electronics

MDPI

Article

Vehicle–Pedestrian Detection Method Based on Improved YOLOv8

Bo Wang, Yuan-Yuan Li, Weijie Xu, Huawei Wang and Li Hu *

School of Automotive and Transportation Engineering, Wuhan University of Science and Technology, Wuhan 430065, China; wangbo66@wust.edu.cn (B.W.); lyy@wust.edu.cn (Y.-Y.L.); xuweijie@wust.edu.cn (W.X.); wanghw@wust.edu.cn (H.W.)
* Correspondence: huli@wust.edu.cn

Abstract: The YOLO series of target detection networks are widely used in transportation targets due to the advantages of high detection accuracy and good real-time performance. However, it also has some limitations, such as poor detection in scenes with large-scale variations, a large number of computational resources being consumed, and occupation of more storage space. To address these issues, this study uses the YOLOv8n model as the benchmark and makes the following four improvements: (1) embedding the BiFormer attention mechanism in the Neck layer to capture the associations and dependencies between the features more efficiently; (2) adding a 160 × 160 small-scale target detection header in the Head layer of the network to enhance the pedestrian and motorcycle detection capability; (3) adopting a weighted bidirectional feature pyramid structure to enhance the feature fusion capability of the network; and (4) making WIoUv3 as a loss function to enhance the focus on common quality anchor frames. Based on the improvement strategies, the evaluation metrics of the model have improved significantly. Compared to the original YOLOv8n, the *mAP* reaches 95.9%, representing an increase of 4.7 percentage points, and the *mAP*50:95 reaches 74.5%, reflecting an improvement of 6.2 percentage points.

Keywords: deep learning; vehicle and pedestrian detection; target detection; YOLOv8

Citation: Wang, B.; Li, Y.-Y.; Xu, W.; Wang, H.; Hu, L. Vehicle–Pedestrian Detection Method Based on Improved YOLOv8. *Electronics* **2024**, *13*, 2149. https://doi.org/10.3390/electronics13112149

Academic Editor: Zhenyi Liu

Received: 18 April 2024
Revised: 24 May 2024
Accepted: 28 May 2024
Published: 31 May 2024

1. Introduction

According to statistics, there are about 200,000 traffic accidents in China every year, resulting in more than 60,000 deaths and 200,000 injuries, and property losses due to traffic accidents exceed CNY 1 billion every year. Relevant studies indicate that only 10% of traffic accidents are attributed to mechanical failures of the vehicle itself. In contrast, the remaining 90% are caused by human factors, including driver fatigue, drunk driving, overloading, speeding, and distracted driving [1]. If an automated driving system can take over the vehicle instead of the driver before a traffic accident occurs and make a correct response, the accident rate can be effectively reduced.

Autonomous driving systems utilize advanced sensors, cameras, and artificial intelligence algorithms to sense the surrounding environment and assist the driver in responding, when necessary, which can help reduce traffic accidents caused by driver inattention, delayed reaction time, or errors in judgment. In terms of environment perception, the onboard camera provides a large amount of environmental information, and the automatic driving system can carry out vehicle–pedestrian detection, lane line detection, road traffic sign recognition, and other tasks through the video image information collected by the camera.

In real-world scenarios, the complexity of the traffic environment frequently results in considerable variations in the sizes of vehicle and pedestrian targets depicted in photos and videos. Moreover, these targets often overlap with each other and are subject to background occlusion. These challenges significantly heighten the difficulty of deploying vehicle and pedestrian detection techniques, thereby posing a stringent test to their accuracy and robustness [2]. The accuracy of the vehicle detection model is critical to the precision of vehicle information extraction. Consequently, research on vehicle–pedestrian detection holds significant theoretical and practical values.

2. Related Work

2.1. Optimizing Autonomous Driving Target Detection

In the realm of computer vision research for autonomous driving, the KITTI dataset [3] has become a vital benchmark due to its extensive scene coverage and diverse target categories, including small and fuzzy objects. Our study addresses the shortcomings of the existing algorithms in handling occlusion, detecting small objects, and recognizing omissions and proposes a new model to overcome these limitations. By thoroughly analyzing the existing research, we aim to demonstrate the innovativeness of our work and its contributions to the field.

Liu et al. [4] proposed an enhanced vehicle detection algorithm for intelligent transportation applications, achieving significant results in coping with the dense distribution of targets and scale variations in images, especially in small object detection, which was improved over the baseline model YOLOX. However, the algorithm still faces challenges for the detection of very small objects. Lou et al. [5], on the other hand, proposed the DC-YOLOv8 model that focuses on small target detection. They combined down-sampling techniques to preserve contextual features and effectively integrated shallow and deep information through an enhanced feature fusion network. These innovations enable DC-YOLOv8 to perform well in complex scenes with improved detection accuracy compared to YOLOv8. Mahaur B et al. [6] further advanced feature extraction and localization classification by proposing the HIC-YOLOv5 model. The model employs a small target detection head to process high-resolution feature maps and utilizes convolutional blocks for channel information enhancement and incorporates CBAM to emphasize important features. The detection accuracy of HIC-YOLOv5 reached 36.95% *mAP*.

In addition, Iqra Nosheen [7] proposes a visual tracking method that combines speckle detection and KCF to enhance vehicle detection through preprocessing techniques. Validated using the KITTI dataset, the method significantly improves the detection accuracy, achieving 52% detection accuracy and 86% tracking accuracy. The technique helps to improve traffic monitoring and flow management.

Our research builds on this series of existing studies, and by deeply analyzing the limitations of these algorithms, we propose a vehicle–pedestrian detection method based on an improved YOLOv8, which effectively improves the detection ability of vehicle–pedestrian targets.

2.2. Two-Stage and One-Stage Target Detection Algorithms

According to the different detection strategies, deep learning-based target detection methods can be divided into two-stage target detection algorithms and single-stage (one-stage) target detection algorithms.

Two-stage target detection algorithms first obtain the candidate region by selective search and feature extraction, then feature extraction is performed on the candidate region to obtain the feature matrix, and finally, feature matrix is fed into the classifier for prediction. The R-CNN model proposed by Girshick et al. [8] in 2014 first applies a convolutional neural network to the target detection task, and the PASCAL VOC 2012 dataset achieved a detection accuracy of about 53%, which is about 30% better than the previous best target detection algorithms and far better than traditional target detection algorithms. However, thousands of candidate regions need to be extracted for each image, resulting in a computationally intensive model and slow detection. He et al. [9] proposed a Spatial Pyramid Pooling Network (SPP-Net), which drastically reduces the computational effort of the network by adding an SPP layer between the convolutional layer and the fully connected layer 2015. Girshick was inspired by the SPP-Net network to propose Fast R-CNN, which uses an ROI pooling layer to normalize the feature maps at different scales and employs a Softmax function instead of an SVM classifier, which effectively improves the detection speed of the network. Ren et al. [10] proposed Faster R-CNN by further investigating the Fast R-CNN. CNN, which introduces a Region Proposal Network (RPN) instead of selective search, greatly improves the detection efficiency of the network. Later,

many scholars improved and extended the Faster R-CNN from different perspectives. Dai et al. [11] proposed the R-FCN (Region-based Fully Convolutional Network) in 2016, which enhanced detection efficiency by generating a location-sensitive distribution map. Cai et al. [12] proposed the Cascade R-CNN in 2018 effectively improves the detection performance of small targets by cascading multiple detection heads. Overall, although the two-stage target detection algorithm performs well in terms of detection accuracy, the requirement to generate multiple candidate regions and classify each of them imposes limitations on the detection efficiency of two-stage target detection algorithms.

Single-stage target detection algorithms do not need to extract candidate regions but directly feed the image into the detection network to extract features and make predictions, and the representative models include SSD, YOLO series, etc. YOLO [13] (You Only Look Once), proposed by Joseph Redmon in 2016, uses regression ideas to perform target detection, which significantly improves the target detection speed. YOLOv1 refers to the GoogLeNet [14] classification network structure, which consists of 24 convolutional layers and 2 fully connected layers, and its core idea is to partition the image to be detected into S $\times$ S small grids, and the grid in which the centroid of the object to be detected is responsible for predicting this object, and each small grid can only generate at most two prediction frames, which results in low accuracy of small-scale object detection and easy to miss detection. Later, on the basis of YOLOv1, the original authors proposed YOLOv2 and YOLOv3 and introduced the K-means clustering algorithm, DarkNet network, etc., in order to improve the detection accuracy and speed of the YOLO algorithm. In recent years, the YOLO series of algorithms have been updated and iterated rapidly, and up to now, it has been developed to YOLOv8. YOLO has been widely used in the field of vehicle and pedestrian detection due to its fast and accurate characteristics.

3. Vehicle–Pedestrian Detection Model

To improve the environment sensing ability of self-driving cars, we proposed a high-precision vehicle–pedestrian detection algorithm. To address the challenges of large-scale differences and target occlusion in road traffic scenes, we propose several enhancements. Firstly, by embedding the BiFormer attention mechanism in the Neck layer and adding an additional detection head in the Head layer, the network's detection capabilities for small targets such as pedestrians and motorcycles are significantly improved. Secondly, the weighted bidirectional feature pyramid structure is utilized to enhance the network's feature fusion ability. Additionally, WIoUv3 [15] is employed as the loss function to improve the network's focus on high-quality anchor frames. Our proposed network model is illustrated in Figure 1.

3.1. BiFormer Attention Mechanism

The attention mechanism aims to enhance the model's focus on input data for weight allocation in computer engineering. It automatically identifies the relevance and importance of data, enabling the model to concentrate more on various inputs. By dynamically assigning weights between elements, the model can prioritize the most critical parts of the input data. However, the introduction of the attention mechanism leads to an increase in computation and memory usage.

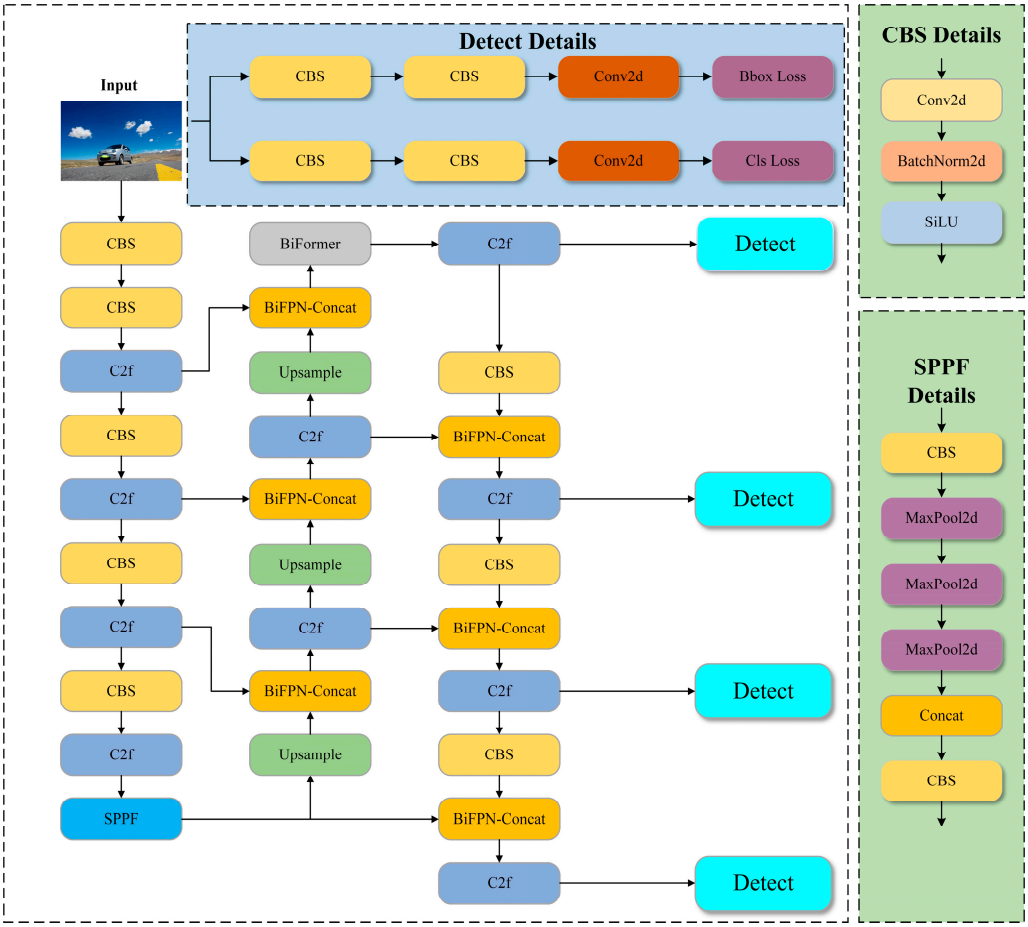


Figure 1. Diagram of the proposed network structure.

Although our task is to detect vehicles and pedestrians in road traffic images, which may be applied to environmental sensing for autonomous vehicles in the future, the autonomous vehicle chip is highly sensitive to the model's computation and memory usage for each task, as it needs to process multiple tasks simultaneously. To tackle the limitations posed by computational and memory constraints, this paper adopts the BiFormer Attention mechanism proposed by Zhu et al. [16], which utilizes Bi-level Rounding Attention (BRA) with a two-level rounding path as its main building block. Compared to other attention mechanisms, BRA employs an area-to-area routing index matrix with fine-grained labeling of each query token and attention to all key-value pairs within the set of routing areas. In other words, BRA has a more precise selection of sensing regions, which is very beneficial for detecting small objects such as pedestrians, and it improves model performance while reducing computational load. It is well suited for the vehicle and pedestrian detection task in this paper.

BiFormer utilizes a four-stage pyramid structure with BRA as its core module. In the first stage, overlapping patch embedding is employed, while in the second to fourth stages, a patch merging module is used to decrease the input spatial resolution while increasing the number of channels. Then, N consecutive BiFormer blocks are employed to transform the features. Indicates a convergence between the two. In each BiFormer Block, a 3×3 deep convolution is first used to implicitly encode the relative position information. Then, the BRA module and the 2-layer MLP module are applied sequentially for cross-position relation modeling and position-by-position embedding, respectively. The overall structure of BiFormer is depicted in Figure 2.

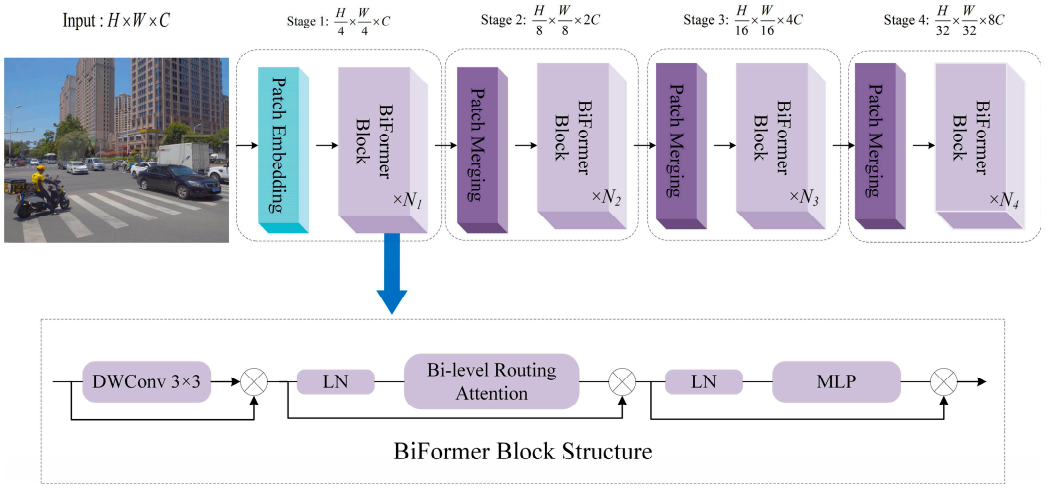


Figure 2. BiFormer's overall structure.

Theoretically, the directed graph routing between regions enables the feature graph to optimize the selection of associated content part by part, which is the core of the BiFormer attention mechanism. BiFormer attention mechanism based on two-layer routing has a finer sensory field selection ability, which can not only select the target more accurately as a whole but also capture the sensory field content more accurately in the local region.

3.2. Multi-Scale Prediction Network

After feature extraction and feature fusion, YOLOv8 outputs three scale feature maps, 80 × 80, 40 × 40, and 20 × 20, where 80 × 80 feature maps are used to predict small targets, 40 × 40 feature maps are used to predict medium targets, and 20 × 20 feature maps are used to predict large targets. However, in real road scenarios, the differences in target scales can be very large. For example, the volume of pedestrians is much smaller than the volume of trucks, and when pedestrians are far away, they occupy fewer pixel points in the image, the features are not obvious after multiple down-sampling, and they are easily missed during detection.

To address this issue, this paper proposes a four-scale prediction network, which adds a 160 × 160 prediction head on the basis of the original three-scale prediction network and is specifically used to improve the efficiency of small target detection. An up-sampling operation is added to the Neck layer of YOLOv8, which outputs a feature map with a scale of 160 × 160 and connects it to the 160 × 160 feature map output from the Backbone layer. By incorporating this high-resolution feature map, the network gains the ability to capture finer details, thereby enhancing the detection accuracy for small targets.

3.3. Weighted Bidirectional Feature Pyramid Structure

In YOLOv8, the feature maps are divided into five scales, denoted as B1-B5 for the backbone, P3-P4 for the FPN, and N4-N5 for the PAN. The original model adopts the PAN-FPN structure [17], which is an optimized version of the traditional FPN structure. The traditional FPN structure conveys deep semantic information through a top–down approach. However, in YOLOv8, the fusion of B3-P3 and B4-P4 is executed to enhance the semantic features of the feature pyramid, but this may lead to a part of the localized information being lost. To solve this problem, PAN-FPN introduces a bottom–up PAN structure at the top of the FPN to compensate for the lost localization information. In YOLOv8, P4-N4 and P5-N5 are fused to enhance the learning of localization features to achieve a complementary effect. The structure of YOLOv8 is shown in Figure 3.

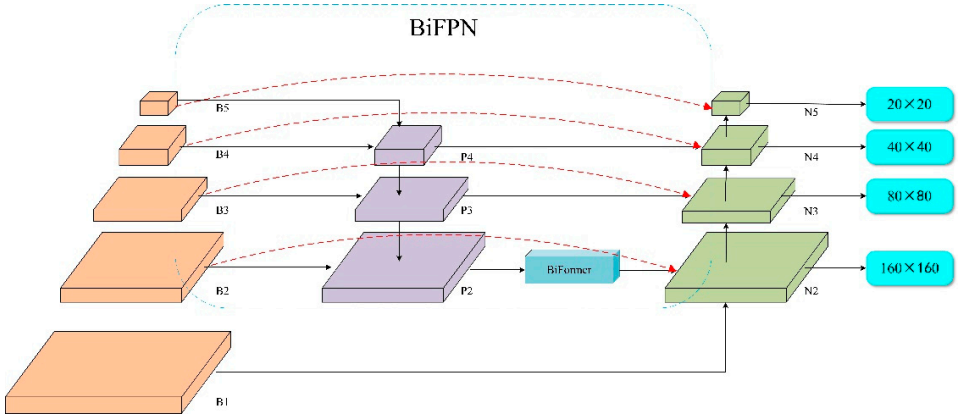


Figure 3. Structure of YOLOv8 using BIFPN.

Although the PAN-FPN structure enriches semantic and localization information, there is still room for improvement. First, the PAN-FPN structure may not be able to adequately handle large-scale feature maps, which may ignore some valuable information and lead to a degradation of detection quality. In addition, after up-sampling and down-sampling, the feature map loses some original information, leading to a relatively low reuse rate.

To more effectively address the aforementioned challenges, this study introduces an advanced feature fusion component reconstruction strategy grounded in the principles of the Bidirectional Feature Pyramid Network (BiFPN) [18]. The BiFPN structure, initially introduced by Google in the Efficient Det object detection algorithm, is specifically devised to augment semantic information within features via efficient bidirectional cross-scale connectivity and weighted feature fusion.

In vehicle–pedestrian target detection, the limited feature information extracted from small-scale targets often leads to low detection accuracy. BiFPN helps to overcome this challenge by extending the sensory field of the model by fully utilizing high-resolution features. No additional processing is applied for feature maps with a single input path due to their low contributive value. When fusing feature maps with two input paths, given that the feature maps have the same scale, cross-level fusion requires the introduction of new paths from the main feature map, as shown by the red line in Figure 3. This approach improves the spatial information of the feature maps and enhances the detection accuracy of the network for small targets.

3.4. WIoU Loss Function

In road traffic scenarios, the design of the loss function is crucial to the detection performance of the model due to the high percentage of small targets and significant differences in target scales (e.g., pedestrians, cars, trucks, etc.).

*WIoUv*3 employs a dynamic non-monotonic mechanism to evaluate the quality of anchor frames and designs a reasonable gradient gain allocation strategy, which reduces the occurrence of large or harmful gradients in extreme samples. In this way, WIoUv3 gives more attention to the anchor frames of ordinary quality, thus improving the detection accuracy and generalization performance of the model. The *WIoU* calculation formula is as follows:

$$L_{WIoU} = R_{WIoU} \times L_{IoU} \tag{1}$$

$$R_{WIoU} = \exp \frac{\left(b_{c_x}^{gt} - b_{c_x}\right)^2 + \left(b_{c_y}^{gt} - b_{c_y}\right)^2}{\left(c_w{}^2 + c_h{}^2\right)} \tag{2}$$

$$L_{IoU} = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

$b_{c_x}^{gt}$; $b_{c_y}^{gt}$—coordinates of the center point of the real frame;
b_{c_x}; b_{c_y}—coordinates of the center point of the prediction frame;
c_w—difference between the lengths of the prediction frame and the real frame;
c_h—height difference between the prediction frame and the real frame;

WIoUv3 defines an outlier on the basis of *WIoU* to measure the quality of the anchor frame and constructs the non-monotonic focus factor r based on the outlier, and the formula of *WIoUv3* is as follows:

$$L_{WIoUv3} = L_{WIoU} \times r \tag{4}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \tag{5}$$

where $\alpha = 1.9; \beta = 3$

4. Experiment and Analysis

4.1. Correlation Dataset

The experiments are conducted on the merged KITTI dataset, which is currently the world's largest computer vision evaluation dataset for autonomous driving scenarios, and the dataset has a total of nine classes of targets. The experiments only detect road vehicles and pedestrians, among them, Cars (small cars), Van, and Truck are extracted from the original dataset, Cyclists (Motorcycle, Bicycle), four classes of targets, and combined Pedestrian, Person sitting into Person (Pedestrian) class, while deleting the Tram, Misc, Do not Care classes, and finally obtaining 7481 images with labels, which were divided into a training set and a test set according to the ratio of 8:2 for training and evaluation of the performance of the model. The dataset label distribution is shown in Table 1.

Table 1. Dataset label distribution.

Dataset Partitioning	Pedestrian	Car	Cyclist	Van	Truck	Total
Training set	3840	22,823	1304	2368	859	31,194
Test set	869	5919	323	546	235	7892
total	4709	28,742	1627	2914	1094	39,086

In the training process, the Mosaic data enhancement method is used. Specifically, nine images are randomly selected, randomly arranged, and spliced together to generate new training samples. Mosaic data enhancement is an effective method, as the spliced images contain more objects and backgrounds, providing richer contextual information for the model, this helps to improve the model's performance and generalization ability.

4.2. Experimental Platform and Evaluation Indexes

Vehicle–pedestrian detection experiments use precision (*P*), recall (*R*), *mAP*50, *mAP*50:95, and model size as evaluation metrics. *mAP*50 denotes the average precision when the IoU is 0.5, and *mAP*50:95 denotes the average precision when the IoU threshold ranges from 0.5 to 0.95, an increment of 0.05.

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$AP = \int_0^1 p(t)dt \tag{7}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{8}$$

AP—average precision;
N—total number of categories;
AP_i—the *i* category accuracy rate.

The experimental parameters are configured as follows: the initial learning rate is set to 0.01, while the final learning rate is adjusted to 0.001. The cosine annealing algorithm is employed, and the warm-up epochs are designated to be threefold. The input image size is 640 × 640, the size is 16, the number of iterations is 200 epochs, the SGD optimizer is used, the momentum is set to 0.937, and the Mosaic data enhancement is turned off for the last 10 epochs.

4.3. Incorporating BiFormer Effects at Different Locations

Given the considerable variance in the impact of integrating the attention mechanism at various network locations, this study endeavors to integrate BiFormer at different junctures within YOLOv8n and subsequently conducts experiments. The specific integration points are delineated in Figure 4, while the resultant experimental outcomes are presented in Table 2.

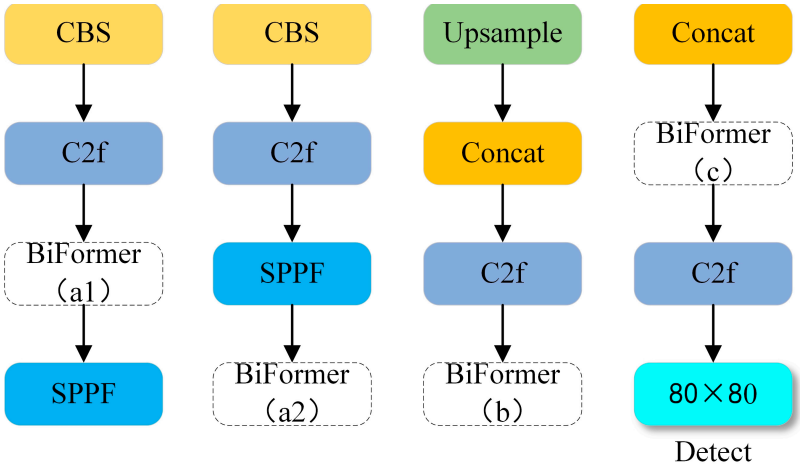


Figure 4. Attention mechanism incorporation modes: (**a1**) before incorporating SPPF module; (**a2**) after incorporating SPPF module; (**b**) after incorporating each C2f module of the Neck layer; and (**c**) before incorporating the C2f module connected to the first detection layer.

Table 2. Experiments of fusing BiFormer at different locations.

Class	Location	P (%)	R (%)	*mAP*50 (%)	*mAP*50:95 (%)
BiFormer	/	93.3	83.7	91.2	68.3
	a1	92.5	83.1	90.4	67.5
	a2	93.8	83.6	91.8	68.9
	b	90.2	81.1	89.7	66.5
	c	94.4	85.4	92.5	70.1
	a2 + c	93.5	83.4	92.3	69.7

Upon examining the experimental findings in Table 2, it becomes evident that the BiFormer attention mechanism fails to enhance the detection performance of the network across all positions.

Incorporating BiFormer at position b, i.e., after each C2f module in the Neck layer, is the least effective, with the average detection accuracy decreasing by 1.5 percentage points; adding a separate layer of BiFormer at position a1, i.e., in front of the SPPF module in the Backbone layer, is also less effective.

Fusing BiFormer at positions a2 and c has some improvement in network performance, and experiments are conducted by simultaneously fusing BiFormer at these two positions. From the experimental results, it can be seen that, when the BiFormer attention mechanism is fused at the same time, the average detection accuracy reaches 92.3%, which is up by 0.5 percentage points compared to fusing the attention at position a2 alone, but not as good as fusing the attention mechanism at position c alone.

In summary, incorporating the BiFormer attention mechanism before the C2f module, which is connected to the first detection layer of YOLOv8n, and after the Concat module is the most effective. It can be seen that BiFormer can effectively enhance the network sensory field and improve the network's ability to detect vehicle–pedestrians after feature maps of different scales have been spliced by the Concat module.

4.4. Comparison Effect of Different Attention Mechanisms

In order to verify the advantages and disadvantages of CA attention mechanism compared with other attention mechanisms on the original network performance enhancement, this paper conducts comparison experiments on the best experimental effect of c position and fusion of SE [19], CBAM [20], and ECA [21], the three popular attention mechanisms, and the experiment's results are shown in Table 3.

Table 3. Comparison experiment of different attention mechanisms.

Model	Attention Mechanisms	P (%)	R (%)	*mAP*50 (%)	*mAP*50:95 (%)
YOLOv8n	/	93.3	83.7	91.2	68.3
	SE	93.3	83.9	91.3	68.3
	CBAM	92.9	83.1	90.8	67.9
	ECA	93.8	84.5	91.9	68.5
	BiFormer	94.4	85.4	92.5	70.1

From Table 3, it can be seen that not all attention mechanisms can improve the detection performance of the model, for example, imposing the CBAM attention mechanism leads to a decrease in the *mAP* of the network. SE and ECA exhibit a slight improvement in the detection performance of the model, while BiFormer is more advantageous in terms of detection accuracy and recall.

4.5. Experiments of Different Loss Functions

To verify the effect of applying different loss functions on the detection accuracy of the model, four versions of the improved model, CIoU, DIoU, SIoU, and WIoU, were applied and compared experimentally. The experiment's results are shown in Table 4.

Table 4. Comparison experiment of different loss functions.

	Loss Function	Map50 (%)	Map50:95 (%)
Improved YOLOv8n	CIoU	94.6	73.5
	DIoU	94.4 (−0.2)	73.3 (−0.2)
	SIoU	95.1 (+0.5)	73.8 (+0.3)
	WIoU v1	94.3 (−0.3)	73.3 (−0.2)
	WIoU v2 (γ = 0.5)	95.2 (+0.6)	74.1 (+0.6)
	WIoU v3 (α = 1.9, γ = 3)	95.9 (+1.3)	74.5 (+1.0)

The experimental results show that the map50 when improving YOLOv8n by applying the original loss function CIoU is 94.6%, and the map50:95 is 73.5, and after replacing the loss function with SIoU, the map50 and map50:95 are improved by 0.5 and 0.3 percentage points, respectively, which is because SIoU introduces the angle of the predicted frame and the real frame based on the CIoU vector information and adds it to the penalty criterion, which effectively solves the problem of the possible direction mismatch between the target frame and the real frame. DIoU and WIoUv1 reduce the detection accuracy of the model, and the highest accuracy is achieved when WIoUv3 is used as the loss function. map50 reaches 95.9, which is improved by 1.3% compared to CIoU, and map50:95 reaches 74.5, which is improved by 1.0%.

4.6. Comparison Experiment with Other Models

To verify that the model proposed in this paper has better performance compared with other classical models, the model in this paper is compared with other models on the same dataset. The model in this paper is compared with SSD, Faster R-CNN, YOLOv3Tiny, YOLOv5n, YOLOv7tiny, and YOLOv8n on the same dataset. All models are trained on the merged KITII dataset from Section 4.1, and the experimental results are shown in Table 5. As depicted in Table 5, the Map50 of the model presented in this paper stands at 95.9%, representing a notable improvement of 4.7 percentage points compared to the pre-improvement of YOLOv8n, and 5.0 and 5.8 percentage points when compared to SSD300 and Faster R-CNN, respectively. Moreover, compared to YOLOv3 tiny, YOLOv5n, and YOLOv7 tiny, the improvement is even more substantial, with gains of 9.4, 6.4, and 6.8 percentage points, respectively, showcasing enhancements of 9.4, 6.4, and 10.0 percentage points, respectively. Comparison experiments in Table 5 show that the improved model is more capable of detecting vehicle–pedestrian features, and the algorithm is more robust in the face of partial occlusion.

Table 5. Comparison experiment of different algorithms.

Model	FLOPs (10^9)	Model Size (M)	P (%)	R (%)	Map50 (%)
SSD300	60.9	100	91.3	82.1	90.9
Faster R-CNN	105	108	90.5	80.9	90.1
YOLOv3 Tiny	12.9	17.0	87.6	70.4	86.5
YOLOv5n	4.1	3.7	91.6	76.5	90.5
YOLOv7tiny	13.1	12.0	86.1	68.9	85.9
YOLOv8n	8.1	6.1	93.3	83.7	91.2
Ours	12.5	13.3	95.5	86.1	95.9

4.7. Ablation Experiments

In rigorously demonstrating the effectiveness of the enhanced BiFormer attention mechanism, the four-scale feature detection network, the weighted bidirectional pyramid structure, and the WIoU loss function presented in this paper, ablation studies are conducted to evaluate the impact of each component on the performance of the detection algorithm under identical experimental conditions. These ablation studies utilize the original YOLOv8n network's experimental results as the baseline and explicitly indicate the employment of the respective improvement strategies.

From the results of the ablation experiments as shown in Table 6, it can be seen that the use of the four-scale feature prediction network improves the detection accuracy of the model greatly, with the *mAP*50 improved by 2.6 percentage points and the *mAP*50:95 improved by 4.5 percentage points, which indicates that the addition of a detection head can significantly improve the detection accuracy of the small targets, and the use of the weighted BiFPN and the WIoU loss function can, respectively, improve the *mAP* by 0.6 and 1.0 percentage points. The weighted bidirectional feature pyramid structure can effectively fuse features of different scales, thus improving the detection ability of targets of different sizes; the WIoU function can better measure the degree of alignment between the prediction frame and the real frame, which helps the model to locate the target more accurately. The simultaneous use of the four improvement strategies makes the model's *mAP* reach 95.9%, which is 4.7 percentage points higher compared to the pre-improvement period, and the *mAP*50:95 reaches 74.5%, which is 6.2 percentage points higher compared to the pre-improvement period. The improved YOLOv8n 's P-R curve is shown in Figure 5.

Table 6. Ablation experiment. (As indicated by the checkmark, the model in this paper adopts the aforementioned improvement strategy).

Model	BiFormer	Four-Scale	BiFPN	WIoU	*mAP*50 (%)	*mAP*50:95 (%)
YOLO v8n					91.2	68.3
	✓				92.5	70.1
		✓			93.8	72.8
			✓		91.8	69.0
				✓	92.2	69.8
	✓	✓			94.3	73.6
	✓	✓	✓		94.6	73.8
	✓	✓		✓	95.4	74.1
	✓	✓	✓	✓	95.9	74.5

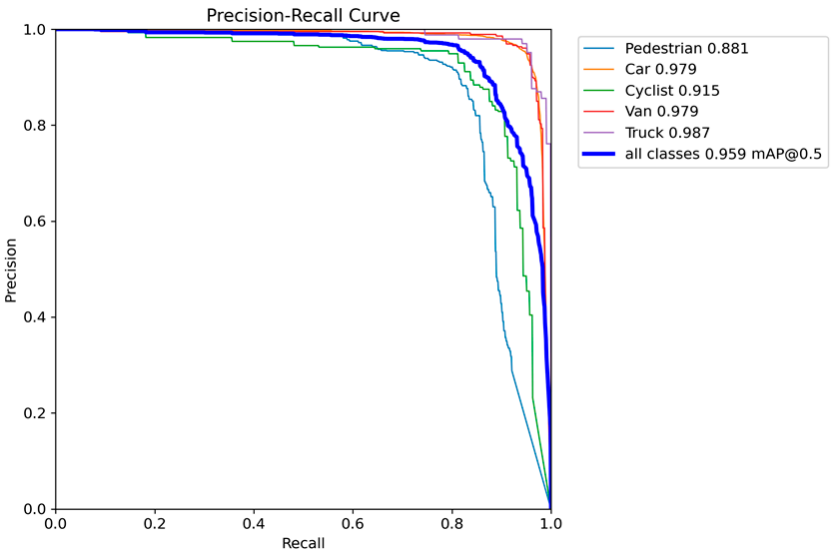
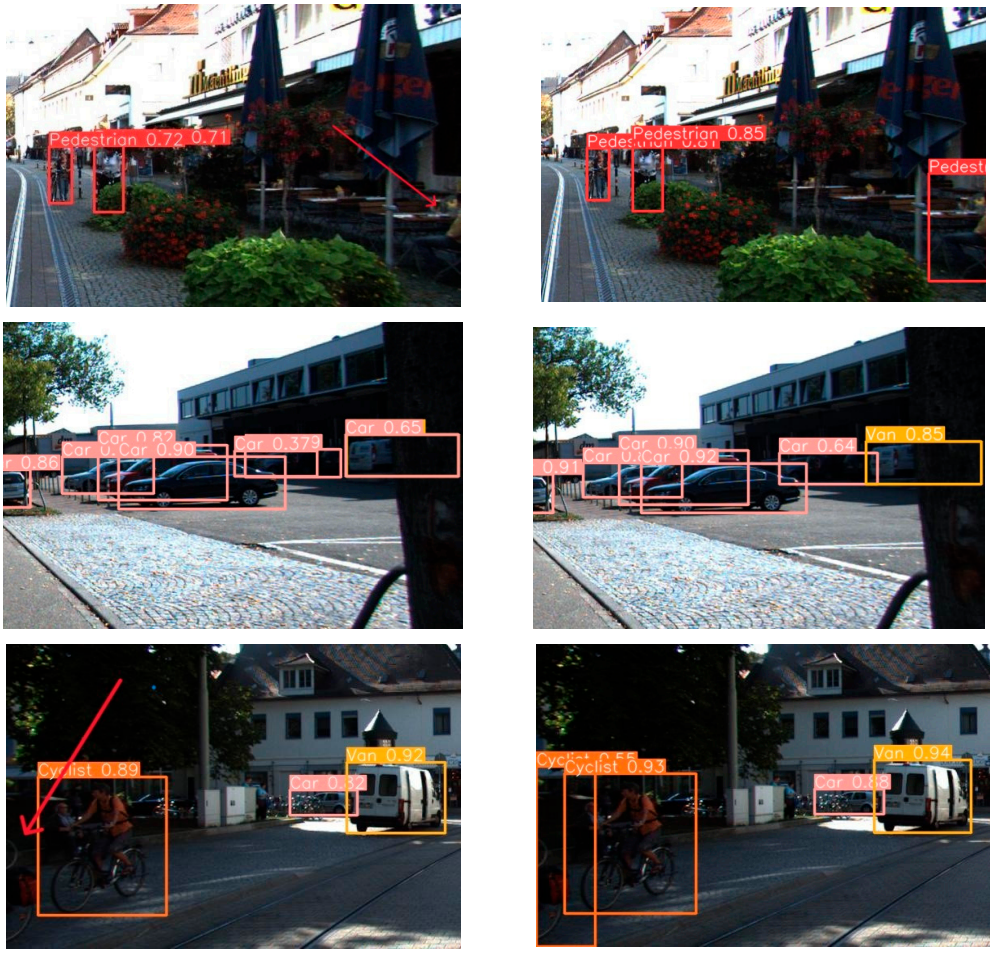


Figure 5. P-R curve of this paper's model.

4.8. Visualization of Results

As evident in Figure 6, in the first row of images, YOLOv8n fails to detect the pedestrian on the lower right, despite only a portion of their body being visible in the frame. However, the improved model successfully detects this pedestrian. In the second row of images, the target on the right is a partially occluded van, which is incorrectly identified as a Car by YOLOv8n, while the improved model correctly identifies it as a Van, and the confidence level of the improved model in detecting other vehicles is also higher than that of YOLOv8n.



(**a**) YOLOv8n (**b**) Improved YOLOv8n

Figure 6. Comparative effect of the algorithm before and after improvement.

5. Conclusions

In this paper, due to the shortcomings of YOLOv8n in vehicle–pedestrian detection scenarios, a vehicle–pedestrian detection method based on improved YOLOv8 is proposed, which effectively improves the detection capability of vehicle–pedestrian targets. The specific improvement measures encompass the following:

(1) Embedding the BiFormer attention mechanism in the Neck layer to capture the association and dependency between features more effectively.
(2) Adding a 160 × 160 small-scale target detection head at the network Head layer to enhance the network's detection capability for pedestrians and motorcycles.
(3) Using a weighted bidirectional feature pyramid structure to enhance the feature fusion capability of the network.
(4) Making WIoUv3 a loss function to enhance the network's focus on common quality anchor frames.

The detection performance of the improved network is validated on the merged KITTI dataset, and the results of the ablation experiments show that the use of the four-scale feature prediction network improves the average detection accuracy of the model by 2.6%, and the mAP50:95 by 4.5 percentage points; the usage of the weighted bi-directional feature pyramid structure and the WIoU loss function improves the mAP by 0.6 and 1.0 percentage points, respectively. The map50 of the model for vehicle–pedestrians reaches 95.9%, an improvement of 4.7%, and map50:95 reaches 74.5%, an improvement of 6.2%.

Author Contributions: B.W. first proposed the experimental idea, and together with Y.-Y.L., he formulated the overall research goal. In the subsequent research and investigation process, Y.-Y.L. and W.X. participated in the experiment or data collection together. Y.-Y.L. is responsible for collating the collected data, while W.X. is responsible for running the code and generating the images processed by the optimization algorithm. After completing the data analysis and image processing, Y.-Y.L. and W.X. collaborated on the first draft. Subsequently, Y.-Y.L. translated the first draft. After the manuscript was completed, B.W. reviewed and commented on the manuscript and raised a series of academic questions for further discussion and improvement. Throughout the research process, L.H. and H.W. took on the role of supervision and leadership, responsible for managing and coordinating the operation of the entire experimental process, ensuring that the planning and execution of research activities were carried out smoothly. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 52375260 and in part by the National Natural Science Foundation of China under Grant 51905389.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Enguo, Q.I.N. Research on the current situation and countermeasures of road traffic accidents in China. *Intern. Combust. Engine Accessories* **2018**, *16*, 184–185.
2. Jia, X.; Tong, Y.; Qiao, H.; Li, M.; Tong, J.; Liang, B. Fast and accurate object detector for autonomous driving based on improved YOLOv5. *Sci. Rep.* **2023**, *13*, 9711. [CrossRef] [PubMed]
3. Geiger, A.; Lenz, P.; Stiller, C.; Stiller, C.; Urtasun, R. Vision meets robotics: The kitti dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [CrossRef]
4. Liu, Z.; Qiu, S.; Chen, M.; Chen, M.; Han, D.; Qi, T.; Li, Q.; Lu, Y. CCH-YOLOX: Improved YOLOX for Challenging Vehicle Detection from UAV Images. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Queensland, Australia, 18–23 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–9.
5. 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]
6. Mahaur, B.; Mishra, K.K. Small-object detection based on YOLOv5 in autonomous driving systems. *Pattern Recognit. Lett.* **2023**, *168*, 115–122. [CrossRef]
7. Nosheen, I.; Naseer, A.; Jalal, A. Efficient Vehicle Detection and Tracking using Blob Detection and Kernelized Filter. In Proceedings of the 2024 5th International Conference on Advancements in Computational Sciences (ICACS), Lahore, Pakistan, 19–20 February 2024; pp. 1–8. [CrossRef]
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 142–158. [CrossRef] [PubMed]
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *39*, 1137–1149. [CrossRef] [PubMed]
11. Wang, Y.; Ji, X.; Zhou, Z.; Wang, H.; Li, Z. Detecting faces using region-based fully convolutional networks. *arXiv* **2017**, arXiv:1709.05256.
12. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
13. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
14. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
15. Du, X.; Cheng, H.; Ma, Z.; Lu, W.; Wang, M.; Meng, Z.; Jiang, C.; Hong, F. DSW-YOLO: A detection method for ground-planted strawberry fruits under different occlusion levels. *Comput. Electron. Agric.* **2023**, *214*, 108304. [CrossRef]
16. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. Biformer: Vision transformer with bi-level routing attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 10323–10333.
17. Wang, G.; Chen, Y.; An, P.; Hong, H.; Hu, J.; Huang, T. UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios. *Sensors* **2023**, *23*, 7190. [CrossRef] [PubMed]

18. Chen, J.; Mai, H.S.; Luo, L.; Chen, X.; Wu, K. Effective feature fusion network in BIFPN for small object detection. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; IEEE: New York, NY, USA, 2021; pp. 699–703.
19. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
20. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Springer: Munich, Germany, 2018; pp. 3–19.
21. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.