Enhanced Object Detection for Autonomous Vehicles Using Modified Faster R-CNN with Attention and Multi-Scale Feature Fusion

Yunsheng Deng*, Aimin Qiao, Yinghui Huang, Zhangbao Chen
College of Electronics and Electrical Engineering, Bengbu University, Bengbu 233030, China
E-mail: dys1217051@126.com
*Corresponding author

Keywords: self-driving, object detection, faster R-CNN, attention, feature fusion

Received: May 13, 2025

The progress of self-driving technology necessitates more stringent demands on object detection systems, and traditional methods are difficult to meet real-time and high-precision requirements in dynamic scenes. Therefore, this study proposes an improved Faster R-CNN model tailored for vehicle object detection in autonomous driving scenarios. Specifically, an enhanced Convolutional Block Attention Module (CBAM) is integrated into the backbone network to strengthen feature representation. The Region of Interest Align (ROI-Align) is employed to improve localization accuracy, especially for small or occluded targets. Moreover, Soft Non-Maximum Suppression (Soft-NMS) is adopted to reduce false negatives in dense object scenarios. Additionally, a multi-scale feature fusion mechanism is introduced to enhance detection performance across varied object sizes. The experiment outcomes indicate that the detection accuracy of the improved model reaches 98.13%, with a miss rate of less than 1.00%. In dense target scenes, the retained accuracy is 94.16%, and the standardized mean square error of target localization is 0.014.In complex environments, the average accuracy of the model in lighting changes, severe weather, and dynamic interference scenarios is 80.45%, 77.83%, and 75.11%, respectively, which is superior to the comparison methods and demonstrates higher robustness. This study enhances the detection performance of faster region-based convolution neural network in automatic driving through technical modifications, solves the problem of feature extraction and target location in complex scenes, and provides important support for the perception reliability of auto drive system.

Povzetek: Članek predstavi izboljšano Faster R-CNN metodo za avtonomna vozila, ki združuje pozornostni modul, ROI-Align, Soft-NMS in večsklopno fuzijo značilk za boljše zaznavanje objektov v zahtevnih prometnih razmerah.

1 Introduction

In recent times, the swiftadvancement of self-driving technology has driven the innovation of intelligent transportation systems, allowing vehicles to autonomously perceive, plan paths, and perform driving operations in highly complex and dynamically changing road environments [1]. Among them, object detection (OD), as the core task of the self-driving perception system, directly affects the vehicle's understanding and decision-making of the surrounding environment, including accurate recognition of key targets such as pedestrians, vehicles, traffic signs, and road markings ahead, as well as real-time tracking and prediction of target motion status [2-3]. High precision OD not only improves the adaptability of the auto drive system to complex traffic scenarios, but also effectively reduces the risk of traffic accidents, enhances driving safety, and provides accurate perception information for advanced driving assistance systems [4]. Conventional approaches to OD depend on hand-crafted feature extraction (FE) techniques, which can achieve certain detection results in static and regularized environments, but their adaptability and generalization ability are limited. Under complex road conditions such as lighting changes, target occlusion, and dynamic interference, traditional methods significantly reduce detection accuracy and are difficult to meet real-time requirements [5]. In addition, single-stage and two-stage detection algorithms that have emerged in recent times have found extensive application in self-driving scenarios, among which the Faster Region-Based Convolutional Neural Network (Faster R-CNN), one of the two-stage detection algorithms, has emerged as one of the predominant approaches for self-driving OD due to its high detection accuracy [6]. However, traditional Faster R-CNN also has limitations in complex scenes, such as insufficient global FE, limited target localization accuracy, and easy missed detection of dense targets [7]. In view of this, the research is based on Faster R-CNN, combined with Convolutional Block Attention Module (CBAM), Region of Interest Align (ROI-Align), and Soft Non-Maximum Suppression (Soft-NMS) to improve it and propose an autonomous vehicle OD and recognition model. The research aims to improve the accuracy of OD in complex traffic environment, enhance the adaptability of the model to dense targets, occluded targets and light changing scenes, optimize the generalization ability of the detection framework, and improve the perception reliability of the auto drive system.

The innovation and contribution of the research focus on the key issues of low detection accuracy of small targets and insufficient robustness in complex scenarios. A task-driven structural integration strategy is devised, which modularly links key components like feature attention, spatial alignment, target screening, and scale perception. Additionally, it establishes an end-to-end collaborative framework between the feature response layer, regional location layer, and candidate screening layer. The directional enhancement of the Faster R-CNN structure in the detection path has been achieved instead of simply stacking the existing methods.

2 Related works

The development of self-driving technology relies on efficient and accurate OD and recognition systems to ensure safe driving of vehicles in complex dynamic environments. As the core task of the self-driving perception system, OD is mainly responsible for identifying and tracking key targets such as vehicles, pedestrians, and traffic signs on the road. In recent years, researchers have conducted extensive research on OD in self-driving. Yang M proposed an optimized object recognition algorithm that combined vehicle perception technology to address the issues of OD accuracy and safety in the process of vehicle self-driving. The superiority of the multi-strategy region recommendation network algorithm was verified, thereby optimizing the performance of OD and recognition and making it more suitable for self-driving environments [8]. Mahaur B et al. proposed a systematic comparative study to address the lack of multidimensional comparisons in detection speed, accuracy, model size, and energy efficiency of existing deep learning OD algorithms in self-driving. The study analyzed the performance of five mainstream deep learning algorithms on large-scale datasets, thereby optimizing the understanding of the advantages and disadvantages of different algorithms and providing reference for practical deployment [9]. Ashqar H I et al. proposed an image processing technology combining computer vision and artificial intelligence to meet the requirements of vehicles for environment perception and intelligent decision-making ability in self-driving, covering camera and sensor technology, image pre-processing, FE and OD, thus optimizing the application of auto drive system in lane maintenance, obstacle detection, traffic signal and sign recognition [10]. Arora N et al. raised a region-based deep learning approach to address the recognition difficulties caused by insufficient data, low lighting, long shadows, and static frame testing in vehicle detection during day and night modes. The method utilized Faster R-CNN to optimize detection performance and improve target recognition ability in complex environments [11].

Faster R-CNN, as an important algorithm in the field of OD, has demonstrated excellent performance in detection accuracy and robustness, and is widely used in various computer vision tasks. In recent times, researchers have conducted various improvement studies on Faster R-CNN. Rani S et al. raised a detection approach based on wireframe features combined with Faster R-CNN to tackle the challenges of computational efficiency and FE accuracy in OD. This approach employs cell logic array processing to extract image wireframes, which are then fed into the detection model. By doing so, it not only accelerates the detection process but also enhances the capability to recognize geometric features, ultimately leading to greater detection accuracy and optimized recognition of both two-dimensional and three-dimensional objects [12]. Yusro et al. proposed an OD method based on Faster R-CNN to address the difficulty of classifying overlapping targets. By optimizing feature separation through dedicated layer filters, the detection ability of overlapping targets was improved, achieving effective recognition of overlapping targets in complex scenes [13]. Güney E et al. proposed a real-time detection system based on Faster R-CNN to address the impact of traffic sign and road OD on driving safety. By training a dataset containing multiple traffic signs and targets, the system improved the accuracy and robustness of detection, and achieved high recognition accuracy in experiments, which can be effectively used in actual driving environments [14]. Siripatatanadilok et al. proposed an OD approach based on Faster R-CNN to address the high-intensity operation problem of relying on manual detection of molting status in soft shell crab farming. By optimizing FE and bounding box confidence screening, the accuracy of crab detection was improved, and automatic recognition of occluded environments was achieved, reducing manual monitoring work and improving production efficiency [15]. The summaries of each literature are shown in Table 1.

Table 1: Comparative analysis of OD methods in related studies.

Study (Author & Year)	Method Framework	Application Scenario	Advantages	Limitations or Drawbacks
[8] Yang M (2022)	Multi-strategy Region Proposal Network	Autonomous driving, vehicle sensing	Improves detection accuracy and safety, suitable for dynamic environments	Lacks analysis for occlusion handling and small object cases
[9] Mahaur B et al. (2022)	Comparative analysis of deep models	General OD in AD	Systematically evaluates five DL models, helpful for deployment decisions	No new method proposed, lacks unified metric settings
[10] Tan K et al. (2024)	CV + AI image processing pipeline	Environmental perception, planning	Enhances lane keeping, obstacle and sign recognition capabilities	Fails to address dense object or illumination interference
[11] Arora N et al. (2022)	Faster R-CNN	Day/night vehicle detection	Improves recognition in low-light and data-sparse scenarios	Focused on dataset, lacks structural improvements

[12] Rani S et al. (2022)	Faster R-CNN + Line Feature Logic	General OD	Strengthens geometry-based features, increases speed and accuracy	Not suitable for small objects or occluded targets
[13] Yusro M M et al. (2023)	Faster R-CNN + Feature Filter Layers	Overlapping OD	Enhances feature separation, improves overlapping target identification	Increased model complexity and computational cost
[14] Güney E et al. (2022)	Faster R-CNN	Traffic sign and road OD	High accuracy and robustness in real environments	Limited performance under multi-target dense scenarios
[15] Siripattanadilok W et al. (2024)	Faster R-CNN + Grad-CAM	Occlusion detection in aquaculture	Accurate under occlusion, reduces manual labor for shell detection	Limited application scope, generalization not yet verified

Existing research has made significant progress in self-driving OD and Faster R-CNN optimization, covering optimization of object recognition algorithms, improvement of detection speed and accuracy, and other aspects. Nevertheless, current methods continue to grapple with issues like false alarms, missed detections, and inadequate real-time processing capabilities in intricate environments. This is particularly evident in scenarios involving lighting fluctuations, target occlusions, and dense OD situations, indicating that there is significant potential for improvement and optimization. Therefore, the research proposes an improved Faster R-CNN-based self-driving OD method. The novelty of this research resides in the integration of an enhanced CBAM to amplify the model's capacity for representing target features. It further refines target localization precision by leveraging ROI-Align and employs Soft-NMS to boost the detection efficacy of densely packed targets. Collectively, these measures significantly enhance the detection robustness and accuracy of the Faster R-CNN model in the challenging context of autonomous driving scenarios.

3 Methods and materials

This section offers an in-depth overview of the proposed autonomous vehicle OD and recognition model based on improved Faster R-CNN. Firstly, an improved CBAM is introduced to enhance feature expression, and ROI-Align is used to optimize target localization. Secondly, Soft-NMS is used to improve target screening accuracy, and multi-scale feature fusion is combined to optimize small OD. Finally, a complete model is constructed.

3.1 FE and target localization optimization based on attention mechanism

The OD and recognition of autonomous vehicles is the core task of perception systems, which directly affects the vehicle's understanding and decision-making in complex environments [16]. Faster R-CNN has become one of the mainstream detection frameworks due to its high detection accuracy and robustness [17]. However, in the complex and ever-changing self-driving environment, Faster R-CNN still has certain limitations. Therefore, the study focuses on Faster R-CNN and proposes an improved model for OD and recognition in self-driving vehicles, thereby enhancing the adaptability and stability of detection and recognition in the field of self-driving. The framework of Faster R-CNN is in Figure 1.

As presented in Figure 1, Faster R-CNN is a two-stage OD framework that mainly includes convolutional FE network, region recommendation network, and Faster R-CNN detection network. Convolutional FE networks are used to extract multi-level depth features from input images, and then output feature maps (FMs) that preserve the spatial structure information of the target. The regional recommendation network plays a pivotal role by generating potential target regions, also known as candidate regions, on the FM. Subsequently, it meticulously filters out regions that are likely to contain the target while simultaneously eliminating irrelevant background information. Faster R-CNN detects the candidate regions generated by the network for receiving area suggestions, further classifies and accurately locates each region, and finally completes the detection.

In the OD and recognition of autonomous vehicles, it is necessary to accurately extract target features and efficiently locate the target area to ensure detection stability in complex environments. However, the FE and target localization modules in Faster R-CNN have problems such as insufficient recognition ability for small and occluded targets, and large alignment errors in the target area [18]. Therefore, the study first introduces an improved CBAM, the goal of which is to enhance the detection ability and positioning accuracy of small targets and occluded targets in complex scenarios. The structure of the improved CBAM is in Figure 2.

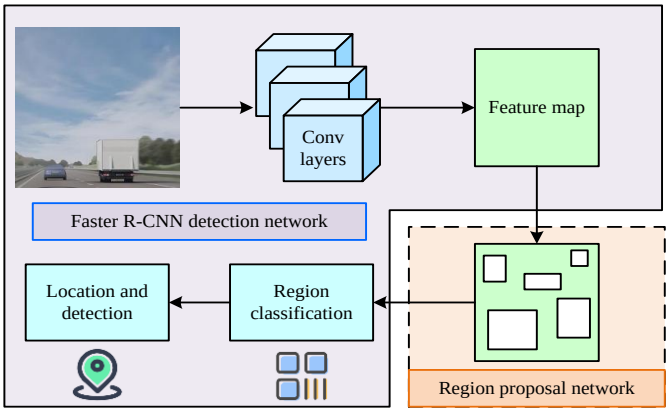


Figure 1: Structure diagram of Faster R-CNN.

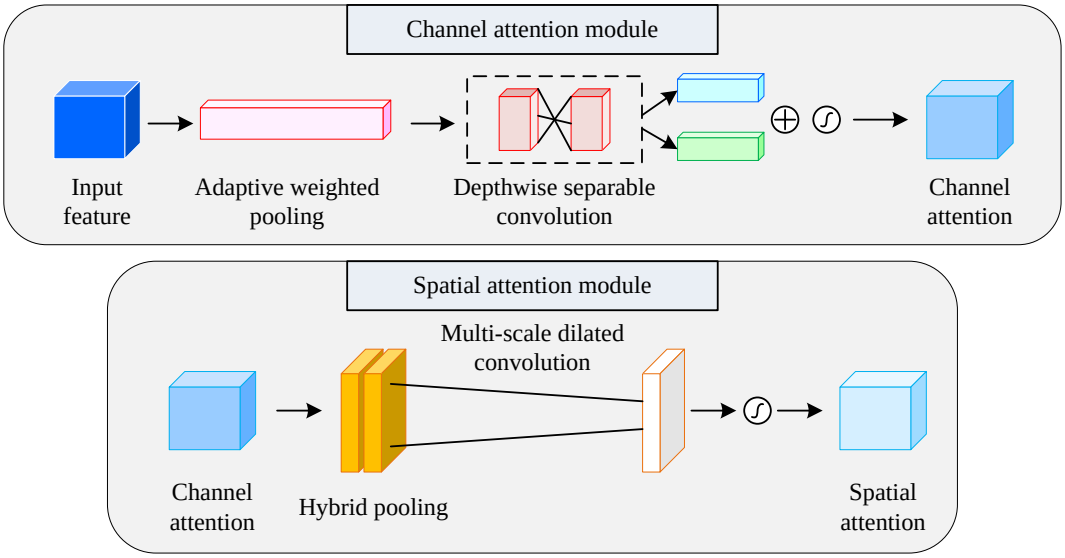


Figure 2: Schematic diagram of the structure of the improved CBAM.

As shown in Figure 2, the specific improvement studied for the channel attention (CA) module in CBAM is the introduction of adaptive weighted pooling, which integrates three strategies: global average pooling, global maximum pooling, and adaptive pooling, to enhance the expression ability of channel dimension features, as shown in equation (1).

$$M_c = \sigma\left(MLP\left(\omega_1 \cdot AvgPool\left(F\right) + \omega_2 \cdot MaxPool\left(F\right) + \omega_3 \cdot AdaptivePool\left(F\right)\right)\right) \quad (1)$$

In equation (1), M_c represents the CA weight, F represents the input FM, $AvgPool\left(F\right)$, $MaxPool\left(F\right)$, and $AdaptivePool\left(F\right)$ respectively represents the global average pooling, global maximum pooling, and adaptive pooling operations. Among them, the adaptive pooling operation can dynamically adjust the size of the pooling window, making FE more flexible. ω_1, ω_2, and ω_3 represent the learned weight parameters. In addition, the study uses depth-wise separable convolution instead of traditional fully connected layers, as shown in equation (2).

$$MLP\left(x\right) = Conv_{1\times1}\left(ReLU\left(BN\left(Conv_{d\times d}^{dw}\left(x\right)\right)\right)\right) \quad (2)$$

In equation (2), $Conv_{1\times1}$ represents a 1×1 convolution operation, BN is the batch normalization, $ReLU$ is the non-linear activation function, and $Conv_{d\times d}^{dw}$ represents the depth-wise convolution on a per-channel basis. Compared to the original fully connected layer, depth-wise separable convolution can reduce the number of parameters, improve computational efficiency, and make the network more efficient in FE. For the spatial attention module in CBAM, a multi-scale dilated convolution is used to replace the original single convolutional layer, allowing the features of different receptive fields to be more fully fused. The calculation formula is in equation (3).

$$M_s = \sigma\left(\left[Conv_{3\times3}^{d=1}\left(F'\right), Conv_{3\times3}^{d=3}\left(F'\right), Conv_{3\times3}^{d=5}\left(F'\right)\right]\right) \quad (3)$$

In equation (3), M_s represents the spatial attention weight. $Conv_{3\times3}^{d}$ represents the 3×3 convolution operation with dilation rate d. $Conv_{3\times3}^{d=1}$, $Conv_{3\times3}^{d=3}$, and $Conv_{3\times3}^{d=5}$ represent the dilated convolutions with dilation rates of 1, 3, and 5, respectively. F' indicates the enhanced features of the channel. In addition, to further improve the accuracy of FE, a hybrid pooling strategy is introduced, which integrates global pooling, local pooling, and adaptive pooling to make spatial attention calculation more flexible. The final calculation method is in equation (4).

$$\begin{cases} M_s = \sigma\left(\left[P, Conv_{3\times3}^{d=3}(F'), Conv_{3\times3}^{d=5}(F')\right]\right) \\ P = Conv_{3\times3}^{d=1}\left(\omega_1 \cdot AvgPool(F') + \omega_2 \cdot MaxPool(F') + \omega_3 \cdot AdaptivePool(F')\right) \end{cases} \quad (4)$$

After improvement, CBAM enables Faster R-CNN to extract target features more accurately in complex environments. The study adopts ROI-Align to optimize the target positioning accuracy in Faster R-CNN. The primary objective is to address the quantization error issue stemming from traditional RoI Pooling. This aims to enhance the alignment precision and stability of bounding boxes, especially when handling small-scale targets and those undergoing complex deformations. Consequently, the goal is to bolster the detection system's robustness across diverse scales and within challenging environmental conditions. The schematic diagram of ROI-Align process is presented in Figure 3.

As presented in Figure 3, ROI-Align optimizes the feature alignment accuracy of candidate regions during the FE process in the target area by maintaining floating-point coordinate information and using bilinear interpolation calculation. Firstly, the input image is processed by a convolutional neural network to extract an FM. The red box represents the mapping position of the candidate region on the FM. Subsequently, the candidate region is divided into fixed sized grids, with several uniformly distributed sampling points selected within each grid, rather than aligning integer pixels like ROI pooling. At each sampling point position, feature

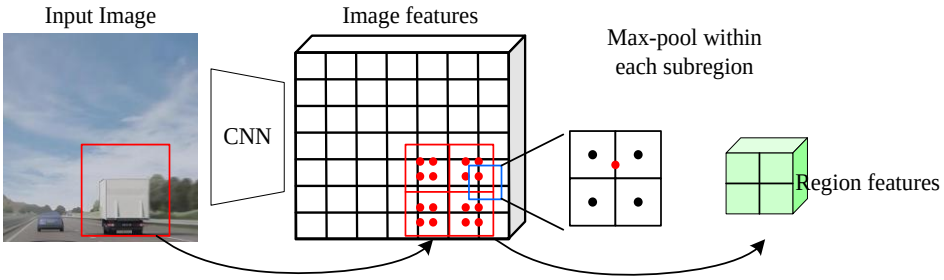


Figure 3: Schematic diagram of ROI-Align process.

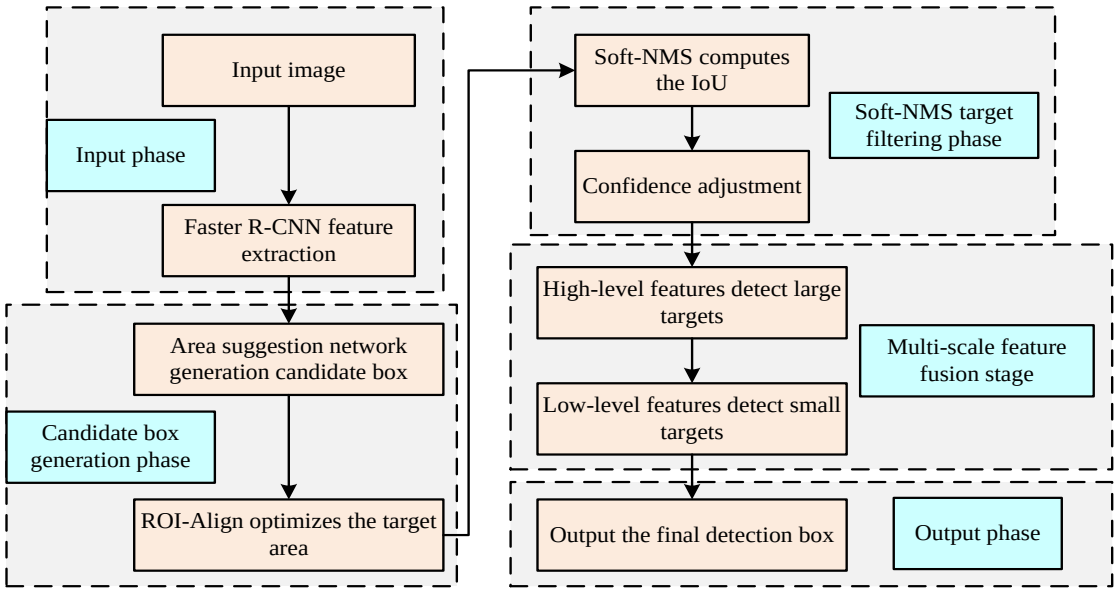


Figure 4: Optimization process of target screening and detection accuracy.

values are calculated through bilinear interpolation to avoid quantization errors caused by discretization. Ultimately, the features of all pooled windows are integrated to form a precisely aligned target feature representation. Through this method, ROI-Align can improve the localization accuracy of target bounding boxes while maintaining feature space continuity. In self-driving, it can effectively reduce target box jitter and information loss caused by feature alignment errors when detecting small and dense target vehicle groups, thereby improving the detection performance of Faster R-CNN.

3.2 Faster R-CNN target screening and detection accuracy optimization

After optimizing the FE and target localization of Faster R-CNN, the model can improve the accuracy of target feature expression and region alignment. However, in complex traffic environments, traditional target screening methods can easily lead to missed or false detections of dense and occluded targets, affecting detection stability [19-20]. In pursuit of the objective of elevating the target retention rate and ensuring recognition integrity within dense environments, the study incorporates Soft-NMS to

refine the target screening strategy. Additionally, it integrates multi-scale feature fusion to augment the detection capacity for targets of varying sizes. These measures collectively contribute to a further enhancement of the target screening capability and detection accuracy of the Faster R-CNN model in autonomous driving scenarios. The optimization process is in Figure 4.

As shown in Figure 4, the process consists of five stages, among which the two most critical stages are the Soft-NMS target screening stage and the multi-scale feature fusion stage. Soft-NMS first calculates the Intersection over Union (IoU) between candidate boxes and the highest confidence target box, and then uses a confidence adjustment strategy to weaken the influence of some low confidence candidate boxes, avoiding the problem of target deletion caused by hard suppression in traditional NMS. The confidence adjustment formula is in equation (5).

$$S_i' = S_i \times e^{-\frac{IoU(i,j)^2}{\sigma}} \tag{5}$$

In equation (5), S_i' and S_i represent the confidence levels of candidate box i before and after the update, $IoU(i,j)$ represents the IoU ratio between candidate box i and the highest confidence target box j, and σ is the hyper-parameter that controls the degree of confidence decay. This method ensures that in dense OD tasks, even if some target boxes have a high IoU, Soft-NMS will not directly remove them, but dynamically reduce their confidence based on IoU, thereby improving the integrity of target screening and reducing missed detections caused by occlusion or dense targets. In the stage of multi-scale feature fusion, this study combines high-level and low-level features to enable effective feature expression for targets of different scales. High level features contain rich semantic information and are suitable for detecting large targets, while low-level features retain more spatial details and are more sensitive to small targets. Multi-scale feature fusion is in equation (6).

$$F_{fusion} = \sum_{s=1}^{N} \alpha_s \cdot G(F_s) + \beta_s \cdot H(F_s) \tag{6}$$

In equation (6), F_{fusion} is the fused FM, N represents the number of feature layers of different scales fused, with a value of 3. The FMs at three different levels, C3, C4, and C5, in the backbone network ResNet-50 are fused respectively. The fusion structure is designed in reference to FPN. The shallow high-resolution features and deep semantic features are upsampled and aligned through horizontal connections, and integrated by convolution and attention weighting mechanisms. F_s represents the FM at the s thlayer, $G(F_s)$ and $H(F_s)$ represent the transformation operations of features at different scales. $G(F_s)$ uses 1×1 convolution for dimensionality reduction and batch normalization to reduce scale differences. $H(F_s)$ uses CA mechanism to enhance the weights of important feature channels. α_s and β_s are learned weight parameters, adaptively controlling the contribution of different hierarchical features in fusion. Feature fusion enables small targets to fully utilize the spatial details of low-level features, while large targets can use high-level features for precise classification and localization. Ultimately, by combining Soft-NMS with multi-scale feature fusion, Faster R-CNN can more accurately detect dense targets, occluded targets, and small targets in self-driving environments. Therefore, the final Faster R-CNN structure after improvement is in Figure 5.

As illustrated in Figure 5, the improved Faster R-CNN model is primarily composed of an enhanced backbone network, a Region Proposal Network (RPN), an ROI-Align module, a Soft-NMS target filtering mechanism, and a multi-scale feature fusion module. Specifically, ResNet-50 is adopted as the backbone network, integrated with a modified CBAM to enhance feature expression. This attention mechanism re-calibrates weights along both channel and spatial dimensions, enabling the model to focus more effectively on salient features of small and occluded objects, while maintaining low parameter overhead and ensuring efficient gradient flow. The RPN is responsible for generating high-quality candidate object proposals from the FM through a sliding-window mechanism, performing foreground/background classification and bounding box regression. To further improve spatial alignment accuracy of object regions, the ROI-Align module is introduced. By eliminating the quantization operations of traditional ROI Pooling and using bilinear interpolation, it accurately extracts features from candidate regions, significantly reducing localization errors and enhancing fine-grained recognition. In the target selection phase, Soft-NMS replaces traditional NMS, adaptively adjusting the confidence decay of overlapping detections. This approach proves especially advantageous in dense and occluded situations, effectively curbing the occurrence of missed detections and enhancing the system's sensitivity. The multi-scale feature fusion module integrates features from different levels of the backbone, strengthening the model's ability to detect targets of varying sizes. Overall, the proposed architecture forms a tightly coupled system in terms of functional flow and information transmission. It ensures precise FE while optimizing bounding box regression and filtering strategies. This comprehensive design balances detection accuracy, target sensitivity, and computational efficiency, demonstrating superior performance in complex driving environments. Therefore, the process of OD and recognition for autonomous vehicles based on improved Faster R-CNN is in Figure 6.

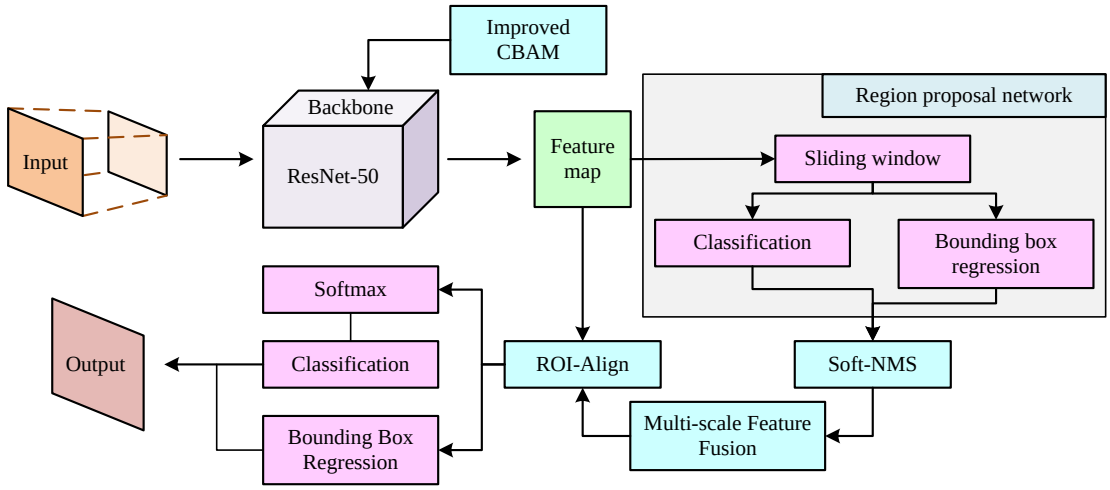


Figure 5: Schematic diagram of the improved Faster R-CNN structure.

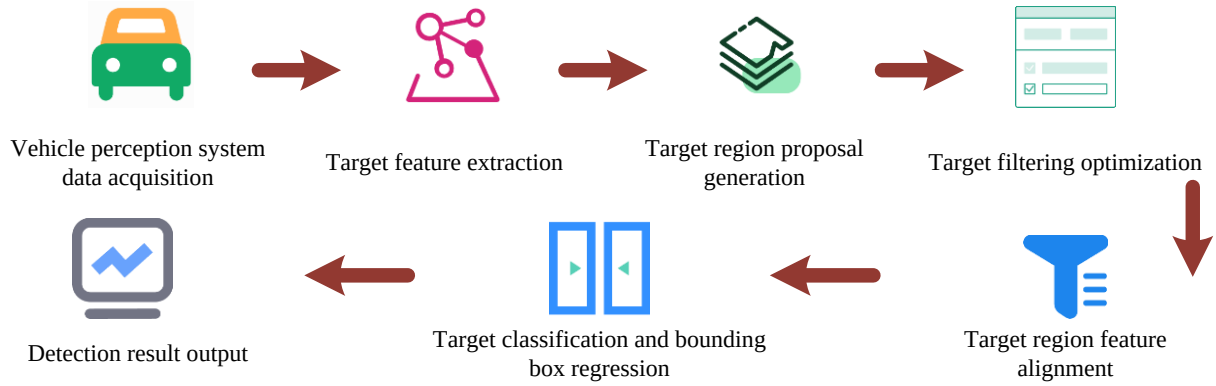


Figure 6: OD and recognition of autonomous vehicles based on improved Faster R-CNN.

As shown in Figure 6, autonomous vehicles first collect road data through cameras or LiDAR, and perform preprocessing to optimize image quality. Subsequently, the ResNet-50 backbone network is combined with improved CBAM to extract key target features, and the region recommendation network is used to generate candidate target boxes for classification and regression. Next, Soft-NMS filters the target boxes to reduce accidental deletion of dense targets, while ROI-Align aligns the features of the target area to improve localization accuracy. Scale feature fusion combines high-level and low-level features to optimize the detection capability of targets of different sizes. Finally, the Softmax classifier completes the target classification, the boundary box regression optimizes the target box position, outputs the detection results, and is used for decision-making of the auto drive system.

4 Results

To confirm the effectiveness and superiority of the raised autonomous vehicle OD and recognition method based on improved Faster R-CNN, the KITTI dataset and nuScenes dataset were selected as experimental data sources. The KITTI datasetranks among the most extensively utilized datasets within the domain of self-driving research, which includes targets such as vehicles, pedestrians, and cyclists in real road scenes. The nuScenes dataset is a large-scale self-driving perception dataset that contains complete 360° perception data, including sensor information such as cameras, LiDAR, and millimeter wave radar, providing richer environmental perception capabilities for self-driving OD.

For data partitioning, the KITTI dataset was split into training and validation sets at an 8:2 ratio, ensuring a balanced distribution of object categories and occlusion levels. For the nuScenes dataset, the official split standard was adopted, using 7,000 frames for training and 1,500 frames for validation to support model training and evaluation. To simulate diverse environmental conditions and enhance model robustness, consistent data augmentation techniques were applied to both datasets, including brightness variation, rotation and cropping, scale transformation, and noise perturbation, thereby improving adaptability to varying scenes and object sizes. In addition, to address the low proportion of occluded and small objects, the training sets were supplemented by oversampling instances with medium to high occlusion levels and object sizes smaller than 50 pixels. This strategy improved the model's detection capability and generalization performance under complex conditions. For the nuScenes dataset specifically, multi-modal data synchronization and annotation consistency checks were conducted to ensure temporal integrity and labeling accuracy, further enhancing model stability and adaptability in dynamic traffic environments. The

experimental environment and basic parameter settings are in Table 2.

Table 2: Experiment environment and basic parameter settings.

Configuration item	Parameter	Configuration item	Parameter
CPU	Intel Xeon W-2295	Optimizer	Adam
GPU	NVIDIA RTX 3090	Initial learning rate	0.0001
Memory	128GB DDR4	Learning rate adjustment strategy	Cosine annealing strategy
Storage	2TB NVMe SSD	Batch size	16
Operating system	Ubuntu 20.04	Training epochs	100
Deep learning framework	PyTorch 1.10.0 (Python 3.8)	Anchor-box scales	32×32, 64×64, 128×128
CUDA version	CUDA 11.3	Anchor-box aspect ratios	1:1, 1:2, 2:1
cuDNN version	cuDNN 8.2	ROI-Align pooling window	7×7

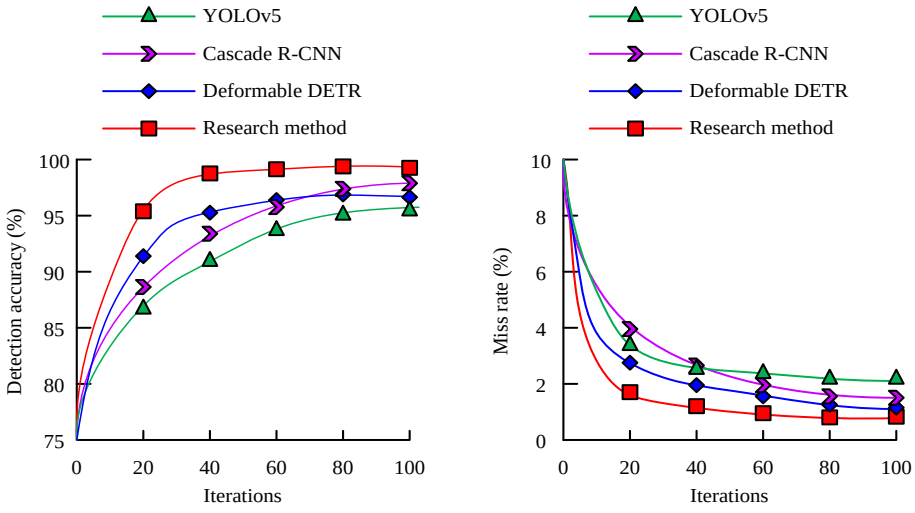
Table 3: Ablation study results.

Model Structure	mAP (%)	Miss Rate (%)	NMSE	Inference Time (ms/img)
Baseline Faster R-CNN	93.84	4.73	0.031	32.2
Faster R-CNN + CBAM	95.62	3.68	0.025	33.8
Faster R-CNN + CBAM + ROI-Align	96.97	2.54	0.017	34.7
Faster R-CNN + CBAM + ROI-Align + Soft-NMS	98.13	0.96	0.014	34.5

Based on Table 2, the ablation experiment was conducted first in the study. The alterations in detection performance were examined separately after sequentially integrating the CBAM attention module, ROI-Align region alignment module, and Soft-NMS screening mechanism into the benchmark Faster R-CNN architecture. An assessment was then carried out to evaluate the distinct contributions of each module towards enhancing target detection accuracy and robustness. The results are shown in Table 3 as follows.

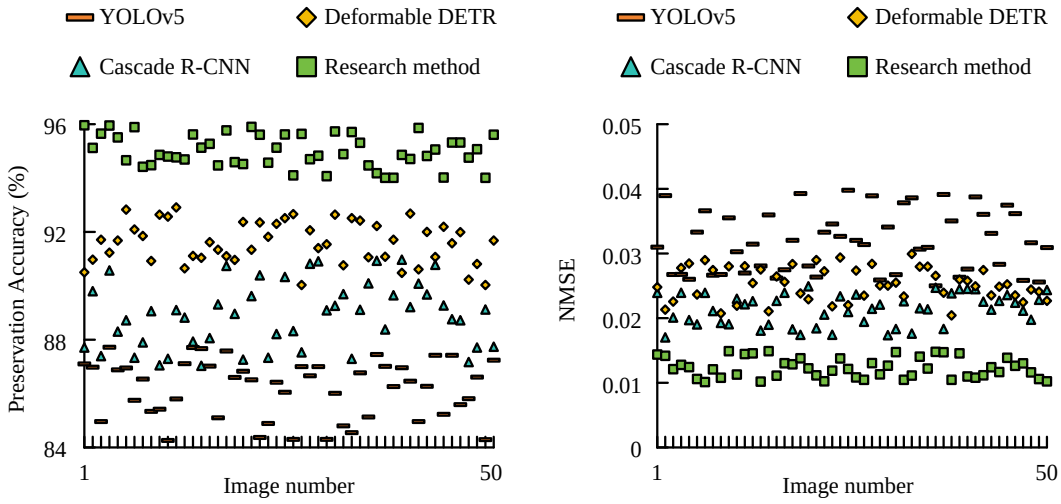
From Table 3, after adding CBAM, the mAP of the model increased by approximately 1.8%, and the missed detection rate decreased by approximately 1%, indicating that the attention mechanism enhanced the model's focus on the key target features. After further adding ROI-Align, the Normalized Mean Squared Error (NMSE) decreased to 0.017, indicating that the regional alignment module effectively reduced the bounding box offset and improved the target positioning accuracy. Finally, after introducing the Soft-NMS screening mechanism, the model further reduced the accidental deletion of targets in the dense target scenario, achieved the optimal comprehensive performance, with the mAP reaching 98.13% and the missed detection rate dropping to 0.96%. Further, You Only Look Once v5 (YOLOv5), Cascade Region-Based Convolutional Neural Network (Cascade R-CNN), and Deformable Detection Transformer (Deformable DETR) were selected as comparison methods. Among them, YOLOv5 has efficient and real-time detection capabilities, making it suitable for self-driving applications. Cascade R-CNN improves target localization accuracy through cascaded multi-level object box regression and is suitable for detecting dense and occluded targets. Deformable DETR combines deformable attention and deformable convolution to enhance the detection ability of complex environments and dynamic targets. The experimental results of the overall detection performance evaluation are in Figure 7.

As presented in Figure 7 (a), the improved Faster R-CNN maintained high detection accuracy in all iteration stages, with a detection accuracy of over 95.00% after 20 iterations, ultimately reaching 98.13%, which was better than Cascade R-CNN's 96.68%, Deformable DETR's 95.52%, and YOLOv5's 94.62%. This indicated that the research method had more advantages in FE and target localization. As shown in Figure 7 (b), the miss rates of all methods gradually decreased with the increase of iteration times. The improved Faster R-CNN reduced the miss rate to below 2.00% after 20 iterations, and finally stabilized at below 1.00%. The miss rates of Cascade R-CNN and Deformable DETR iterations were both between 1.50% and 2.00%, while YOLOv5 was above 2.50%, indicating that the overall detection performance of the research method was more excellent. On this basis, the study selected a self-driving image set containing dense targets in the dataset and tested the target screening performance and target area localization accuracy of various methods. The results are in Figure 8.



(a) Comparison of detection accuracy (b) Comparison of miss rate

Figure 7: Overall test performance evaluation experimental results.

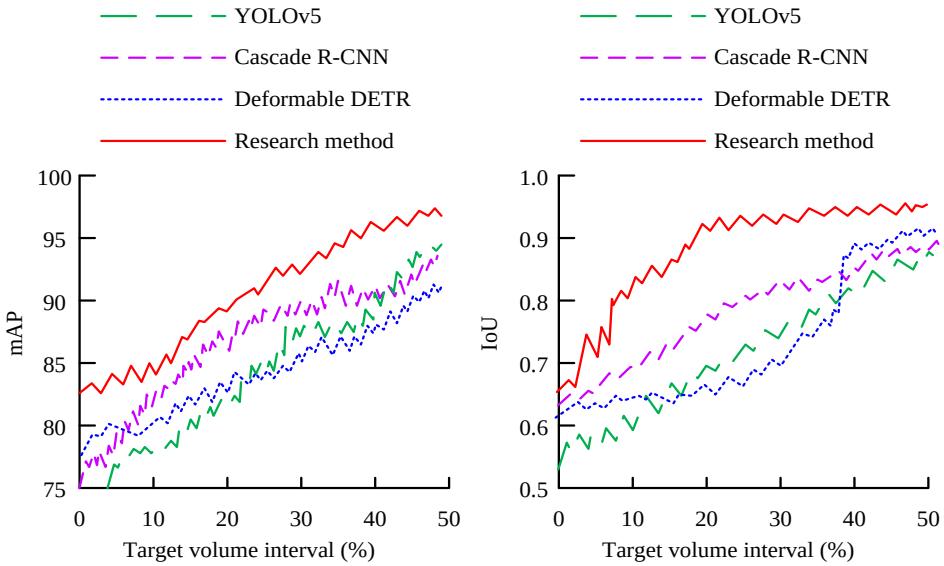


(a) Comparison of target screening performance (b) Comparison of positioning accuracy of target area

Figure 8: Performance comparison under dense targets.

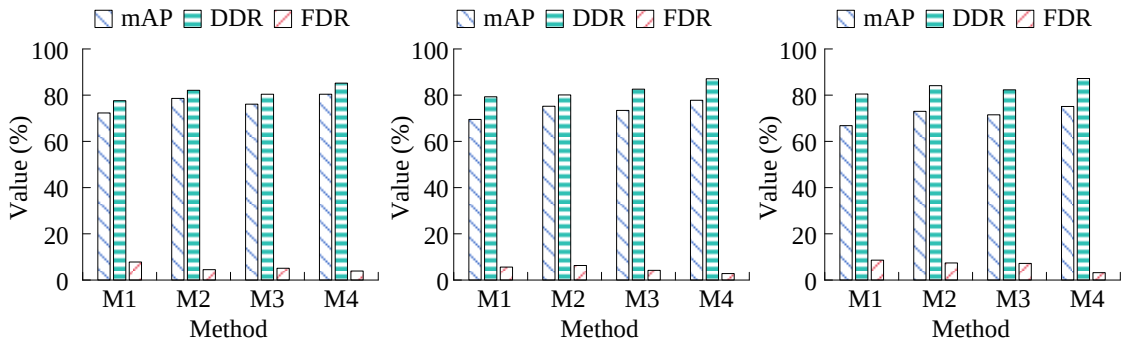
From Figure 8 (a), the improved Faster R-CNN achieved a preservation accuracy of 94.16%, which was 7.80% higher than YOLOv5's 86.36%. This indicated that Soft-NMS effectively reduced the false deletion of dense targets, and also had advantages compared to Cascade R-CNN's 91.52% and Deformable DETR's 89.83%. In Figure 8 (b), the average Normalized Mean Squared Error (NMSE) of the improved Faster R-CNN was 0.014, which was lower than YOLOv5's 0.032 and better than Cascade R-CNN's 0.020 and Deformable DETR's 0.023. This indicated that ROI-Align significantly reduced the target box localization error in target feature alignment and improved the stability and accuracy of OD. Furthermore, different scale OD capability tests were conducted, and the results are in Figure 9.

In Figure 9 (a), the improved Faster R-CNN maintained high detection accuracy across all target sizes. When the target volume interval reached 50%, the Mean Average Precision (mAP) reached 97.01%, while YOLOv5, Cascade R-CNN, and Deformable DETR only had 94.86%, 93.92%, and 92.64%, respectively. From Figure 9 (b), for IoU, the improved Faster R-CNN still performed the best in different target volume intervals, reaching a maximum of 0.93, while the IoU of the three comparison methods was all below 0.90.From this, improving the multi-scale feature fusion of Faster R-CNN had excellent optimization effects on OD of different sizes. Furthermore, the study conducted robustness testing on complex traffic environments, selecting three scenario data: changes in lighting (daytime, nighttime), severe weather (rainy, foggy), and dynamic interference (occlusion of targets, target deformation). The results are in Figure 10.



(a) Comparison of mAP (b) Comparison of IoU

Figure 9: Detection ability test of different scales.



(a) Illumination variation scenario (b) Adverse weather scenario (c) Dynamic interference scenario

Figure 10: Robustness test in complex traffic environment.

M1-M4 in Figure 10 represent YOLOv5, Cascade R-CNN, Deformable DETR, and Improved Faster R-CNN, respectively. As shown in Figure 10 (a), the mAP of the improved Faster R-CNN under illumination changes was 80.45%, which was higher than other methods. The Deformation Detection Rate (DDR) of the target was 85.26%, and the False Detection Rate (FDR) was 3.94%, which was better than YOLOv5 and Cascade R-CNN, indicating stronger adaptability to changes in lighting. According to Figure 10 (b), in rainy and foggy environments, the mAP of the improved Faster R-CNN was 77.83%, and the DDR was 87.12%, both of which were superior to other methods. The FDR was only 2.81%, the lowest, indicating higher detection stability in low contrast environments. From Figure 10 (c), in occlusion and target deformation scenarios, the mAP of the improved Faster R-CNN was 75.11%, DDR was 87.28%, and FDR was only 3.27%, all of which were better than other methods, indicating its stronger robustness and lower false detection rate in dynamic environments. Finally, a comparison of computational resource consumption was conducted, and the outcomes are in Table 4.

As shown in Table 4, the improved Faster R-CNN demonstrated superior performance in terms of inference time, memory usage, frame rate, and GPU utilization compared to Cascade R-CNN and Deformable DETR. Specifically, the improved model achieved an inference time of 34.5 ms/img, memory usage of 5.8 GB, frame rate of 29.0 FPS, and GPU utilization of 82.5%, striking a good balance between detection accuracy and computational cost. YOLOv5 had the fastest inference time of just 18.7 ms/img and the highest FPS at 53.5, with the lowest memory usage at 3.2 GB. However, its parameter count was only 23.8M, resulting in a smaller model that struggles to maintain stability in complex scenarios. Cascade R-CNN and Deformable DETR consumed more resources, with inference times of 42.3 ms/img and 50.1 ms/img, memory usage of 6.4 GB and 7.1 GB, FPS of 23.6 and 19.9, and GPU utilization both exceeding 85%, making them less suitable for deployment in real-time systems. Overall, the improved Faster R-CNN maintained high detection accuracy while achieving efficient computational performance, making it suitable for autonomous driving scenarios that demand both real-time processing and robustness. Future work could explore

model pruning and lightweight optimization techniques to enhance deployment on edge platforms.

Table 4: Comparison of computing resource consumption.

Method	Inference Time (ms/img)	Memory Usage (GB)	Model Parameters (M)	FPS (Frames/sec)	Avg. Layer Latency (ms)	GPU Utilization (%)
Improved Faster R-CNN	34.5	5.8	45.6	29.0	2.3	82.5
YOLOv5	18.7	3.2	23.8	53.5	1.1	69.2
Cascade R-CNN	42.3	6.4	52.1	23.6	3.1	87.4
Deformable DETR	50.1	7.1	60.3	19.9	3.5	90.8

5 Discussion

The improved Faster R-CNN model proposed in this study demonstrated superior performance over YOLOv5, Cascade R-CNN, and Deformable DETR in terms of detection accuracy, object localization, dense target recognition, and adaptability to complex environments. Compared with YOLOv5, the improved model consistently maintained high accuracy throughout the training process, ultimately reaching 98.13%, significantly higher than YOLOv5's 94.62%. In dense target scenarios, the retention accuracy improved from 86.36% to 94.16%. This improvement was mainly attributed to the two-stage detection architecture combined with the CBAM module for enhanced feature expression, and the Soft-NMS mechanism, which dynamically adjusted the confidence scores of overlapping boxes during target screening, effectively reducing missed detections in dense scenes. Compared with Cascade R-CNN, the proposed model also showed clear advantages in dense OD and bounding box localization, with a normalized mean squared error (NMSE) of 0.014, significantly better than Cascade R-CNN's 0.020. ROI-Align eliminated quantization errors inherent in traditional RoI Pooling through floating-point sampling during the alignment process, thus improving boundary fitting precision and ensuring stable performance under complex target distributions. In comparison with Deformable DETR, although the latter enhanced spatial adaptability through deformable attention mechanisms, its accuracy in dense and small OD was inferior. In multi-scale detection experiments, the improved Faster R-CNN achieved a mean average precision (mAP) of 97.01%, clearly surpassing Deformable DETR's 92.64%, and reached a maximum IoU of 0.93, higher than the sub-0.90 levels of the baseline methods. Additionally, in robustness tests involving lighting variation, adverse weather, and dynamic occlusion, the improved model outperformed all three baselines in terms of mAP, deformation detection rate (DDR), and false detection rate (FDR), particularly under nighttime and foggy conditions. Error trend analysis revealed that YOLOv5 often misses small or occluded objects, Cascade R-CNN suffered from bounding box instability under low contrast, and Deformable DETR tended to generate false positives in cluttered backgrounds. In contrast, the proposed model, with the integration of Soft-NMS and ROI-Align, effectively mitigated suppression errors and boundary misalignment, thereby maintaining higher detection accuracy and object consistency in complex environments. Overall, the improved Faster R-CNN exhibited stronger FE and alignment capabilities across various scenarios and object types, showing high practical value and strong potential for real-world applications.

6 Conclusion

In response to the challenges of low OD accuracy, missed detections of densely packed targets, and the difficulty in recognizing small-sized targets within the context of autonomous driving, this study presents an enhanced Faster R-CNN model. The model incorporates and optimizes the CBAM to boost FE capabilities. It employs ROI-Align to elevate positioning precision, utilizes Soft-NMS to refine the screening of dense targets, and integrates multi-scale feature fusion to enhance the detection of targets of varying sizes. The model performance was validated on the KITTI and nuScenes datasets through experiments. The overall detection accuracy reached 98.13%, higher than YOLOv5's 94.62%, Cascade R-CNN's 96.68%, and Deformable DETR's 95.52%. The miss rate was less than 1.00%, which was better than the comparative methods' 1.50% -2.50%.In dense target scenes, the preservation accuracy was 94.16%, and the NMSE was only 0.014, which was lower than the 0.020-0.032 of other methods, indicating more accurate target localization. In OD at different scales, when the target volume interval was 50%, the mAP reached 97.01% and the highest IoU was 0.93, which was better than the comparison method's 94.86% and IoU below 0.90, demonstrating excellent detection ability for small targets. In the robustness test of complex environments, the mAP of the model under lighting changes, severe weather, and dynamic interference scenarios were 80.45%, 77.83%, and 75.11%, respectively. The FDRs were 3.94%, 2.81%, and 3.27%, respectively, which were better than other methods, proving its stronger adaptability. In terms of computational performance, the inference time was 34.5ms/img and the FPS was 29.0, balancing efficiency and accuracy.

The theoretical contribution of the research lies in constructing a Faster R-CNN OD framework that

integrates the improved CBAM, ROI-Align, Soft-NMS and multi-scale feature fusion mechanisms, and systematically enhances the detection ability of the model in dense occlusion, small object recognition and complex environments. The empirical results showed that this method not only achieved better detection accuracy and robustness than the existing mainstream models on the KITTI and nuScenes datasets, but also took into account the inference speed and computing resource overhead. It has strong practical deployment value and is especially suitable for real-time perception tasks in autonomous driving.

7 Limitations and future research directions

However, in this study, a fixed Gaussian attenuation parameter σ (σ=0.5) was adopted in the Soft-NMS module, and no systematic sensitivity analysis was conducted. The influence of different σ values on the performance of dense target detection might were ignored. Subsequent studies can further explore the influence mechanism of the change of σ parameters on the detection accuracy to enhance the adaptability of the model in multiple scenarios. Secondly, this study mainly focused on quantitative experiments and did not present the visualization results of detection in complex scenarios in the manuscript. Subsequent studies will further demonstrate the adaptive performance of the model by combining typical image examples. In addition, the research mainly focused on the verification of the improved structure in terms of overall detection performance and adaptability to complex environments. No more fine-grained performance analysis was conducted on the detection effects of various types of targets and the distribution of false detection types. Further supplementation can be made in the future to enhance the comprehensiveness of the evaluation. Finally, the study only selected two mainstream datasets, KITTI and nuScenes, for model training and testing. Although typical urban roads and multi-modal perception scenarios were covered, there are still certain generalization limitations. It has not yet been tested on larger-scale or diversified autonomous driving datasets such as Waymo Open Dataset and BDD100K. In the future, the experimental scope can be further expanded to more comprehensively verify the adaptability and robustness of the model in different traffic environments and scenarios.

Fundings

The research is supported by: Key research Project of Anhui Provincial Department of Education: Research on Key Technologies of Ultra-High Precision Load Data Acquisition System (No. 2023AH052946); Key research project of Bengbu University: "License Plate Detection and Recognition System Based on Intelligent Image Processing" (No.2021ZR03zd).

References

[1] Jules Karangwa, Jun Liu, and Zixuan Zeng. Vehicle detection for autonomous driving: a review of algorithms and datasets. IEEE Transactions on Intelligent Transportation Systems, 24(11):11568-11594, 2023. https://doi.org/10.1109/TITS.2023.3292278

[2] Shuncheng Tang, Zhenya Zhang, Yi Zhang, Jixiang Zhou, Yan Guo, Shuang Liu, Shengjian Guo, Yan-Fu Li, Lei Ma, Yinxing Xue, and Yang Liu. A survey on automated driving system testing: Landscapes and trends. ACM Transactions on Software Engineering and Methodology, 32(5):1-62, 2023. https://doi.org/10.1145/3579642

[3] Shahbaz Khan, Muhammad Tufail, Muhammad Tahir Khan, Zubair Ahmad, Javaid Iqbal, and Arsalan Wasim. A novel framework for multiple ground target detection, recognition and inspection in precision agriculture applications using a UAV. Unmanned Systems, 10(1):45-56, 2022. https://doi.org/10.1142/S2301385022500029

[4] Yang Sun, Yuhang Zhang, Haiyang Wang, Jianhua Guo, Jiushuai Zheng, and Haonan Ning. SES-YOLOv8n: Automatic driving object detection algorithm based on improved YOLOv8. Signal, Image and Video Processing, 18(5):3983-3992, 2024. https://doi.org/10.1007/s11760-024-03003-9

[5] Zhenhua Tao and Wai Keng Ngui. A review of automatic driving target detection based on camera and millimeter wave radar fusion technology. International Journal of Automotive and Mechanical Engineering, 22(1):11965-11985, 2025. https://doi.org/10.15282/ijame.22.1.2025.3.0920

[6] Jameer Kotwal, Ramgopal Kashyap, and Mr Mohd Shafi Karim Pathan. Artificial driving based EfficientNet for automatic plant leaf disease classification. Multimedia Tools and Applications, 83(13):38209-38240, 2024. https://doi.org/10.1007/s11042-023-16882-w

[7] Srikanta Pal, Ayush Roy, Palaiahnakote Shivakumara, and Umapada Pal. Adapting a swin transformer for license plate number and text detection in drone images. Artificial Intelligence and Applications, 1(3):145-154, 2023. https://doi.org/10.47852/bonviewAIA3202549

[8] Min Yang. Research on vehicle automatic driving target perception technology based on improved MSRPN algorithm. Journal of Computational and Cognitive Engineering, 1(3):147-151, 2022. https://doi.org/10.47852/bonviewJCCE20514

[9] Bharat Mahaur, Navjot Singh, and K. K. Mishra. Road object detection: a comparative study of deep learning-based algorithms. Multimedia Tools and Applications, 81(10):14247-14282, 2022. https://doi.org/10.1007/s11042-022-12447-5

[10] Huthaifa I. Ashqar, Taqwa I. Alhadidi, Mohammed Elhenawy, and Nour O. Khanfar. Leveraging multimodal large language models (MLLMs) for enhanced object detection and scene understanding in thermal images for autonomous driving systems.

Automation, 5(4):508-526, 2024. https://doi.org/10.3390/automation5040029

[11] Nitika Arora, Yogesh Kumar, Rashmi Karkra, and Munish Kumar. Automatic vehicle detection system in different environment conditions using fast R-CNN. Multimedia Tools and Applications, 81(13):18715-18735, 2022. https://doi.org/10.1007/s11042-022-12347-8

[12] Shilpa Rani, Deepika Ghai, and Sandeep Kumar. Object detection and recognition using contour based edge detection and fast R-CNN. Multimedia Tools and Applications, 81(29):42183-42207, 2022. https://doi.org/10.1007/s11042-021-11446-2

[13] Muhamad Munawar Yusro, Rozniza Ali, and Muhammad Suzuri Hitam. Comparison of faster r-cnn and yolov5 for overlapping objects recognition. Baghdad Science Journal, 20(3):0893-0893, 2023. https://doi.org/10.21123/bsj.2022.7243

[14] Emin Güney and Cuneyt Bayilmis. An implementation of traffic signs and road objects detection using faster R-CNN. Sakarya University Journal of Computer and Information Sciences, 5(2):216-224, 2022. https://doi.org/10.35377/saucis...1073355

[15] Wanit Siripattanadilok and Thitirat Siriborvornratanakul. Recognition of partially occluded soft-shell mud crabs using Faster R-CNN and Grad-CAM. Aquaculture International, 32(3):2977-2997, 2024. https://doi.org/10.1007/s10499-023-01307-0

[16] V. Nisha Jenipher, and S. Radhika. Lung tumor cell classification with lightweight mobileNetV2 and attention-based SCAM enhanced faster R-CNN. Evolving Systems, 15(4):1381-1398, 2024. https://doi.org/10.1007/s12530-023-09564-3

[17] Abdulghani Abdulghani and Gonca Gökçe Menekşe Dalveren. Moving object detection in video with algorithms YOLO and faster R-CNN in different conditions. European Journal of Science and Technology, (33):40-54, 2022. https://doi.org/10.31590/ejosat.1013049

[18] Rosa Gonzales-Martínez, Javier Machacuay, Pedro Rotta, and César Chinguel. Hyperparameters tuning of faster R-CNN deep learning transfer for persistent object detection in radar images. IEEE Latin America Transactions, 20(4):677-685, 2022. https://doi.org/10.1109/TLA.2022.9675474

[19] Mohamed Othmani. A vehicle detection and tracking method for traffic video based on faster R-CNN. Multimedia Tools and Applications, 81(20):28347-28365, 2022. https://doi.org/10.1007/s11042-022-12715-4

[20] Yan Zhang and Qianjun Tang. Accelerating autonomy: an integrated perception digital platform for next generation self-driving cars using faster R-CNN and DeepLabV3. Soft Computing, 28(2):1633-1652, 2024. https://doi.org/10.1007/s00500-023-09510-0