Sensing and Imaging (2021) 22:19
https://doi.org/10.1007/s11220-021-00342-6

ORIGINAL PAPER

Modified Cascade RCNN Based on Contextual Information for Vehicle Detection

Xiaofei Han[1]

Received: 29 November 2020 / Revised: 26 January 2021 / Accepted: 29 March 2021 /
Published online: 7 April 2021

Abstract

In the process of traditional vehicle detection, there are some problems such as the fault detection and missing detection for small objects and shielded objects. Therefore, we propose a modified Cascade region-based convolutional neural network (RCNN) based on contextual information for vehicle detection. Firstly, the feature pyramid is improved to integrate the shallow information into the deep network layer by layer to enhance the features of small objects and occlusion objects. In here, we introduce the predictive optimization module and combine the context information of the region of interest (ROI), which makes the feature information have stronger robustness. Meanwhile, the multi-scale and multi-stage prediction is realized through the multi-threshold prediction network of internal cascade. Under the premise that the network parameters are basically unchanged, the accuracy rate is improved. Secondly, the multi-branch dilated convolution is introduced to reduce the feature loss during the down-sampling process. Finally, the region of interest and context information are fused to enhance the object feature expression. Experimental results show that the new Cascade RCNN method can better detect small and shielded vehicles compared with other state-of-the-art vehicle detection methods.

Keywords Vehicle detection · Cascade RCNN · Contextual information · ROI · Feature pyramid

1 Introduction

Vehicle detection is the research basis of unmanned driving, vehicle positioning, vehicle tracking and other fields, which has important research significance and broad application prospect. In the detection process, the detection accuracy is low

✉ Xiaofei Han
ljnan127@163.com

[1] Department of Architecture, Henan Technical College of Construction, Zhengzhou 450000, China

due to the different vehicles, the occlusion, deformation and other problems in the driving process [1–3]. Therefore, it has become a research hotspot in the field of vehicle detection to solve the problems of missing detection for small objects and occlusion objects and improve the detection accuracy during vehicle detection.

Kim et al. [4] integrated SPP network into YOLO-V3 and added multiple prediction layers, ultimately improved the detection accuracy of occlusion objects. Nguyen [5] improved the feature extraction network of Faster RCNN and used context-attention pooling to replace ROI pooling, which improved the detection accuracy of small vehicles and shielded vehicles. Zhou [6] combined the adversarial nets with the cascaded Faster RCNN to detect small objects and shielded objects, which had better robustness. Cai [7] proposed Cascade RCNN network, which improved the IOU (Intersection over Union) threshold of candidate frame layer by layer through Cascade detector, and finally improved the detection accuracy of small objects and shielded objects. Gong [8] proposed an improved panoramic traffic monitoring object detection method based on YOLOV3, and improved the accuracy of object detection through multi-scale fusion and k-means cluster analysis for object frame. Du [9] proposed a multi-object vehicle tracking algorithm based on YOLOV3 and KCF (Kernel Correlation Filter) in traffic scenes to achieve vehicle trajectory acquisition in traffic scenes. Li [10] adopted the idea of MobileNet to construct a multi-scale fully convolutional network structure combining feature pyramid, which improved the detection performance of objects in multi-task scenarios. Chen [11] improved the detection accuracy for small objects effectively by cascading the context information of ROI. Li [12] used ResNet structure to improve Darknet-53 feature extraction network and added a scale detection layer, it effectively improved the detection accuracy of small and medium objects in complex scenes. And some researchers also propose the deep learning-based vehicle detection methods [13–16].

The above methods cannot use the context information of the ROI sufficiently and ignore the importance of the shallow location information for the small object and the shielded object. Therefore, we propose a new cascade RCNN based on contextual information in this paper. Our main contributions are as follows.

1. Multi-scale feature fusion from shallow to deep.
2. Multi-branch dilated convolution down-sampling.
3. Context information fusion.
4. The new cascade RCNN can make full use of the shallow network location information that is beneficial to the detection of small objects and shielded objects, which can reduce the effective information loss of feature map caused by down-sampling in the process of feature fusion, improve the detection accuracy of shielded objects and small objects.

The organization of this paper is as follows. Section 2 introduces the Cascade RCNN. In Sect. 3, the new vehicle detection method is proposed. Section 4 displays the experiments and analysis. There is a conclusion in Sect. 5.

2 Cascade RCNN

The structure of Cascade RCNN is shown in Fig. 1, which is composed of feature extraction network ResNet101 [17], Feature Pyramid Networks (FPN) [18] and Cascade detector. After feature extraction by ResNet101, multi-scale fusion is conducted from the deep layer to the shallow layer for the output feature map of each layer. Then the fused feature map {P2, P3, P4, P5} is input to the recommended network RPN (Region Proposal Network) [19] to obtain the candidate object region.

In the detection stage, the cascade RCNN uses a cascade detector, each detector contains ROI Align, full-connection layer (FC), classification score (C) and border regression (B). During the detection process, the candidate object areas are re-sampled through the border regression B output from the previous detector. The new classification score C and border regression B are obtained by gradually improving the IOU threshold to finally improve the sample quality and network training effect.

Although the detection accuracy of cascade RCNN has been improved nearly by 2.8% than RNN based on experiments, the following problems still exist: FPN of cascade RCNN performs multi-scale feature fusion from deep to shallow, which makes each layer of feature map retain the information of the current layer and also fuse semantic features. But in the process of object detection, the shallow position information is more important to the detection of small objects and shielded objects. In the detection phase, due to the lack of contextual information of the ROI surrounding areas, it is easy to mis-detect objects when cascade RCNN detecting the occlusion object. To solve the above problems, this paper proposes a new Cascade RCNN vehicle detection method, which will be detailed explained in Sect. 3.

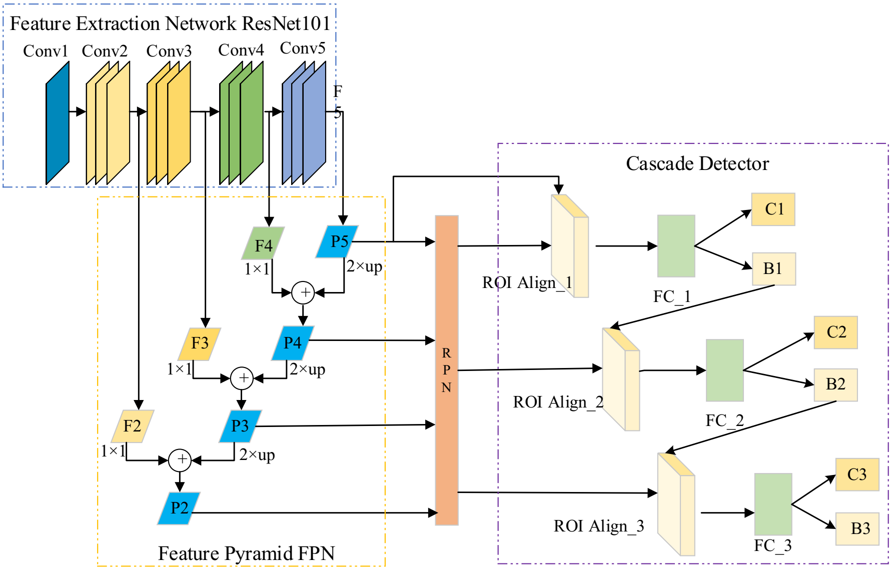


Fig. 1 Cascade RCNN structure

3 Proposed Cascade RCNN Method

The proposed vehicle detection method in this paper is shown in Fig. 2. Proposed cascade RCNN fuses the feature map output by ResNet101 network layer by layer from the shallow to the deep layer as the feature pyramid and exported to RPN. During the fusion process, multi-branch dilated convolution is used to conduct down-sampling for the shallow feature map, and then fuse with the output deep feature map after dimensionality reduction. After the selection of candidate regions by RPN, the ROI maps to the original image and its surrounding context information is fused through the unified size of the ROI Align layer [20]. Then the full connection layer is input for classification and regression. Through the border regression output from the previous detector, the candidate object area is re-sampled and the IOU threshold is gradually increased. Finally, the new classification score and border regression are obtained through training.

3.1 Improved Feature Pyramid

The feature extraction network is convolved and pooled for many times. The output feature graph of each layer is shown in Fig. 3.

Figure 3b–f are the feature graphs output from layer 1 to layer 5 in the feature extraction network ResNet101. It can be seen from the figure that the output feature map of the shallow network has a higher resolution and contains more location and detailed information, while the output feature map of the deep network has more global semantic information, but the resolution is lower and the perception of details

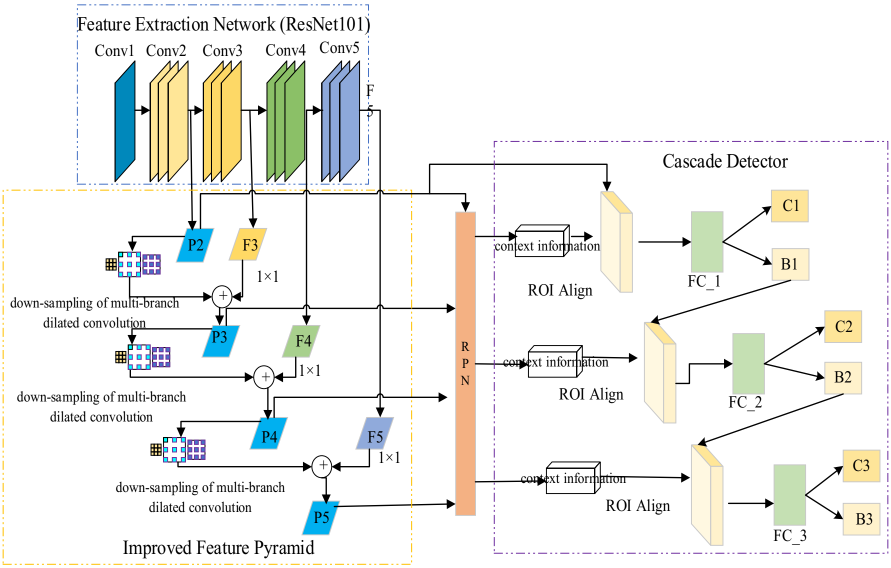


Fig. 2 Proposed cascade RCNN structure

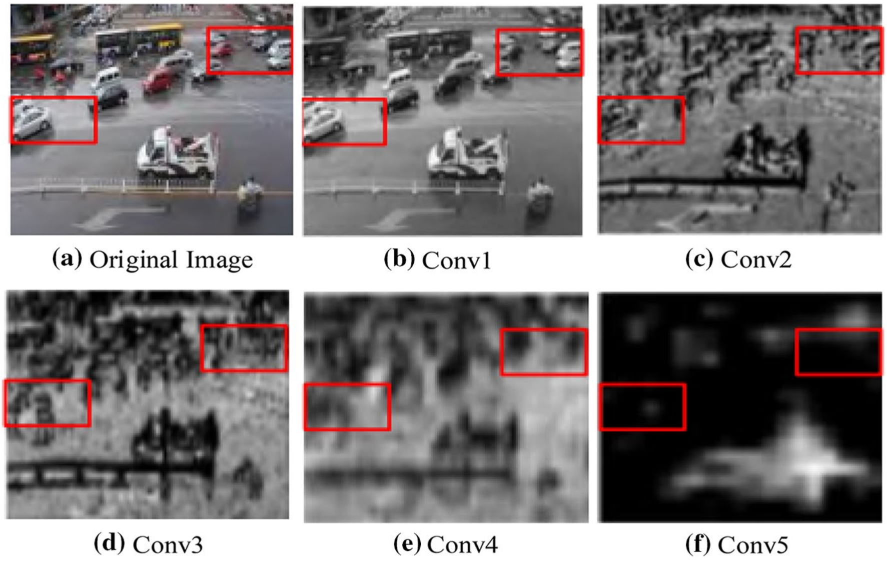

(a) Original Image
(b) Conv1
(c) Conv2
(d) Conv3
(e) Conv4
(f) Conv5

Fig. 3 Feature map visualization

is poor. In the figure, each image marked with red box is a small object with fewer features and occlusion object. After multiple convolution and pooling, the feature and its position information in the red box gradually decrease in Fig. 3b, and disappear in Fig. 3c–f. It can be seen that the features and location information about small objects and occlusion objects contained in shallow features are gradually lost after feature extraction.

The traditional feature pyramid fuses the deep features layer by layer into the shallow network, making the shallow network retain location information while fusing more semantic information. For example, Fig. 3f, e are fused. However, in the detection process of small objects and occlusion objects, the shallow features contain more detailed information. In order to enhance the features of small objects and occlusion objects, this paper improves the traditional feature pyramid and fuses the output feature graph of the shallow network layer by layer into the deep network. The improved feature pyramid is shown in Fig. 4. Because the output feature graph of the first layer contains much noise information with sliding window approach. The resolution of this output feature graph is very large, which is easy to increase the network operation cost. Therefore, from the feature extraction network Conv2 output feature graph P2, an improved feature pyramid is constructed. First, the size of P2 is reduced through the down-sampling operation. In order to ensure that the depth and shallow feature map can be combined with the addition operation, 1×1 convolution is needed to change the number of channels in the deep feature map, so that the number of channels is the same as that in the shallow feature map after the down-sampling. Then, the down-sampled P2 is fused with the Conv3 output feature map with a change of channel number to obtain P3. Since the direct superposition of features is likely to cause discontinuity and feature chaos, 3×3 convolution is used

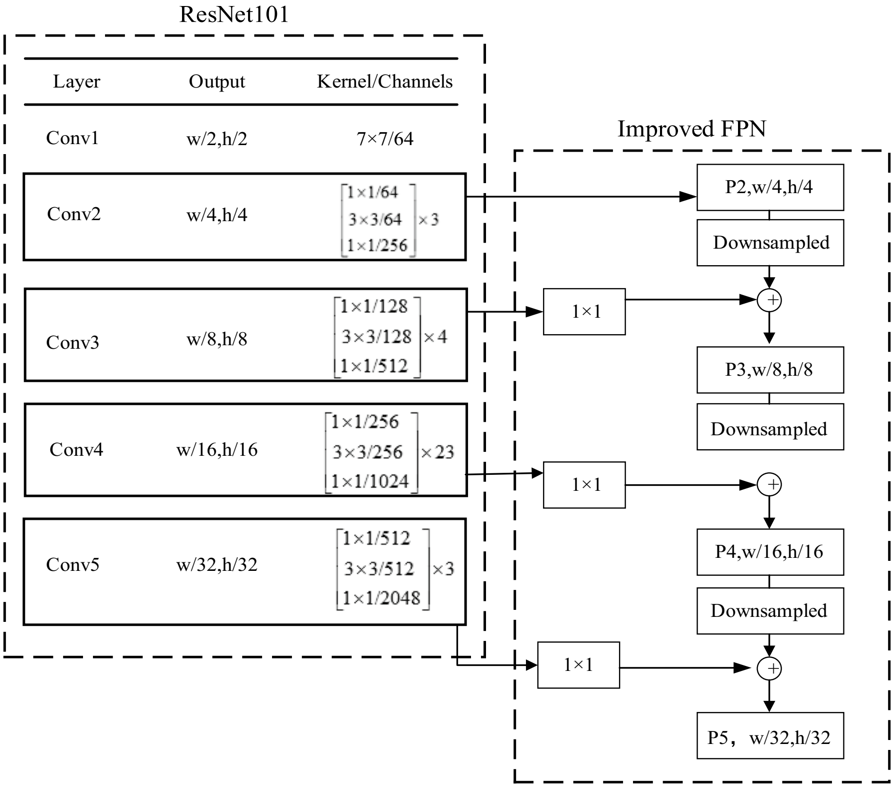


Fig. 4 Improved feature pyramid

to conduct convolution operation on the fused feature map to eliminate the difference in feature distribution between different feature maps and ensure the stability of features. Similarly, P4 and P5 can be obtained, and the specific fusion formula is as follows:

$$F_i = conv(I_i) + \dim(I_{i+1}), \quad i \in [2, n] \tag{1}$$

where F_{i} is the feature graph after the i-level fusion. I_i represents the ith feature graph output by ResNet101 network, $conv(\cdot)$ is the down-sampling operation for the feature graph of the i-layer. $\dim(\cdot)$ represents the 1×1 convolution dimensionality reduction operation for the feature graph of the $i + 1$ layer. n denotes the ResNet101 network hierarchy. The above fusion method not only provides semantic information of the current layer, but also fuses rich location information of the shallow layer, which is beneficial to the detection of small objects and occlusion objects.

In the pyramid structure, we introduce a prediction optimization module, its basic structure is shown in Fig. 5. This module mainly combines the context information of the ROI to obtain the high-semantic feature expression, and improves its positioning capability through the cascade multi-threshold prediction network.

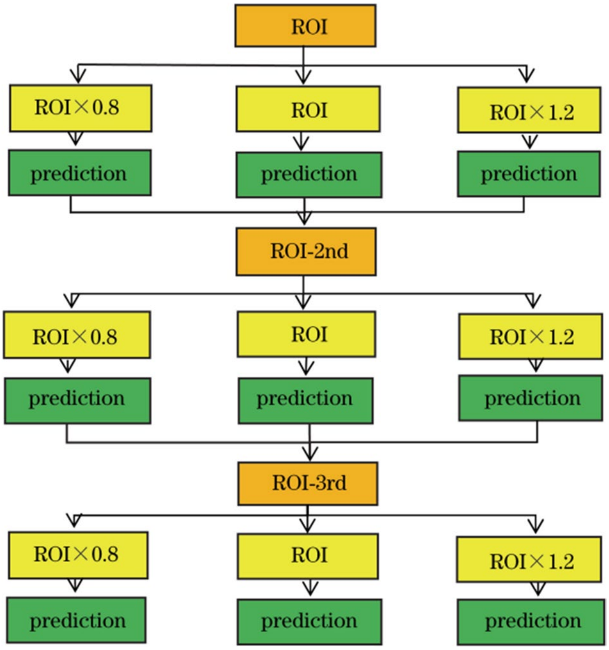


Fig. 5 Prediction optimization module in feature pyramid

This module is divided into three phases. In the first phase, the ROI region generated in the feature pyramid is first scaled (×0.8, ×1, ×1.2). Thus, it is divided into three branches, each of which is followed by the same prediction network for calculating classification and bounding box losses. After the prediction network, the bounding box losses of each branch are added up, and the average is taken as the final loss for the bounding box regression. The generated new ROI is then put into the second stage to do the same. In each stage, the threshold of prediction network is different, and the threshold of the latter stage is larger than that of the former stage. Finally, the prediction results of three stages can be obtained. In network prediction, the boundary box regression algorithm with different cascaded thresholds is used.

3.2 Multi-branch Dilated Convolution Down-Sampling

In the process of feature fusion as described in 3.1, the shallow feature map is sampled and then fused with the reduced dimension deep feature map. The down-sampling can enlarge the receptive field effectively, which makes a single pixel represent a wide range of features [21]. But the features of small objects and occlusion objects are gradually reduced or even lost in the mapping process.

In order to reduce the loss of feature, we use dilated convolution [22] for down-sampling. Compared with the common convolution, the dilated convolution can expand the receptive field and preserve the image resolution without additional computation. Because different scale objects correspond to different receptive fields, three kinds of dilated convolutions with different dilated ratios can be used in parallel to obtain different range and size of information around the object. At the same time, the convolution range of different dilated convolution is different, which can retain features with different range after convolution and reduce feature loss. The operation is shown in Fig. 6.

As shown in Fig. 6, the input feature map is processed uniformly by parallel dilated convolution with multiple different dilated rates. The specific method is as follows:

1. using three types of 3 × 3 dilated convolution with step size 2 to down-sampling, and the dilated rates are 1, 2, 3, respectively.
2. Then, the feature map after convolution is fused with concat fusion method containing batch normalization.
3. Finally, the dimensionality reduction operation is carried out with 1 × 1 convolution to facilitate subsequent multi-scale feature fusion.

This process can be expressed as the following formula:

$$F = H_{3,1(x)} + H_{3,2(x)} + H_{3,3(x)} \tag{2}$$

where $H_{k,r(x)}$ represents dilated convolution, k represents the size of the convolution kernel. r represents the dilated rate. F is the fused feature. The feature graph of single convolution down-sampling and multi-branch dilated convolution down-sampling are compared as shown in Fig. 7.

The down-sampled feature map is presented visually in the thermal form. The marked rectangular box in each image is the area of obvious feature. Where m

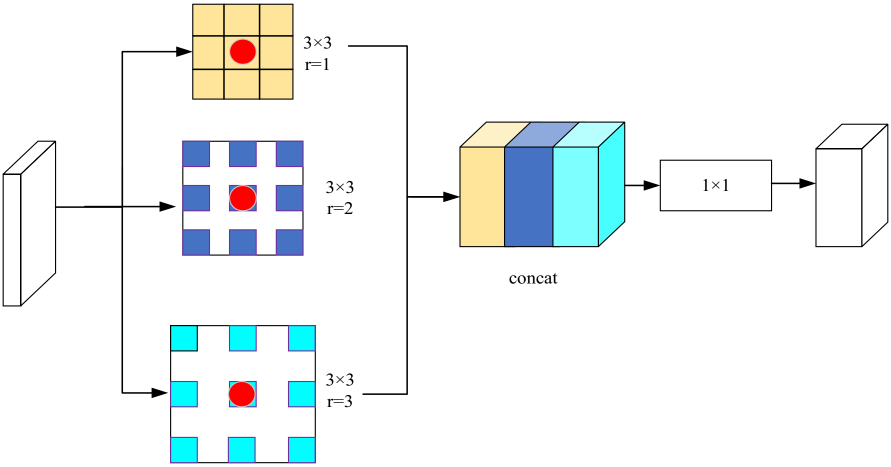
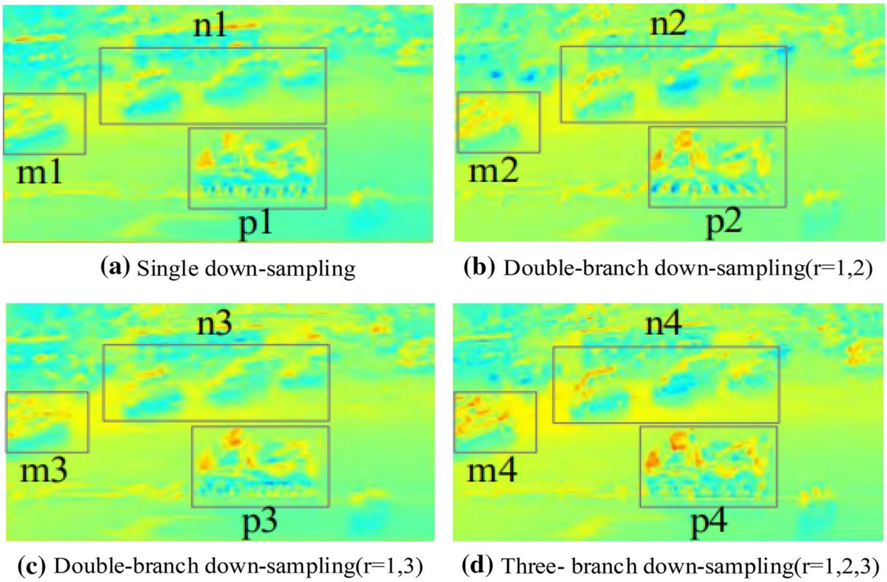


Fig. 6 Multi-branch dilated convolution down-sampling



(a) Single down-sampling **(b)** Double-branch down-sampling(r=1,2)

(c) Double-branch down-sampling(r=1,3) **(d)** Three- branch down-sampling(r=1,2,3)

Fig. 7 Thermal comparison chart of sampling

is the small object region or occlusion object region with fewer features. n is medium object region. p is large object region.

By comparing Fig. 7a–c, it can be seen that m_2 and m_3 have more features than m_1. n_2, p_2 and n_3, p_3 have richer features than n_1 and p_1. The object contours are more clear. Therefore, the down-sampling of parallel double-branch dilated convolution reduces the feature loss of small objects and occlusion objects, while it enriches the feature of medium objects and large objects. When the dilated rate is 3, the receptive field of the dilated convolution is larger and contains a wide range of features, so that more global object features can be obtained. After connected with a single convolution, the fused global features can make the object contour clearer. When the dilated rate is 2, the receptive field of cavity convolution is small, and more delicate features can be extracted with a fused convolution to enrich the features of small objects and occlusion objects. By comparing Fig. 7b, c, it can be seen that m_2 has more features than m_3, but n_3 has clearer object contour features than n_2. Therefore, it can be concluded that the parallel dilated convolution with different dilated rates can enhance the object features. The smaller dilated rate can retain more detailed information, and the larger dilated rate can make the object contour feature clearer.

It can be seen from Fig. 7d that m_4 retains the most features, while p_4 and n_4 have richer features and clearer object contour feature compared with the other three graphs. Therefore, parallel three-branch down-sampling can obtain multi-scale receptive field information, which effectively enhances the features of small objects and occlusion objects, reduces feature loss, and enriches the features of medium objects and large objects.

Through the above analysis, compared with single convolution down-sampling, multi-branch dilated convolution down-sampling enables the final fused feature map to retain the features after single convolution down-sampling and integrate the surrounding information of the object. Finally, it can enrich the object features and reduce feature loss.

3.3 Context Information Fusion

Because ROI contains limited features, there is little information about the object occlusion when the objects are highly overlapped. When the occlusion vehicles have the same or similar features with surrounding vehicles, it is easy to cause uncertainty in the subsequent classification and regression resulting in low detection accuracy for occlusion objects. In this paper, local context information will combine the surrounding area of ROI to enrich the feature expression of the object. The detailed process is shown in Fig. 8.

The ROI region and its context information are mapped to a rectangular box with same size by ROI Align. Then they are fused by addition operation. The context information is obtained by magnifying the ROI area. Assuming that the width and height of each ROI area are w and h. The magnification scale factor of the ROI area is set as 1.5, the width and height of the area containing the context information are 1.5w and 1.5 h with the same center as ROI. The specific fusion formula is as follows:

$$f = R[r(w,h)] + R[r(w',h')] \tag{3}$$

where f is the fused feature graph. $R[\cdot]$ represents the ROI Align. $r(\cdot)$ denotes the object region, w is the width of the region, h is the height of the region. $w'$and h' are the width and height multiplied by the magnification factor. The fusion of ROI region and its context information can enlarge the perception region of the feature map, so the improved method can learn more abundant occlusion object information.

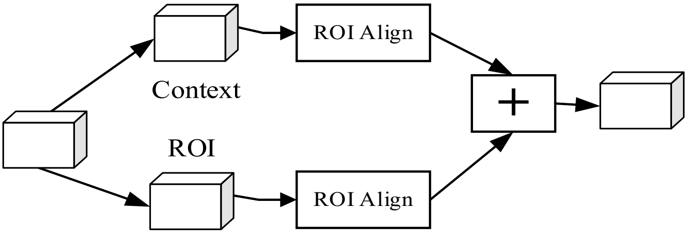


Fig. 8 Contextual information fusion

4 Experiment and Result Analysis

4.1 Parameter and Data Setting

The experiment environment is as follows: operating system Ubuntu16.04, NVIDIA GeForce RTX 1060Ti, Xeon(R) Silver 4110 CPU @2.10 GHz, Python3.6.

In this paper, KITTI [23] and UA-DETRAC [24] data sets will be used for experiments comparison analysis.

The KITTI data set includes 15 vehicles and 30 pedestrians in each image at most with a lot of vehicle occlusion and interception. Because this data set contains a variety of data categories, the categories are irrelevant to the vehicles that will be deleted during the experiment process. The different types of vehicles are merged into one category.

The UA-DETRAC dataset consists of vehicle video sequences including various scenarios such as sunny days, cloudy days, rainy days, nighttime. When conducting experiments on this data set, the effectiveness of proposed method will be tested on different scenarios. The division of KITTI and UA-DETRAC data set is shown in Table 1.

4.2 Evaluation Index

In this paper, AP (Average Precision) and FPS (transmission frame rate per second) are adopted as evaluation indexes for the vehicle object detection. Recall-Precision curve is drawn to evaluate the performance of the detection method. The area enclosed by RP curve is the AP value. The larger area denotes the greater AP value and the higher detection accuracy. Where, AP, Recall and Precision are calculated as follows:

$$AP = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

Table 1 Data set division

Data set	KITTI	UA-DETRAC
Number of images	7500	140,000
Image size	1242×375	960×540
Training set	6000	84,000
Test set	15,000	56,000

where *TP* is the probability of detected positive sample. *FP* is the probability of detected negative sample. *FN* is the probability of undetected correct sample. *TN* is the probability of detected false sample.

4.3 Training Parameters

During the training, the momentum is set as 0.9. The initial learning rate is set as 0.001. The iteration number is set as 180,000. The attenuation coefficient is set as 0.0005. When the iteration number is set as 70,000 and 120,000, the learning rate is adjusted to 0.0001 and 0.00001. The loss convergence of proposed method on KITTI and UA-DETRAC data set is shown in Fig. 9.

In Fig. 9, after 180,000 iterations, the loss value tends to be stable, and the model reaches to the optimal state.

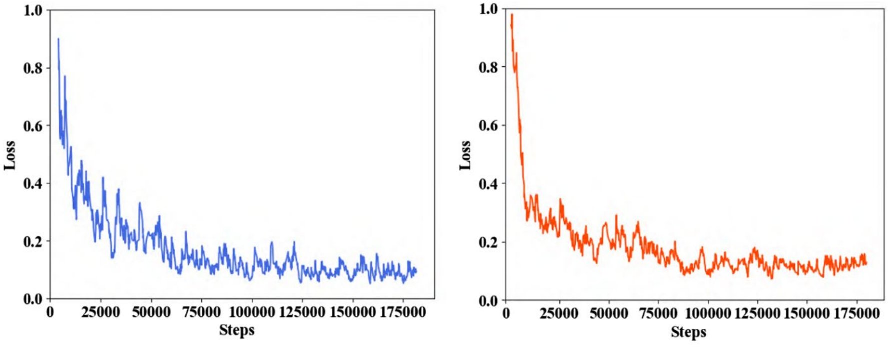
4.4 Experiment Results

We select four state-of-the-art object detection methods to make comparison: Faster RCNN [25], R-FCN [26] and CoupleNet [27], Cascade RCNN. Faster RCNN is a two-stage network based on RPN. R-FCN is a fully convolution network based on RPN. Since it does not contain the full connection layer, the detection speed is faster. CoupleNet is based on R-FCN, which combines global information and local information, so the detection speed is lower than R-FCN.

Experiment on KITTI data set.

In this subsection, comparison will be conducted on KITTI to verify the effectiveness of the proposed method. The IOU and confidence thresholds are set to 0.5. The results on KITTI are shown in Table 2. The highlighted denotes the best value.

In Table 2, "F" denotes the improved pyramid. "C" denotes multi-branch dilated convolution down-sampling. "Context" represents the context information fusion.



(a) KITTI **(b)** UA-DETRAC

Fig. 9 Loss function curve

Table 2 Results on KITTI

Method	AP (%)	FPS
Faster RCNN	78.96	12.8
R-FCN	83.27	19.6
CoupleNet	85.77	15.3
Cascade RCNN	89.72	7.2
Cascade RCNN + F	91.05	7.0
Cascade RCNN + F + C	91.38	6.6
Cascade RCNN + Context	91.74	6.1
Proposed	**93.85**	**5.9**

The bold values denote the best results

In KITTI data set, the AP value of proposed method increases by 4.13%, 8.08%, 10.58%, 14.89% than Cascade RCNN, CoupleNet, R-FCN, Faster RCNN respectively. FPS value of proposed method reduces by 1.3, 9.4, 13.7, 6.9 than Cascade RCNN, CoupleNet, R-FCN, Faster RCNN respectively.

The ablation results show that the improved feature pyramid improves the detection accuracy by 1.33% and reduces FPS by 0.1. The detection accuracy is increased by 1.65% and FPS decreased by 0.4 with adding the multi-channel dilated convolution down-sampling. Contextual information fusion results in a better results compared with other methods.

In order to verify the effectiveness of the down-sampling model of multi-branch cavity convolution, this paper uses single convolution, parallel double-branch cavity convolution and three-branch cavity convolution to conduct down-sampling for the shallow feature map, and the final detection results are shown in Table 3.

According to Table 3, the following conclusions can be drawn. Compared with the single convolution down-sampling, the detection accuracy of the dilated convolution with dilated rates 1, 2 and 1, 3 is only improved by 0.23% and 0.32% respectively. Therefore, only two kinds of dilated convolutions are connected to carry out the down-sampling, the detection accuracy is improved slightly due to the lack of comprehensive information around the object. Compared with the double branch and single branch, the detection accuracy is improved by 2.66% with three different dilated rates in parallel.

Table 3 Experiment results of dilated convolution

Method	Dilated rate	AP (%)	Detection speed (ms)
Single convolution	1	92.13	124
	1, 2	92.36	126
Multi-branch dilated convolution	1, 3	92.45	128
	1, 2, 3	**94.79**	**129**

The bold values denote the best results

In order to ensure the same size of feature map after down-sampling with different dilated convolution, the original feature map needs to be padded before down-sampling. The larger dilated rate denotes the larger padding value. Bigger padding value can increase network computation. The detection speed slows down with multi-branch dilated convolution. Compared with the single convolutional down-sampling, the detection time of each image increases from 124 to 129 ms, which adds 5 ms in total and increases the network operation cost.

The experiment data above verifies the effectiveness of the proposed innovation in this paper. The combination of multi-branch dilated convolution and context information fusion module reduces the detection speed to some extent, but it is within the acceptable range. The PR curve of each method is shown in Fig. 10. Figure 10 shows that proposed method has a higher accuracy rate with the same recall rate. The RP curve is larger.

Experiment on UA-DETRAC dataset.

In this section, comparison experiments will be conducted under different weather conditions in UA-DETRAC data set to verify the effectiveness of the proposed method in this paper. Meanwhile, the accuracy of small, medium and large objects detected with different methods is analyzed. In this paper, MS COCO data set is adopted to distinguish and define small object, medium object and big object [28] as shown in Table 4. It sets the same IOU and confidence thresholds as KITTI. The final results are shown in Table 5.

As shown in Table 5, the AP values of proposed method in the comprehensive scenario increases by 2.7%, 8.69%, 12.64%, and 24.06% compared with Cascade RCNN, CoupleNet, R-FCN and Faster RCNN, respectively. FPS is reduced by 0.9, 9.4, 13.3, 6.2, respectively. In addition, the detection accuracy of proposed method

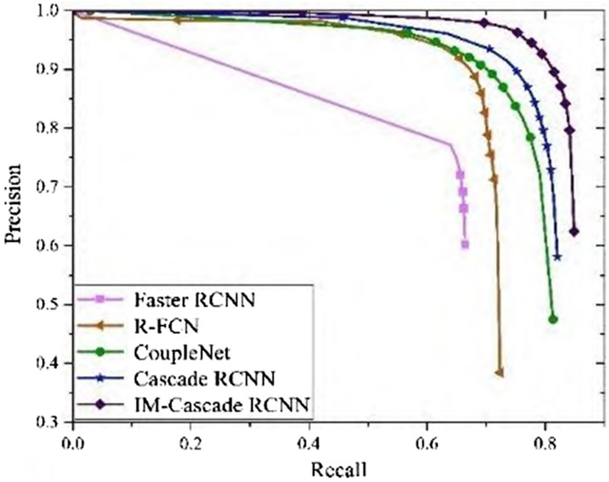


Fig. 10 RP curve on KITTI data set

Table 4 Object size division method

Object size	Minimum object box area	Maximum object box area
Small object	0× 0	32× 32
Medium object	32× 32	96× 96
Large object	96× 96	∞× ∞

Table 5 AP test results on UA-DETRAC data set/%

Method	Faster RCNN	R-FCN	CoupleNet	Cascade RCNN	Proposed
Sunny	62.35	84.09	85.24	89.13	**90.46**
Cloudy	66.31	74.38	77.19	85.53	**88.45**
Rainy day	45.17	56.23	63.45	68.97	**72.38**
Nighttime	69.86	75.12	76.12	78.18	**81.49**
Small	39.27	48.85	53.24	61.53	**64.35**
Medium	62.61	72.16	77.29	83.76	**86.74**
Large	71.48	79.94	83.64	88.19	**90.57**
Comprehensive	58.46	69.88	73.83	79.82	**82.52**
FPS	11.2	18.3	14.4	5.9	**5.1**

The bold values denote the best results

is the highest in the four different scenarios including sunny day, cloudy day, rainy day and night, which proves that the proposed method in this paper has good detection effect under different weather conditions and it is applicable to vehicle detection in the complex scenarios. Meanwhile, proposed method has a good detection effect for small objects, medium objects and large objects. Compared with the original Cascade RCNN method, the detection accuracy of the proposed method for small objects and medium objects has been improved by 2.82% and 2.89%, respectively. The detection accuracy of large objects improves to 90.57%. We can draw a conclusion that the occlusion situation is more serious, it is difficult to detect the small and medium vehicle. Therefore, the increasing detection accuracy of proposed method for small and medium targets is slightly higher than that of large targets, which proves that the proposed method in this paper can improve the problem of wrong detection and missing detection for small targets and shielded targets.

Since the UA-DETRAC data set contains more small objects and occlusion objects, the effect is more obvious than KITTI data set. The PR curve of each method in UA-DETRAC is shown in Fig. 11.

It can be seen from Fig. 11 that the PR curve of proposed method is the highest, and the accuracy is greatly improved compared with other methods.

Detection results.

Figure 12a–e shows the detection results with the five methods. a green box represents a correctly detected vehicle, a yellow dashed box represents an undetected

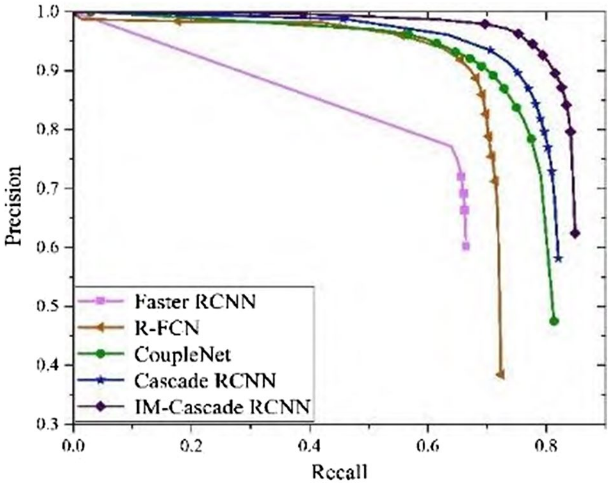


Fig. 11 RP curve of UA-DETRAC data set

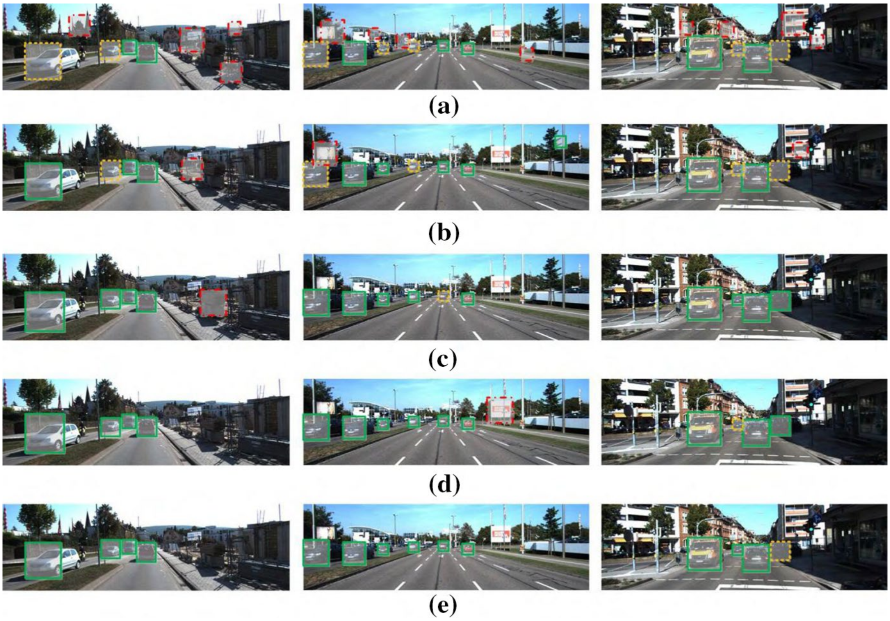


Fig. 12 Detection results of different methods. **a** Faster RCNN; **b** R-FCN; **c** CoupleNet; **d** Cascade RCNN; **e** Proposed method

vehicle and a red dashed box represents a falsely detected vehicle. Figure 13 is the examples of detection results with proposed method, the data source is from reference [29].

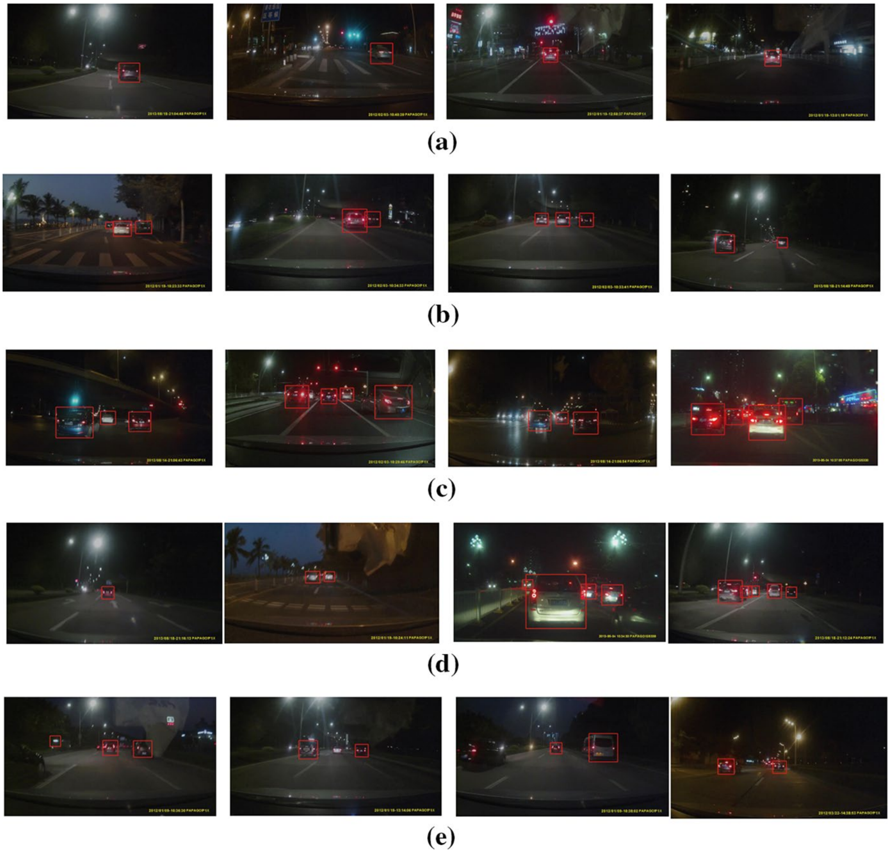


Fig. 13 Examples of detection results with proposed method. **a** Single and **b** multiple vehicles detected with the proposed method. **c** Vehicles detected accurately under complex backgrounds. **d** Blurred and **e** partly occluded vehicles are also detected, although **f** some results missed vehicles or produced false positives

By comparing (a), (b), (c), (d) and (e), it can be seen that when using Cascade RCNN for vehicle detection, due to the serious loss of detail information in the process of feature extraction and the insufficient using of context information, small objects and occlusion objects are not well detected. Under the condition of the same image resolution and detection threshold, proposed method adopts the feature fusion method to retain the position information of small objects, so that small objects can be better detected. Due to the introduction of multi-branch dilated convolution and context information module, the proposed method integrates more information around the object, which makes the occlusion object be better detected, as shown in Fig. 12e. Therefore, the proposed method has good detection effect compared with other methods.

5 Conclusion

The proposed vehicle object detection method in this paper fuses abundant location information and context information. And it reduces the feature loss caused by down-sampling in the feature map. The experiment results show that proposed method can effectively improve the detection accuracy, reduce the wrong detection and missing detection for small objects and occlusion objects under different weather conditions. However, due to the introduction of multi-branch dilated convolution and context information extraction module, the detection speed is reduced to a certain extent. How to balance the detection accuracy and speed is our next research direction. And we will research the advanced deep learning networks to conduct vehicle detection tasks.

References

1. Nastaran, Y. E., Menéndez, J. M., Jiménez, D., et al. (2018). Robust vehicle detection in different weather conditions: Using MIPM. *PLoS ONE, 13*(3), e0191355.
2. Zhou, H., Wei, L., Lim, C. P., et al. (2018). Robust vehicle detection in aerial images using bag-of-words and orientation aware scanning. IEEE *Transactions on Geoscience and Remote Sensing*.
3. Yin, S., Zhang, Ye., & Karim, S. (2018). Large scale remote sensing image segmentation based on fuzzy region competition and Gaussian mixture model. *IEEE Access, 6*, 26069–26080.
4. Kim, K. J., Kim, P. K., Chung, Y. S., et al. (2019). Multi-scale detector for accurate vehicle detection in traffic surveillance data. *IEEE Access, PP*(99), 1–1.
5. Nguyen, H. (2019). Improving faster R-CNN framework for fast vehicle detection. *Mathematical Problems in Engineering, 2019*(3), 1–11.
6. Zhou, T., Li, Z., & Zhang, C. (2019). Enhance the recognition ability to occlusions and small objects with Robust Faster R-CNN. *International Journal of Machine Learning and Cybernetics, 2019*(9).
7. Cai, Z., & Vasconcelos, N. (2018). Cascade R-CNN: Delving into high quality object detection. *Computer Vision and Pattern Recognition,* 6154—6162.
8. Gong, J., Zhao, J., Li, F., et al. (2020). Vehicle detection in thermal images with an improved yolov3-tiny. *2020 IEEE international conference on power, intelligent computing and systems (ICPICS)*. IEEE.
9. Du, K., Song, J., Wang, X., et al. (2020). A multi-object grasping detection based on the improvement of YOLOv3 algorithm. In *2020 Chinese control and decision conference (CCDC)*.
10. Li, Y., & Huang, H., et al. (2018). Research on a surface defect detection algorithm based on MobileNet-SSD. *Applied Sciences*.
11. Jingming, C., Jie, J., & Weifeng, W. (2019). Improved algorithm based on feature pyramid networks. *Laser & Optoelectronics Progress, 56*(21), 211505.
12. Li, S., Lin, J., Li, G., et al. (2018). Vehicle type detection based on deep learning in traffic scene. *Procedia Computer Science, 131*, 564–572.
13. Yahya, M. A., Abdul-Rahman, S., & Mutalib, S. (2020). Object detection for autonomous vehicle with LiDAR using deep learning. *2020 IEEE 10th International conference on system engineering and technology (ICSET)*. IEEE.
14. Tan, Q., Ling, J., Hu, J., Qin, X., & Hu, J. (2020). Vehicle detection in high resolution satellite remote sensing images based on deep learning. *IEEE Access, 8*, 153394–153402. https://doi.org/10.1109/ACCESS.2020.3017894.
15. Kausar, A., Jamil, A., Nida, N., et al. (2020). Two-wheeled vehicle detection using two-step and single-step deep learning models. *Arabian Journal for Science and Engineering, 45*, 10755–10773. https://doi.org/10.1007/s13369-020-04837-4.

16. Sudha, D., & Priyadarshini, J. (2020). An intelligent multiple vehicle detection and tracking using modified vibe algorithm and deep learning algorithm. *Soft Computing, 24*, 17417–17429. https://doi.org/10.1007/s00500-020-05042-z.
17. Yin, S., Liu, J., & Li, H. (2018). A self-supervised learning method for shadow detection in remote sensing imagery. *3D Research, 9*(4).
18. Shahid, K., Ye, Z., Shoulin, Y., & Muhammad Rizwan, A. (2018). An efficient region proposal method for optical remote sensing imagery. *IGARSS 2018—2018 IEEE international geoscience and remote sensing symposium*, pp. 2455–2458.
19. Songjiang, L. I., Ning, W. U., Peng, W. A. N. G., et al. (2020). Vehicle target detection method based on improved Cascade RCNN. *Computer Engineering and Applications*. https://doi.org/10.3778/j.issn.1002-8331.2005-0416.
20. Yin, S., Li, H., & Teng, L. (2020). Airport detection based on improved faster RCNN in large scale remote sensing images. *Sensing and Imaging*. https://doi.org/10.1007/s11220-020-00314-2.
21. Li, Y., Chen, Y., Wang, N., & Zhang, Z. (2019). Scale-aware trident networks for object detection. *2019 IEEE/CVF international conference on computer vision (ICCV)*, Seoul, Korea (South), 2019, pp. 6053–6062. https://doi.org/10.1109/ICCV.2019.00615.
22. Yamashita, T., Furukawa, H., & Fujiyoshi, H. (2018). Multiple skip connections of dilated convolution network for semantic segmentation. *2018 25th IEEE international conference on image processing (ICIP)*, Athens, 2018, pp. 1593–1597. https://doi.org/10.1109/ICIP.2018.8451033.
23. Chu, J., Guo, Z., & Leng, L. (2018). Object detection based on multi-layer convolution feature fusion and online hard example mining. *IEEE Access,* 19959–19967.
24. Han, G., Su, J., Zhang, C., et al. (2019). A method based on multi-convolution layers joint and generative adversarial networks for vehicle detection. *Ksii Transactions on Internet and Information Systems, 13*(4), 1795–1811.
25. Ren, S., He, K., Girshick, R., et al. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems,* 91–99.
26. Dai, J., Li, Y., He, K., et al. (2016). R-FCN: Object detection via region-based fully convolutional networks. *arXiv: Computer Vision and Pattern Recognition*.
27. Zhu, Y., Zhao, C., Wang, J., et al. (2017). CoupleNet: Coupling global structure with local parts for object detection. *International conference on computer vision,* pp. 4146–4154.
28. Kisantal, M., Wojna, Z., Murawski, J., et al. (2019). Augmentation for small object detection. *arXiv: Computer Vision and Pattern Recognition*.
29. Kuang, H., Chen, L., Gu, F., et al. (2016). Combining region-of-interest extraction and image enhancement for nighttime vehicle detection. *IEEE Intelligent Systems, 31*(3), 57–65.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.