electronics

Article

Enhancing YOLOv8's Performance in Complex Traffic Scenarios: Optimization Design for Handling Long-Distance Dependencies and Complex Feature Relationships

Bingyu Li [1,2], Qiao Meng [1,2,*], Xin Li [1,2], Zhijie Wang [1,2], Xin Liu [1,2] and Siyuan Kong [1,2]

[1] School of Computer Technology and Application, Qinghai University, Xining 810016, China; ys230854040300@qhu.edu.cn (B.L.); ys220854100361@qhu.edu.cn (X.L.); ys230854040325@qhu.edu.cn (Z.W.); ys230854040328@qhu.edu.cn (X.L.); ys230854100347@qhu.edu.cn (S.K.)

[2] Intelligent Computing and Application Laboratory of Qinghai Province, Xining 810016, China

[*] Correspondence: 2010990037@qhu.edu.cn

Abstract: In recent years, the field of deep learning and computer vision has increasingly focused on the problem of vehicle target detection, becoming the forefront of many technological innovations. YOLOv8, as an efficient vehicle target detection model, has achieved good results in many scenarios. However, when faced with complex traffic scenarios, such as occluded targets, small target detection, changes in lighting, and variable weather conditions, YOLOv8 still has insufficient detection accuracy and robustness. To address these issues, this paper delves into the optimization strategies of YOLOv8 in the field of vehicle target detection, focusing on the EMA module in the backbone part and replacing the original SPPF module with focal modulation technology, all of which effectively improved the model's performance. At the same time, modifications to the head part were approached with caution to avoid unnecessary interference with the original design. The experiment used the UA-DETRAC dataset, which contains a variety of traffic scenarios, a rich variety of vehicle types, and complex dynamic environments, making it suitable for evaluating and validating the performance of traffic monitoring systems. The 5-fold cross-validation method was used to ensure the reliability and comprehensiveness of the evaluation results. The final results showed that the improved model's precision rate increased from 0.859 to 0.961, the recall rate from 0.83 to 0.908, and the mAP50 from 0.881 to 0.962. Meanwhile, the optimized YOLOv8 model demonstrated strong robustness in terms of detection accuracy and the ability to adapt to complex environments.

Keywords: vehicle object detection; yolov8; attention mechanism; focal modulation; complex dependency handling

check for updates

Citation: Li, B.; Meng, Q.; Li, X.; Wang, Z.; Liu, X.; Kong, S. Enhancing YOLOv8's Performance in Complex Traffic Scenarios: Optimization Design for Handling Long-Distance Dependencies and Complex Feature Relationships. *Electronics* **2024**, *13*, 4411. https://doi.org/10.3390/electronics13224411

Academic Editors: Xin Geng and Manohar Das

Received: 12 September 2024
Revised: 22 October 2024
Accepted: 7 November 2024
Published: 11 November 2024

1. Introduction

Currently, the field of deep learning and computer vision is experiencing a growing interest in vehicle target detection. This technology plays a pivotal role in enhancing work efficiency and minimizing human resource consumption, particularly in applications like autonomous driving systems and intelligent traffic monitoring. As society increasingly prioritizes safety and efficiency, achieving high-precision detection has become a significant challenge in contemporary research [1].

In the realm of object detection, various methodologies exhibit distinct characteristics. For instance, Faster R-CNN combines a Region Proposal Network with ROI Pooling to achieve accurate detections, while YOLO employs a single-stage direct prediction strategy to enhance processing speed. However, both approaches face common challenges. Two-stage detectors like Faster R-CNN and Mask R-CNN are typically slower in inference times, which limits their utility in real-time applications. In contrast, single-stage detectors such as YOLO and SSD often struggle to accurately identify small and densely packed objects. Furthermore, advancements like RetinaNet, which utilizes feature pyramids, and

CenterNet, which adopts an anchor-free approach, introduce increased model complexity and vulnerability to noise, further impeding their widespread adoption.

Modern deep learning models significantly enhance vehicle target detection accuracy by managing complex reasoning; however, this capability also introduces its own set of challenges. Accurate target detection is crucial for navigating increasingly complex and diverse traffic scenarios. For instance, urban environments are characterized by a high volume of vehicles, traffic congestion, abrupt lighting changes, and variable weather conditions, all of which require detection models to exhibit heightened precision and robustness [2]. Furthermore, autonomous vehicles must recognize and differentiate various traffic participants—such as pedestrians, bicycles, and motorcycles—as well as different traffic signs and signals, which further elevates the demands on the models' recognition and generalization abilities.

Vehicle detection faces several key challenges, including difficulties in detecting occluded and small targets. In real-world scenarios, vehicles may be partially obscured by other vehicles or environmental objects, leading to the loss or distortion of crucial information [3]. This complicates the task for detection algorithms, which struggle to capture complete vehicle features, thereby affecting accurate identification and positioning. Similarly, detecting small targets presents significant challenges; vehicles can appear very small in complex scenes or at a distance, and are characterized by sparse pixel information and blurred boundaries [4]. Accurately extracting features and positioning these small targets requires high algorithmic precision and robust image processing capabilities to ensure reliable detection results.

Despite these challenges, researchers are actively exploring optimizations in network architectures and algorithms to enhance the precision and robustness of vehicle target detection. This endeavor is not only academically significant but also has a direct impact on the safety and efficiency of future intelligent transportation systems. Progress in this field is crucial for the development of autonomous driving technologies, traffic management, and overall road safety, making it a pivotal area of study with far-reaching practical implications [5].

Moreover, the application of these technologies on embedded systems and mobile devices is particularly critical. Efficient object detection must strike a balance between accuracy and the constraints of computational resources, as well as the demand for real-time performance. Deploying detection systems in resource-constrained environments requires careful equilibrium between algorithmic complexity and device capabilities, ensuring that detection processes remain both swift and accurate. This balance is essential for implementing intelligent transportation solutions that rely on real-time data processing and decision-making [6]. Therefore, high-precision vehicle target detection acts as both a catalyst for technological innovation and a crucial safeguard for the safe and sustainable development of society. This drives researchers to continuously pursue more advanced algorithms and solutions to tackle the increasingly complex challenges posed by modern traffic environments.

2. Related Work

The development of vehicle target detection technology has transitioned from traditional feature design to deep learning models. Traditional methods rely on manually crafted features, which are often insufficient for addressing the complexity and variability of real-world scenarios and have limited generalization capabilities. In contrast, deep learning techniques can automatically learn more effective feature representations from vast amounts of data, thereby improving detection accuracy across diverse scenarios.

Currently, mainstream deep learning vehicle target detection algorithms are primarily divided into two major categories: two-stage methods and one-stage methods. Two-stage methods (such as the R-CNN series [7]) first generate candidate regions using a Region Proposal Network (RPN) [8], and then each candidate region undergoes classification and bounding box regression. This approach can achieve a high level of detection accuracy, albeit at the expense of speed [9]. In contrast, one-stage methods (such as the YOLO series [10,11] and SSD [12]) directly generate bounding boxes across the entire image for object detection, resulting in faster performance but potentially sacrificing some accuracy. The advancements in deep learning technology have not only propelled the development of intelligent transportation systems but also demonstrated broad application prospects in both academic research and practical implementations. In the field of autonomous driving, accurate vehicle detection is a crucial step for achieving environmental perception and decision-making, significantly enhancing the performance of driver assistance systems. In intelligent traffic monitoring, effective object detection technology can improve the efficiency and safety of traffic management, reducing the likelihood of accidents.

Among these algorithms, YOLOv8 is built upon the architecture of YOLOv5 and introduces new structures to further enhance performance and scalability [13]. According to the criteria of depth and breadth, YOLOv8 is categorized into five versions: YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, and YOLOv8x. With each iteration, there is a notable increase in the number of model parameters and computational load, aimed at improving detection accuracy to meet the demands of various application scenarios [14]. YOLOv8 employs an adaptive image scaling strategy at the input stage, allowing for dynamic adjustment of input dimensions, and incorporates mosaic data augmentation techniques to enhance the model's robustness.

Currently, with the rapid advancement of deep learning and computer vision technologies, various modules in the field of object detection have demonstrated exceptional performance, providing robust support for improving detection accuracy and efficiency. These modules can significantly enhance the capabilities of models like YOLOv8, further optimizing their performance in complex environments. Notably, the following techniques stand out.

Scale-Perceptive Feature Fusion (SPFF) is a feature fusion technique that effectively integrates information across different scales, enhancing the model's ability to detect multi-scale targets. This technology is particularly crucial for improving the detection accuracy of small or distant objects. FocalNet, on the other hand, is a neural network architecture designed to enhance the detection of challenging targets by focusing attention on focal regions, thereby increasing the detection rate of hard-to-detect objects. Additionally, exponential moving average (EMA) is a weight update strategy used during the training of neural networks. By smoothing short-term fluctuations through the exponential moving average of the weights, EMA facilitates model convergence and enhances performance on test data.

In summary, YOLOv8 builds upon the rapid detection and high-precision characteristics of the YOLO series, further enhancing its versatility and practicality by optimizing network architecture and introducing new technologies. As the field of computer vision continues to evolve, the application prospects of YOLOv8 in areas such as intelligent transportation and security monitoring are poised to expand significantly, providing robust support and solutions for the realization of intelligent scenarios. The following section illustrates the network architecture of YOLOv8 (see Figure 1).

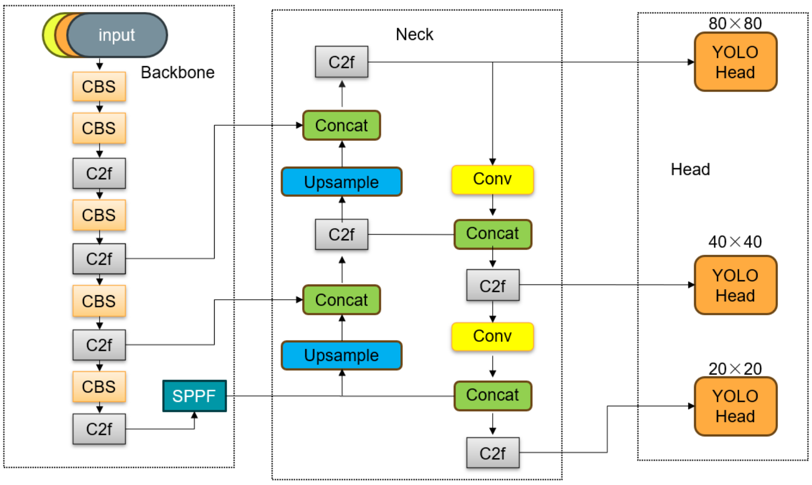

Figure 1. Yolov8 Structural Framework.

3. Optimization of YOLOv8's Multi-Scale Feature Integration and Complex Dependency Handling Methods

3.1. Reconstruction of the Backbone Network

Spatial Pyramid Pooling Fast SPPF and focal modulation networks (FocalNets) are two techniques in the field of object detection designed to handle input images of varying sizes. They exhibit significant differences in their principles, functionalities, and technical implementations.

SPPF is aimed at enhancing the speed and efficiency of the model when processing inputs of different dimensions. It achieves this by performing regional pooling operations at multiple scales, ensuring that input images of diverse sizes can produce output feature maps of uniform dimensions. The primary mechanism involves integrating features across different hierarchical levels through pooling operations, thereby enhancing the model's versatility and processing speed.The network architecture of SPPF is shown in Figure 2.

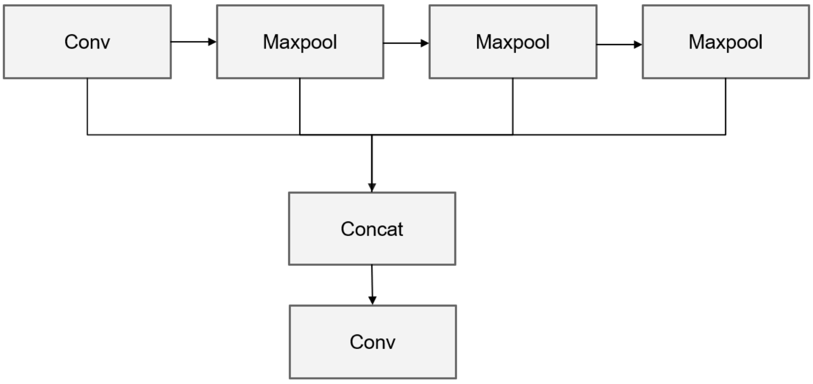

Figure 2. SPPF Structural Framework.

In comparison, FocalNets incorporate a focal modulation mechanism [15], designed to replace traditional self-attention modules (see Figure 3 for the Self-Attention structure). This mechanism, by employing gated aggregation and element-wise affine transformations, effectively captures long-distance dependencies within images and infuses multi-scale visual contextual information into each query token, thereby enhancing the model's perceptual capabilities and its ability to process complex scenes. By stacking deep convolutional layers and implementing focal modulation mechanisms, FocalNets achieve a more efficient integration of contextual information, enabling the network to improve both performance and efficiency when handling complex visual tasks. Below is a comparative diagram of the self-attention and FocalNets modules.

In contemporary computer vision research, the encoding strategies for feature mapping are of paramount importance for understanding the inherent structure and representation of visual data. This discussion delves into two distinct mechanisms for generating feature representations: self-attention and focal modulation (see Figure 4 for the Focal Modulation structure). The self-attention mechanism, as illustrated in Equation (1), involves y_i denoting the i-th output feature vector, with M_1 serving as a mapping function, T_1 as the initial transformation function, x_i representing the i-th input feature vector, and X encompassing the entire input feature mapping. This module relies on the interaction between the query and the comprehensive feature map, followed by an aggregation process. While this approach boasts strong expressive capabilities, it incurs a relatively high computational cost.

$$y_i = M_1(T_1(x_i, X) \cdot X) \tag{1}$$

$$y_i = T_2(M_2(i, X) \cdot x_i) \tag{2}$$

$$y_i = q(x_i) \cdot m(i, X)(3) \tag{3}$$

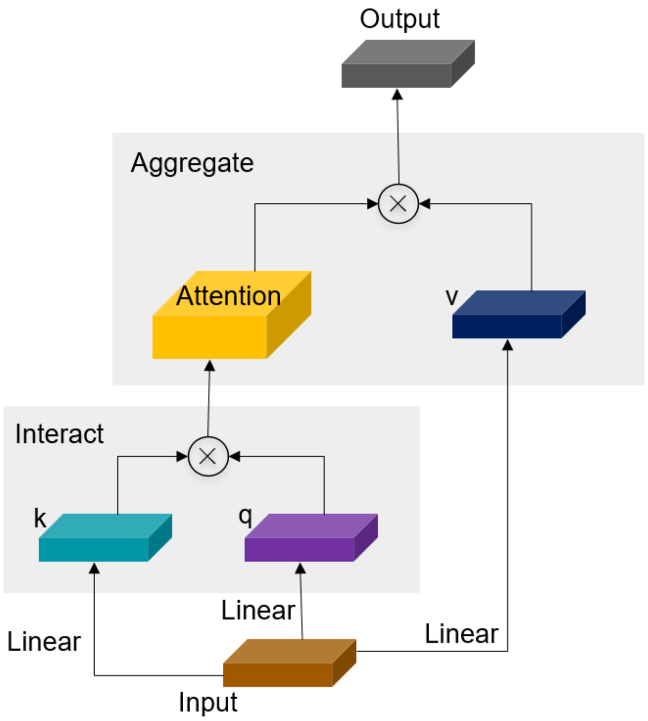

Figure 3. Self-Attention.

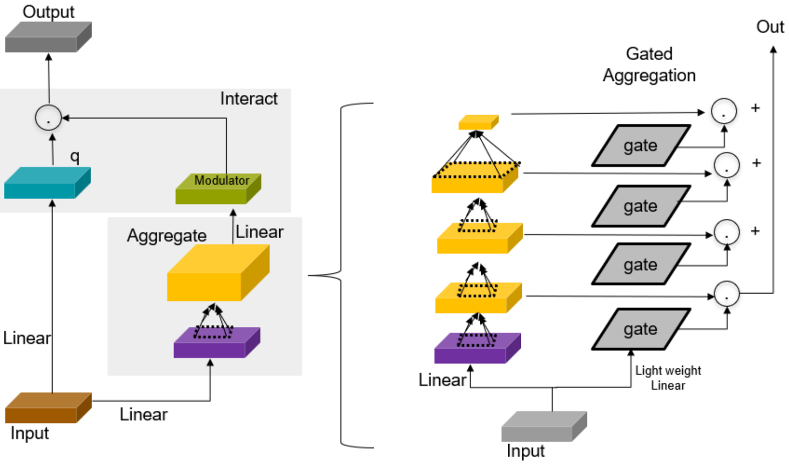

Figure 4. Focal Modulation.

In comparison, focal modulation (as shown in Equation (2)) employs an innovative early aggregation approach, with the specific implementation detailed in Equation (3). Here, y_i represents the i-th output feature vector, q denotes the query function, m signifies the modulation function, and i is the positional index of the feature. Feature representations are generated through query projection and context modulation specific to the feature's location [16]. The advantage of focal modulation lies in its translation invariance, sensitivity to input, and specific modulation in spatial and channel dimensions, while effectively separating and combining fine-grained features with contextual information. This approach not only enhances computational efficiency but also enriches the representation of features, providing new research directions and practical guidance for feature encoding in the field of computer vision.

$$\mathbf{Z}^{\ell} = f_{\alpha}^{\ell}(\mathbf{Z}^{\ell-1}) \triangleq \mathrm{GeLU}\Big(\mathrm{DWConv}(\mathbf{Z}^{\ell-1})\Big) \in \mathbb{R}^{H\times W\times C} \tag{4}$$

$$Z^{out} = \sum_{\ell=1}^{L+1} G^{\ell} \odot Z^{\ell} \in \mathbb{R}^{H \times W \times C} \tag{5}$$

In the FocalNet architecture, Equation (4) employs a hierarchical contextualization method, where $\mathbf{Z}^{\ell}$ represents the feature map of the ℓ layer, obtained by processing the feature map of the $(\ell - 1)$-th layer through depthwise separable convolution and the non-linear activation function GELU. This method combines depthwise separable convolution with the non-linear activation function GELU to recursively extract and modulate feature maps, capturing contextual information across multiple scales. This process not only enhances the model's understanding of image details and structural information but also optimizes computational efficiency. Subsequently, Equation (5) introduces a gated aggregation mechanism, where Z^{out} denotes the final output feature map, obtained by performing a weighted summation of feature maps from all levels. The weighted summation is conducted through gating weights G^{ℓ}, which determine the contribution of each level's feature map in the final output. This mechanism effectively integrates contextual information from local to global by applying spatially aware gating weights to the feature maps of different hierarchies [17]. This aggregation strategy enables the model to maintain fine details while also capturing broader contextual relationships, thereby enhancing the richness and flexibility of feature representation.

$$y_i = q(x_i) \odot \mathrm{h}\left(\sum_{\ell=1}^{L+1} g_i^{\ell} \cdot z_i^{\ell} \right) \tag{6}$$

In the final feature processing stage of the FocalNet architecture, Equation (6) plays a pivotal role, where y_i represents the i-th output feature, obtained by multiplying the query feature with the aggregated feature map. Here, $q(x_i)$ denotes the query feature, a feature vector associated with the input feature x_i; h represents a non-linear function that transforms the aggregated feature map to yield the final output feature y_i; and g_i signifies the modulation weight of the i-th layer, determining the contribution of the i-th layer's feature map in the final output. It combines the query feature with the aggregated contextual information through a context-sensitive modulation process, thereby refining output features with rich semantics. Although the elements of this formula have not been explicitly expanded, its essence lies in implementing an efficient feature fusion strategy, providing the model with a comprehensive and nuanced visual representation, which is considered an innovative and rigorous feature processing technique in the academic field.

3.2. Incorporation of EMA Attention Mechanism

This section will provide a detailed introduction to the efficient multi-scale attention (EMA) module, which holds significant potential for application in computer vision tasks. The EMA module, through innovative reorganization of channel and batch dimensions, as well as cross-dimensional interaction, significantly enhances the model's efficiency and accuracy in handling complex feature relationships (see Figure 5 for the EMA Structural Framework). Specifically, the * symbol in the figure represents element-wise multiplication, where each element of one tensor is multiplied by the corresponding element of another tensor. This operation is commonly used in deep learning models to fuse and weight features, effectively strengthening the relationships between features, while reducing computational complexity and improving operational efficiency.

One notable feature of the EMA module is the reorganization of the input feature's channel dimension. Specifically, the module reshapes some of the input feature channels into the batch dimension, thereby forming multiple sub-feature groups. This reorganization strategy not only effectively preserves the original information of each channel but also significantly reduces computational complexity, thereby improving the model's operational efficiency [18]. This innovative method of dimensional reorganization provides a solid foundation for subsequent feature extraction and optimization.

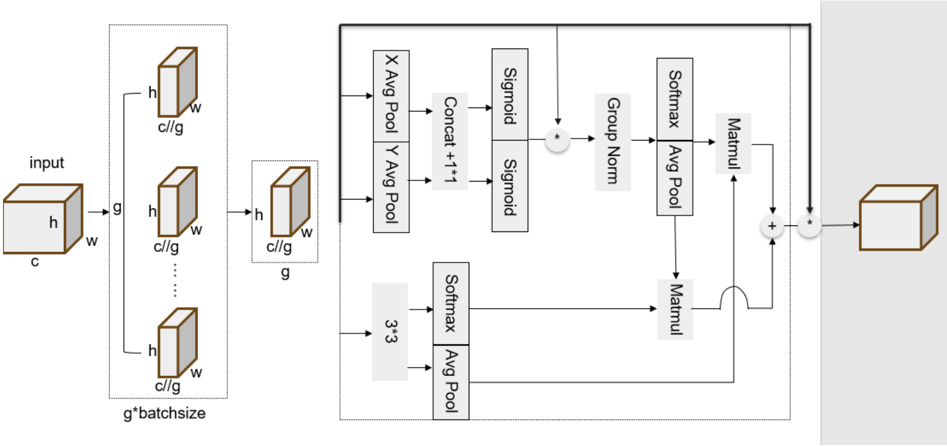

Figure 5. EMA Structural Framework.

$$z_c^H(H) = \frac{1}{W} \sum_{0 \leq i \leq W} x_c(H, i) \tag{7}$$

$$z_c^W(W) = \frac{1}{H} \sum_{0 \leq j \leq H} x_c(j, W) \tag{8}$$

Building upon the channel reorganization, the EMA module encodes global information through one-dimensional global pooling operations, such as horizontal (X Avg Pool) and vertical (Y Avg Pool) average pooling. Equations (7) and (8) describe this process, respectively. In Equation (7), $z_c^H(H)$ represents the global feature value of channel c in the height direction, where $x_c(H, i)$ represents the feature value of channel c at the position with height H and width i in the input feature map. H denotes the height of the input feature map, and W denotes the width. In Equation (8), $z_c^W(W)$ represents the global feature value of channel c in the width direction, where $x_c(j, W)$ represents the feature value of channel c at the position with height j and width W in the input feature map. These operations enable the module to capture the holistic features of various parts within the input feature map, providing essential information for feature modulation and optimization. Furthermore, global information encoding is also utilized to calibrate the channel weights in each parallel branch, enhancing the consistency and expressive power of the feature representation.

$$z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} x_c(i, j) \tag{9}$$

The EMA module is also characterized by its cross-dimensional interaction mechanism. Through this mechanism, the module can conduct an in-depth analysis of the interdependencies at the pixel level within the input feature map. Equation (9) describes the two-dimensional global pooling operation involved in the cross-dimensional interaction, where z_c represents the global feature value of channel c, and $x_c(i, j)$ represents the feature value of channel c at the position with height i and width j in the input feature map. This interaction mechanism optimizes the generation process of the feature mapping, allowing the model to capture complex feature relationships more finely, which is relatively rare in traditional attention models.

The operational process of the EMA module encompasses input grouping, parallel processing, modulation, and fusion, culminating in the output of the final feature mapping. Each input group is segmented according to a predefined number of groups and undergoes feature extraction and modulation through independent processing branches. Some branches perform global pooling operations, while others extract local features via convolutional operations. Ultimately, after modulation through the sigmoid function and normalization, the cross-dimensional interaction module is utilized for feature fusion, resulting in a feature mapping with high discriminative power [19]. The innovative advantage of the EMA module lies in its efficient multi-scale attention mechanism and cross-dimensional interaction capability. This not only significantly reduces computational overhead but also enables the precise capture of feature relationships, addressing the processing needs of multi-scale and complex visual scenes [20]. Global information encoding and channel weight calibration further enhance the model's integration and optimization capabilities, demonstrating significant advantages and application potential across a wide range of computer vision applications.

EMA can be viewed as the market's memory. Unlike simple moving averages that treat all historical data equally, EMA acts more like a selective historical recorder, focusing more on recent events. It functions as a forward-looking observer, giving greater weight to the latest data to help us concentrate on current dynamics and potential trends in the market. In this sense, EMA is not just a smoother of price movements but also a predictor of future trends. Its slope and shape reveal market sentiment and possible future directions. Additionally, this weighted nature makes EMA an effective noise filter and trend confirmation tool, helping us find order amidst the chaos of market information.

3.3. Two Modules Address the Issues

The following figure illustrates the reconstructed backbone structure (see Figure 6 for the Backbone Network).

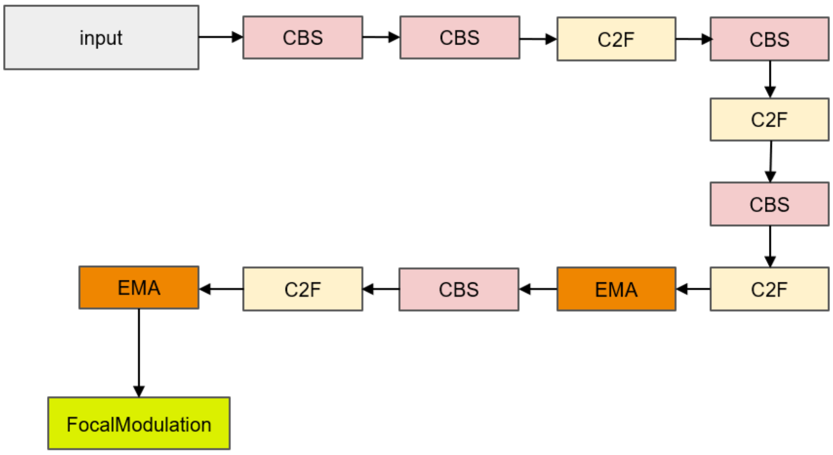

Figure 6. Reconstructing the Backbone Network.

The primary challenge addressed by FocalNets is the difficulty in handling long-range dependencies and complex contextual information. Traditional attention mechanisms, such as self-attention modules, often face issues of computational complexity and memory consumption when capturing these dependencies. FocalNets introduce a focal modulation mechanism that effectively captures long-range dependencies in images through gated aggregation and element-wise affine transformations, injecting multi-scale visual contextual information into each query token. This mechanism significantly enhances the model's perceptual ability and performance in complex scenes.

The EMA module primarily addresses the efficiency and accuracy of handling complex feature relationships. It improves the model's capability to process multi-scale features and intricate visual scenes through innovative reorganization of channel and batch dimensions, global information encoding, and cross-dimensional interaction. Specifically, the EMA module reshapes some channels into the batch dimension and employs a cross-dimensional interaction mechanism following global information encoding, effectively analyzing interdependencies at the pixel level in the input feature map. This operational process not only reduces computational costs but also enhances the model's sensitivity and expressive power for complex feature relationships.

4. Ablation Study

4.1. Dataset and Experimental Framework

This experiment utilizes the UA-DETRAC dataset, a challenging collection designed for multi-object traffic target detection. The dataset was captured using a Canon EOS 550D camera at 24 locations in Beijing and Tianjin, China, aiming to realistically reflect complex traffic scenarios. It comprises 8250 manually annotated vehicle targets and 1.21 million detailed bounding boxes [21]. Unlike traditional vehicle datasets, the UA-DETRAC dataset adopts a high-angle top-down perspective, aligning more closely with the needs of actual traffic monitoring and safety surveillance, contrasting sharply with the common horizon-view perspective of in-vehicle equipment. Each image may contain over 30 vehicles and encompasses a variety of weather conditions from day to night, including clear and rainy days. The scenarios in the dataset include not only regular roads but also complex traffic junctions such as crossroads and three-way intersections, fully demonstrating the diversity of real-world traffic situations.

In this experiment, YOLOv8n serves as the base model for training, with a learning rate set to 0.001, a batch size of 16, and 50 epochs to ensure adequate training and convergence of the model. The training data from the last epoch is selected for comparison.

4.2. Evaluation and Analysis of Experimental Results

This experiment evaluates the impact of introducing the EMA attention mechanism and replacing the SPPF module with focal modulation on the performance of target detection in the YOLOv8 model (see Table 1 for the Ablation Experiment Results Comparison).

Table 1. Ablation Experiment Results Comparison.

Replace Sppf	Backbone+EMA	Head+EMA	Precision	Recall	Map50
no	no	no	0.61	0.528	0.576
yes	no	no	0.627	0.523	0.577
yes	no	yes	0.594	0.529	0.559
yes	yes	yes	0.616	0.551	0.57
yes	yes	no	0.681	0.557	0.604

The experiment initially assessed the impact of incorporating the EMA attention mechanism and substituting the SPPF module with focal modulation on the performance of the YOLOv8 model for target detection (see Figure 7 for the comparison of original and our data). The baseline model, without modifications, achieved an accuracy of 0.61, a recall rate of 0.528, and an mAP50 of 0.576. In the experiment where SPPF was replaced with focal modulation, accuracy increased to 0.627, although recall slightly decreased to 0.523, with mAP50 remaining at 0.577. When SPPF was replaced with focal modulation and EMA was added to the head part, precision dropped to 0.594, while recall slightly increased to 0.529, and mAP50 decreased to 0.559. Concurrently replacing SPPF with focal modulation and adding EMA to both the backbone and head parts led to a slight increase in precision to 0.616, with recall at 0.551 and mAP50 at 0.570. Furthermore, substituting SPPF with focal modulation and inserting the EMA module into the backbone resulted in a significant enhancement in accuracy to 0.681, with corresponding improvements in recall rate and mAP50 to 0.557 and 0.604, respectively.

The experimental results indicate that applying EMA in the backbone section can effectively enhance the model's precision and mAP50 metrics. Additionally, replacing the SPPF module with focal modulation can improve the model's performance indicators. However, introducing the attention mechanism in the head part has limited effects on enhancing model performance and may even lead to a decrease, necessitating cautious consideration of modifications to different parts of the model to avoid adverse impacts on overall performance.

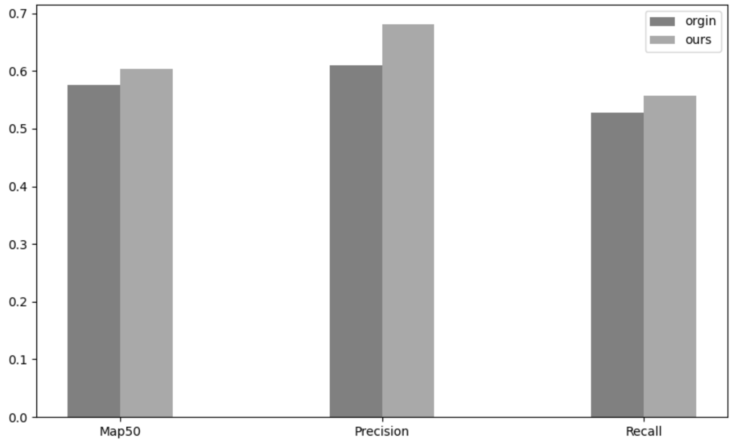

Figure 7. Comparison of Original and Our Data.

From the aforementioned experimental outcomes and analysis, the following conclusions can be drawn: In the head part, YOLOv8 has already considered the requirements of the target detection task in its design, including effective feature extraction and adaptation to object detection at various scales. Adding an EMA module to the head part might overlap with YOLOv8's existing feature processing mechanisms, leading to redundant functions or counterproductive effects rather than benefits. Introducing the EMA module at an earlier stage of the model can optimize feature representation in advance, aiding in faster convergence to a better solution. This reduces the burden on subsequent layers and ensures the training efficiency and stability of the entire model. In summary, the primary reason for the performance enhancement of the model when the EMA module is incorporated into the backbone part of YOLOv8 is its ability to enhance the expressiveness of features, optimize feature fusion and spatial relationships, aid in model training and convergence, and be compatible with the overall architecture of YOLOv8. The combined effect of these factors makes the introduction of the EMA module in the backbone part of YOLOv8 effective in improving the performance of the target detection model.

5. Model Performance Analysis

5.1. Model Performance Evaluation Using 5-Fold Cross-Validation

Due to the imbalance in the number of samples of different categories in the UA-DETRAC dataset, we have decided to use k-fold cross-validation to evaluate the trained model. This method can effectively assess the performance of the model, ensuring accurate evaluation results across all categories (see Figure 8 for the Five-Fold Cross-Validation process).

We selected a dataset containing 10,870 images and employed a 5-fold cross-validation method for evaluation. This approach divides the dataset into five equal subsets, each containing approximately 2174 images. This method is widely recognized for assessing a model's generalization and robustness [22]. By training and validating on multiple subsets, it mitigates the randomness associated with a single data partition and provides more stable performance metrics, such as accuracy and precision [23]. The UA-DETRAC dataset exhibits significant class imbalance, and 5-fold cross-validation helps ensure that the model encounters diverse samples, thereby improving its generalization ability.

We chose the UA-DETRAC dataset due to its variety of complex traffic scenarios, allowing for thorough testing under different conditions, including day, night, and adverse weather. The dataset is well annotated, ensuring accuracy in both training and validation. Additionally, it is widely used in traffic monitoring, facilitating comparisons with existing research. Finally, its range of vehicle types and perspectives aids in assessing the model's adaptability across different environments. The following figure illustrates the evaluation results for one of the folds.

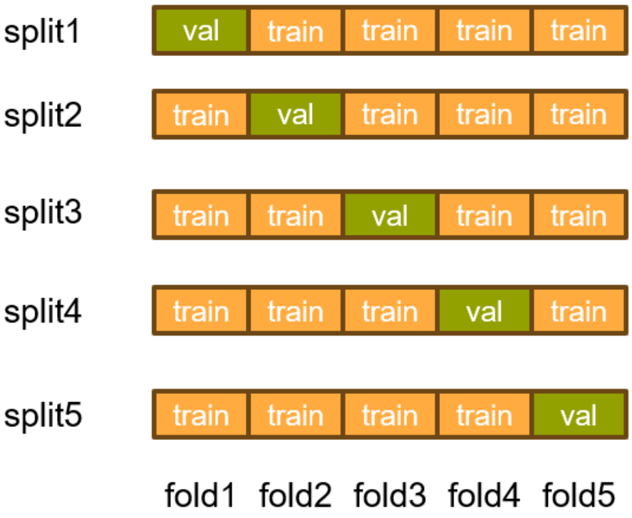
split1
val train train train train
split2
train val train train train
split3
train train val train train
split4
train train train val train
split5
train train train train val
fold1 fold2 fold3 fold4 fold5

Figure 8. Five-Fold Cross-Validation.

As shown in Figure 9, this chart demonstrates the outstanding performance of our trained model in object detection tasks. The results graph (Figure 9) lists several key metrics, including mAP, recall, and precision. These metrics indicate the model's strong effectiveness in accurately identifying and classifying targets, showing its ability to maintain high detection performance in various complex scenarios.

Meanwhile, the P-curve graph (Figure 10) further reinforces this performance evaluation, clearly indicating that as confidence increases, precision also rises. This trend not only reflects the model's reliability and stability at higher confidence levels but also suggests its effectiveness in distinguishing between true positives and false positives. This capability is crucial for ensuring the accuracy of object detection, particularly in real-time applications that require swift decision-making. Therefore, these results not only highlight the robustness and effectiveness of our model but also demonstrate its applicability in various real-world scenarios, such as autonomous driving, video surveillance, and smart security systems.

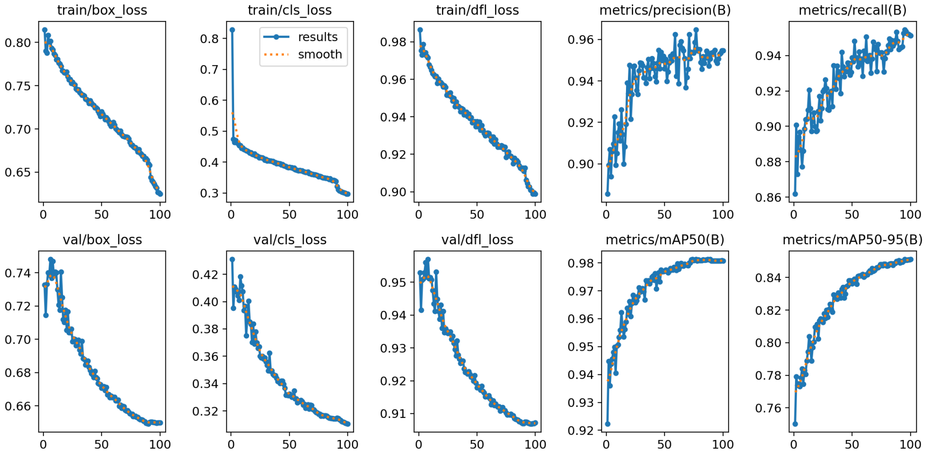

Figure 9. Results.

In video surveillance and analysis, high-precision object detection is crucial, as it reduces false positives, enhances security, and effectively tracks dynamically changing targets. Compared to YOLOv8, which achieves a precision of 0.859 and a recall of 0.830 but tends to miss detections in complex scenarios, our model demonstrates superior performance with a precision of 0.961 and recall of 0.908. This indicates that our model can more accurately capture targets in complex scenes and rapid movements, particularly excelling in situations with occlusions and lighting variations. Furthermore, an mAP50 score as high as 0.962, compared to YOLOv8's 0.881, highlights our model's reliability and comprehensive detection performance in videos (see Table 2 for a comparison of different experimental results).

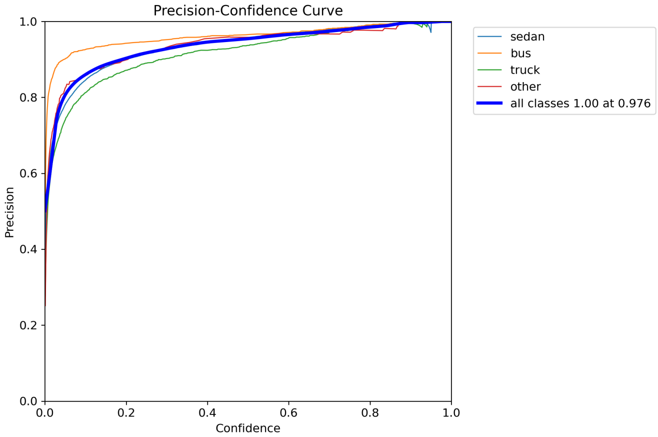

Figure 10. P_curve.

Table 2. Comparison of Different Experimental Results.

Model	Precision	Recall	Map50	Params/10^6	GFLOPs
Faster-Rcnn [7]	0.645	0.901	0.878	28.48	939.60
FCOS [24]	0.825	0.875	0.879	32.16	161.6
Yolov8	0.859	0.83	0.881	3.01	8.1
improved Yolov8 [25]	0.893	0.834	0.912	11.87	38.7
EfficientDet [26]	0.935	0.915	0.920	3.9	2.54
Rank-DETR [27]	0.940	0.895	0.930	28	4.5
Ours	0.961	0.908	0.962	3.5	11.8

Analyzing Rank-DETR, which achieves a precision of 0.940 and a recall of 0.895, it also shows strong performance in handling occlusions and complex scenes, making it a competitive alternative to our model. However, while Rank-DETR performs well, our model still surpasses it in terms of precision and overall mAP50, indicating its enhanced capability in various detection tasks.

Further analyzing the model's computational complexity is essential. Our model maintains a computational complexity of 11.8 GFLOPs while increasing its parameter count to 3.5 million. This means that although our model has significant performance advantages, its higher demand for computational resources may limit its application in certain hardware environments. In contrast, EfficientDet has a lower complexity of 2.54 GFLOPs, making it more suitable for resource-constrained environments, such as embedded devices or mobile platforms. While EfficientDet also performs well in terms of accuracy (0.935) and recall (0.915), its performance may not match that of our model or Rank-DETR in complex scenarios. Therefore, when selecting a model, it is important to weigh computational resources against detection performance. If the application prioritizes high precision and reliability, such as in security monitoring and autonomous driving, our model is more appropriate; whereas, in resource-limited situations, EfficientDet may be the more practical choice. Ultimately, these differences in computational complexity allow each model to excel in different application scenarios.

Increasing the complexity of our model allows it to better capture features and learn intricate relationships within complex datasets, effectively reducing bias and avoiding underfitting. This leads to improved performance on validation and test sets. Moreover, advancements in hardware and optimization algorithms mean that using our more complex model does not incur excessive computational costs. In practical applications, the direct impact of improved accuracy on decision-making and reliability justifies this moderate increase in complexity. Therefore, the benefits gained from enhancing our model's performance and feature learning capabilities outweigh the associated costs, enabling us to choose the most effective model for optimal detection performance in various scenarios.

5.2. Visual Analysis of Detection Results

To validate the effectiveness of the improved model in complex traffic scenarios, we selected four typical conditions—congestion, nighttime, rain, and haze—for testing and compared them with the original model. This comparative analysis not only allows us to comprehensively evaluate the model's performance under various traffic conditions but also provides deep insights into its accuracy, robustness, and adaptability. Through these experimental results, we can clearly identify the model's strengths and limitations, guiding practical applications and revealing potential directions for optimization.

In assessing the impact of these scenarios, various indicators can be quantified to evaluate the influence of each condition on model performance. For instance, congestion levels can be assessed through vehicle density, speed, and lane occupancy rates, which directly affect traffic flow and detection accuracy. As shown in the congestion scene (Figure 11), dense traffic conditions make vehicle detection more challenging due to the overlap and occlusion between vehicles. Nighttime scenarios can be evaluated based on light intensity, uniformity, and the type of light sources to ensure the model remains effective in low-light conditions. The nighttime scene (Figure 12) demonstrates how reduced visibility and uneven lighting impact detection performance. The influence of rain can be measured through raindrop density, road slickness, and glare reflection, which increase visual complexity and challenge the model's detection capabilities. In the rainy scene (Figure 13), the rain affects the image quality through glare and reflection, making vehicle detection harder. Finally, haze conditions can be evaluated based on particulate concentration, visibility, and color distortion, placing higher demands on the reliability of computer vision algorithms. The haze scene (Figure 14) illustrates the reduced visibility and color distortion caused by particulate matter, which further complicates detection.

By integrating computer vision algorithms with image processing techniques, we can precisely quantify these indicators, facilitating a comprehensive evaluation of traffic scene complexity. Furthermore, employing advanced data analysis methods, such as machine learning and deep learning techniques, allows for a deeper understanding of the model's performance across different scenarios, thereby enhancing its adaptability and generalization capabilities. This multidimensional analysis not only provides reliable technical support for traffic monitoring and management but also points the way for future research directions.

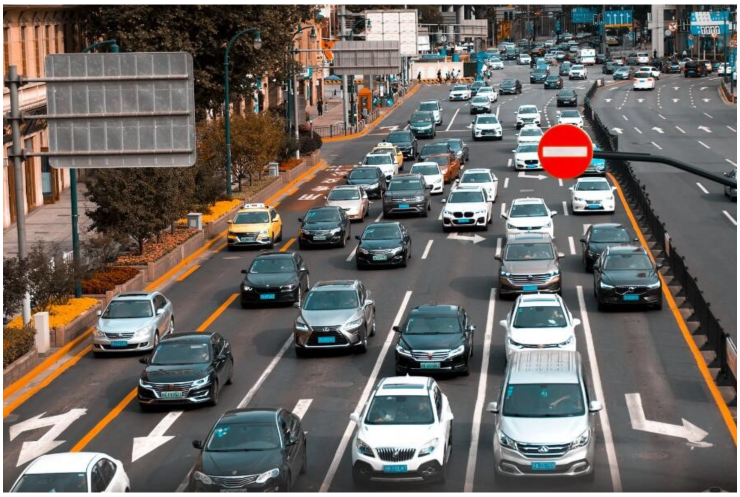

Figure 11. Congested Scene.

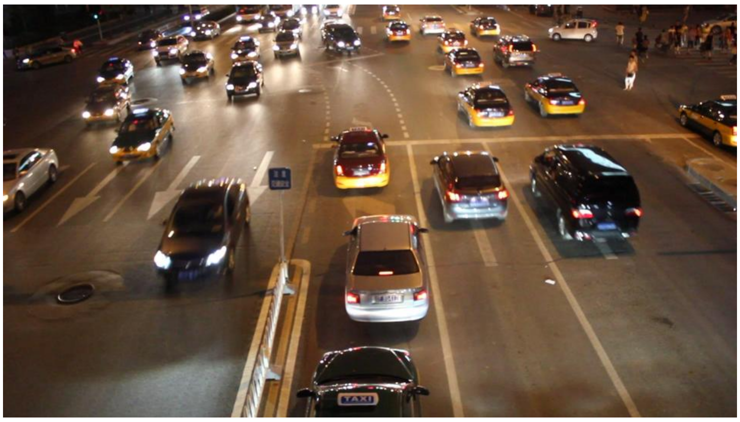

Figure 12. Night Scene.

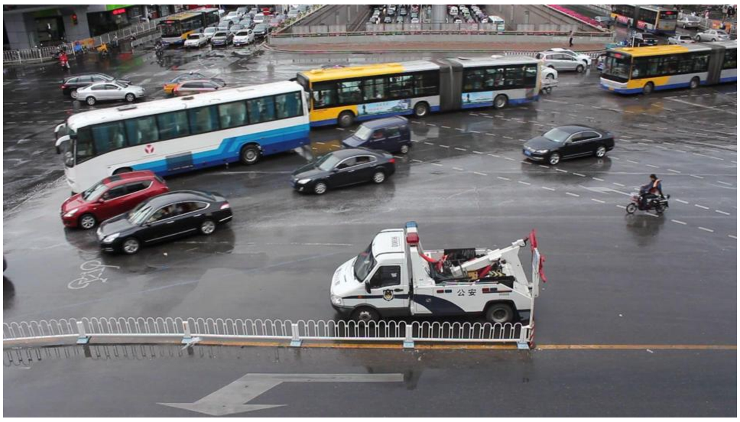

Figure 13. Rainy Scene.

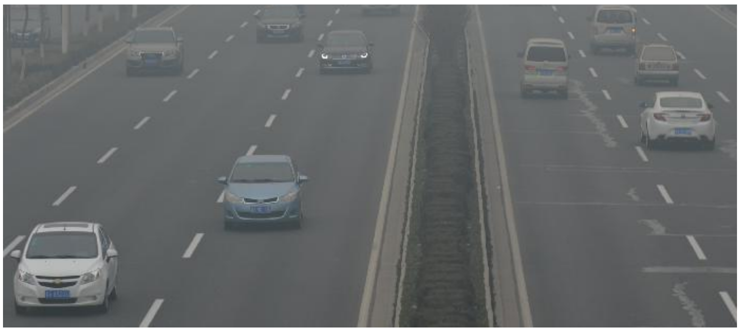

Figure 14. Haze Scene.

In congested scenarios, where complex and dense traffic flows are present, the original model (Figure 15) faced challenges with false positives and missed detections, which impacted both the accuracy and completeness of the detection. Through targeted improvements, especially optimizations for handling increased traffic volume and complexity, the enhanced model (Figure 16) significantly improved detection performance, reducing both false positives and missed detections. However, in typical congested scenarios, EfficientDet (Figure 17) outperformed our improved model, and Rank-DETR (Figure 18) demonstrated superior performance overall. Rank-DETR showed particular advantages in managing complex traffic flows, thanks to its sorting mechanism and better handling of occlusions between vehicles.

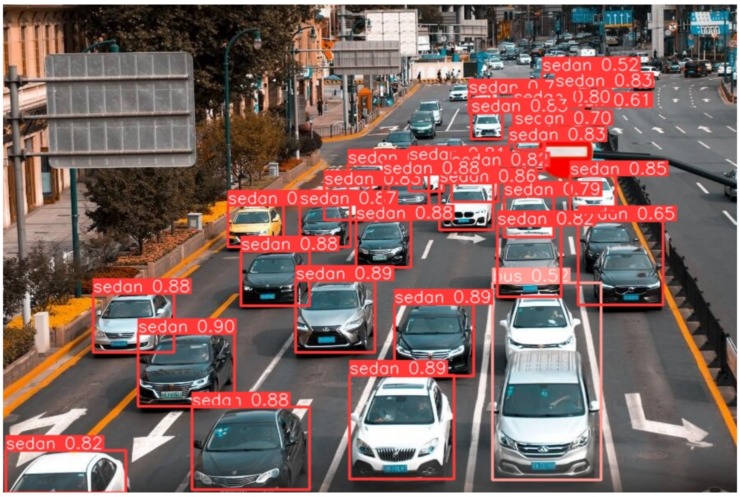

Figure 15. YOLOv8 Detection1.

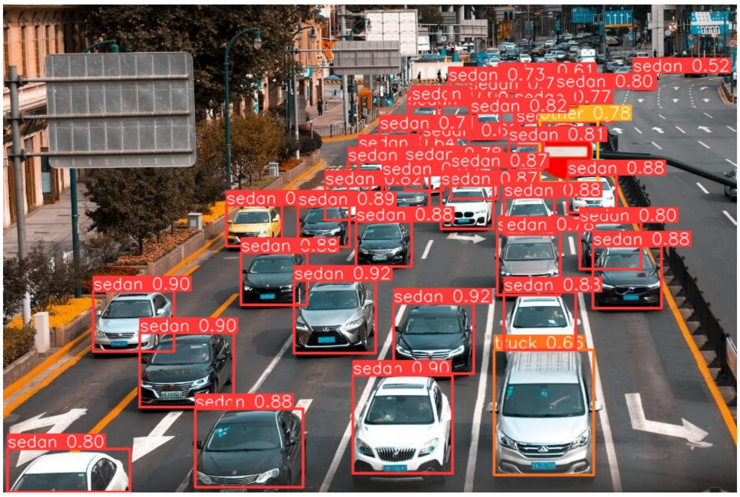

Figure 16. Our Detection1.

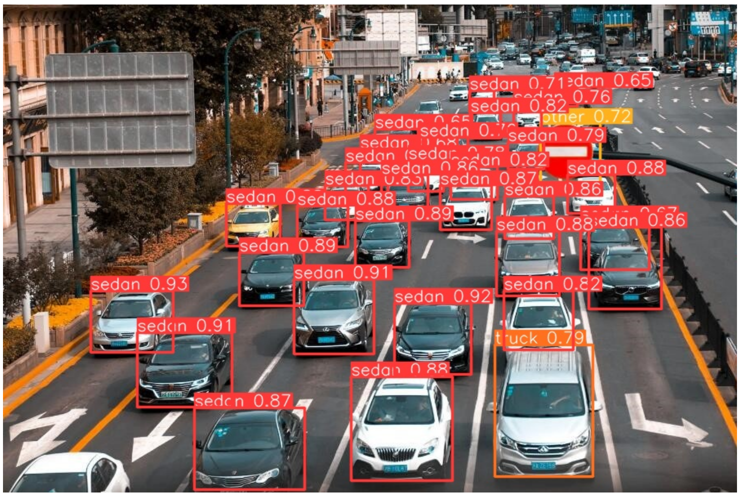

Figure 17. EfficientDet Detection1.

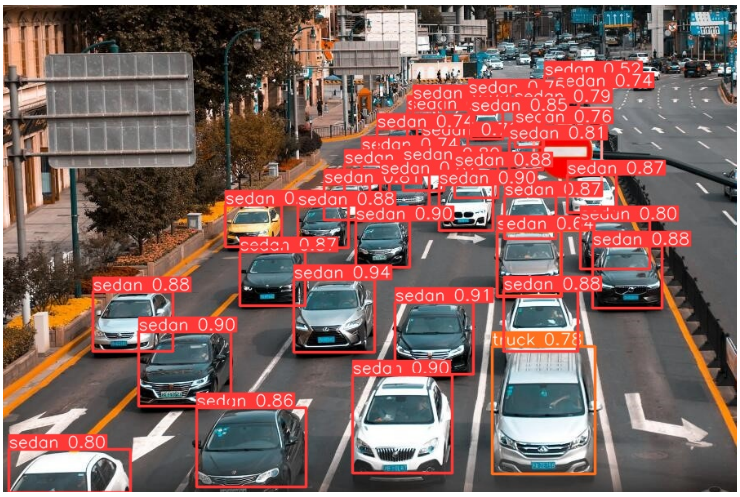

Figure 18. Rank-DETR Detection1.

Under nighttime conditions, the original model (Figure 19) may not only produce false detections due to insufficient lighting or uneven illumination but may also result in missed detections. The improved model (Figure 20) made finer adjustments and optimizations for lighting conditions, employing more advanced image processing techniques and deep learning algorithms to effectively improve detection accuracy and stability in nocturnal environments. Although both EfficientDet (Figure 21) and Rank-DETR (Figure 22) performed reasonably in this aspect, they lacked mechanisms specifically designed to handle nighttime conditions, resulting in detection performance that fell short of our improved model.

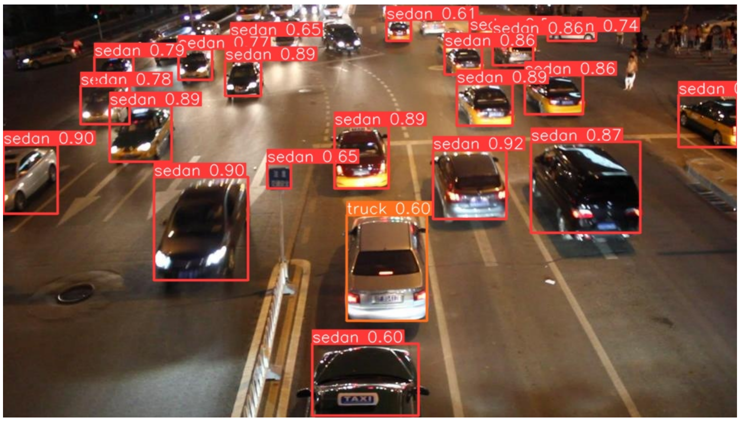

Figure 19. YOLOv8 Detection2.

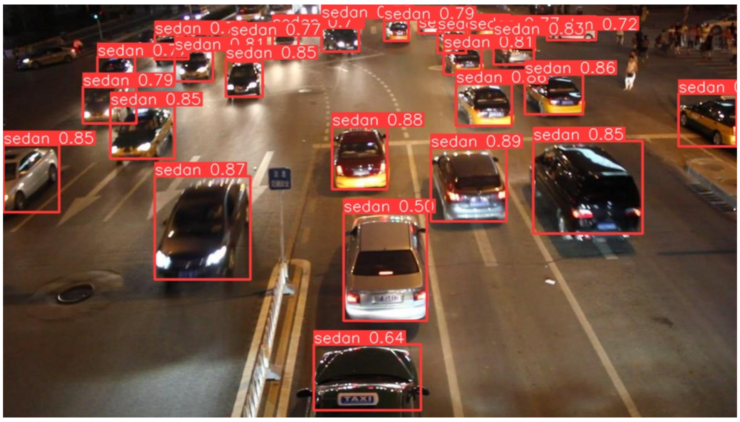

Figure 20. Our Detection2.

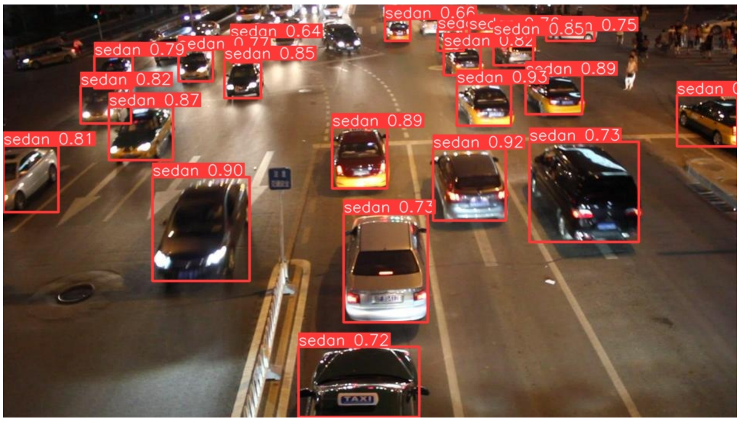

Figure 21. EfficientDet Detection2.

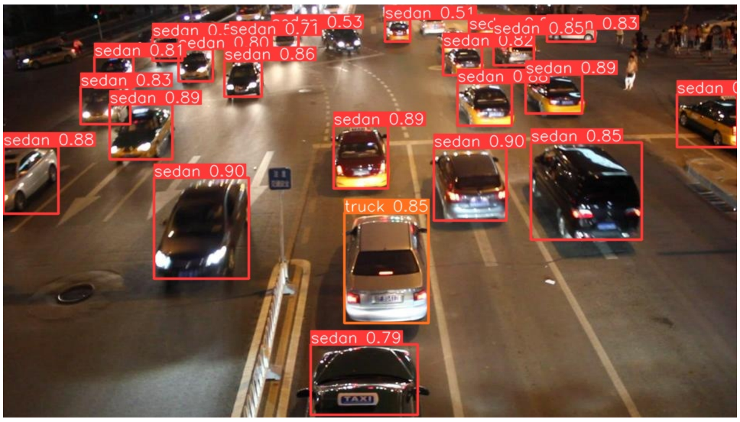

Figure 22. Rank-DETR Detection2.

Similarly, in rainy conditions, the original model (Figure 23) faced challenges with false detections due to rain, causing interference in vehicle detection. The improved model (Figure 24) addressed these issues by conducting a more detailed analysis and processing of image data affected by rain, significantly enhancing vehicle detection capabilities under adverse weather conditions. However, the performance of both EfficientDet (Figure 25) and Rank-DETR (Figure 26) was also limited and did not reach the level of our improved model. Despite these optimizations, some misdetections still occurred under rainy conditions, highlighting the need for further improvements. Factors such as raindrop interference, reflections and glare, background clutter, and variations in ambient light continue to affect the model's performance. This underscores the importance of continued research to ensure the model's accuracy and reliability in complex and challenging weather conditions.

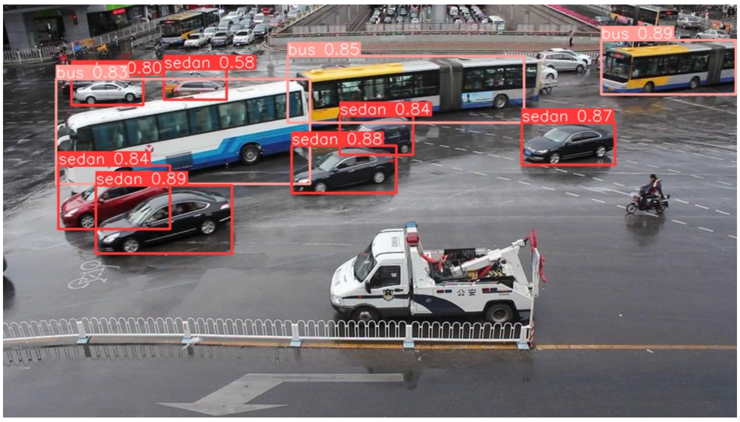

Figure 23. YOLOv8 Detection3.

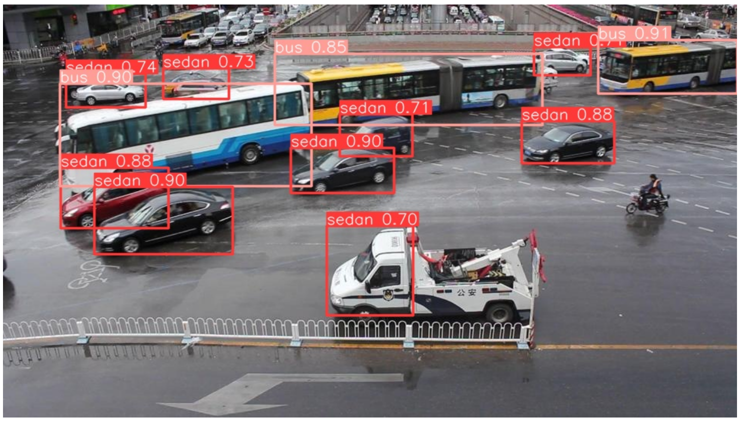

Figure 24. Our Detection3.

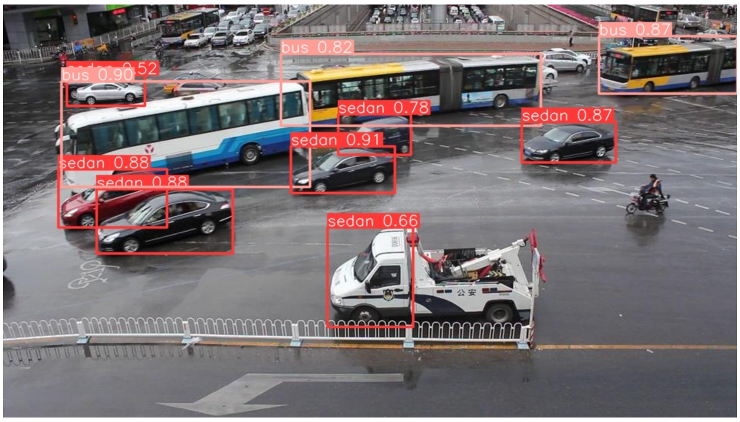

Figure 25. EfficientDet Detection3.

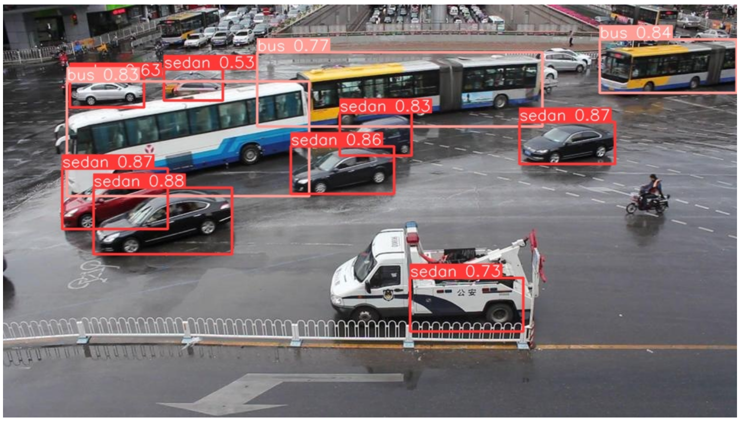

Figure 26. Rank-DETR Detection3.

In hazy conditions, while the original model (Figure 27) exhibited some limitations, the improved model (Figure 28) performed better, though some missed detections still occurred. This can be attributed to reduced visibility and image blur caused by haze, which affected the clarity and contour recognition of vehicles. Both EfficientDet (Figure 29) and Rank-DETR (Figure 30) performed less effectively under these conditions, especially since the specific optimizations made for haze improved the adaptability of our enhanced model, giving it a performance edge over the others.

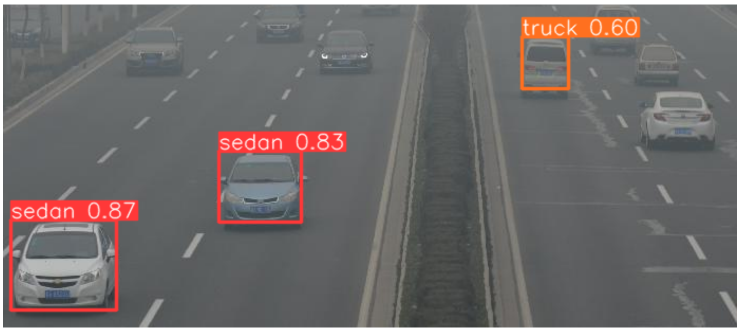

Figure 27. YOLOv8 Detection4.

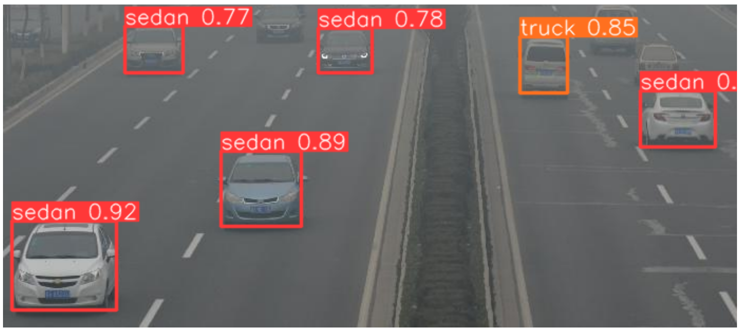

Figure 28. Our Detection4.

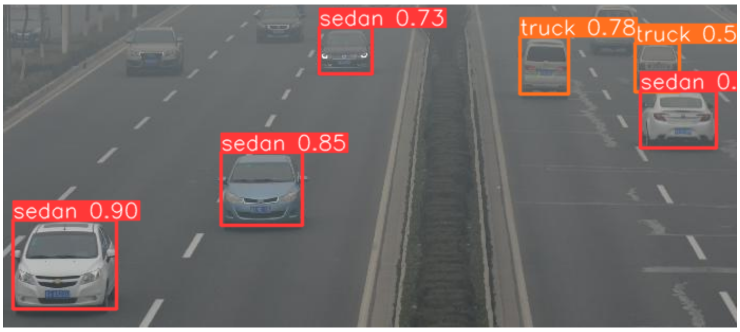

Figure 29. EfficientDet Detection4.

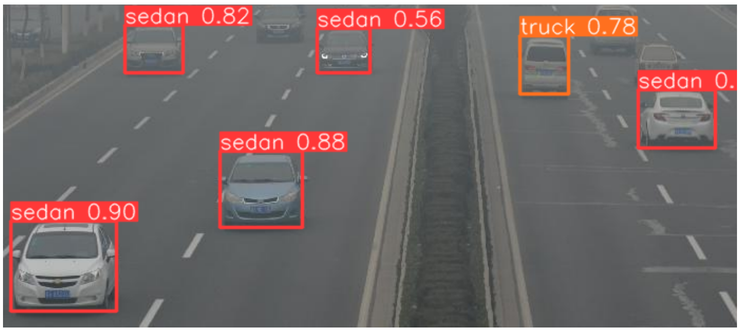

Figure 30. Rank-DETR Detection4.

In various challenging conditions, including congested, nighttime, rainy, and hazy scenarios, the original model faced issues such as false positives and missed detections. However, the improved model significantly enhanced detection performance through targeted optimizations, especially excelling in adverse weather conditions and demonstrating greater adaptability. While Rank-DETR outperformed EfficientDet and even surpassed our improved model in typical congested scenarios, both Rank-DETR and EfficientDet struggled under nighttime, rainy, and hazy conditions. In contrast, our improved model effectively addressed the detection challenges in these environments through specific optimization mechanisms, highlighting its unique advantages.

Subsequently, to visually demonstrate the enhanced recognition capability of the target area by the improved model, we compared the heat maps of the last layer of the backbone section between the baseline and the improved models. Heat maps can intuitively reflect the model's attention to different areas within an image, with higher heat values indicating greater attention to that area.We selected Figures 31–33 for testing.

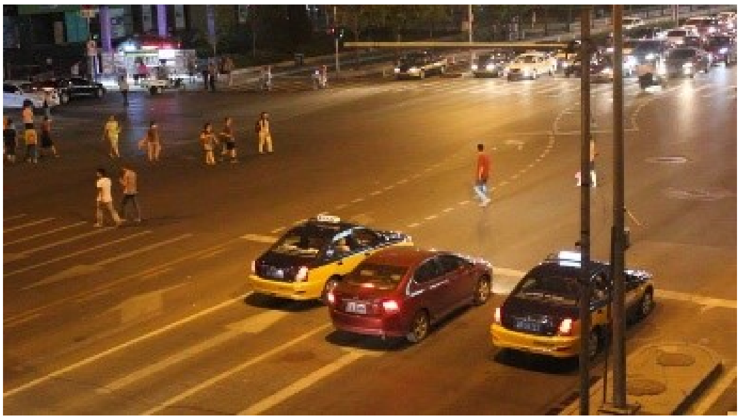

Figure 31. Original Image 1.

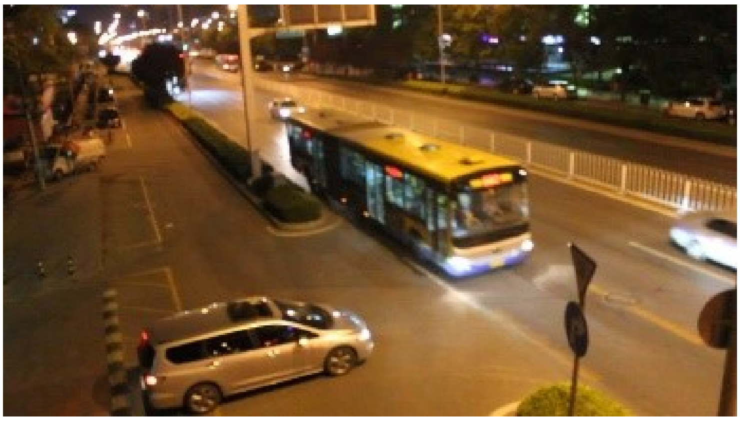

Figure 32. Original Image 2.

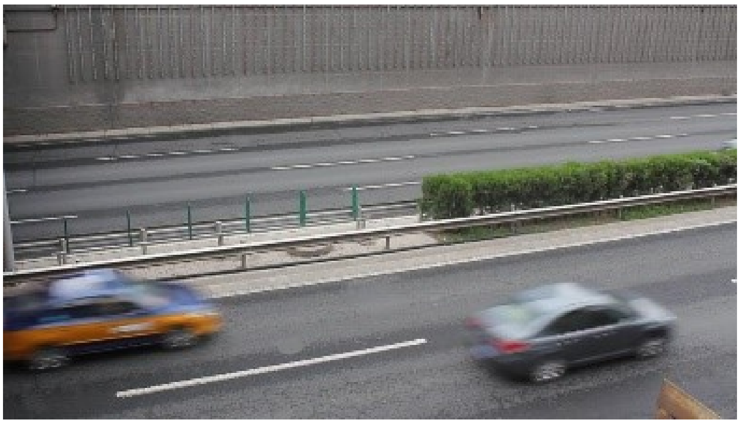

Figure 33. Original Image 3.

As shown in Figure 34–36, the heatmaps of the baseline model reveal a divergent trend at the final layer of the backbone, which indicates insufficient focus on the target regions. This divergence is a result of the model's inability to effectively differentiate between the target and background, which leads to a decrease in accuracy during object detection and localization. This issue is particularly pronounced in complex scenarios such as multi-object occlusions or drastic lighting changes, where the baseline model's feature extraction capabilities at the final layer are notably inadequate.

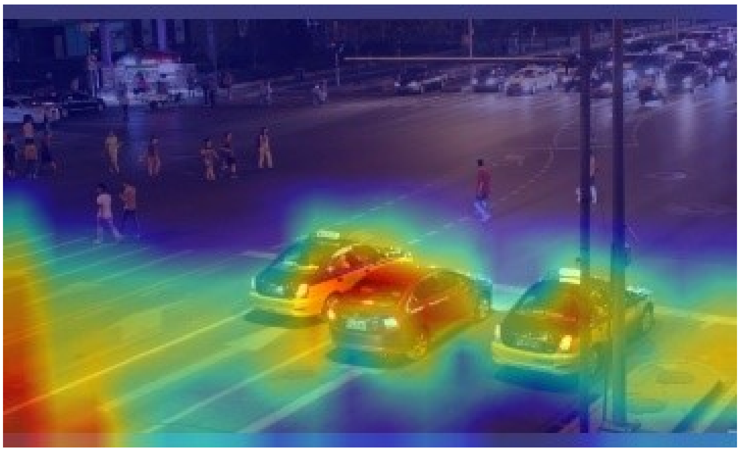

Figure 34. YOLOv8 Heatmap1.

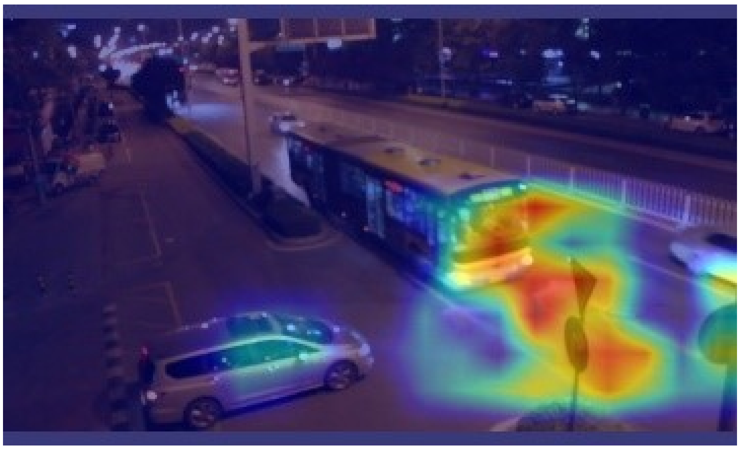

Figure 35. YOLOv8 Heatmap2.

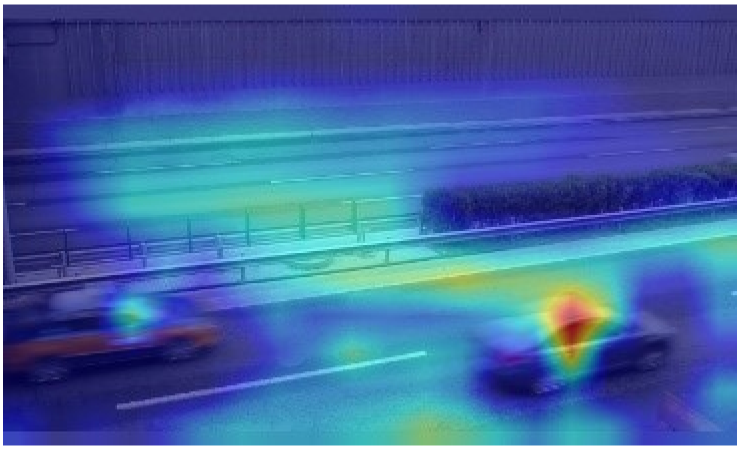

Figure 36. YOLOv8 Heatmap3.

In contrast, as depicted in Figures 37–39, the heatmaps of the improved model show a significant focusing effect. The values are highly concentrated and form distinct peaks that accurately correspond to the target regions. This suggests that the improved model is more effective at extracting target features at the final layer of the backbone, while also suppressing background noise and interfering elements. This enhanced focusing ability is attributed to optimizations made to the backbone, including the introduction of the EMA attention mechanism and focal modulation technology. These techniques improve the model's sensitivity and accuracy in processing complex contextual information.

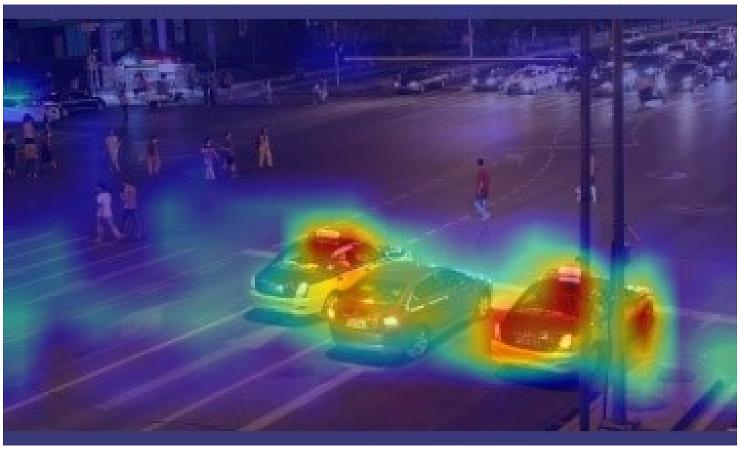

Figure 37. Our Heatmap1.

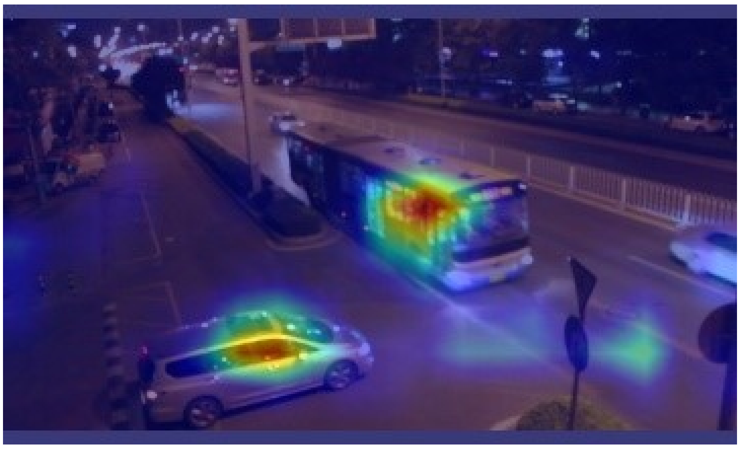

Figure 38. Our Heatmap2.

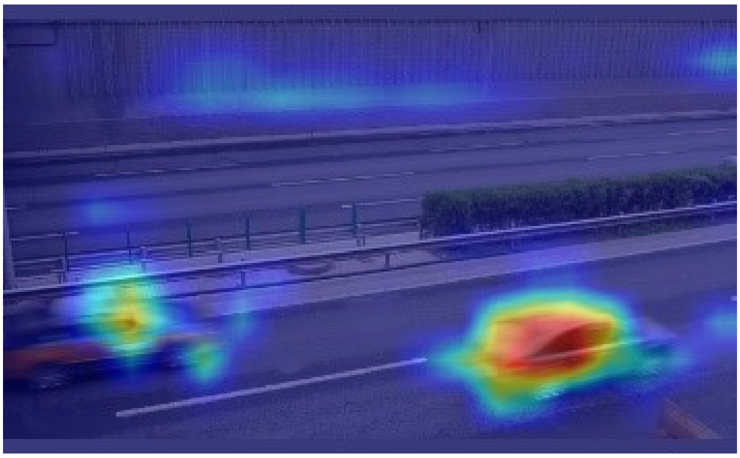

Figure 39. Our Heatmap3.

6. Conclusions

In response to the insufficient accuracy of vehicle target detection in complex traffic monitoring scenarios, this study aims to enhance detection performance by innovatively proposing an improved YOLOv8 vehicle target detection method based on self-attention and focal modulation techniques. This method addresses challenges such as occlusions, lighting variations, and the difficulty in detecting small targets in complex traffic monitoring scenes. By leveraging the self-attention mechanism to deeply analyze the intrinsic relationships between features and the focal modulation technique to improve the model's adaptability to scene changes, the method effectively increases the accuracy of vehicle target detection. Experimental results indicate that the improved YOLOv8 model has achieved significant performance improvements in accuracy evaluation metrics precision (P) and mean average precision (mAP). The enhancement in detection accuracy is particularly pronounced in complex situations, such as occlusions and lighting changes, demonstrating that this method effectively resolves the accuracy limitations of traditional vehicle target detection methods in challenging scenarios. Theoretically, this method offers a new perspective for improving the detection efficiency of traffic monitoring systems and the safety of autonomous driving systems. Its performance has shown promising preliminary results in simulation experiments, laying the foundation for subsequent practical exploration and theoretical deepening. Future research will continue to build on this foundation, exploring and implementing additional strategies to enhance detection accuracy and adaptability to meet the evolving demands of traffic monitoring, further advancing the development of intelligent traffic monitoring and autonomous driving technologies.

Currently, most surveillance cameras are edge devices with low computational power. Effectively deploying efficient vehicle target detection models on these devices is also a major direction for future research exploration [28]. Based on current research, we will focus on optimizing model architecture to more effectively handle multi-scale and multi-angle vehicle target detection tasks [29]. Simultaneously, we will explore how to utilize lightweight model architectures and compression techniques to deploy efficient vehicle target detection models on edge devices with limited computational power [30].

Author Contributions: Conceptualization, B.L. and Q.M.; methodology, B.L. and Q.M.; software, B.L., X.L. (Xin Li) and Z.W.; validation, B.L., Q.M. and X.L. (Xin Liu); formal analysis, B.L.; investigation, B.L., Q.M. and X.L. (Xin Li); resources, B.L. and X.L. (Xin Li); data curation, B.L. and S.K.; writing—original draft preparation, B.L.; writing—review and editing, B.L. and Q.M.; visualization, B.L. and Z.W.; supervision, B.L. and Q.M.; project administration, B.L. and X.L. (Xin Li); funding acquisition, Q.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Foundation of Qinghai Province, grant number: 2023-ZJ-989Q. The APC was funded by Natural Science Foundation of Qinghai Province.

Data Availability Statement: This study utilized the public dataset UA-DETRAC, which can be accessed via the following link: [UA-DETRAC dataset] (https://paperswithcode.com/dataset/ua-detrac (accessed on 7 March 2024)). The dataset is publicly available. For more details on data availability statements, please refer to the MDPI Research Data Policies page: https://www.mdpi.com/ethics (accessed on 3 September 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Djenouri, Y.; Belhadi, A.; Lin, J.C.-W.; Djenouri, D.; Cano, A. A survey on urban traffic anomalies detection algorithms. *IEEE Access* **2019**, *7*, 12192–12205. [CrossRef]
2. Wang, Z.; Zhan, J.; Duan, C.; Guan, X.; Lu, P.; Yang, K. A review of vehicle detection techniques for intelligent vehicles. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 3811–3831. [CrossRef] [PubMed]
3. Meng, C.; Bao, H.; Ma, Y. Vehicle detection: A review. *J. Phys. Conf. Ser.* **2020**, *1634*, 012107. [CrossRef]
4. Ghahremannezhad, H.; Shi, H.; Liu, C. Object detection in traffic videos: A survey. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 6780–6799. [CrossRef]
5. Abbas, A.F.; Sheikh, U.U.; Al-Dhief, F.T.; Mohd, M.N.H. A comprehensive review of vehicle detection using computer vision. *TELKOMNIKA (Telecommun. Comput. Electron. Control* **2021**, *19*, 838–850. [CrossRef]
6. Jain, N.K.; Saini, R.K.; Mittal, P. A review on traffic monitoring system techniques. In *Soft Computing: Theories and Applications, Proceedings of the SoCTA 2017*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 569–577.

7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*; Springer: Berlin, Gremany, 2015; Volume 28, pp. 91–99.
8. Girshick, R. Fast R-CNN. *arXiv* **2015**, arXiv:1504.08083.
9. Kantor, C.; Rauby, B.; Boussioux, L.; Jehanno, E.; Talbot, H. Over-CAM: Gradient-Based Localization and Spatial Attention for Confidence Measure in Fine-Grained Recognition using Deep Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021.
10. Liu, Q.; Ye, H.; Wang, S.; Xu, Z. YOLOv8-CB: Dense Pedestrian Detection Algorithm Based on In-Vehicle Camera. *Electronics* **2024**, *13*, 236. [CrossRef]
11. Dai, Y.; Kim, D.; Lee, K. An Advanced Approach to Object Detection and Tracking in Robotics and Autonomous Vehicles Using YOLOv8 and LiDAR Data Fusion. *Electronics* **2024**, *13*, 2250. [CrossRef]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot Multibox Detector. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part I*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
13. Li, A.; Sun, S.; Zhang, Z.; Feng, M.; Wu, C.; Li, W. A Multi-Scale Traffic Object Detection Algorithm for Road Scenes Based on Improved YOLOv5. *Electronics* **2023**, *12*, 878. [CrossRef]
14. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics* **2023**. *12*, 2323. [CrossRef]
15. Yang, J.; Li, C.; Dai, X.; Gao, J. Focal Modulation Networks. *arXiv* **2022**, arXiv:2203.11926.
16. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
17. Hendrycks, D.; Gimpel, K. Gaussian error linear units (GELUs). *arXiv* **2016**, arXiv:1606.08415.
18. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient multi-scale attention module with cross-spatial learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York City, NY, USA, 2023; pp. 1–5.
19. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; IEEE: New York City, NY, USA, 2021; pp. 13713–13722.
20. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: New York City, NY, USA, 2018; pp. 7794–7803.
21. Lyu, S.; Chang, M.-C.; Du, D.; Wen, L.; Qi, H.; Li, Y.; Wei, Y.; Ke, L.; Hu, T.; Del Coco, M. UA-DETRAC 2017: Report of AVSS2017 & IWT4S challenge on advanced traffic monitoring. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; IEEE: New York City, NY, USA, 2017; pp. 1–7.
22. Jung, Y. Multiple predicting K-fold cross-validation for model selection. *J. Nonparametric Stat.* **2018**, *30*, 197–215. [CrossRef]
23. Yadav, S.; Shukla, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In Proceedings of the 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, India, 27–28 February 2016; IEEE: New York City, NY, USA, 2016; pp. 78–83.
24. Detector, A.-F.O. FCOS: A simple and strong anchor-free object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 1922–1933.
25. Fei, Z.; Guo, D.; Wang, Y.; Wang, Q.; Qin, Y.; Yang, Z.; He, H. Vehicle detection algorithm based on improved YOLOv8 in traffic surveillance. *J. Comput. Eng. Appl.* **2024**, *60*, 110–120.
26. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
27. Pu, Y.; Liang, W.; Hao, Y.; Yuan, Y.; Yang, Y.; Zhang, C.; Hu, H.; Huang, G. Rank-DETR for high quality object detection. *arXiv* **2023**, arXiv:2310.08854.
28. Boukerche, A.; Hou, Z. Object detection using deep learning methods in traffic scenarios. *ACM Comput. Surv.* **2021**, *54*, 1–35. [CrossRef]
29. Razi, A.; Chen, X.; Li, H.; Wang, H.; Russo, B.; Chen, Y.; Yu, H. Deep learning serves traffic safety analysis: A forward-looking review. *Iet Intell. Transp. Syst.* **2023**, *17*, 22–71. [CrossRef]
30. Liu, C.; Li, S.; Chang, F.; Wang, Y. Machine vision based traffic sign detection methods: Review, analyses and perspectives. *IEEE Access* **2019**, *7*, 86578–86596. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.