Improved Faster R-CNN-Based Anomaly Target Detection for Truck Driving Environment

Long Li*

Zhejiangwanli University, Ningbo, Zhejiang, 315000, China
**Corresponding author*

Abstract: *With the development of the economy and the improvement of people's living standards, the volume of goods transported is constantly on the rise. At present, the dominant position in the transport industry is still road transport, and at the same time the problem of road traffic accidents is becoming increasingly prominent. Generally speaking, the truck is relatively large, with a wide blind spot and high cab height, which makes the driver unable to have a good observation and control of the environment around the vehicle, so it is extremely easy to have all kinds of traffic accidents. In such a context, an improved Faster R-CNN algorithm is proposed for the detection of abnormal targets in the truck driving environment. First, the ResNet-50 network is chosen to replace the VGG network, which reduces the training difficulty and effectively improves the gradient disappearance problem; then, in order to enhance the feature extraction ability of the network, the Squeeze-and-Excitation attention mechanism is introduced in the residual structure to strengthen the feature extraction ability; finally, the original feature pyramid network structure is improved by adding a self Finally, the original feature pyramid structure was improved by adding a bottom-up channel enhancement route to improve the propagation of lower-level features and further enhance the propagation of feature information. The experimental results show that the improved mAP value improves by 2.05% compared to the original algorithm.*

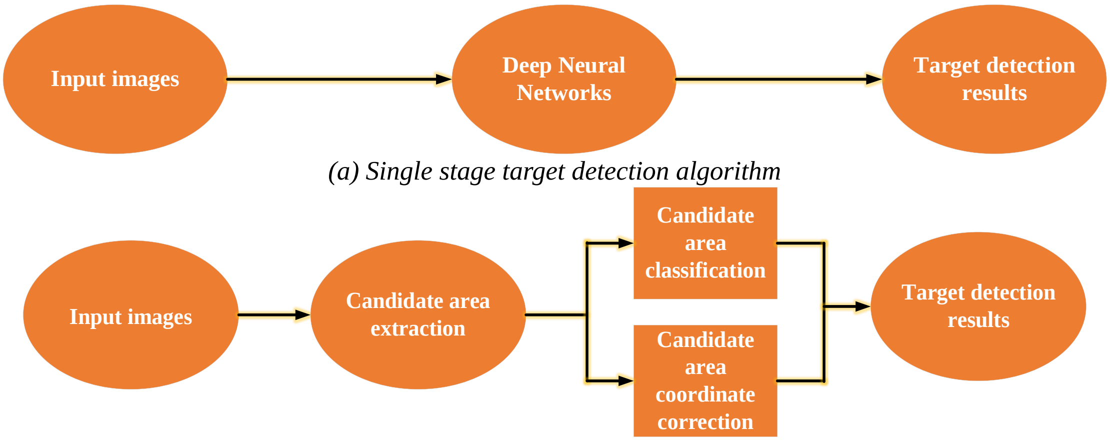
Keywords: *Faster R-CNN, Driving environment detection, Attention mechanisms*

1. Introduction

With the rapid development of the economy, the volume of goods transported is on the rise, and road transport still dominates the transport industry at present. The most important feature of such vehicles is that they have long bodies, large wheels and high cabs, which directly one of the greater disadvantages of transport vans [1]. Large blind spots in the field of vision. Larger blind spots mean that accidents are more likely to occur and, in addition, if this type of vehicle is involved in an accident, the outcome is usually more serious and the fatality rate is higher. This has taken a huge toll on society, families and individuals. Therefore, the introduction of machine vision into the truck driving environment anomaly target detection and analysis, which will have a self-evident significance to pedestrian personal safety and reduce the occurrence of truck vehicle traffic accidents.

Traditional detection methods are built on hand-crafted features and shallow trainable architectures [2], some typical algorithms include DPM (Deformable Parts Model) [3], Selective Search [4], Oxford-MKL [5] and NLPR-HOGLBP [6]. These approaches combine a large number of low level image features and high level semantic information from target detectors and scene classifiers to build complex systems that do not perform well. Secondly, traditional object detection use a sliding window approach, which reduces the speed of detection; the feature extraction method is based on manual design, and this process is highly susceptible to subjectivity, which can further reduce the accuracy of detection.

Deep learning-based target detection algorithms are currently divided into two-stage detection algorithms and single-stage detection algorithms in the field of deep learning, depending on the detection idea, as shown in Figure 1: The two-stage detection algorithm, also known as the candidate region based target detection algorithm, decomposes the target detection process into three steps: candidate region extraction, candidate region classification and candidate region coordinates correction. Single-stage detection algorithms, also called regression analysis-based target detection algorithms, treat the target detection problem as a regression analysis of target location and category information, with a neural network model that directly outputs detection results. In comparison, the two-stage target detection algorithm is more accurate and has a greater advantage for scenarios requiring high accuracy detection.



(a) Single stage target detection algorithm

(b) Two-stage target detection algorithm

Figure 1: Two target detection algorithms based on deep learning

In 2014, Girshick proposed the R-CNN algorithm [7]. In 2015, the Spatial Pyramid Pooling (SPP) structure was proposed [8]. This was followed by the Fast R-CNN algorithm proposed by Girshick [8], finally to the Faster R-CNN proposed by Ren [9], which is based on the excellent classical algorithm of the two-stage detection algorithm.

The main focus of this article is to investigate the application of the improved Faster R-CNN to scenarios where anomalous targets are detected in real van driving environments. Firstly, the ResNet-50 architecture was adopted for the backbone network to improve the gradient disappearance problem of the algorithm; secondly, the SE attention mechanism was introduced in the residual network structure to further enhance the feature extraction capability; finally, in the feature fusion phase, a bottom-up channel enhancement route is added to improve the propagation of low-level features, further enhancing the propagation of feature information and strengthening the detection capability of the algorithm.

2. Improved Faster R-CNN based truck driving environment anomaly target detection

2.1 Improved backbone network

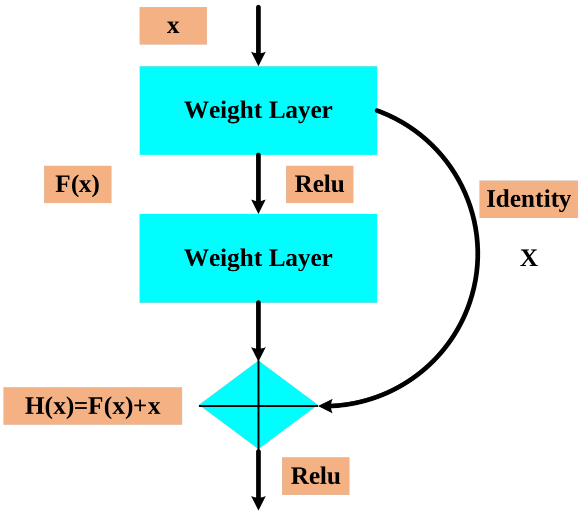


Figure 2: Residual structure diagram

The feature extraction backbone network used for the Faster R-CNN algorithm in this paper is ResNet50. ResNet (residual network) is now one of the commonly used feature extraction backbones, it was proposed in 2015 by Kai-Ming He, Xiang-Yu Zhang, Shao-Qing Ren, and Jian Sun [10] to solve the gradient explosion problem generated by CNNs (Convolutional Neural Networks). As shown in Figure 2: The residual structure introduces jump connections for constant mapping, which allows deeper networks to still learn key information at shallower levels without creating network degradation problems.

ResNet18 and ResNet34 in the ResNet network use the two-layer residual structure of the a-plot in Figure 3; ResNet50, ResNet50, ResNet101 and ResNet152 use the three-layer residual structure of the b-plot in Figure 3. The difference between the two is that the three layers are convolved using two types of convolution kernels, 1×1 and 3×3. A 1×1 convolution is used to adjust the dimensionality of the features, and this design deepens the network structure while reducing the computational effort.

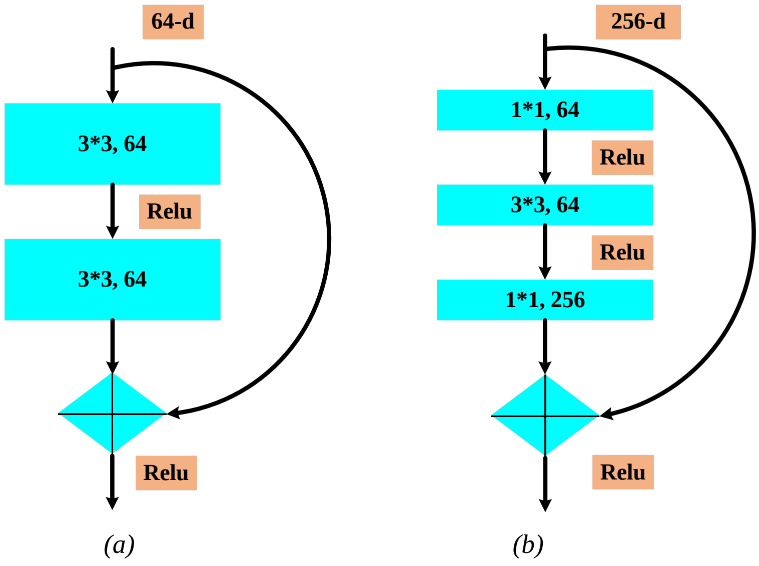


Figure 3: Two- and three-layer residual structure network diagram

The ResNet50 network structure is shown in Figure 4. In order to enhance the feature extraction capability of the network, the SE (Squeeze-and-Excitation) attention mechanism is introduced in ResNet50. In traditional convolutional pooling, each channel of the default feature layer is equally important. Nevertheless, only a few of the features extracted for the objects in the graph are the key features required. The SE attention mechanism is a simulation of the idea that the human eye pays different levels of attention to a thing. It proposes a channel vector to weight the feature channel dimensions, which can be trained to learn, and as the network is trained, the SE attention module is capable of augmenting important channel features and suppressing irrelevant ones. This enables key features in the image to be extracted more effectively.

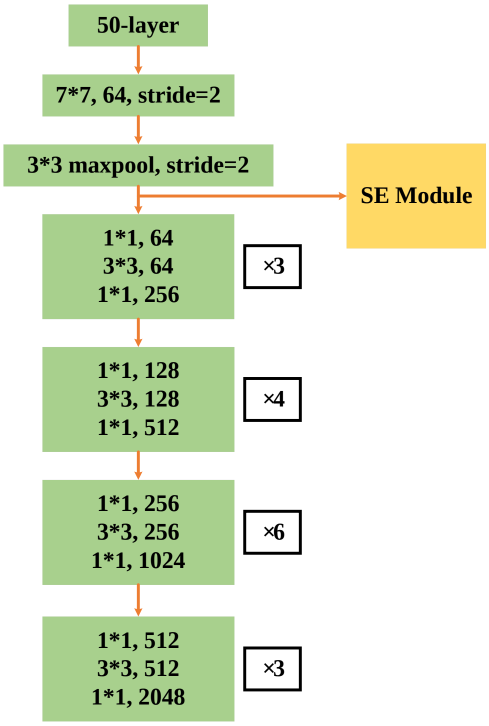


Figure 4: ResNet50 network structure

The SE attention module is shown in Figure 5: Suppose the input is a $h \times w \times c$ feature map. The first step is to perform a global average pooling (pooling size $h \times w$) to obtain a $1 \times 1 \times c$ feature vector, this operation fuses the feature maps of each channel into a single point; Then, after two fully connected layers, the first fully connected layer is dimensioned down and the second is dimensioned up, where r = 16 in the figure 5, which has the advantage of adding more non-linear processing to fit the complex correlations between channels; Subsequently, a Sigmod layer is added to obtain the final feature vector. Finally, the channel attention operation is performed by multiplying an original $h \times w \times c$ eigenmap with a $1 \times 1 \times c$ eigenvector, this is where the channel vectors are weighted against the channels of the original feature map.

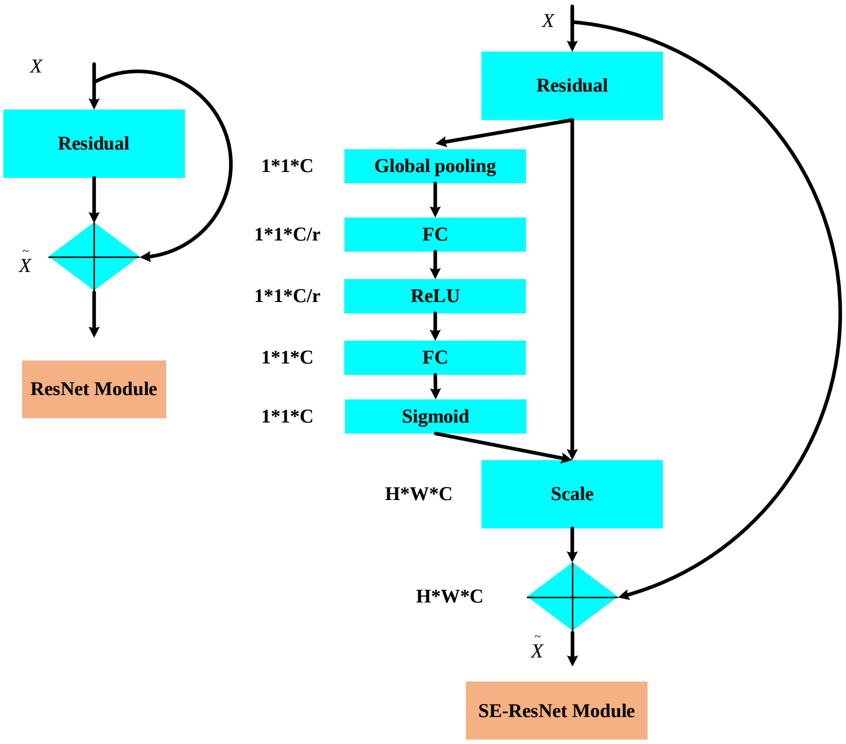


Figure 5: Residual network structure with attention module added

2.2 Improved feature fusion network

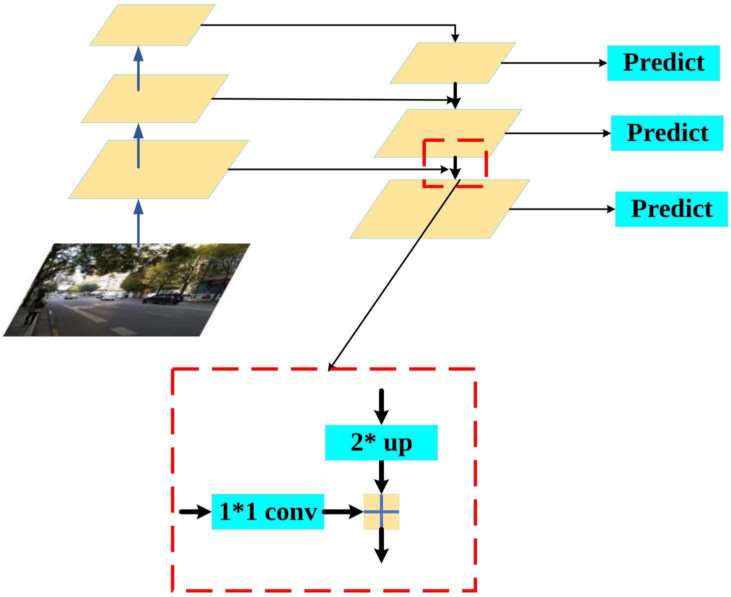


Figure 6: FPN Feature Pyramid

A commonly used feature fusion network is the FPN (Feature Pyramid Networks) [11], the general structure of the algorithm is shown in Figure 6: A bottom-up line, a top-down line, and a lateral connection. The enlarged area in the diagram is the lateral connection, here the number of channels of the feature map is changed by using a 1*1 convolution kernel during the fusion of the features of the different layers, without changing the size of the feature map. FPN is a network proposed in 2017, FPN will fuse high level feature information with low level feature information through a simple change in network connectivity. Generally speaking, the low-level feature semantic information ratio is relatively low, but its target location is more accurate; the high-level feature semantic information is relatively high, but the target location is not as accurate. The fusion between the two has a significant improvement on the localisation accuracy and classification accuracy in target detection tasks.

The feature fusion phase of the original Faster R-CNN uses the same commonly used feature pyramid network, which is improved in this paper by borrowing from PANet [12]. The Improved Pyramid Network structure is shown in Figure 7: The orange arrows indicate that because the feature information in an FPN network needs to have a bottom-up process, the shallow features need to pass through dozens or even hundreds of network layers in the BackBone network to reach the top layer. However, after passing through so many layers, the loss of feature information in the shallow layers becomes more serious. The blue arrow indicates the addition of a Bottom-up Path Augmentation structure, in this way the shallow features pass through the lateral connections in the original FPN to P2 and then from P2 along the Bottom-up Path Augemtation to the top layer, passing through less than 10 layers and preserving the shallow feature information better. Note that N2 and P2 here represent the same feature map. However, N3, N4, N5 is not the same as P3, P4, P5; in fact, N3, N4, N5 is the result of the fusion of P3, P4, P5. Its improved propagation of low-level features further enhances the propagation of feature information.

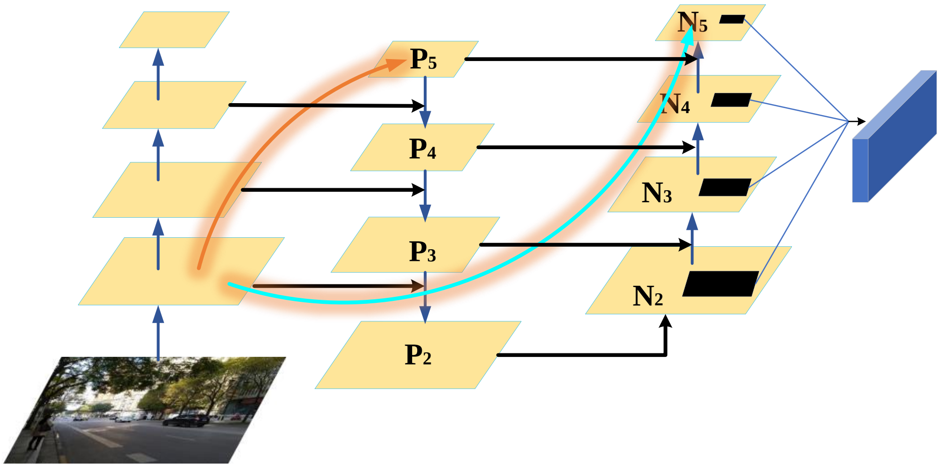


Figure 7: Improved FPN Pyramid Structure

3. Experimental results and analysis

3.1 Introduction to the dataset

Training samples for warp networks usually require at least a few thousand, and if the sample data is too small, the accuracy and reliability of the detection will be affected. In this paper, a dataset of 1500 images (1200 images in the training set and 300 images in the test set) was created by collecting images from the Internet, images provided by logistics and transportation companies, as well as by taking the images ourselves. We defines three categories of driving environment anomaly targets: pedestrians, fallen pedestrians and cars. There are 500 sheets for each of the three categories. As there is no publicly available dataset of fallen pedestrians on the internet, we basically relied on our own filming of simulated fall scenarios and online collection to build it. The newly created samples contain images of people, vehicles or people and vehicles at the same time, and contain different lighting, different shooting perspectives, different resolutions, different road environments and road conditions, etc. to meet the requirements of sample diversity. The different illumination enables the detection model to be able to detect in daylight or low light conditions. The different camera views allow for better detection of targets in different environments. Multiple angles of the picture improve the generalization of the detection model, allowing it to perform well when it comes to real detection. The different road environments take into account the fact that transport vehicles cannot travel only on a single highway, it also appears in some more complex scenarios such as urban roads.

The optimisation of samples is important for the robustness of the algorithm's detection. If the newly created dataset does not take into account the lighting factor and only has a single sample under illumination, it will to the point where the trained model may have significant false or missed detections when detecting poorly lit images or videos. The same holds true for putting in images from different road environments and different resolutions. The significance of dataset optimisation is to achieve better detection results with fewer training samples, to improve the generalisation ability of the detection model and to make it perform consistently when detecting images outside the training set. An example of a sample dataset is shown in Figure 8:

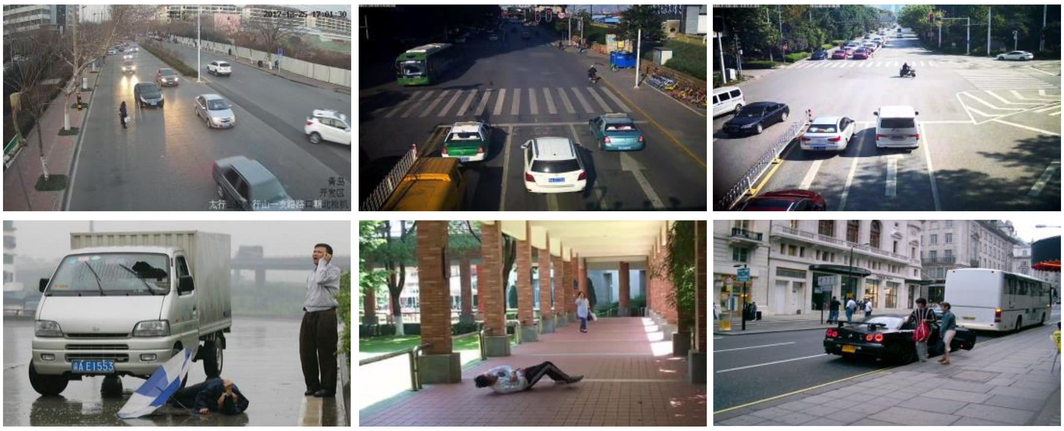
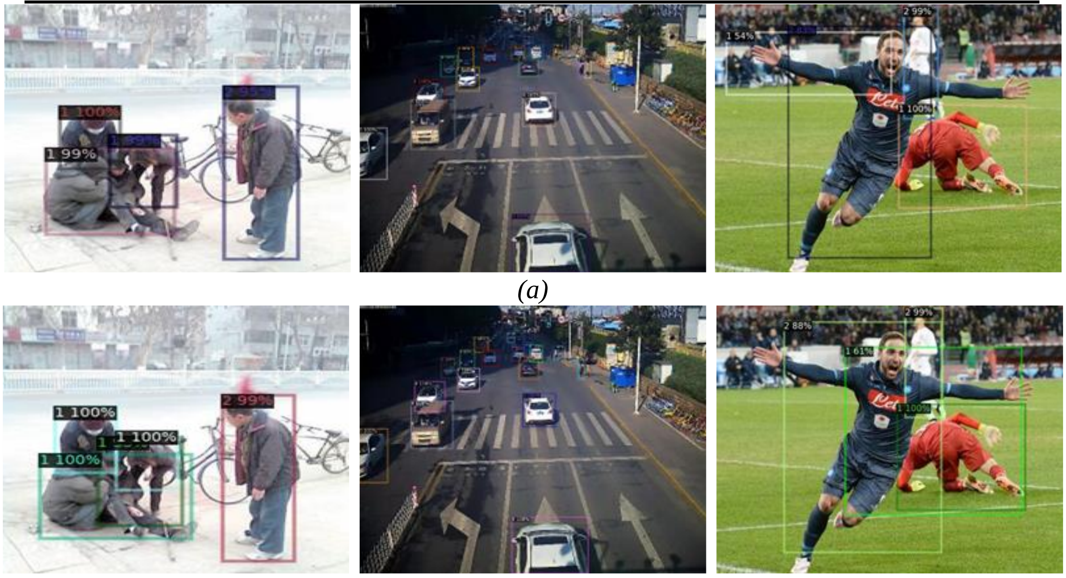


Figure 8: Sample datasets

3.2 Introduction to the dataset

Table 1: Comparison of algorithm performance before and after improvement.

Methods	Accuracy(AP50)	FPS	Test data set (sheets)
Faster R-CNN	93.27%	20.96	300
Improved Faster R-CNN	95.31%	18.75	300



(a)

(b)

Figure 9: Visualization results of the algorithm before and after improvement

Experimental environment: The experiments in this article are based on the Ubuntu operating system. Desktop computer with NVIDIA TITAN RTX graphics card with 8G of video memory; Intel(R) Xeon(R) Sliver 4214 processor at 2.20GHz and 128GB of disk memory.

The Faster R-CNN and the improved Faster R-CNN were compared experimentally by examining the dataset. From Table 1 we can see that the improved Faster R-CNN algorithm improves in accuracy by about 2.05 percentage points with essentially the same speed. This demonstrates the effectiveness of the improved model. Some representative visualisation results of the comparison experiments are shown in Figure 9, where (a) is the visualisation result of the original algorithm and (b) is the visualisation result of the improved algorithm.

4. Conclusion

In this paper, the original backbone feature extraction network is replaced by improving the target detection algorithm Faster R-CNN and introducing an attention mechanism module. The feature extraction capability of the network is improved by weighting the individual channels of the original feature map with trainable channel feature vectors to highlight important features. In addition, drawing on the idea of feature fusion in PANet, a bottom-up channel enhancement path is added to the FPN to retain the high-level information while introducing more low-level feature information, further optimising the feature information extracted by the network and enabling the detection accuracy of the Faster R-CNN to be improved. The experimental results on the self-built data and show that the improved algorithm completes the target detection of the truck driving environment in the actual scene and effectively improves the accuracy of the abnormal target detection.

References

[1] Zhao Lei. Analysis of heavy-duty truck design trends in the context of new transportation[J]. China Aviation Weekly, 2021(36):54-55.

[2] Dong Changqing, Liu Yongxian, Zhao A. Research on vehicle vision detection method based on deep learning algorithm[J]. Manufacturing Automation, 2019, 41(03):119-122.

[3] DOLLAR P,WOJEK C,SCHIELE B,et al. Pedestrian detection:a benchmark [C]. //Proceedings of IEEE Conference on Computer Vision and Pattern Recognition,2009:304-311.

[4] UIJLINGS J R,VAN DE SANDE KE A,GEVERS T,et al. Selective search for object recognition[J]. International Journal of Computer Vision, 2013,104(2):154-171.

[5] VEDALD A, GULSHAN V, VARMA M, et al. Multiplekernels for object detection [C]. //IEEE 12th International Conference on Computer Vision,2009:606-613.

[6] YU Y,ZHANG J,HUANG Y,et al. Object detection by context and boosted HOG-LBP[C]. //ECCV Workshop onPASCAL VOC,2010.

[7] Ra M, Jung H G, Suhr J K. Part-based Vehicle Detection in Side-rectilinear Images for Blind-Spot Detection [J]. Expert Systems with Applications, 2018, 101.

[8] Jae Kyu Suhr, Ho Gi Jung. Rearview Camera-Based Backover Warning System Exploiting a Combination of Pose-Specific Pedestrian Recognitions [J]. IEEE Transactions on Intelligent Transportation Systems, 2017, (99): 1-7.

[9] Tang Shi. In-vehicle video-based road vehicle and pedestrian detection [D]. Chengdu: University of Electronic Science and Technology,2018.

[10] Li Jun, Wei Minxiang. Research on large vehicle collision avoidance system based on "fly-eye" sensing network [J]. Agricultural Equipment and Vehicle Program, 2010(3): 3-6.

[11] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. //Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[12] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C]. //Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[13] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]. //Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8.