Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5237568, 11 pages
https://doi.org/10.1155/2021/5237568

*Research Article*

# A Vehicle Detection Model Based on 5G-V2X for Smart City Security Perception

**Teng Liu,[1] Cheng Xu,[1] Hongzhe Liu,[1] Xuewei Li,[1] and Pengfei Wang[2]**

[1]*Beijing Key Laboratory of Information Service Engineering, College of Robotics, Beijing Union University, Beijing, China*
[2]*Communication and Information Center of Ministry of Emergency Management of the People's Republic of China, Beijing, China*

Correspondence should be addressed to Hongzhe Liu; liuhongzhe@buu.edu.cn

Received 23 September 2021; Accepted 10 November 2021; Published 27 November 2021

Academic Editor: Deepak Gupta



Security perception systems based on 5G-V2X have become an indispensable part of smart city construction. However, the detection speed of traditional deep learning models is slow, and the low-latency characteristics of 5G networks cannot be fully utilized. In order to improve the safety perception ability based on 5G-V2X, increase the detection speed in vehicle perception. A vehicle perception model is proposed. First, an adaptive feature extraction method is adopted to enhance the expression of small-scale features and improve the feature extraction ability of small-scale targets. Then, by improving the feature fusion method, the shallow information is fused layer by layer to solve the problem of feature loss. Finally, the attention enhancement method is introduced to increase the center point prediction ability and solve the problem of target occlusion. The experimental results show that the UA-DETRAC data set has a good detection effect. Compared with the vehicle detection capability before the improvement, the detection accuracy and speed have been greatly improved, which effectively improves the security perception capability based on the 5G-V2X system, thereby promoting the construction of smart cities.

## 1. Introduction

With the continuous advancement of the construction of smart cities, the technology of intelligent networked vehicles has been greatly developed. They are equipped with sensors, controllers, actuators, and other devices on the basis of traditional vehicles. With the support of modern communication technology and network technology, information sharing and exchange between cars and everything is realized. To a certain extent, the intelligent networked vehicle realizes the organic unity of complex environment perception and intelligent decision-making and effectively realizes the control and execution management of the vehicle system. Improving the overall security of the smart city system has become a focal issue. V2X technology has different delay and coverage requirements in different business scenarios. For example, in active safety management such as collision avoidance and warning, the delay is required to be maintained at 20-100 ms. In business scenarios such as navigation, traffic lights, and road conditions, the delay should be controlled to 500 ms. Although the size of this type of data packet is small, its coverage is relatively large. Generally, the communication coverage of V2X technology under this type of business is as high as 1000 m, standardized control of V2X technology, and reduce the time for perception and decision-making; expand the control range of navigation, traffic lights, etc., can effectively ensure the efficiency and safety of road traffic. At the perception level, it effectively improves the detection time of vehicles and other terminals on pedestrians, vehicles, traffic signs, and other targets. It can greatly reduce the processing time of the overall network, thereby improving the communication efficiency of the 5G-V2X network. The 5G-V2X-based technology effectively realizes the complementary advantages of each vehicle-mounted sensor system. Therefore, the improvement based on 5G-V2X perception technology provides an important guarantee for the safety of the entire system [1, 2].

In recent years, deep learning technology has continued to develop and has made huge breakthroughs, using convolutional neural networks to automatically extract target features. Thanks to the powerful feature extraction capabilities of the convolutional neural network, the detection accuracy

of the target detection algorithm is greatly improved, and it has stronger robustness and can adapt to more complex recognition scenarios. In the process of vehicle target detection, there are problems such as intervehicle occlusion, target deformation, and small target size, which make the detection accuracy not high. Therefore, the detection problem of small targets and occluded targets in vehicle detection is solved, and the detection accuracy and speed are improved. It has become a hot issue in the field of vehicle target detection. In response to the above problems, researchers have proposed feature extraction methods such as feature pyramids [3–5] and target detection algorithms without anchor frame [6, 7]. Some scholars [8] proposed a spatiotemporal event interaction model (STEIM) on this basis to solve the problem of time and data interaction in the V2X environment.

The improved method based on the feature pyramid can extract the features of small-scale targets well, but it requires a lot of computational overhead. The improved method based on the anchorless frame can directly extract the target features and classify them. The detection speed is faster than the previous method, but it cannot detect small-scale targets and adjacent targets well.

To solve the above problems, this paper proposes an improved algorithm based on the single-stage target detection algorithm CenterNet. Three improved methods are mainly used:

(1) An adaptive feature extraction method is proposed to increase multiscale target features and improve the feature extraction capability of small targets

(2) An adaptive feature fusion method is proposed, adding a feature branch to effectively fuse high-level and low-level features

(3) A center point enhancement method is proposed to increase spatial information and effectively enhance the prediction ability of the target center point

The improved network I-CenterNet (Improved CenterNet) can fully extract low-level network location information, reduce the loss of feature map information in feature fusion, increase the attention of small targets, and improve the detection accuracy and speed of small targets and occluded targets.

The second chapter is related work, the third chapter is technical method, the fourth chapter is experiments and data set description, the fifth chapter is the analysis of experimental results, and the sixth chapter is the conclusion.

## 2. Smart City Security Perception Related Work

In recent years, the rapid development of deep learning has made significant breakthroughs in environmental perception technology. The proposal of AlexNet [9] in 2012 opened the curtain on the development of deep learning, and the proposal of VGGNet [10] in 2014 made the realization of deep neural networks possible. In 2015, ResNet [11] proposed to solve the gradient explosion problem through the residual connection method and reduce the model convergence time. Shi et al. [12] propose a one-stage, anchor-free detection approach to detect arbitrarily oriented vehicles in high-resolution aerial images. The vehicle detection task is transformed into a multitask learning problem by directly predicting high-level vehicle features via a fully convolutional network. Nowadays, target detection algorithms are mainly divided into two categories: one-stage method and two-stage method. The two-stage method generates a series of candidate frames through algorithms and then performs regression and classification on the candidate frames. It is characterized by high accuracy but relatively low recognition speed.

In order to overcome the above problems, the researchers proposed a one-stage method, which mainly cancels the step of candidate frame generation, directly uses a convolutional neural network to perform convolution operations on image data, and detects and classifies the extracted features. In 2016, the YOLO (you only look once) [13] series of algorithms was proposed, which solved the real-time problem of the algorithm while ensuring the recognition accuracy. The SSD [14] detection algorithm combines the advantages of the Fast R-CNN series [15, 16] algorithm and the YOLO algorithm and realizes the detection of targets of various sizes by generating candidate frames of different sizes on the multiscale feature inspection map. Such as literature [17, 18] is used for vehicle target detection. Zheng and Chen [19] improved the cascading region of interest to increase the context information, which effectively improved the detection accuracy of small targets. Liang et al. [20] used an extra scaling branch of the deconvolution module with an average pooling operation to form a feature pyramid. The original feature fusion branch is adjusted to be better suited to the small object detection task. Chen [21] and others improved the feature extraction network by adding ResNet, deconvolution, and other methods to increase the detection ability of small target vehicles. Wang et al. [22] proposed a soft weighted average method, which "punishes" the detection result of the corresponding relationship through confidence attenuation, which improves the detection accuracy of road vehicles. Xu et al. [23] used deep learning models and blockchains in combination with blockchains and supporting intelligent hardware to achieve true recording and tracking of the entire fruit process. Improve the transparency and efficiency of the supply chain and reduce the cost of the supply chain.

In recent years, there has been an anchorless frame method [24, 25], which directly detects and locates the target through key points, which greatly reduces the network parameters and calculations, improves the detection speed, and its detection accuracy is also higher than that of the traditional one-stage and two-stage method. The one-stage method slides the complex arrangement of possible bounding boxes (anchors) on the image and then directly classifies the boxes without specifying the contents of the boxes. The two-stage method recalculates image features for each potential frame and then classifies those features. Postprocessing, namely, nonmaximum suppression (NMS), deletes duplicate detection frames of the same target by calculating the IOU between the bounding boxes. The method of the

anchorless target detection network is different from other networks. For example, the ConerNet [26] algorithm uses two corner points to predict the target, and the CenterNet [27] uses the target center point to present the target, and the image needs to be transferred to the volume. In the product neural network, a heat map is obtained, and the peak center point of the heat map is the center point. Then, return to the target's size, positions, and other attributes at the center point, thus turning the target detection problem into a standard key point estimation problem.

This type of algorithm is different from the traditional one-stage method. The anchor point of CenterNet is placed in a position, which can be regarded as an anchor of a shape and position. It does not need to manually set a threshold to distinguish between the foreground and the background, so the network does not need to prepare the anchor in advance. Each target has only one positive anchor, so there is no need for NMS operation to screen candidate frames, which greatly reduces network parameters and calculations. Its detection accuracy is also higher than the traditional one-stage and two-stage methods, and the detection speed meets the requirements of real-time detection, but there are still insufficient multiscale feature extraction, and the recognition of small-scale targets and occluded targets is not accurate. Insufficient context information at the time leads to the problem of false detection and missed detection of adjacent targets. Therefore, this article is based on the CenterNet network to improve, overcome the above problems in the small-scale vehicle detection problem, and propose the I-CenterNet vehicle target detection method.

## 3. Vehicle Detection Model Based on 5G-V2X

In order to solve the problem of insufficient low-dimensional feature extraction in the vehicle target detection problem in the 5G-V2X scene, adaptive context feature extraction is adopted. In order to overcome the problem that the network is more sensitive to high-dimensional features than low-dimensional features, the feature fusion method is improved, and the weight of small target features is increased. Aiming at the problem of inaccurate prediction of the target center point position in the detection method based on the anchorless framework, a center point position enhancement method is proposed. Improve the ability of the improved network to detect small-scale targets and occluded targets in vehicle small target detection, effectively improve the vehicle perception efficiency in the Internet of Vehicles environment, and reduce the operating pressure of the 5G-V2X system. The overall structure is shown in Figure 1.

*3.1. Adaptive Context Feature Extraction.* In the process of vehicle target detection, there are problems that the target is occluded, and the target is too small. During the detection process, a large amount of feature information will be lost after convolution and pooling operations, which will reduce the detection accuracy. And CenterNet only uses ResNet50/101 as the backbone network for feature extraction, which is prone to insufficient feature extraction.

In response to the above problems, this paper uses an adaptive context feature extraction method to improve the input layer of the network as follows. As shown in Figure 2, the input feature map of the Conv3-3 layer is pooled to $3 \times 3$, $7 \times 7$, and $9 \times 9$, three different scales, to get different contextual information. Each pooled feature uses $1 \times 1$ convolution for channel integration and then uses the deconvolution operation to separate each the feature map is upsampled to the same size as the shape.

The input traffic scene picture contains vehicles of various scales, and context features cannot be simply combined. Therefore, a scale fusion unit is added after the context feature extraction network, and the weights of each feature are added to increase the weight of small-scale targets. And use the jump connection method to fuse the original features into the upsampling features, the operation is as follows:

$$y_{ff} = a \cdot f^1 + b \cdot f^2 + c \cdot f^3 + d \cdot f^4. \tag{1}$$

Among $y_{ff}$ is the output feature of the adaptive feature extraction network, which $f^k(k \in \{1, 2, 3, 4\})$ represents the context feature map extracted at different levels after upsample and using the point multiplication operation to fuse the original features. For example, the $f^1$ is as follows:

$$f^1 = f^{2\times 2} \cdot f_2. \tag{2}$$

Among $f^1$, as mentioned above, $f^{2\times 2}$ is the original $2 \times 2$ convolution feature, and $f_2$ is the feature after upsampling.

The parameters $a$, $b$, and $c$, $d$ represents scale weights, and the network can automatically learn these parameters, set $a + b + c + d = 1$, $a, b, c, d \in (0, 1)$, and take the calculation formula $a$ as an example, as shown below:

$$a = \frac{at\left(f^4\right)}{at\left(f^4\right) + at\left(f^3\right) + at\left(f^2\right) + at\left(f^1\right)}. \tag{3}$$

Among $at$ is the composition of average pooling and sigmoid activation function, which can be calculated by the same calculation method to obtain $b$, $c$, and $d$.

*3.2. Improved Feature Fusion Module.* After the context feature is extracted, it is integrated through $1 \times 1$ convolution, followed by an improved feature fusion module. It can adaptively select important spatial location information and semantic information from the context feature extraction network by weighting, and complete the information fusion after fusing each feature. Among them, the features from the bottom layer contain a large amount of spatial information, which is suitable for target positioning. The high-level features contain a lot of semantic features, which are suitable for target classification. However, the original network cannot effectively use the spatial information of the underlying network and the semantic information of the high-level features, so this paper proposes an improved feature extraction module.

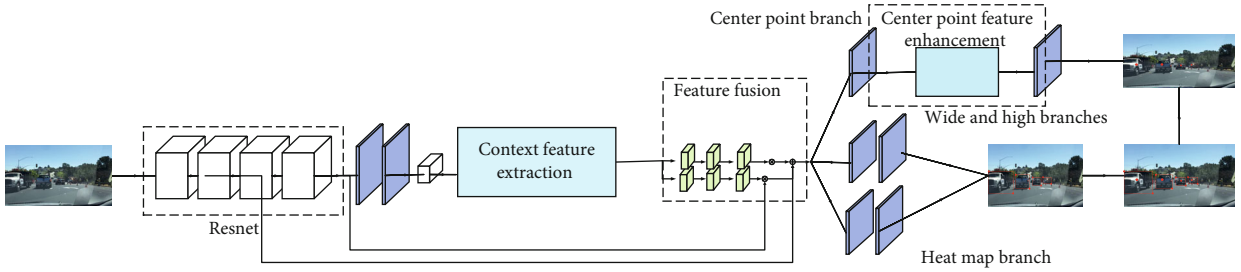



Figure 1: The overall network framework. The image has been improved feature extraction and feature fusion module, and the center point feature enhancement module of the attention mechanism is added when the center point and anchor frame are generated, and the obtained anchor frame and center point are matched to obtain the final result.

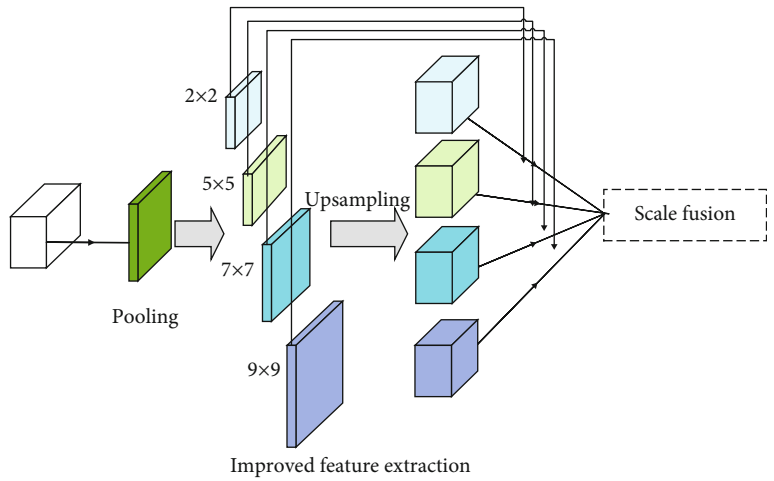



Figure 2: Feature fusion module. By pooling to different scales and using treaty links to get more features.

The improved feature extraction module proposed in this paper can adaptively perform feature fusion, as shown in Figure 3. Among them are the basic feature maps of each layer of feature extraction. Since the low- and high-level feature maps have different sizes of resolution and number of channels, bilinear interpolation is used to unify them to the same size. The input $f_{\text{in}}$ is the original input and then enters the feature selection layer, uses $1 \times 1$ convolution to continue feature smoothing, adjusts the resolution and the number of channels after a $3 \times 3$ convolution layer, and then connects to the sigmoid activation function for output. Among them, the learning weight parameters are a and b, and feature fusion is performed in the manner shown in equation (4). Since low-dimensional and high-dimensional features mainly exist in the bottom and top layers of the network, this paper simply takes only the bottom features as the input low-dimensional features, and the highest-level output is taken as high-dimensional features.

$$y = a \otimes f_l + b \otimes f_h, \tag{4}$$

where $y$ represents the final output feature of the feature fusion module, $f_1$ represents the processed low-level feature, and $f_h$ represents the processed high-level feature. $\bigotimes$ means the corresponding position is multiplied, and $\bigoplus$ means the corresponding position is added. The improved feature fusion module weights the features of different layers through the learned weights and performs the screening and fusion of feature information, which not only strengthens the semantic features in the low-level features but also adds more spatial locations to the high-level feature information.

*3.3. Center Point Feature Enhancement.* In order to solve the problem that when the original network generates the heat map to predict the target center point, the center point position does not match the true center point position. The original network believes that the peak of the heat map is the center point, and the width and height information of the target at the peak point of each feature map is used. But in the actual picture, the geometric center of the object is not necessarily the peak point of the heat map, and there is a certain deviation. And when the network faces obstructed objects, it is easy to show that the predicted heat map has only one peak point. This leads to deviations between the predicted center point of the heat map and the actual target center point. In this paper, the problem of center point position offset is solved by the way of center point feature enhancement.

Like the CBAM (Convolutional Block Attention Module) module, the channel attention module structure in this

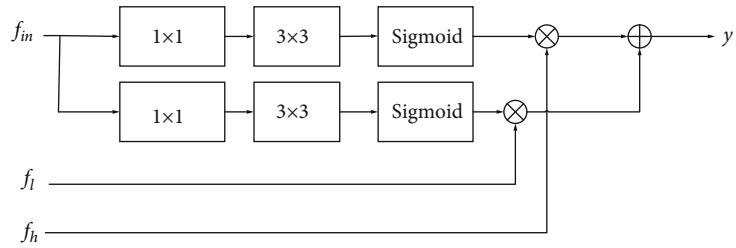



FIGURE 3: Feature fusion module. The input features go through two identical convolution modules and are, respectively, fused with the underlying and high-level semantic features, and the final fused feature is obtained.

paper is shown in Figure 4. First, the features are, respectively, passed through the maximum pooling and average pooling operations to obtain two one-dimensional vectors, and then the two channels' attention of the feature is obtained by fusing the features, which can reduce the complexity of the operation and maintain a high channel attention. The calculation can be expressed by the following formula.

$$M_c(F) = \sigma \text{AvgPool}(F) + \omega \text{MaxPool}(F). \quad (5)$$

Among them, $F$ denotes the input feature map, AvgPool and MaxPool denote average pooling and maximum pooling, respectively, and $\sigma$ and $\omega$ denote the weights of the two operations, taking 1 and 0.5, respectively.

The spatial attention structure is shown in Figure 5. First, the input features are pooled to the maximum, and then, the pooled features are averagely pooled, followed by a convolution operation with a convolution kernel of $3 \times 3$ and a skip connection. Fuse the original features of the input and the pooled features to increase the spatial feature attention, and finally output through the sigmoid function. The calculation formula is as follows:

$$M_s(F) = \partial\left(f^{3\times3}([\text{AvgPool}(F); \text{MaxPool}(F)]) \cdot F\right), \quad (6)$$

where $\partial$ represents the sigmoid activation function and represents the input feature map, and *AvgPool* and *MaxPo ol* represent average pooling and maximum pooling, respectively.

In this paper, the improved channel and spatial attention are connected in series. Since the position of the center point is sensitive to spatial information, a spatial attention module is added, as shown in Figure 6.

By introducing the center point feature enhancement module, the accuracy of center point prediction is increased, and the problem that the position of the predicted target center point in the original network does not match the true center point is solved. In the vehicle detection, the accuracy of the center point prediction of obstructed vehicles and smaller vehicles in the distance is increased.

In summary, this paper proposes an adaptive feature extraction network, which can not only extract multiscale context features but also adaptively perform weighted fusion of features according to the different scale distributions of potential targets in the input image. The improved feature

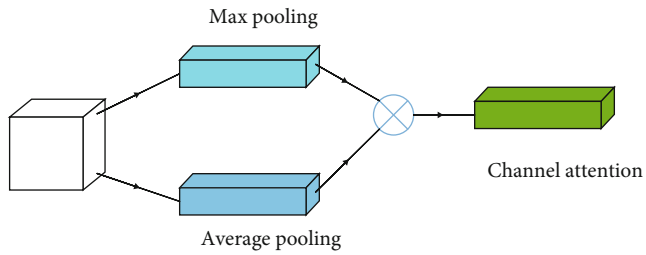



FIGURE 4: The channel attention module passes through the maximum pooling and average pooling layers, respectively, and the features obtained by the pathogenesis are multiplied to obtain the final channel attention.

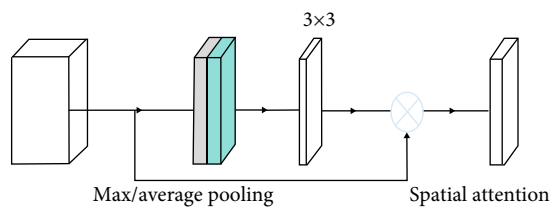



FIGURE 5: The spatial attention module adopts $3 \times 3$ convolution after two maximum pooling and average pooling and superimposes the original features to obtain the final spatial attention output.

fusion module can combine low-level and high-level features. Features are weighted and fused to increase the weight of small targets. On this basis, the central point feature enhancement module is added, which effectively improves the accuracy of the target central point position. In summary, it effectively improves the network's ability to detect small vehicle targets and occluded targets.

## 4. Analysis of Experimental Results

*4.1. Experiment Environment.* The experimental platform of this article is as follows: Intel(R) Xeon E5@1.5 GHz, 32 G memory, Ubuntu 18 system, graphics card Nvidia GTX 1080ti, the program running python environment is Python 3.6, using PyTorch 1.5, CUDA 10.1, and the data set using UA-DETRAC data. When training, the following data enhancement methods are adopted for the original data during training, and the data is amplified to increase the diversity of training samples, including random angle rotation, brightness change, noise interference, and moderate transformation.

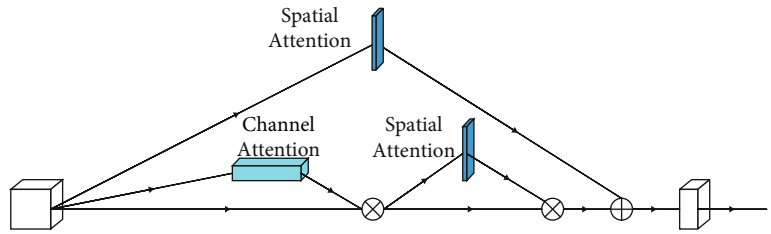



Figure 6: The center point feature enhancement passes through one spatial attention convolution module; the other uses the channel attention module first and then uses the spatial attention module. Output the two features through $1 \times 1$ convolution.

*4.2. Results and Evaluation Indexes.* This paper uses $P$ (precision), $R$ (recall), and mean average precision (mAP) to test the performance of the model.

The accuracy rate $P$ is actually the proportion of samples that are actually positive and predicted to be positive to all samples that are predicted to be positive. The formula is as follows:

$$P_{\text{pre}} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{7}$$

Among TP (true positives) refers to samples that were originally positive and classified as positive; FP (false positives) refers to samples that were originally negative but were classified as positive.

Recall rate $R$ is the proportion of samples that are actually positive and predicted to be positive to all samples that are actually positive. The formula is as follows:

$$R_{\text{re}} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

Among FN (false negative) refers to samples that were originally positive but classified as negative. The area enclosed by the P-R curve is the average accuracy mean mAP. The index to measure recognition accuracy in target detection is mAP. In multiple categories of object detection, each category can draw a curve based on recall and precision. AP is the area under the curve, and mAP is the average of multiple categories of AP.

Using the trained model to test the test set, the average accuracy can reach 89.9%, the accuracy rate $P$ reaches 94.3%, the recall rate $R$ reaches 93.7%, and the detection speed reaches 0.185/s. At the same time, the mainstream detection models are compared. As shown in Table 1, it can be seen from the table that the average accuracy of the method proposed in this paper is 7.7% higher than that of the original network, and the speed is relatively faster, and the detection speed is basically the same. In this case, the accuracy is higher than that of YOLOv4. The experimental effect is better and faster than Faster-RCNN.

The experimental results are shown in Figure 5. The results compare the I-CenterNet, Faster-RCNN, and CenterNet of this article. It can be seen from the figure that I-CenterNet effectively recognizes the smaller vehicles in the distance and successfully detects the occluded vehicles. Faster-RCNN and CenterNet cannot accurately identify small vehicles in the distance and recognize the two blocked vehicles as one target, and there are cases where individual vehicles cannot be detected.

Table 1: Comparison of multiple detection algorithms.

| Model | $P$/% | $R$/% | mAP/% | Time/s |
|---|---|---|---|---|
| Faster-RCNN [16] | 93.2% | 90.3% | 91.5% | 0.548 |
| SSD [14] | 85.6% | 79.1% | 82.3% | 0.173 |
| CenterNet [27] | 87.8% | 86.4% | 87.2% | 0.195 |
| YOLOv3 [13] | 86.3% | 83.1% | 84.8% | 0.192 |
| YOLOv4 [28] | 89.2% | 86.6% | 87.4% | 0.188 |
| Ours | 95.3% | 92.7% | 94.9% | 0.185 |

The experimental results are shown in Figure 7. The results compare the I-CenterNet, Faster-RCNN, and CenterNet of this article. It can be seen from the figure that I-CenterNet effectively recognizes the smaller vehicles in the distance and successfully detects the occluded vehicles. Faster-RCNN and CenterNet cannot accurately identify small vehicles in the distance and recognize the two blocked vehicles as one target, and there are cases where individual vehicles cannot be detected.

*4.3. Comparison of Various Modules.* In this paper, an ablation experiment is performed on each module, and the detection method is the same as above. The original CenterNet network, CenterNet+ improved feature extraction and feature fusion (CenterNet$*$), CenterNet+ center point feature enhancement (CenterNet$**$), and CenterNet+ improved feature extraction and feature fusion + center point feature enhancement (I-CenterNet) were compared, respectively. And draw the PR curve of Bus, Truck, Car, and others category, as shown in Figure 8; it can be seen from the figure that under this data set, the detection effect of the "Bus" category has been improved, when $R = 0.5$ and $P = 0.16$ improved. The accuracy of the latter algorithm is 32% higher than that of the CenterNet network. And from the figure, it can be seen that P-R curve of the improved algorithm has more surrounding area, and the AP value of the proposed model (I-CenterNet) in Bus recognition is 87.6%, which is an increase of 1.4% compared to the original model. The model proposed in this paper has an AP value of 91.5% in the identification of the Truck class, which is an increase of 0.8% compared to the original model, and the detection effect is better.

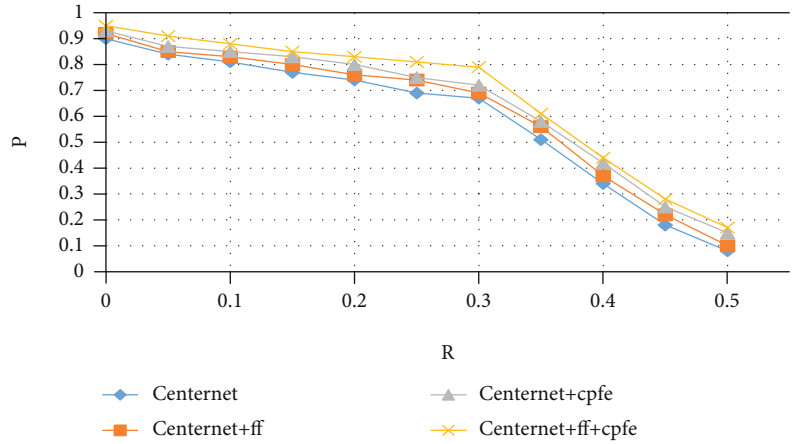
(a) Image (b) I-CenterNet

(c) Faster-RCNN (d) CenterNet

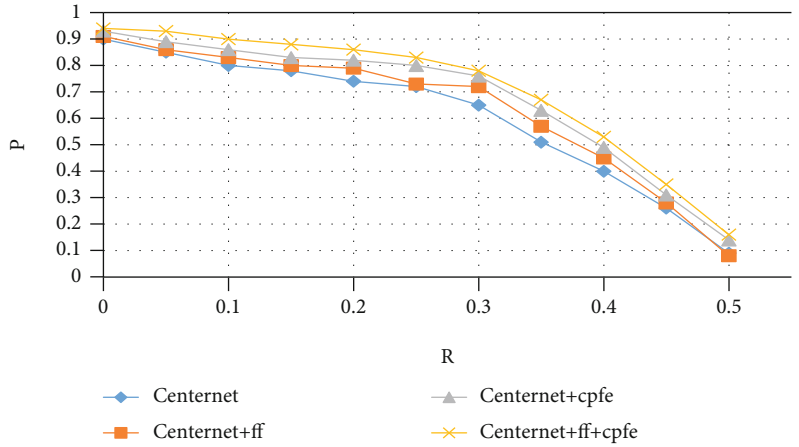
Figure 7: Experimental results of this article. Among them, (a) is the original input image, (b) is the improved result proposed in this paper, (c) is the output result of Faster-RCNN, and (d) is the output result of CenterNet. The model in this paper has identified more vehicles and effectively identified blocked vehicles.

The feature extraction and feature fusion methods proposed in this paper strengthen the extraction operation of the underlying features through skip connection and increase the spatial information of the target. On this basis, the low-level and high-level features are effectively integrated through feature fusion methods. The spatial and location information in the feature is improved, and the network's ability to extract small targets is improved. It can be seen from the comparative experiment that the improved method has greatly improved the detection performance of the network. Finally, the center point feature enhancement method is used to correct the center point position to improve the network's ability to detect current objects. The contribution rate of this part to network performance is close to 1%.

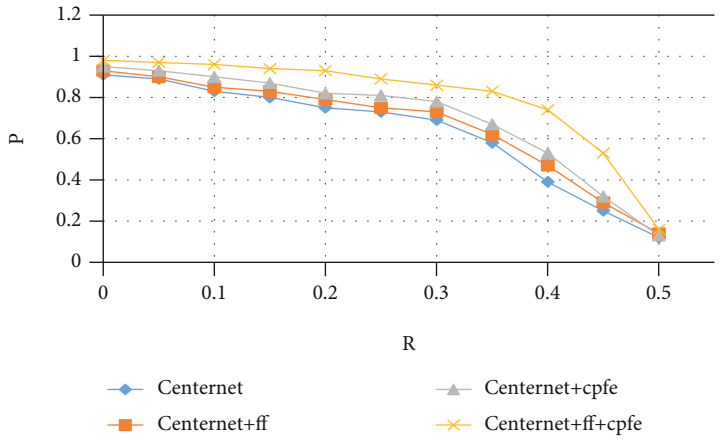
We also compared the results of center point enhancement. As shown in Figure 9, (a) represents the original image, (b) represents the result of applying center point enhancement, and (c) represents the result of the original network. The point on the vehicle in the figure is the predicted center point. It can be seen that compared with the center point predicted by the improved network, the




(a)

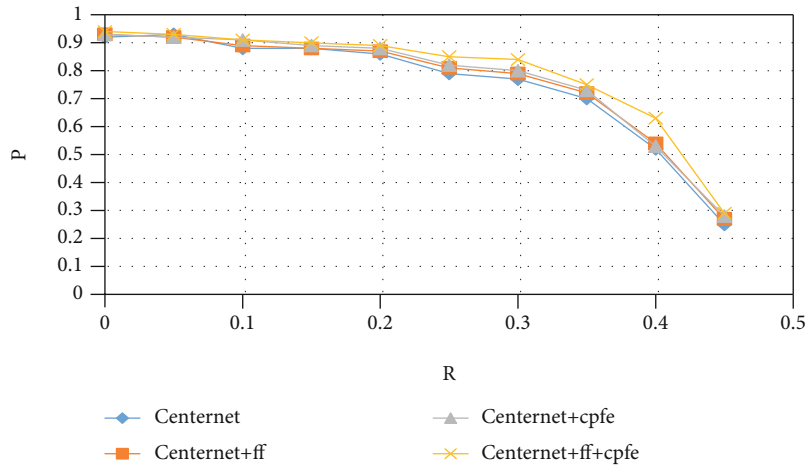



(b)




(c)

Figure 8: Continued.




(d)

Figure 8: Comparison of the accuracy-call rate curves of different algorithms in this data set: (a) Bus; (b) Truck; (c) Car; (d) others. The results show that the curve enclosing area of the model in this paper is more.

(a) Image

(b) I-CenterNet

(c) CenterNet

Figure 9: Comparison of modules with enhanced center points: (a) the input image; (b) the result of the model in this paper; (c) the output result of CenterNet. It can be seen from the results that the center point enhancement model is used, and the predicted midline point is closer to the true center point of the target.

predicted center point in the original network is deviated from the obvious predicted point in individual vehicles, and it is not in the geometric center of the vehicle. In the smaller vehicles in the distance, the block and target are smaller. The center point of the original network prediction is often close or not predicted, which will cause the predicted target frame to be inaccurate, and two vehicles that are very close together are recognized as one. The method in this paper avoids the problems of occlusion and inaccurate detection of small targets and accurately predicts the center point.

Table 2 compares and improves the detection effects of each module. It can be seen from the table that on this data

Table 2: Comparison and improvement of the detection effect of each module.

| Model | *P*/% | *R*/% | mAP/% |
|---|---|---|---|
| CenterNet [27] | 87.8% | 86.4% | 87.2% |
| CenterNet* | 90.3% | 87.3% | 89.3% |
| CenterNet** | 93.9% | 89.6% | 91.8% |
| I-CenterNet | 95.3% | 92.7% | 94.9% |

Table 3: Different backbone networks use the improved feature extraction method, and the results show that the method in this paper has a more obvious improvement effect on the ResNet network.

| Model | Backbone | Feature extraction | Box AP/% |
|---|---|---|---|
| Faster-RCNN [16] | R-101 | *N* | 39.4% |
| | R-101 | *Y* | 41.5% |
| Dynamic R-CNN [29] | R-50 | *N* | 38.9% |
| | R-50 | *Y* | 40.3 |
| ConerNet [26] | HourglassNet-104 | *N* | 40.4 |
| | HourglassNet-104 | *Y* | 40.8 |

set, the improved model is compared with the original model in the same category of detection, and the model average of the feature extraction and feature fusion modules is added. The accuracy has increased by 1.5%, and the average accuracy of the model with the attention mechanism has increased by 1.0%. And the average frame rate meets the requirements of real-time detection. According to the experimental results, the improved CenterNet algorithm performs the best overall, which not only guarantees the real-time detection but also ensures the detection ability of small vehicle targets and occluded targets. We use the feature extraction method in this article for different backbone networks, where "*N*" means not using the method in this article and "*Y*" means using the method in this article. It can be seen from Table 3 that the method in this paper has an improvement effect on the mainstream backbone network, and the improvement of the backbone network of the ResNet series is more obvious. Using ResNet 50 network, Box AP value increased by 1.4%.

## 5. Conclusion

Aiming at the 5G-V2X-based smart city security perception vehicle detection problem, this paper proposes an improved vehicle target detection algorithm based on a deep learning target detection network. An adaptive feature extraction module is proposed to increase multiscale feature extraction capabilities for small vehicle targets. The feature fusion method is improved, and low-level features and high-level features are adaptively fused through weighting, which overcomes the problem that the network is more sensitive to high-dimensional features than low-dimensional features. A center point feature enhancement method is proposed to improve the prediction accuracy of the center point position, which can improve the useful feature weight suppression and suppress the invalid weight and solve the problem of inaccurate prediction of the center point position of similar targets. The test results show that the overall performance of the model proposed in this paper is better than the original CenterNet network. The average accuracy rate is 94.9%, and the effectiveness of each module is verified through experiments. On the premise of ensuring the detection speed, the network's ability to detect small vehicles and obstructed vehicles is improved. It can realize the detection of vehicles and other targets at the vehicle terminal, improve the operating efficiency of the perception layer, reduce the communication pressure of the 5G-V2X network in the network layer, reduce the transmission and processing pressure of cloud data, which can leave more computing resources for other tasks, and improve the operating efficiency and safety of the overall system.

## Data Availability

The data sets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the Beijing Municipal Commission of Education Project (Nos. KM202111417001 and KM201911417001), the National Natural Science Foundation of China (Grant Nos. 61871039, 62102033, 62171042, 62006020, and 61906017), the Collaborative Innovation Center for Visual Intelligence (Grant No. CYXC2011), and the Academic Research Projects of Beijing Union University (Nos. BPHR2020DZ02, ZB10202003, ZK40202101, and ZK120202104).

## References

[1] C. Xu, H. Liu, Z. Pan, W. Li, and Z. Ye, "A group authentication and privacy-preserving level for vehicular networks based on fuzzy system," *Journal of Intelligent & Fuzzy Systems*, vol. 39, no. 2, pp. 1547–1562, 2020.

[2] C. Xu, H. Liu, Y. Zhang, and P. Wang, "Mutual authentication for vehicular network in complex and uncertain driving," *Neural Computing and Applications*, vol. 32, no. 1, pp. 61–72, 2020.

[3] T. S. Sharan, S. Tripathi, S. Sharma, and N. Sharma, "Encoder modified U-net and feature pyramid network for multi-class segmentation of cardiac magnetic resonance images," *IETE Technical Review, Early Access*, pp. 1–13, 2021.

[4] Q. Zhou, J. Wang, J. Liu, S. Li, W. Ou, and X. Jin, "RSANet: towards real-time object detection with residual semantic-guided attention feature pyramid network," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 77–87, 2021.

[5] S. Chen, Z. Zhang, R. Zhong, L. Zhang, H. Ma, and L. Liu, "A dense feature pyramid network-based deep learning model for road marking instance segmentation using MLS point clouds," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 784–800, 2021.

[6] S. Fan, F. Zhu, S. Chen et al., "FII-CenterNet: an anchor-free detector with foreground attention for traffic object detection," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 121–132, 2021.

[7] Z. He, Z. Ren, X. Yang, Y. Yang, and W. Zhang, "MEAD: a mask-guidEd anchor-free detector for oriented aerial object detection," *Applied Intelligence*, pp. 1–16, 2021.

[8] C. Xu, H. Luo, H. Bao, and P. Wang, "STEIM: a spatiotemporal event interaction model in V2X systems based on a time period and a raster map," *Mobile Information Systems*, vol. 2020, Article ID 1375426, 20 pages, 2020.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Computational and Biological Learning Society*, pp. 1–14, 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, Nevada, 2016.

[12] F. Shi, T. Zhang, and T. Zhang, "Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5221–5233, 2021.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Las Vegas, Nevada, 2016.

[14] W. Liu, D. Anguelov, D. Erhan et al., "October. SSD: single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, Cham, Amsterdam, Netherlands, 2016.

[15] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, Santiago, Chile, 2015.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[17] J. Cao, C. Song, S. Song et al., "Front vehicle detection algorithm for smart car based on improved SSD model," *Sensors*, vol. 20, no. 16, p. 4646, 2020.

[18] M. A. A. al-qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, "An improved YOLO-based road traffic monitoring system," *Computing*, vol. 103, no. 2, pp. 211–230, 2021.

[19] Q. Zheng and Y. Chen, "Feature pyramid of bi-directional stepped concatenation for small object detection," *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 20283–20305, 2021.

[20] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1758–1770, 2020.

[21] S. Chen, "A traffic scene target detection algorithm with dual attention module," *World Scientific Research Journal*, vol. 6, no. 9, pp. 99–107, 2020.

[22] H. Wang, Y. Yu, Y. Cai, X. Chen, L. Chen, and Y. Li, "Soft-weighted-average ensemble vehicle detection method based on single-stage and two-stage deep learning models," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 100–109, 2021.

[23] C. Xu, K. Chen, M. Zuo, H. Liu, and Y. Wu, "Urban fruit quality traceability model based on smart contract for Internet of Things," *Wireless Communications and Mobile Computing*, vol. 2021, Article ID 9369074, 10 pages, 2021.

[24] L. Jiao, S. Dong, S. Zhang, C. Xie, and H. Wang, "AF-RCNN: an anchor-free convolutional neural network for multi-categories agricultural pest detection," *Computers and Electronics in Agriculture*, vol. 174, article 105522, 2020.

[25] Z. Sun, M. Dai, X. Leng et al., "An anchor-free detection method for ship targets in high-resolution SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7799–7816, 2021.

[26] H. Law and J. Deng, "CornerNet: detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, no. 3, pp. 642–656, 2020.

[27] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6569–6578, Seoul, Korea, 2019.

[28] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020, https://arxiv.org/abs/2004.10934.

[29] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic R-CNN: towards high quality object detection via dynamic training," in *European Conference on Computer Vision*, pp. 260–275, Springer, Cham, Seoul, Korea, 2020.