Proceedings of CONF-MLA 2026 Symposium: Learning and Decision Making in Multi Agent Software Systems
DOI: 10.54254/2755-2721/2026.BA37311

Progress in Target Detection Technology for Road Traffic Vehicles

Zhenghao Chi

School of Computer Science, University of Nebraska at Kearney, Kearney, USA
73862517@nebraska.edu

Abstract. With the rapid implementation of intelligent transportation systems and autonomous driving technology, vehicle object detection, as the core technology of environmental perception, is widely used in road monitoring, traffic flow statistics, violation identification, and on-board forward perception. Real road traffic scenarios present problems such as large vehicle scale variation, mutual occlusion of targets, complex illumination conditions, and severe weather interference, which can easily lead to missed detection and false detection. This paper divides deep learning-based vehicle detection algorithms into three technical routes: two-stage detection, one-stage detection, and Transformer end-to-end detection, and systematically reviews the core principles, technical evolution paths, and scenario adaptation characteristics of each algorithm. Taking the lightweight YOLO11-n model as the research carrier, a comparative experiment is designed based on the SODA10M large-scale autonomous driving public dataset to quantitatively verify the improvement effect of the multi-scale data augmentation strategy on vehicle detection performance, and multi-angle result analysis is completed using statistical charts, precision curves, and feature heatmaps. Finally, the core challenges faced by current vehicle detection technology are summarized, and future optimization directions are discussed, which can provide theoretical reference for the selection and optimization of intelligent transportation perception equipment. Experimental results show that the multi-scale augmentation strategy improves the mAP@0.5 of YOLO11-n from 74.3% to 78.9% on the SODA10M dataset, demonstrating its effectiveness in addressing vehicle scale variation.

Keywords: vehicle object detection, deep learning, YOLO11, multi-scale augmentation, SODA10M dataset

1. Introduction

In the process of smart city construction, the accuracy of intelligent traffic management and control systems directly determines urban traffic efficiency and road safety. Vehicle object detection is the core foundational technology for applications such as traffic flow statistics, violation capture, congestion warning, and intelligent parking lot management. In the field of autonomous driving, vehicle detection is the basic perception module for forward collision warning, adaptive cruise control, and automatic lane change functions, and its detection accuracy and inference latency are directly related to driving safety.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BA37311

Early vehicle detection relied on manually designed features and shallow classifiers, which could achieve basic recognition in ideal scenarios with simple backgrounds and uniform illumination. However, in real road environments, factors such as day-night illumination changes, rain, snow, and fog weather, mutual occlusion of vehicles, and far-near scale differences lead to significant performance degradation of traditional methods, which cannot meet deployment requirements. With the breakthrough of convolutional neural networks and deep learning technology, data-driven detection algorithms have greatly improved recognition ability in complex scenarios and have become the mainstream technical solution for current vehicle detection.

The development of vehicle object detection technology can be divided into three stages. The first stage is the traditional manual feature stage, represented by HOG+SVM, Haar feature+Adaboost, and background subtraction methods, which rely on manually defined edge and texture features for classification, with weak generalization ability and low accuracy ceiling. The second stage is the rise of two-stage deep learning algorithms, represented by the R-CNN series, which transfers feature extraction, candidate region generation, and classification regression to neural networks, achieving a qualitative leap in detection accuracy, but the inference speed is difficult to meet real-time requirements. The third stage is the parallel development stage of one-stage algorithms and Transformer algorithms, represented by the YOLO series, SSD, and DETR, which continuously improve inference speed while ensuring accuracy, promoting detection technology to evolve from offline analysis to real-time online deployment, and gradually penetrating into low-computing scenarios such as edge devices and vehicle terminals.

This paper is divided into 5 chapters. Chapter 1 is the introduction, which elaborates the research background, significance, and technology development status. Chapter 2 introduces the detail. Chapter 3 carries out comparative experiments based on YOLO11-n, introducing the dataset, experimental setup, baseline results, and optimization effect of multi-scale augmentation. Chapter 4 summarizes the current technical challenges and future research directions. Chapter 5 is the conclusion of the full paper.

2. Mainstream vehicle object detection algorithms

2.1. Two-stage detection algorithms

Two-stage detection algorithms split the detection process into two independent steps: candidate region generation and target classification regression, with the core advantage of high detection accuracy and precise localization.

R-CNN is the foundational algorithm for deep learning detection. It generates approximately 2000 candidate regions through selective search, and inputs them into convolutional networks one by one to extract features before classification with SVM, but it suffers from redundant computation and slow inference speed. On this basis, Fast R-CNN proposes Region of Interest Pooling (RoI Pooling), which extracts global features from the entire image at once, greatly improving computational efficiency, but it still relies on selective search to generate candidate boxes.

Faster R-CNN is the mature benchmark framework of two-stage algorithms, and its core innovation is the proposal of the Region Proposal Network (RPN), which completely replaces selective search. RPN generates candidate boxes directly on the shared feature map based on the anchor mechanism, which presets reference boxes with different sizes and aspect ratios, and directly predicts the offset of candidate boxes and target confidence, realizing full-process neural network operation. In vehicle detection scenarios, Faster R-CNN has high localization accuracy and good stability in congested scenarios, and is often used in scenarios with high accuracy requirements such

DOI: 10.54254/2755-2721/2026.BA37311

as traffic violation forensics and offline traffic flow statistics. However, its two-stage calculation mode results in a low inference frame rate, which is difficult to adapt to real-time monitoring and on-board perception scenarios. Subsequent studies often optimize its multi-scale detection capability by introducing Feature Pyramid Network (FPN)and attention modules, but cannot fundamentally overcome the speed limitation. Feature Pyramid Networks further enhance the multi-scale detection capability of two-stage detectors by fusing semantic and spatial features across different levels [1, 2].

2.2. One-stage detection algorithms

One-stage detection algorithms skip the independent candidate region generation step, and can output target categories and bounding box coordinates simultaneously through one forward inference of the input image. The core advantages are fast inference speed, simple structure, and low deployment cost, making it the mainstream solution for current engineering deployment.

SSD is an early classic one-stage algorithm that adopts a multi-scale feature map detection strategy. Shallow feature maps are responsible for small targets and deep feature maps for large targets, taking into account the detection effect of vehicles of different scales, but the problem of insufficient semantic information for small targets persists. Aiming at the problem of imbalance between positive and negative samples in one-stage training, RetinaNet proposes Focal Loss, which reduces the weight of easily classified background samples and makes the model focus on hard samples, enabling the accuracy of one-stage algorithms to catch up with two-stage algorithms for the first time. RetinaNet addresses the extreme class imbalance problem in one-stage training through Focal Loss, which down-weights easy negative samples and prioritizes hard examples [3, 4].

The YOLO series is the most widely used one-stage algorithm in the field of vehicle detection, which has been continuously optimized through multiple generations of iterations. YOLOv3 introduces a feature pyramid structure and realizes multi-scale prediction for the first time, greatly improving the missed detection problem of small target vehicles. YOLOv4 and YOLOv5 introduce the CSP backbone structure and mosaic data augmentation, achieving simultaneous improvement of accuracy and speed at the engineering level. YOLOv7 and YOLOv8 optimize the feature fusion structure and sample allocation strategy, further strengthening detection performance in occlusion scenarios and high-speed scenarios. A comprehensive benchmark review traces the architectural evolution from YOLOv3 through YOLO11, analyzing the accuracy-speed tradeoffs across model sizes [5, 6].

YOLO11 is the latest stable version, providing multi-scale models such as n, s, m, and l. Among them, YOLO11-n is a lightweight version that compresses the number of parameters through channel pruning and reparameterization technologies, with significantly improved inference speed, and can be directly deployed on low-computing hardware such as embedded on-board equipment and edge surveillance cameras. Aiming at the problem of large vehicle scale variation on roads, multi-scale data augmentation is a low-cost and high-benefit optimization method: randomly scaling the input image resolution during training allows the model to learn vehicle features at different sizes, effectively alleviating the missed detection problem of long-distance small vehicles. This paper selects YOLO11-n as the basic experimental model [7].

2.3. Transformer end-to-end detection algorithms

DETR (Detection Transformer) introduces the Transformer architecture into object detection, completely abandoning manually designed components such as anchors and candidate boxes. Its

DOI: 10.54254/2755-2721/2026.BA37311

architecture consists of an encoder and a decoder. The encoder performs global feature modeling on the image through the self-attention mechanism, capturing long-distance dependencies between pixels. The decoder directly outputs a fixed number of detection results through learnable query vectors, and realizes end-to-end training through bipartite matching loss, without post-processing steps such as Non-Maximum Suppression (NMS).

Different from the local receptive field of convolutional networks, the global modeling ability of Transformer has unique advantages in dense and congested vehicle scenarios, which can better distinguish overlapping vehicle targets and reduce the missed detection rate under severe occlusion. However, the original DETR has defects of slow training convergence and poor small target detection effect. The subsequently derived Swin-DETR introduces layered window attention, which compensates for the above problems to a certain extent. On the whole, Transformer detection algorithms have a large number of parameters, high computing power requirements, and slow inference speed. At present, they are mostly used for back-end high-precision batch analysis, and it is difficult to realize real-time deployment on ordinary edge devices. Swin-DETR further improves small-target performance by replacing the standard Transformer backbone with a hierarchical window-attention architecture [8, 9].

2.4. Comprehensive comparison of three types of algorithms

There are significant differences in accuracy, speed, hardware requirements, and applicable scenarios among the three types of algorithms, as shown in Table 1.

Table 1. Comprehensive comparison of three types of vehicle detection algorithms

Algorithm type	Representative algorithm	Detection accuracy	Inference speed	Hardware computing requirement	Core application scenario
Two-stage detection	Faster R-CNN	High	Slow	Medium-high	Offline high-precision analysis, traffic violation forensics
One-stage detection	YOLO11 series	Relatively high	Fast	Low-medium	Real-time road monitoring, on-board edge perception
Transformer detection	DETR, Swin-DETR	Relatively high	Slow	High	High-precision analysis of dense traffic flow, background batch processing

From Table 1, two-stage algorithms achieve high detection accuracy but suffer from low inference efficiency, making them suitable for offline high-precision analysis. One-stage YOLO-series algorithms achieve a good balance between accuracy and speed, which are more suitable for real-time vehicle detection tasks on edge devices. Transformer-based detectors own powerful global modeling capability for occluded targets, yet heavy computation limits their practical deployment on low-power hardware.

DOI: 10.54254/2755-2721/2026.BA37311

On the whole, the lightweight one-stage YOLO model achieves the optimal balance among accuracy, speed, and deployment cost, and is the preferred solution for ordinary road real-time monitoring scenarios.

3. Comparative experiment of vehicle detection based on YOLO11

Note: The experimental baseline data and visualization results in this chapter are cited from supporting studies of the SODA10M dataset and public benchmark test results of YOLO11, which are used for comparative analysis of algorithm performance.

3.1. Experimental dataset

This experiment adopts the SODA10M large-scale autonomous driving public dataset, which contains a large number of real road scene images, covering multiple types of traffic targets such as cars, trucks, buses, non-motor vehicles, and pedestrians. The samples cover complex road conditions such as daytime, nighttime, backlight, rain and snow, vehicle occlusion, and long-distance small targets, which can fully reflect the detection difficulties of real roads.

The dataset has significant long-tail distribution and scale imbalance characteristics. In terms of category, the number of car samples is the largest, while the number of samples such as tricycles and trams is very small. In terms of target size, small and medium-sized vehicle bounding boxes account for a very high proportion, while large-size targets account for a low proportion. Vehicle scale variation is the core difficulty of the detection task, and the dataset statistics reflect this scale imbalance [10].

3.2. Experimental setup and evaluation indicators

3.2.1. Experimental environment and parameter setting

Two groups of comparative experiments are designed. The baseline group trains the original YOLO11-n model without any additional optimization strategies, while the improved group keeps all training parameters unchanged but enables multi-scale data augmentation during the training process.

3.2.2. Evaluation indicators

General quantitative indicators for object detection are adopted to evaluate model performance. The mAP@0.5 indicator represents Mean Average Precision at an Intersection over Union threshold of 0.5, which primarily measures overall detection accuracy. Precision describes the proportion of real targets among all detection results, reflecting the false detection rate of the model. Recall measures the proportion of real targets that are successfully detected, indicating the missed detection level of the model.

3.3. Baseline model experimental results

The original YOLO11-n is trained and tested on the SODA10M dataset, and the overall mAP@0.5 of the baseline model is 74.3%.

From the category dimension, the model has high recognition accuracy for large vehicles (buses, trucks), and low recall rate for motorcycles and long-distance small vehicles, with obvious missed

DOI: 10.54254/2755-2721/2026.BA37311

detection. The precision changes of each category can be intuitively analyzed through the precision-confidence curve and precision-recall (PR) curve. The normalized confusion matrix and F1-confidence curve further show the classification errors of each category of the model.

3.4. Comparative experiment with multi-scale augmentation

Keep the dataset and training parameters completely consistent, retrain after turning on multi-scale data augmentation. The quantitative indicators of the two groups of experiments are compared as shown in Table 2.

Table 2. Performance comparison between baseline model and multi-scale augmented model

Model version	mAP@0.5	Average Precision	Average Recall
Baseline YOLO11-n	74.3%	66.1%	70.9%
Multi-scale augmented YOLO11-n	78.9%	70.2%	76.8%

3.5. Experimental summary

The lightweight YOLO11-n has fast inference speed and can meet the computing power requirements of real-time road vehicle detection. Multi-scale augmentation is a data augmentation method in the training stage, which does not need to modify the network structure, has low implementation cost, and can effectively improve the missed detection problem of long-distance small target vehicles. It is a highly cost-effective optimization scheme. However, the improvement effect of this strategy is limited in extreme severe weather and severe occlusion scenarios, which needs to be further optimized combined with image preprocessing and network structure improvement.

This experiment only verifies the single augmentation strategy of multi-scale. Subsequent combined tests with mosaic augmentation and color perturbation can be carried out to further explore model performance.

4. Existing challenges and future development directions

4.1. Core technical challenges

At present, deep learning vehicle detection technology has become relatively mature, but several core challenges remain in real deployment scenarios. First, the model suffers from insufficient robustness under extreme environmental conditions, as severe weather such as low light at night, strong backlight, rain, snow, and fog lead to image quality degradation and feature blur, which significantly attenuates detection accuracy. Second, small target detection remains highly difficult because long-distance vehicles account for a very low proportion of pixels with weak feature information, making this category a high-incidence area of missed detection. Third, the model shows low discrimination ability in occlusion scenarios, where vehicles are densely overlapped at peak congested intersections and it becomes difficult for the model to accurately distinguish each independent target, leading to repeated detection or missed detection. Fourth, cross-scenario generalization ability remains weak, as models trained on specific road sections and specific cameras experience varying degrees of performance degradation after changes in shooting angles, illumination conditions, and regional scenarios.

DOI: 10.54254/2755-2721/2026.BA37311

4.2. Future research directions

In response to the above challenges, future research on vehicle detection technology can be carried out around several key directions. Image enhancement and preprocessing optimization can improve image quality in harsh environments through algorithms such as low-light enhancement, defogging, and deraining, reducing detection difficulty from the input end. Continuous iteration of lightweight networks can further compress model parameters and computation on the premise of ensuring accuracy, adapting to more low-cost edge devices and promoting the popularization of end-side intelligent detection. Multi-strategy fusion optimization combines multiple means such as multi-scale augmentation, mosaic augmentation, and sample resampling to alleviate the long-tail distribution problem of datasets and improve the overall generalization ability of the model. Multi-modal fusion perception fuses visible light images with multi-source data such as infrared, millimeter wave radar, and lidar to compensate for the perception defects of single visual sensors in harsh weather and realize all-weather stable detection, which is also the mainstream technical route of autonomous driving perception [11].

5. Conclusion

Vehicle object detection is the core foundational technology in the field of intelligent transportation and autonomous driving. The development of deep learning has promoted it from the traditional manual feature era to a new stage of high precision and high real-time performance. This paper systematically reviews the technical principles and scenario adaptability of two-stage, one-stage, and Transformer detection algorithms, and carries out comparative experiments based on the YOLO11-n lightweight model and the SODA10M dataset, verifying the improvement effect of the multi-scale augmentation strategy on vehicle detection performance. Considering performance, cost, and deployment difficulty comprehensively, the lightweight one-stage YOLO model is the optimal choice for current intelligent transportation real-time detection scenarios.

At present, vehicle detection technology still faces practical problems such as insufficient robustness in extreme environments, missed detection of small targets, and occlusion misjudgment. In the future, continuous breakthroughs can be made through technical routes such as image enhancement, lightweight network optimization, and multi-modal fusion, to promote the development of vehicle detection towards higher accuracy, stronger generalization, and lower cost.

References

[1] Ren, S., He, K., Girshick, R., et al. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*.
[2] Lin, T. Y., Dollar, P., Girshick, R., et al. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117-2125.
[3] Liu, W., Anguelov, D., Erhan, D., et al. (2016). SSD: Single shot multibox detector. *European Conference on Computer Vision*. Springer, Cham, 21-37.
[4] Lin, T. Y., Goyal, P., Girshick, R., et al. (2017). Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2980-2988.
[5] Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv* preprint arXiv:1804.02767.
[6] Borkovic, A., Segvic, S., Brkic, K., et al. (2024). YOLO evolution: A comprehensive benchmark and architectural review of YOLOv12, YOLO11, and their previous versions. *arXiv* preprint arXiv:2411.00201.
[7] Ultralytics. (2025). YOLO11 official documentation. *Ultralytics Docs*. https://docs.ultralytics.com/models/yolo11/
[8] Carion, N., Massa, F., Synnaeve, G., et al. (2020). End-to-end object detection with transformers. *European Conference on Computer Vision*. Springer, Cham. 213-229.

DOI: 10.54254/2755-2721/2026.BA37311

[9] Liu, Z., Lin, Y., Cao, Y., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012-10022.
[10] Chen, K., Wang, J., Pang, J., et al. (2022). SODA10M: A large-scale 2D self/semi-supervised object detection dataset for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*.
[11] Zhang, H., Deng, L., Wang, X. (2022). Research review of vehicle object detection based on deep learning. *Computer Engineering and Applications, 58*(12), 1-18.