feature extraction efficiency and scale adaptability of the model, especially improving the robustness for detecting weak texture targets. This research has significantly enhanced the accuracy and reliability of road vehicle detection through innovative improvements in key structures, providing a solid technical support for the construction of a more intelligent, efficient and safe traffic perception system.

References

- [1] Taslim Arif, et al. "Vision-based real-time traffic flow monitoring system for road intersections in Dhaka city." *Applied Intelligence* 55.11(2025): 800-800.
- [2] Amir T. Mohamed, et al. "Integrating EnlightenGAN for enhancing car logo detection under challenging lighting conditions." *Multimedia Tools and Applications* preublish(2025): 1-28.
- [3] Prakash P.S., et al. "Accelerating Electric Vehicle (EV) adoption: A remote sensing data driven and deep learning-based approach for planning public car charging infrastructure." *Remote Sensing Applications: Society and Environment* 37.(2025): 101447-101447.
- [4] Aswath S., et al. "Smart Car Parking System with Online Reservation." *Procedia Computer Science* 258.(2025): 2777-2786.
- [5] XinyunFeng, et al. "ADWNet: An improved detector based on YOLOv8 for application in adverse weather for autonomous driving." *IET Intelligent Transport Systems* 18.10(2024): 1962-1979.
- [6] Do Yoon Jung, Yeon Jae Oh, and Nam Ho Kim. "A Study on GAN-Based Car Body Part Defect Detection Process and Comparative Analysis of YOLO v7 and YOLO v8 Object Detection Performance." *Electronics* 13.13(2024): 2598-2598.
- [7] Sonkavde, Gaurang, et al. "Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications." *International Journal of Financial Studies* 11.3 (2023): 94.
- [8] Shrotriya, Lalit, et al. "Cryptocurrency algorithmic trading with price forecasting analysis using PowerBI." *International Journal of Engineering, Science and Technology* 15.4 (2023): 1-8.
- [9] Yuwei Hu. "Improved YOLOv8 algorithm for vehicle image target detection based on learning rate optimisation strategy". Ed.2024,
- [10] Gao, Jie. "Research on stock price forecast based on Arima-GARCH model." *MSIEID 2022: Proceedings of the 4th Management Science Informatization and Economic Innovation Development Conference, MSIEID 2022, December 9-11, 2022, Chongqing, China. European Alliance for Innovation, 2023.*