Dim Target Detection Method Based on Deep Learning in Complex Traffic Environment

Hao Zheng
Pingdingshan University
Jianfang Liu (liu_jianfang@126.com)
PingDingShan University
Xiaogang Ren
Soochow University

Research Article

Keywords: Object detection, Faster R-CNN, Residual Network (ResNet), Deep learning, Complex traffic environment, Structure optimization

Posted Date: February 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-177944/v1

Dim target detection method based on deep learning in complex traffic environment

Hao Zheng[1], Jianfang Liu[1*], Xiaogang Ren[2,3]

1 School of Software, Pingdingshan University, Pingdingshan, Henan, 467000, China

2 Affiliated Changshu Hospital, Soochow University, Changshu, Jiangsu, 215500, China

3 School of Information and Control Engineering, China University of Mining and Technology, Xuzhou, Jiangsu, 221116, China

*. Corresponding author: Jianfang Liu, Email: liu_jianfang@126.com

Abstract Although the current vehicle detection and recognition framework based on deep learning has its own characteristics and advantages, it is difficult to effectively combine multi-scale and multi category vehicle features, and there is still room for improvement in vehicle detection and recognition performance. Based on this, an improved fast R-CNN convolutional neural network is proposed to detect dim targets in complex traffic environment. The deep learning model of fast R-CNN convolutional neural network is introduced into the image recognition of complex traffic environment, and a structure optimization method is proposed, which replaces vgg16 in fast RCNN with RESNET to make it suitable for small target recognition in complex background. Max pooling is the down sampling method, and then feature pyramid network is introduced into RPN to generate target candidate box to optimize the structure of convolutional neural network. After training with 1497 images, the complex traffic environment images are identified and tested.

Keywords Object detection; Faster R-CNN; Residual Network (ResNet); Deep learning; Complex traffic environment; Structure optimization

1 Introduction

Images are an important source of information for human beings, and vision is one of the main ways to receive information from the outside world in daily life. Target detection is to use image processing, pattern recognition, machine learning and other multi-directional knowledge to locate and identify objects of interest from images or videos through calculations, and combine target detection and recognition, which is more difficult than image classification [1, 2]. The detection and recognition of various targets is an important part of the field of computer vision. In recent years, with the upgrading of hardware equipment, the rapid progress of deep learning technology and the development of big data, the universal combination of target detection and recognition technology and the industry has made its application scope wider and wider. It is of great significance in the fields of national security and national defense. At the same time, the precise detection and recognition of various targets by target detection and recognition technology also laid the foundation for the development of video surveillance, unmanned driving, scene semantic understanding, Internet mobile terminals, image retrieval and other fields [3-5].

When the target detection and recognition technology is used in the transportation field, it can detect and recognize various types of targets that are common in road traffic scenes. It can accurately and timely determine the route of the vehicle ahead. It not only improves the intelligence of traffic, but also guarantees the safety of the field of intelligent driving [6, 7]. However, in actual situations, due to the complex road conditions and many goals, there are still some practical problems: （1）There are many similar targets. In the actual detection scene, when the shooting angle is fixed, due to light and non-rigid deformation, there will be many targets with small intra-class similarity and large inter-class similarity. Under this condition, it is easy to interfere with the detection effect and reduce the accuracy of detection [8, 9].（2）There are more redundant information and insufficient use of effective information. In the target detection and recognition of 2D images, due to the insufficient utilization of effective information, the detection of the target to be detected in the actual scene will be missed, which makes it difficult to improve the detection effect. Therefore, target detection and recognition based on the deep learning framework still has great research significance and space under the current research status.

In summary, transportation is a scene closely related to human travel and life. Combining with target detection and recognition technology is the basis for the development of intelligent transportation. It can not only improve the convenience of travel, but also play a very important role in the safety of human travel. At the same time, the existing research work has made some progress, but there are still many difficulties unsolved. There is still a certain gap with the use in real life. Therefore, further detection and recognition of various targets in actual traffic scenes still have certain theoretical research significance and practical application value. Based on this, a method for dim target detection using improved Faster R-CNN convolutional neural network in a complex traffic environment is proposed.

2 Related works

The intelligent transportation system collects road vehicle driving information through cameras, and then the central computer processes the information. So as to realize the tracking and identification of vehicles, the identification of illegal traffic vehicles, and assist in handling various traffic violations. While reducing the work pressure of traffic police, it also increases the utilization rate of roads and reduces the accident rate. In recent years, intelligent driving technology has also made great progress, which is also inseparable from the rapid development of target detection technology [10]. The regular functions of modern vehicles such as cruise control, adaptive cruise, lane keeping and lane departure warning are inseparable from target detection technology. The surrounding information of the vehicle is collected through the cameras installed at various positions of the vehicle, and the current operating environment of the vehicle is analyzed to realize the driving assistance function. It reduces the fatigue of the driver and improves the safety of the vehicle.

In the traditional intelligent transportation system, vehicle detection is mainly realized by special sensors. Reference [11] proposed an algorithm for vehicle detection using ultrasonic data after analyzing the optimal energy-saving method of sensors. Reference [12] built a wireless sensor network to estimate the speed of the vehicle. These methods are not affected by the weather and can quickly detect passing vehicles. However, the installation of sensors tends to temporarily close the traffic there, and the maintenance cost of these sensors is also a considerable expense.

Nowadays, with the development of economy and technology, camera technology has been integrated into every corner of social production and life. With the development of camera technology, video storage, playback and processing have also made considerable progress, laying a solid foundation for the development of computer vision technology. At the same time, with the rapid improvement of computer computing power, the use of computer vision to achieve target detection has become the main development direction of modern scientific research. This also provides effective tools and methods for the progress of target detection. Vision-based vehicle detection system came into being. The visual vehicle detection system can detect vehicle type, traffic volume, vehicle speed and even predict traffic accidents. Reference [13] proposes a vehicle detection algorithm that can adaptively distinguish the front background. But as a background difference method, it has a common disadvantage. That is, it is difficult to detect a stationary or slow moving vehicle. This limits the final detection accuracy of this method.

The development of deep learning has promoted the research of target detection, such as YOLO [14], RCNN [15] and SSD [16]. Reference [14] proposed a road image vehicle detection algorithm based on an improved YOLOv3 network. In order to improve the detection efficiency, a new and improved YOLOv3 network structure with only 16 layers is constructed. Reference [17] proposed a hybrid deep neural network to divide the convolutional layer and the maximum pooling layer of the network into multiple blocks by dividing the final mapping of the two, including the receptive domain and the maximum pooling domain. The network can extract and learn multi-scale features of pictures. In a complex actual traffic scene, the traffic is larger and traffic jams are more likely to occur. The convolutional neural network greatly improves the adaptability of the model to the input image because it does not rely on various artificial features. Reference [18] improves vehicle detection performance through a well-designed convolutional feature map neural network. In the traffic scene of vehicle congestion, reference [19] detects occluded vehicles by training two sets of paired support vector machines. Reference [20] studied a vehicle detection algorithm based on convolutional neural networks that fused color images and depth images. The algorithm is mainly researched by convolutional network multi-scale forward-looking depth imaging positioning model, forward-looking variable-scale vehicle detection pre-positioning algorithm, and typical model recognition algorithm based on transfer learning. Combined with simulation tests, the algorithm performance and the correctness of the model identification algorithm are verified.

According to research hotspots in recent years, the detection and recognition rate of the above model methods for vehicles with small pixel sizes in images is generally low, and it is difficult to meet the accuracy requirements in actual traffic applications [21, 22]. The current vehicle detection and recognition framework based on deep learning has its own characteristics and advantages, but it is difficult to effectively combine multi-scale and multi-category vehicle features. There is still room for improvement in vehicle detection and recognition performance [23]. Based on this, a method for dim target detection using improved Faster R-CNN convolutional neural network in a complex traffic environment is proposed. Aiming at the problem that it is difficult to accurately identify the common dim targets in road traffic scenes, a similar target detection and recognition method based on the improvement of residual network is proposed. The residual network with stronger learning ability is used to obtain more effective feature expression and rich semantic information.

3 Target detection based on improved Faster -RCNN network model

3.1 Proposed target detection network structure

Faster-RCNN unifies feature extraction, candidate region extraction and box regression in one network to improve the efficiency of the network. It is mainly divided into four stages:

(1) Feature extraction stage. In the first step, the convolutional neural network is still used as the basis to perform deep feature extraction on the sample data to obtain a feature map. This feature map can continue to be used by subsequent region proposal generation networks, so it can be called a shared feature map. So that a sample image only needs to go through the convolutional neural network once.

(2) Region proposal generation stage. Use the Region Proposal Network (RPN) to generate multiple anchors for each pixel in the feature map of the previous step. Use SoftMax to determine whether an anchor belongs to the foreground target or the background, and output a category confidence probability value for each anchor. Finally, a bounding box regression method is used to correct the position of the anchor containing the target to obtain a more accurate target region.

(3) Normalize the feature matrix stage. The region of interest pooling layer is used to map each RoI obtained by the RPN layer to the feature map. The operation of pooling the region of interest for region proposal of different sizes is normalized into feature regions of the same size, which are used to input to the fully connected layer to determine the category of the region and calculate the offset of the target position.

(4) Classification and regression stage. Two fully connected layers, a classification layer and a regression layer, are used to classify the category of each target in the feature area. At the same time, bounding box regression is used to correct the target frame to obtain an accurate position offset.

This paper firstly normalizes the image of any size of complex traffic environment to 1000×600 pixels. Then, the feature map is generated through the convolutional layer and the pooling layer in the CNN. The Faster-RCNN algorithm uses deep learning methods for feature extraction, which can effectively reduce the time and space complexity while meeting the accuracy requirements. VGGNet is one of the feature extraction networks in Faster-RCNN. VGGNet is a classic convolutional neural network that emerged during the development of deep learning. It does not use a larger convolution window, but uses a smaller convolution kernel to gradually extract multiple layers of the original image. Using multiple such small convolution kernels in cascade can achieve the same effect as a convolution with a larger window, and the generalization ability of the VGG model is stronger. The superposition of multiple small-scale convolutional layers and pooling layers effectively improves the learning ability of the network structure for image features.

In this study, in order to improve the recognition accuracy of vehicles and dim targets in the image, the VGG16 network was not selected as the basic feature network to extract image features. In other fields such as target detection, image segmentation, video analysis and recognition, replacing VGG16 in Faster-RCNN with a residual network (ResNet) can improve system performance. On the PASCAL VOC2007 data set, by replacing VGG16 with ResNet101, the MAP increased from 73.2% to 76.4%, and on PASACAL VOC2012 from 70.4% to 73.8% [24]. Since there are not many target categories and numbers in traffic environment images, the ResNet50 network is selected to extract image features. Extract foreground Region of Interest (ROI) and region scores through RPN and feature pyramid networks (FPN) networks on all feature maps. The area with the highest score is used as the final vehicle and dim target candidate region.

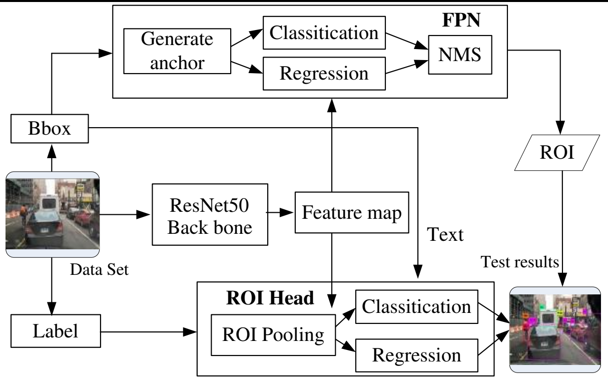


Fig.1 The structure of target detection network is proposed

Any target region proposal is mapped to the corresponding position of the feature map through the ROI Pooling layer, and the area is down-sampled into a 7×7 feature map. Then each input feature map is extracted into a 7×7×256 dimension feature vector through a fully connected layer. Finally, the feature vector is input to two output layers of the same level: One is the classification layer, which determines whether the target is a vehicle or other dim target. The other is the boundary regression layer, which mainly fine-tunes the position and size of the ROI border.

3.2 RPN (Region Proposal Network)

RPN (Region Proposal Network)

RPN is a fully convolutional network. After end-to-end training, high-quality foreground target region proposal for complex traffic environments are generated. Simultaneously complete the target boundary and target score prediction of vehicles and dim targets at each location. The network shares the convolutional features of the image with the vehicle target detection network. The residual network based on ResNet50 and the Faster R-CNN model share the convolutional layers from C2 to C5.

FPN（Feature Pyramid Network）

The FPN algorithm uses both the high-resolution of low-level features and the high-semantic information of high-level features. The prediction effect is achieved by fusing the features of these different layers. In order to improve the accuracy of target detection in traffic environment images, this paper uses FPN to fuse features of different layers in the RPN network to generate target region proposal of interest.

FPN designs the feature map as a multi-scale pyramid structure, and each layer of the pyramid uses a single-scale anchor. Corresponding to each layer of Pyramid {P2, P3, P4, P5, P6} corresponding to the anchor scale of ResNet50 are {32×32, 64×64, 128×128, 256×256, 512×512}. Use 3 types of ratios {1:2, 1:1, 2:1}, and share 15 types of anchor to predict the target object and background in the traffic environment image. Generate region proposal of interested targets (vehicles, dim targets). The RPN framework is shown in Fig. 2.

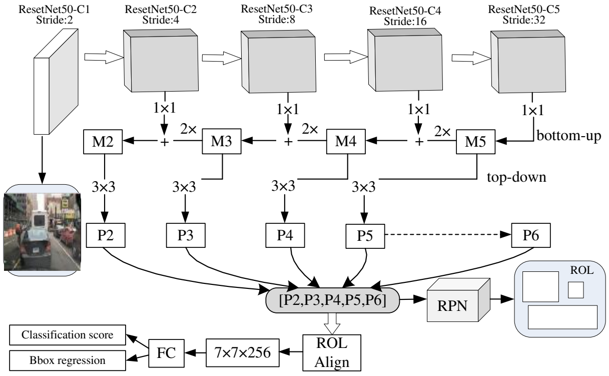


Fig.2 Region proposal networks architecture

ROI Pooling module

ROI Pooling maps the ROI to the position corresponding to the feature map according to the input image. Divide the mapped area into sections of the same size, the number of which is the same as the output dimension. Perform max pooling operation for each section. In this process, two quantization operations are carried out. The result of RPN is input into the ROI Pooling layer and mapped into 7×7 features. Then all the output passes through 2 Fully-connected Layers, then the classification layer and the boundary regression layer to get the final result. The classification layer gives the probability that the object in the region proposal is a vehicle and a weak target, and the boundary regression layer gives the coordinates of the vehicle and the dim target region proposal.

3.3 Network model training

Before training the RPN network, each anchor will be assigned a binary label that is the background or target. The anchors that assign positive labels are:

(1) An anchor that has the largest intersection over union (IoU) with a bounding box ground truth (GT) of a target's true position.

(2) An anchor whose intersection ratio with the true position of any target is greater than 0.7.

The anchors that assign negative labels are:

(1) An anchor whose intersection ratio with the bounding box of all target real positions is less than 0.3.

The process of bounding box regression is the process of fine-tuning anchors. Although 9 different sizes of anchors are used to cover all the targets in the image, they can only cover roughly. It is also necessary to modify each anchor within a certain range to make the anchors containing the target closer to the real target position.

In the process of returning the frame, F represents a certain anchor, and G represents the true position of a certain target. The position coordinates and width and height of the proposal window are x, y, w, h , then the process of bounding box regression is to transform F into f and find a set of offset values G' to make it as close as possible to the real target position G . For transformation f , first do translation and then zoom:

$$G'_x = A_w \cdot d_x \;\; A \;\; + A_x , G'_y = A_h \cdot d_y \;\; A \;\; + A_y \tag{1}$$

$$G'_w = A_w \cdot \exp \;\; d_w \;\; A \;\; , G'_h = A_h \cdot \exp \;\; d_h \;\; A \tag{2}$$

Where, as long as the four transformations of $d_x \;\; A$, $d_y \;\; A$, $d_w \;\; A$ and $d_h \;\; A$ are known, the offset value G' can be obtained. When the positions of GT and anchor are more consistent, the transformation can be approximately regarded as a linear transformation. The above four values can be obtained by linear regression:

$$Y = WX \tag{3}$$

That is, given the input X as the feature vector, W is the parameter to be learned, so that X is infinitely close to the real position Y after linear regression. For the problem of target detection and recognition:

$$d_* \;\; A \;\; = w_*^T \cdot \Phi \;\; A \tag{4}$$

Where, $\Phi \;\; A$ is the feature vector mapped from the anchor to the feature map output by the convolutional neural network. $*$ represents x, y, w, h , and $d \;\; A$ is the final predicted value. The goal is to minimize the difference between the final predicted value and GT. The loss function is:

$$loss = \sum_i^N \;\; t_*^i - \hat{w}_*^T \cdot \Phi \;\; A^i \quad ^2 \tag{5}$$

Where, let $loss$ be the smallest, you can learn a set of transformation values to get the final offset. During the training process, the network is fine-tuned by minimizing the multi-task loss function:

$$L \;\; p_i \;\; , \; t_i \;\; = \frac{1}{N_{cls}} \sum_i L_{cls} \;\; p_i , p_i^* \;\; + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg} \;\; t_i , t_i^*$$

(6)

Where, i represents that the i -th candidate frame is selected during the training process. p_i represents the predicted probability of the i -th candidate frame as the target. p_i^* represents the true probability of the i -th candidate frame target. If the i -th proposal box is a positive label, it belongs to the target, then $p_i = 1$. If the i -th proposal box is a negative label, it belongs to the background, then $p_i = 0$. t_i represents the position of the predicted bounding box. t_i^* represents the position of the real box corresponding to the prediction box. L_{cls} is the classification loss, defined as follows:

$$L_{cls} \;\; p_i , p_i^* \;\; = -\log \;\; p_{i\,p_i^*} \tag{7}$$

L_{reg} is the regression loss, defined as follows:

$$L_{reg} \;\; t_i , t_i^* \;\; = \sum_{i \in \;\; x,y,w,h} smooth_{L1} \;\; t_i , t_i^* \tag{8}$$

$$smooth_{L1} \;\; x \;\; = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & |x| \geq 1 \end{cases} \tag{9}$$

$$\begin{cases} t_x = \;\; x - x_a \;\; / w_a , t_y = \;\; y - y_a \;\; / h_a \\ t_w = lb \;\; w / w_a \;\; , t_h = l_b \;\; h / h_a \\ t_x^* = \;\; x^* - x_a \;\; / w_a , t_y^* = \;\; y^* - y_a \;\; / h_a \\ t_w^* = lb \;\; w^* / w_a \;\; , t_h^* = l_b \;\; h^* / h_a \end{cases} \tag{10}$$

Where, x, y is the coordinate value of the predicted bounding box, and x_a, y_a is the coordinate of the proposal box. x^*, y^* is the real GT coordinate. w and h indicate the width and height of the bounding box. N_{cls} and N_{reg} are the normalized parameters when calculating regression coordinates and classification confidence, respectively.

4 Target detection experiment

4.1 Experimental setup

The data set used in the experiment is the KITTI public dataset, which is currently one of the most commonly used datasets in the field of autonomous driving, and is also one of the internationally common visual evaluation algorithm datasets. The data set contains pictures in multiple scenes, such as urban roads, residential areas, campuses and other common scenes. There are eight categories including Car, Van, Truck, Pedestrian, Person (sitting), Cyclist, Tram and Mis. There are up to 15 cars in each image. This article mainly focuses on the detection of vehicles in motor lanes and other dim targets (pedestrians, pets, etc.) in sidewalks. Therefore, Car, Van, Truck, and Tram in the kitti dataset are merged into one type of vehicle. Combine other categories into one category of dim targets to form two categories of detection. An example image is shown in Fig. 3.

Fig.3 Datasets used in the experiment

In this paper, 7481 images in the data set are formed into a training verification set and a test set at a ratio of 2:8. That is, 1497 training verification sets and 5984 test sets.

This paper chooses the stochastic gradient descent method to train Faster R-CNN in an end-to-end joint manner. A Gaussian distribution with a mean of 0 and a standard deviation of 0.01 is used to randomly initialize the weights of all newly added layers. The remaining layers are initialized with the parameters of the pre-trained ImageNet classification model. Set the learning rate to 0.005, momentum to 0.9, weight decay coefficient to 0.0001, epoch to 1500, and number of iterations to 550,000. The model is saved every epoch, and finally the model with the highest accuracy is selected.

A traffic scene image RPN network gets about 20,000 anchors. Use the NMS algorithm to select the 2000 RoIs with the highest probability, which correspond to regions of different sizes in the feature map. Use Proposal Target Creator to select 128 RoIs, and then use ROI Pooling to pool all these regions of different sizes to the same scale (7×7).

4.2 Evaluation index

In order to evaluate the effectiveness of the proposed method, two indicators, precision and recall, are used for model evaluation, both of which range from [0,1]. At the same time, F1 value is introduced for harmonic average evaluation. The specific evaluation calculation formula is:

$$P=\frac{n_{TP}}{n_{TP}+n_{FP}}\times 100\% \qquad (11)$$

$$R=\frac{n_{TP}}{n_{TP}+n_{FN}}\times 100\% \qquad （12）$$

$$F1=\frac{2PR}{P+R}\times 100\% \qquad （13）$$

Where, P represents precision; R represents recall; $F1$ represents the harmonic average of precision and recall; n_{TP} represents the number of correctly identified vehicles and dim targets; n_{FP} represents the number of misidentified vehicles and dim targets; n_{FN} represents the number of unidentified vehicles and dim targets.

4.3 Training loss

Using the Faster R-CNN structure described above, 1497 training set sample data are used for training. Performing 1500 iterations on the above training set took 20 hours. The training accuracy loss curve is shown in Fig. 4.

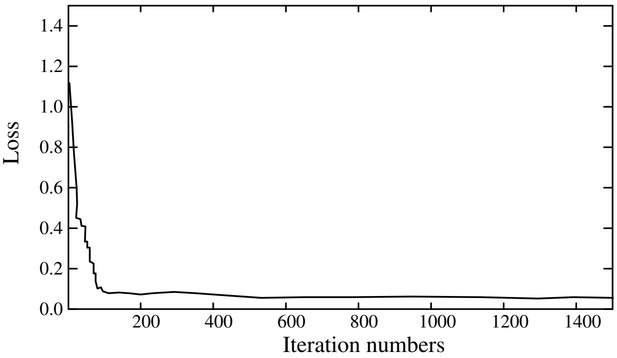


Fig.4 Loss curves of training

It can be seen from Fig. 4 that as the number of iterations continues to increase, the accuracy loss produced by the training set gradually decreases. When the iteration reaches 1200 times, the accuracy loss drops to 3%, indicating that the model training effect is good. The training loss basically converges to a stable value, indicating that the improved Faster R-CNN achieves the expected training effect.

4.4 Model performance verification

In order to verify the reliability and stability of the model, after the training is completed, the 1497 images in the test set are identified. Choose the mean average precision (MAP), average recall, average precision (AP) as the evaluation index of the validity of the test results. Use the average processing time to evaluate the speed of recognition.

$$Average\ processing\ time=\frac{Test\ run\ time}{Number\ of\ test\ pictures} \qquad （14）$$

The experimental results show that the average time for the method in this paper to recognize a single image is 1.55s. Moreover, it was found in the experiment that occlusion and background similarity are the main reasons that affect target recognition. The recognition effect is shown in Fig. 5. Under normal conditions, the algorithm can detect and recognize the objects to be inspected in the road traffic scene respectively. Output the category of each target and the specific location of each target. It can also distinguish objects that are far away, partially obscured, and blurred. It can be seen that the detection and recognition effect of common targets in traffic is improved. In the actual situation, there is more redundant information in the road traffic scene and less effective information of the target, so that the target to be inspected in the image is often interfered by other information, such as insufficient lighting, shadow occlusion, etc. And in the shooting process, some target boundaries are not clear and fuzzy, which makes it difficult to accurately detect and recognize the road traffic targets to be inspected. The accuracy of the trained road traffic target detection and recognition model is reduced. Replacing VGG16 in Faster-RCNN with a residual network (ResNet) can simultaneously improve the utilization of effective information from the channel and space. Use the shallow detail information in the feature map to achieve better detection and recognition results.

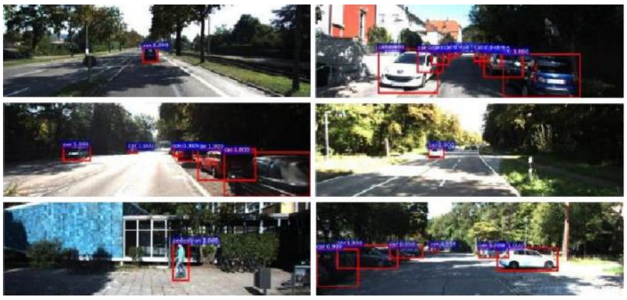


Fig.5 Recognition effect picture

4.4.1 Model precision comparison

In order to demonstrate the performance of the proposed method in terms of precision indicators, it is compared with the methods in reference [13], [14], and [20]. The result is shown in Fig. 6.

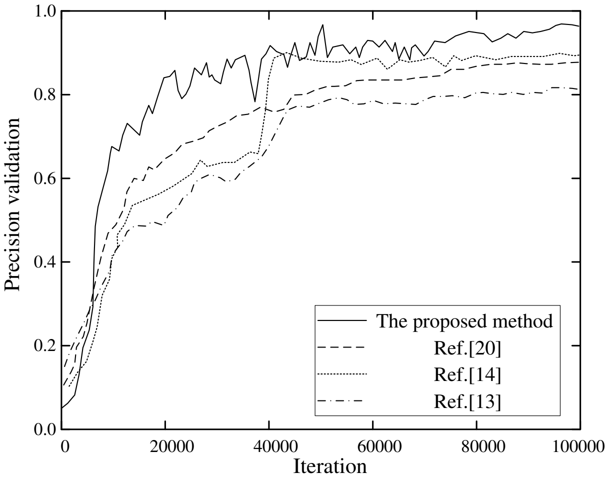


Fig.6 Model precision comparison

It can be seen from Fig. 6 that as the number of iterations increases, the precision of various methods is also increasing. The precision of the proposed method is better than other comparison methods, the highest precision reaches 94.7%, showing certain advantages. Because the improved Faster R-CNN deep network model deeply integrates RPN, it can generate high-quality region proposal boxes and improve the precision of recognition. Reference [20] model is used for global feature extraction and classification, but it is difficult to adapt to the complex road traffic environment. Therefore, the recognition accuracy needs to be further improved. However, the accuracy of reference [13] and [14] is low, and it is difficult to identify dim targets.

4.4.2 Model recall comparison

In order to demonstrate the performance of the proposed method in terms of recall index, it is compared and analyzed with the methods in reference [13], [14], and [20]. The result is shown in Fig. 7.

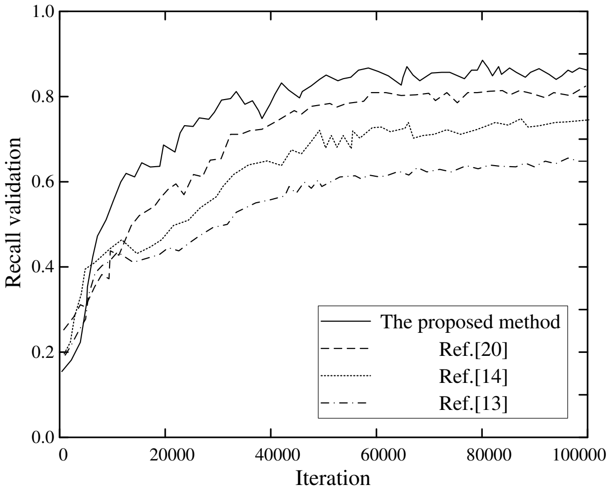
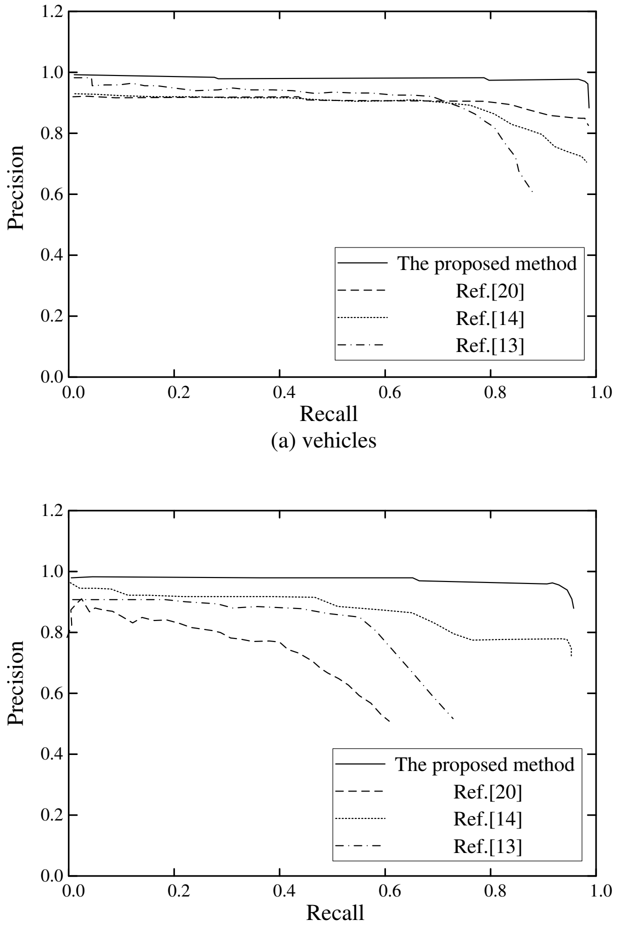


Fig.7 Model recall comparison

As can be seen from Fig. 7, as the number of iterations increases, the recall of various methods is also rising. The recall of the proposed method is better than other comparison methods. Because the improved Faster R-CNN network can adapt to various complex environments, the highest recall reaches 85.6%. The recall of reference [20] and reference [14] are similar, and both are lower than the proposed method. Among them, the reference [14] is optimized through the YOLOv3 network and used for vehicle target recognition, but the influence of factors such as illumination is not considered, and the recall is low.

4.4.3 Comparison of precision recall curves

When performing target object detection, IoU is used to define the matching degree between the real object and the predicted object, and the PR (precision-recall) precision recall curve is drawn through calculation. The precision recall curves of vehicles and dim targets in different methods are shown in Fig. 8.



(a) vehicles

(b) Small target

Fig.8 Precision recall curve of vehicle and dim target

It can be seen from Fig. 8 that the precision recall curve of the vehicle is better than the precision recall curve of the weak target. Because the vehicle has more targets, the characteristics are clearer. There are many types of dim targets, and some of them are similar in type, so it is difficult to identify them. In addition, the precision recall curve of the proposed method is better than other comparison methods for the recognition of vehicles or dim targets.

5 Conclusion

Aiming at the problem that it is difficult to accurately identify similar dim targets in road traffic scenes, an improved detection and identification method based on residual network is proposed. The residual network with stronger learning ability is used to obtain more effective feature expression and rich semantic information. This paper establishes a ResNet50 network model to extract features of vehicles and dim targets from the original image. The model does not rely on image preprocessing and data conversion, and can autonomously extract feature expressions of vehicles and dim targets through learning. Compared with the various features extracted by manual design, it can more accurately reflect the effective information of the identified target. The test results show that the average target recognition accuracy of this method is 94.7%. It has excellent actual generalization performance and obtains a stable high recognition accuracy rate. The disadvantage is that the Faster R-CNN model requires a long training time. Training requires training data under the condition of GPU memory greater than 8G. But after training, it does not affect the recognition speed of the actual test.

Ethical approval

All authors have read "Ethical Responsibilities of Authors" of Soft Computing, and they all promise to strictly abide by the publication ethics.

Funding details (In case of Funding)

This work was supported by the Youth Scientific Research Fund Project of Pingdingshan University (No. PXYQNJJ2017002), Key scientific research projects of institutions of higher learning of Henan Provincial Department of Education (No. 21B520014), HN-ITSETRC(Henan Intelligent Traffic Safety Engineering Technology Research Center) and HN-MDITSETRC (Henan Multimodal Data Intelligent Traffic Safety Engineering Technology Research Center).

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Informed Consent

This research did not involve human participants.

Author's contribution

The main idea of this paper is proposed by Jianfang Liu. The algorithm design and experimental environment construction are jointly completed by Hao Zheng and Xiaogang Ren. The experimental verification was completed by all the three authors. The writing of the article is jointly completed by Hao Zheng and Xiaogang Ren. And the writing guidance, English polish and funding project are completed by Jianfang

Liu.

References

[1] Lin Y, La N, Lou et al. Robot vision system for 3D reconstruction in low texture environment[J]. Optics & Precision Engineerin, 2015, 23(2):540-549.

[2] Hirasawa T, Aoyaraa K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images[J], Gastric Cancer, 2018, 21(4): 653-660.

[3] Liu M, Shan S, Wang R, et al. Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014:1749-1756.

[4] Tang P, Wang H, Kwong S. G-MS2F: GoogLeNet based rtiulti-stage feature fusion of deep CNN for scene recognition[J]. Neurocomputing, 2017, 225(Feb.l5): 188-197.

[5] Huang B, Huang M, Gao Y, et al. 3D object detection incorporating instance segmentation and image restoration[J], Wuhan University Journal ofNatural Sciences, 2019, 24(4): 360-368.

[6] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[7] 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. Las Vegas, NV, USA, 2016, IEEE, 2016: 770-778.

[8] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C]// Advances in neural information processing systems. Kuching, Malaysia, 2015: 91-99.

[9] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[C]// Advances in neural information processing systems. Barcelona, Spain, 2016,IEEE, 2016: 379-387.

[10] Redmon J, Diwala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA, 2016, IEEE, 2016: 779-788.

[11] Jo Y, Jung I. Analysis of vehicle detection with WSN-based ultrasonic sensors[J]. Sensors, 2014, 14(8): 14050-14069.

[12] Kim D H, Choi K H, Li K J, et al. Performance of vehicle speed estimation using wireless sensor networks: a region-based approach[J]. Journal of super-conaputing, 2015, 71(6): 2101-2120.

[13] Unzueta L, Nieto M, Cortes A, et al. Adaptive multicue background subtraction for robust vehicle counting and classification[J]. IEEE Transactions on Intelligent Transportation Systems, 2012, 13(2): 527-540.

[14] X. Zhang and X. Zhu. Vehicle Detection in the Aerial Infrared Images via an Improved Yolov3 Network[C]// 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), Wuxi, China, 2019: 372-376.

[15] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, realtime object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016: 779-788.

[16] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]// Proceedings of the European Conference on Computer Vision, Amsterdam, Netherlands, 2016: 21-37.

[17] Chen X, Xiang S, Liu C L, et al. Vehicle detection in satellite images by hybrid deep convolutional neural networks[J]. IEEE Geoscience and remote sensing letters, 2014, 11(10): 1797-1801.

[18] Gao Y, Guo S, Huang K, et al. Scale optimization for full-image CNN vehicle detection[C]. Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV), California, USA , 2017: 785-791.

[19] Tian B, Tang M, Wang F Y. Vehicle detection grammars with partial occlusion handling for traffic surveillance[J]. Transportation Research Part C: Emerging Technologies, 2015, 56: 80-93.

[20] W. Dong, Z. Yang, W. Ling, Z. Yonghui, L. Ting and Q. Xiaoliang. Research on vehicle detection algorithm based on convolutional neural network and combining color and depth images[C]// 2019 2nd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 2019: 274-277.

[21] Tang P, Wang H, Kwong S. G-MS2F: GoogLeNet based rtiulti-stage feature fusion of deep CNN for scene recognition[J]. Neurocomputing, 2017, 225(Feb.l5): 188-197.

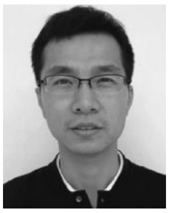
[22] Huang B, Huang M, Gao Y, et al. 3D object detection incorporating instance segmentation and image restorationfJ], Wuhan University Journal ofNatural Sciences, 2019, 24(4): 360-368.

[23] Yang H, Qiu S. Realtime vehicle detection and counting in complex traffic scenes using background subtraction model with lowrank decomposition[J]. IET Intelligent Transport Systems, 2018, 12(l):75-85.

[24] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

Hao Zheng, MA.Eng of Computer Science, Lecturer. Graduated from Wuhan University in 2012. Worked in Pingdingshan University. His research interests include Machine learning, image processing and natural language processing.

Jianfang Liu, MA.Eng of Computer Science, Lecturer. Graduated from Wuhan University in 2011. Worked in Pingdingshan University. Her research interests include image processing, machine learning and data mining.



Xiaogang Ren received the B.S. degree in electronic information engineering from Jiangnan University, in 2003, and the M.S. degree in computer technology from the East China University of Technology, in 2016. He is now a PhD Student in China University of Mining and Technology. His research interests include 3D reconstruction and 3D printing, medical image segmentation, and machine learning.

Figures

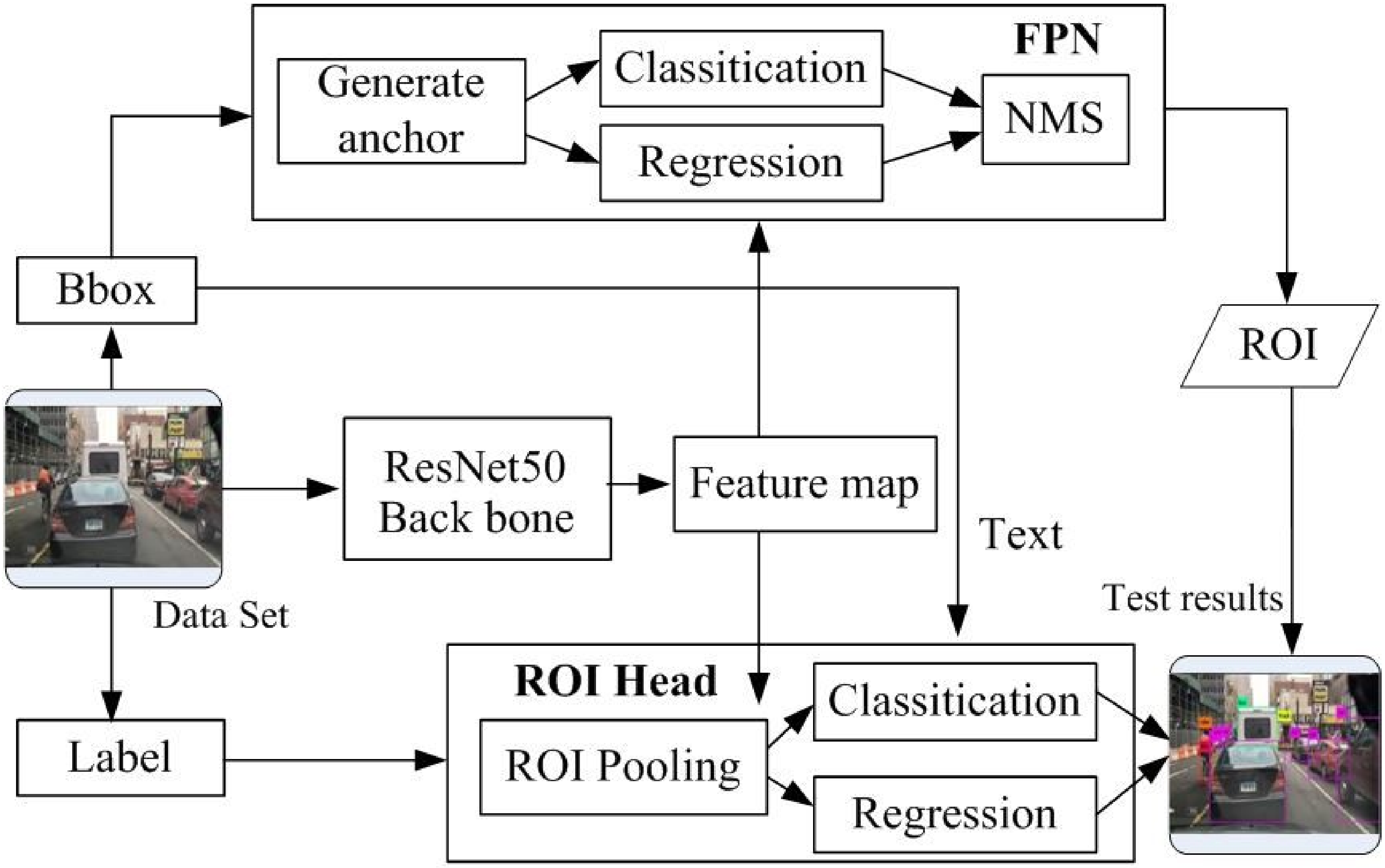


Figure 1

The structure of target detection network is proposed

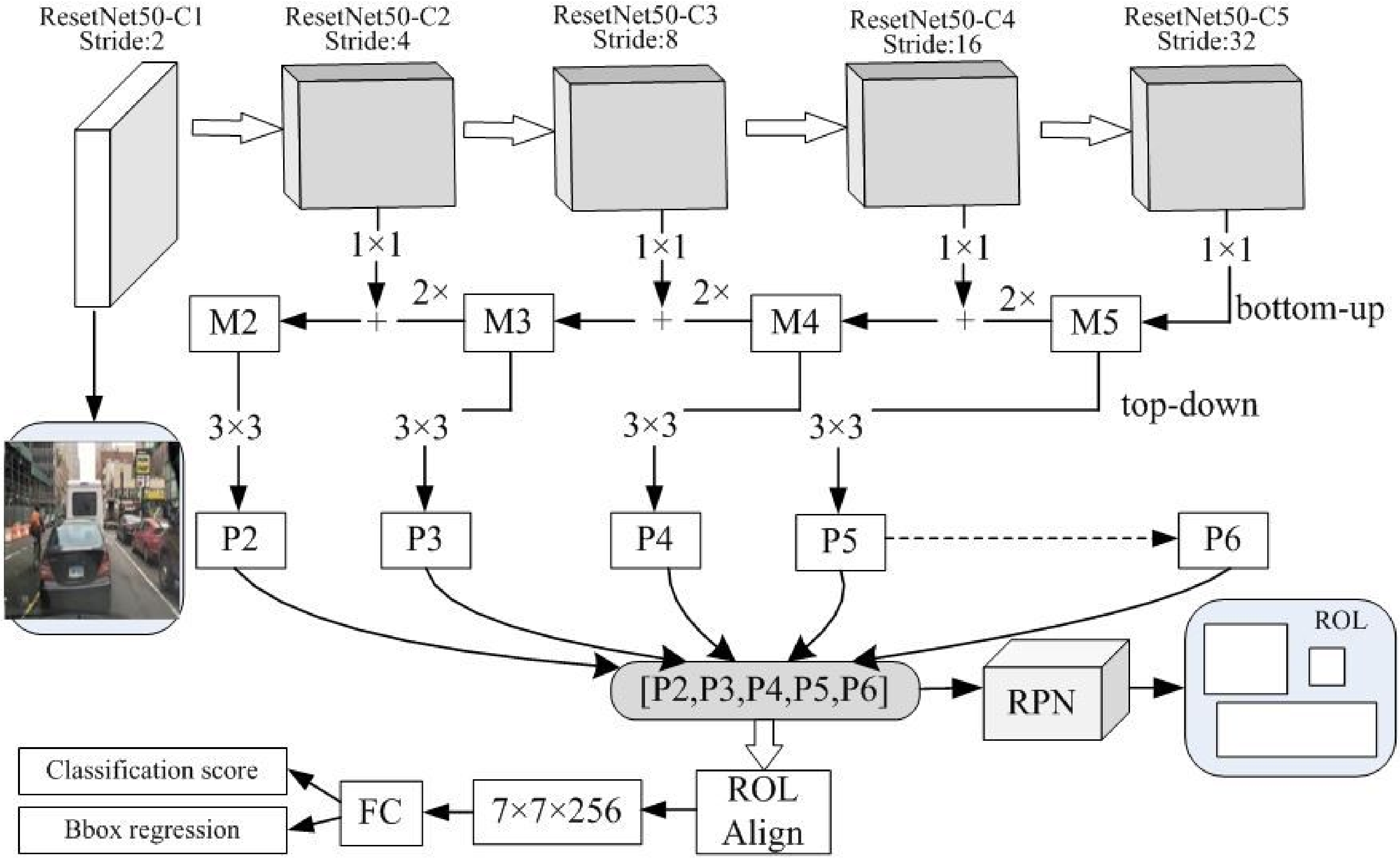


Figure 2

Region proposal networks architecture

Figure 3

Datasets used in the experiment

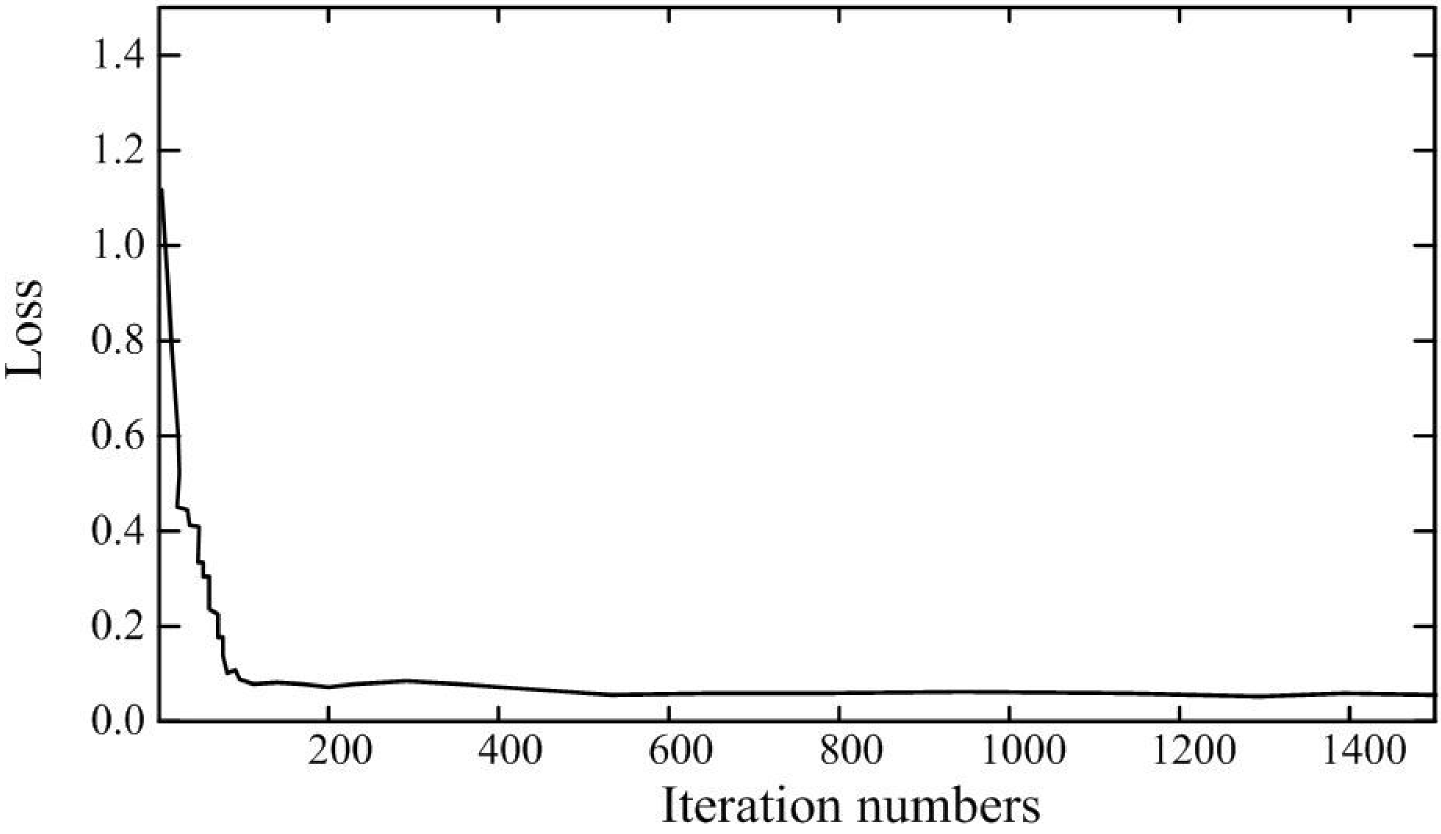


Figure 4

Loss curves of training

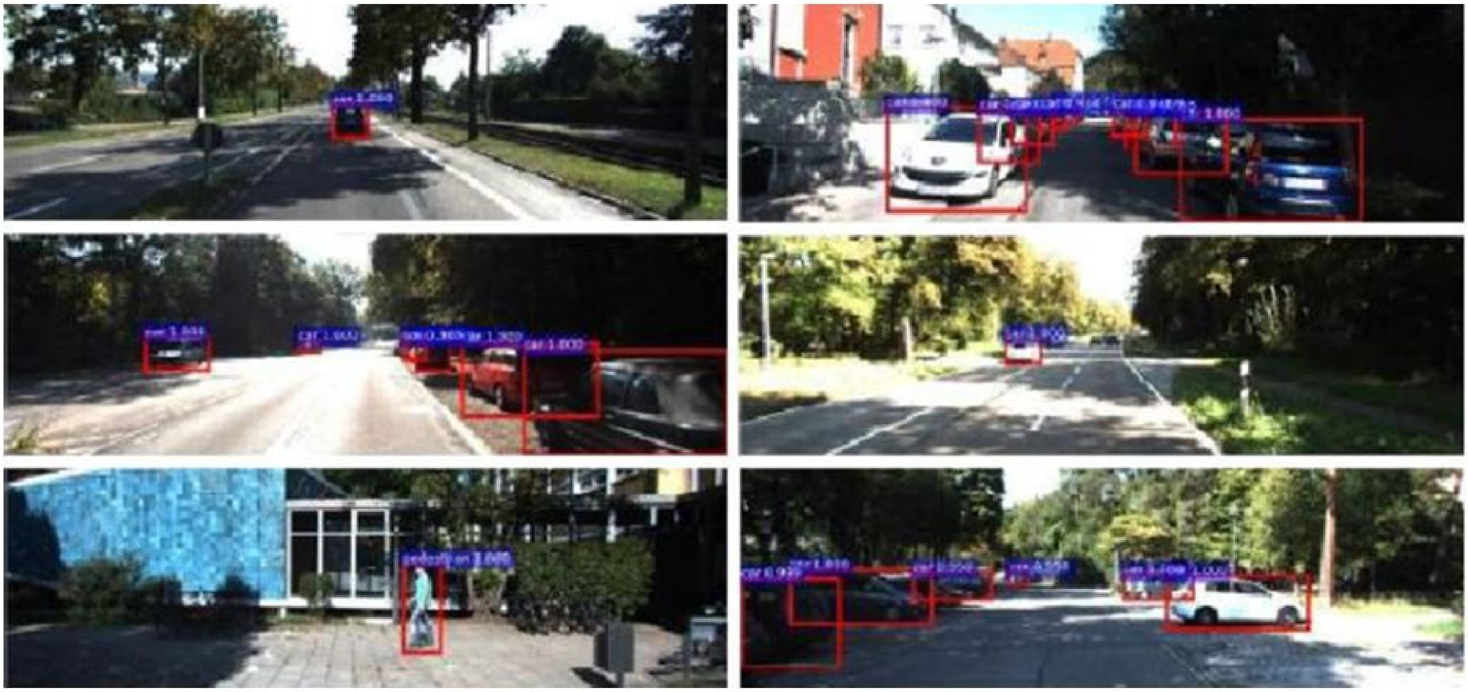


Figure 5

Recognition effect picture

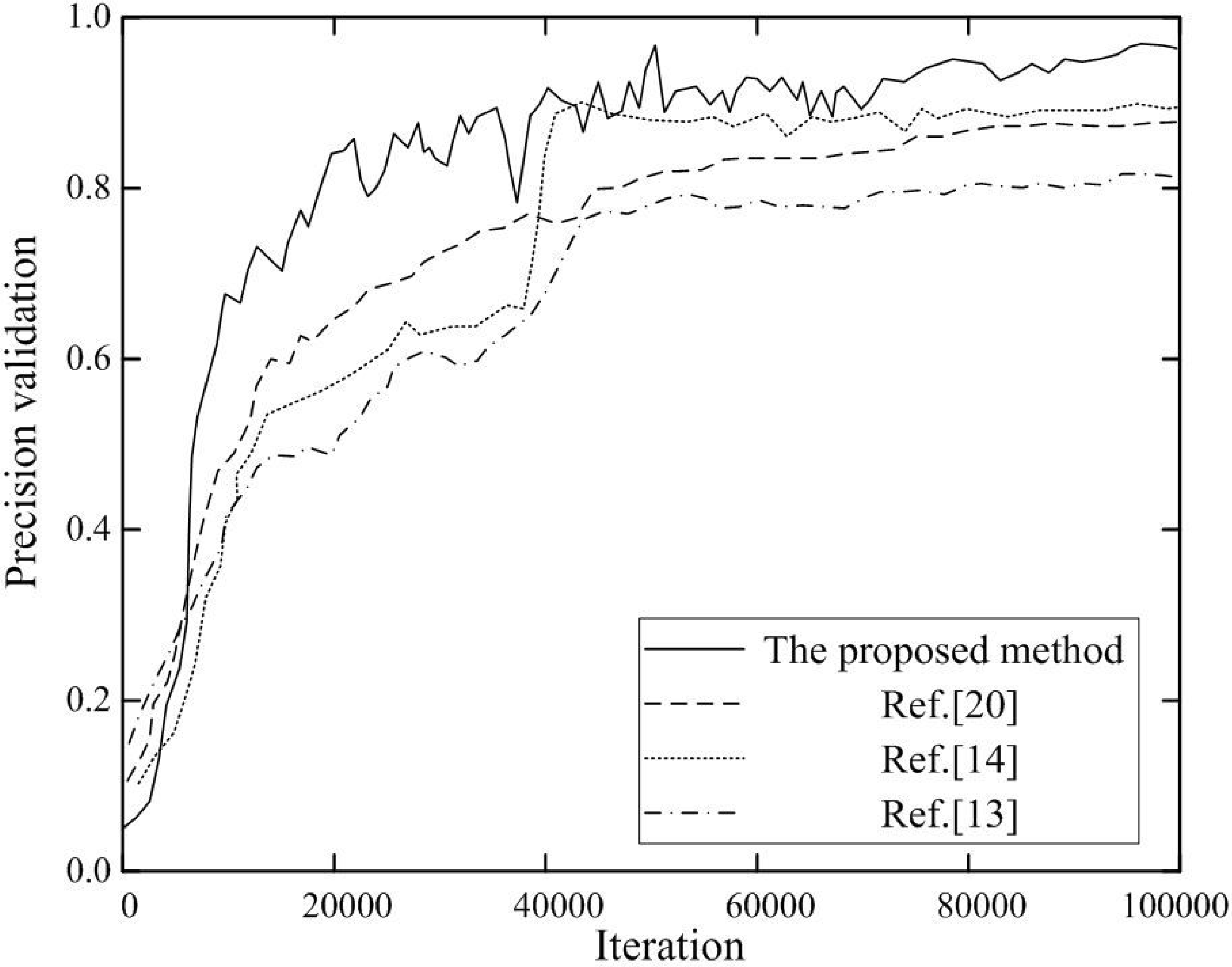


Figure 6

Model precision comparison

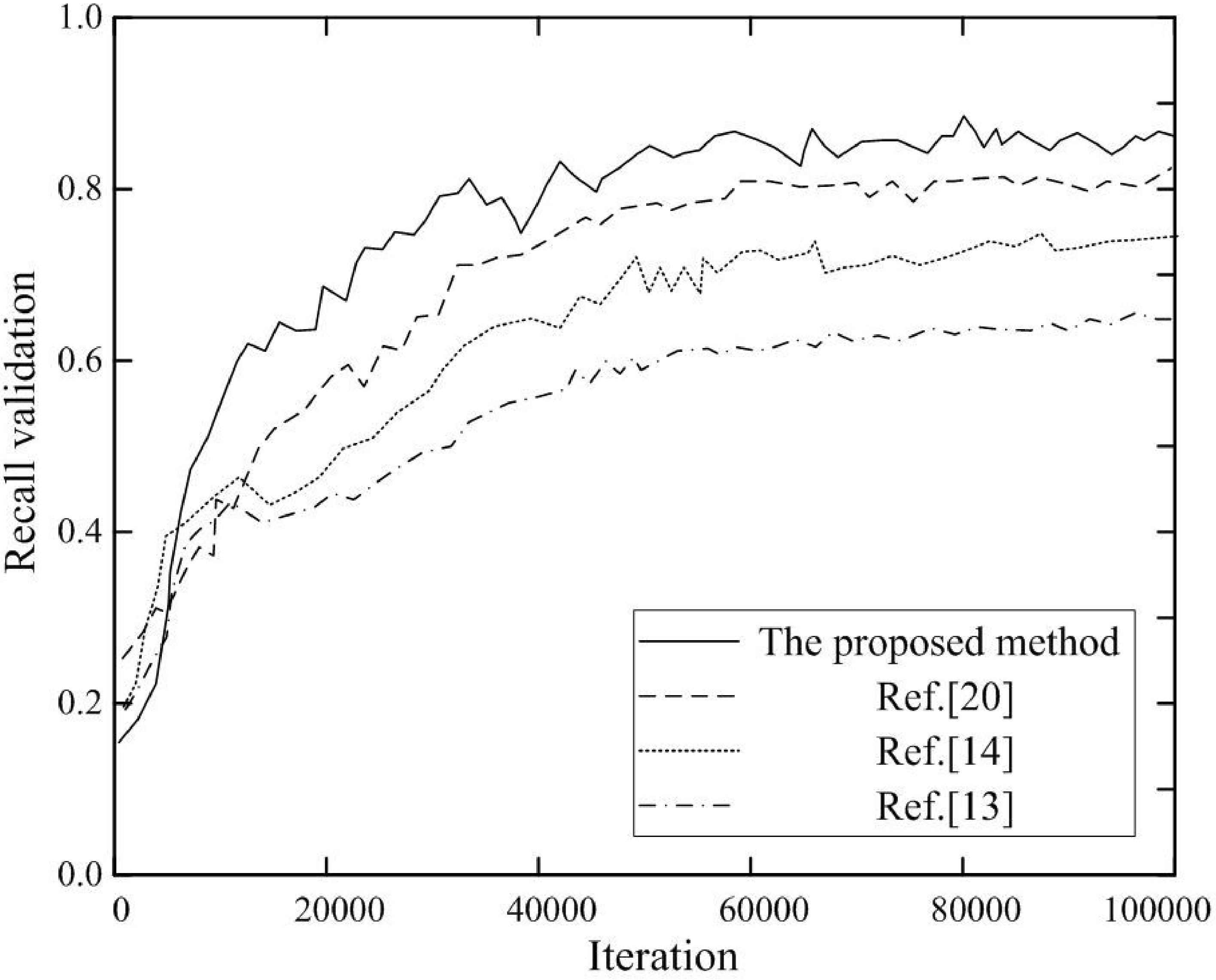
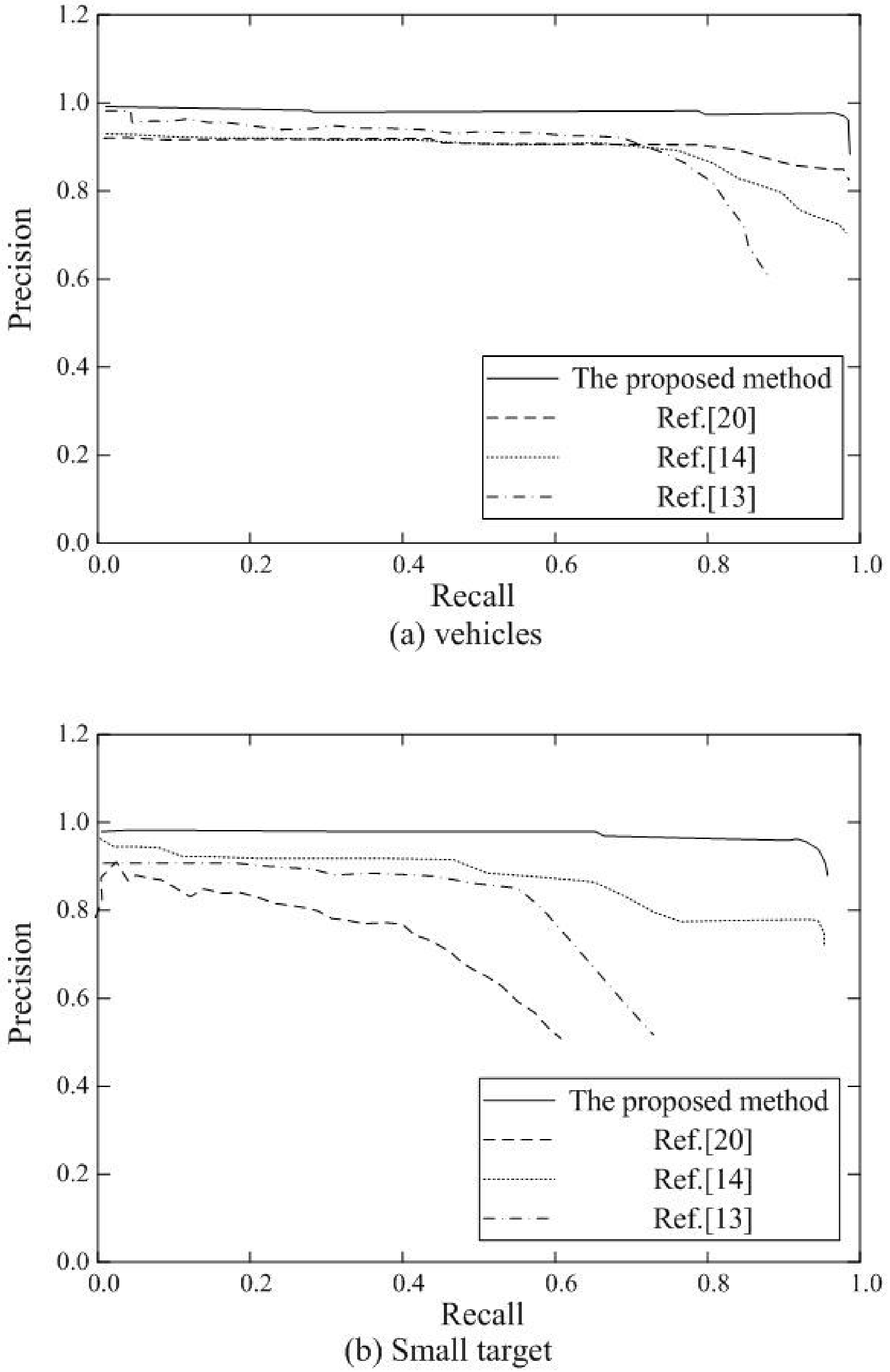


Figure 7

Model recall comparison



(a) vehicles

(b) Small target

Figure 8

Precision recall curve of vehicle and dim target