Sensing and Imaging (2023) 24:17
https://doi.org/10.1007/s11220-023-00425-6

ORIGINAL PAPER

A Novel Faster RCNN Model Based on Multi-scale Feature Fusion and Shape Priori for Dense Vehicle Detection

Yamin Li[1]

Received: 3 July 2021 / Revised: 31 January 2023 / Accepted: 20 May 2023 /
Published online: 9 June 2023

Abstract

Compared with conventional targets, dense targets have less information, and the training data is difficult to be labeled, which leads to the poor detection effect with general target detection methods on dense targets, while the detection methods specially designed for dense targets are often too complex or not universal. To solve the above problems, this paper proposes a novel faster RCNN model based on multi-scale feature fusion and shape priori for dense vehicle detection. This proposed dense vehicle detection model is divided three steps. Firstly, the training strategy of multi-scale network captures the lost detail of target density. Secondly, the anchor point generation method based on shape priori is used to calculate the shape changes of dense targets at different scales. Thirdly, considering that the objects with dense distribution have large appearance differences, different sizes of convolution kernels are used to extract the feature information of different scales in images. It effectively solves the problem of target information loss in the existing detection models. Finally, we conduct experiments on the public CARPK dataset to illustrate the effectiveness of the proposed method. Compared with the state-of-the-art vehicle detection methods, the proposed network model can achieve better detection effect for the dense distributed targets in different scene conditions.

Keywords Dense vehicle detection · Faster RCNN · Multi-scale feature fusion · Shape priori · Anchor point generation

✉ Yamin Li
ljnan127@163.com

[1] School of Electronic and Electrical Engineering, Zhengzhou University of Science and Technology, Zhengzhou 450064, China

1 Introduction

With the acceleration of China's modernization, the construction speed of urban infrastructure cannot meet the requirements of rapid economic growth. The crowed people and congested vehicles often occur in the city, they result in traffic safety, road congestion, environmental pollution and other increasingly prominent problems [1]. One of the main reasons for the frequent accidents is the high volume of traffic on the road. There is a direct relationship between traffic congestion and the degree of vehicle density. As shown in Fig. 1, the dense distributed vehicles in the actual scene will not only directly lead to traffic accidents, but also affect the service capacity of urban infrastructure. If the intelligent traffic analysis system can detect the dense distribution of vehicles in real time, it can further guide the relevant management departments to check and control the abnormal behavior of vehicles and the road flow. And if it can timely remind, warning these behaviors, road congestion and accidents can be avoided to a large extent. How to efficiently analyze the dense distribution of vehicles to ensure travel safety and smooth roads has become the urgent problem to be solved by the intelligent traffic analysis system, which first involves the dense target detection technology.

However, the images or video data collected by surveillance cameras in different places not only have complex and diverse backgrounds, but also dramatically change the scale and perspective of the shooting target [2, 3]. Dense distributed targets are often limited by shooting Angle, occlusion, light, background, image resolution and other factors, and even the same target often presents different appearance details. These factors often lead to the difficulty of feature extraction, which makes the detection of dense targets be a very challenging problem. For these challenges, some detection methods for dense targets have been proposed by scholars. Compared with non-dense targets, densely distributed targets show

Fig. 1 Examples of a dense vehicle target in a real world scenario. Left: Traffic Accident. Right: Illegal parking

larger differences in appearance, scale and visual Angle, and they are easy to block each other [4]. Most of the existing target detection methods do not consider the characteristics of the dense target itself, there are still the following problems that have not been solved.

First of all, dense targets have large scale differences. After the image is input into the network for processing, the edge features of small targets are often not obvious or even missed, which can lead to that targets cannot be recognized and detected normally. Secondly, the detection scene of dense target is relatively complex and the number of targets is large, so the existing methods cannot effectively detect all targets. To address these deficiencies, this paper proposes a novel faster RCNN model based on multi-scale feature fusion and shape priori for dense vehicle detection. The flow chart of the proposed model is shown in Fig. 2. This deep network model is based on Faster region-based Convolutional Neural Network [5], using Resnet50 (Residual Neural Network) [6] to replace the original VGG16(Visual Geometry Group Network) [7] structure. The clustering-based adaptive anchor point generation method, multi-resolution training and Inception [8] network structure are combined to detect the dense targets.

The main contributions of this paper are as follows:

1. In view of the challenge brought by the large scale difference of dense targets, in this paper, the real labeled Windows in the dense target training set are clustered, and an adaptive anchor point generation strategy based on shape priori is adopted to generate candidate Windows that are convenient for network training.
2. Faster RCNN is used to train the network. By using images with different resolutions for training, more robust features can be obtained and the problem of small target information loss in the existing detection model can be solved.

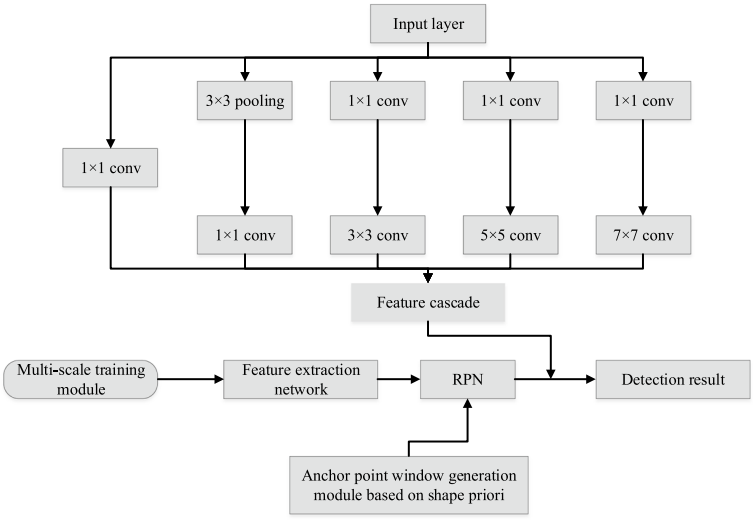

Fig. 2 The proposed detection model

3. In this paper, by using the convolution kernel with different sizes to extract the feature information of different scales, all the features are connected in series, which greatly enriches the content of the feature information and effectively solves the problem of the important feature information loss of dense targets.
4. The proposed method is verified on the open dense vehicle data set CARPK [9]. Experimental results show that the detection accuracy of the proposed method is significantly higher than that of the existing classical target detection methods and small target detection methods, and the problem of dense target detection is solved.

2 Related Work

At present, target detection algorithm based on deep learning has become the mainstream. Compared with traditional target detection methods, the target detection method based on deep learning uses neural network to automatically extract more effective features, which greatly improves the performance of target detection task. This section mainly introduces the relevant work from the following three aspects.

2.1 Two-Stage Target Detection Methods

The two-stage target detection method, also known as candidate window based detection method, mainly divides the whole detection process into two parts. Firstly, the region of interest (ROI) is generated and features are extracted through the deep convolutional network. The position and category of the target are obtained by the detector. Typical two-stage target detection includes Region based Convolutional Neural Network (R-CNN) [10], Spatial PyramidPooling Network (SPP-NET) [11] and fast region-based convolutional Neural Network (Fast-CNN) [12], Faster R-CNN and so on. R-CNN obtains the candidate region by selective search algorithm, and obtains the feature information of each candidate window through a Convolutional Neural Network (CNN). However, the efficiency is low because of the overlapping region calculation between multiple candidate regions. Fast R-CNN inputs the whole image into the convolutional network, designs a new ROI pooling layer and extracts the features of each candidate window. In addition, the classification and border regression are combined to improve the experimental accuracy and training speed. The calculation of traditional methods generating candidate area is complex. Faster R-CNN uses RPN to replace selective search algorithm. The candidate region generation, feature extraction and window classification and regression modules are integrated into an end-to-end deep network model, which not only reduces the computational cost of the algorithm, but also improves the detection performance. The proposed deep network model is based on the Faster RCNN model. Resnet50 feature extraction network is used to replace the original VGG16 network. Because Resnet3 uses residual module, it can effectively optimize the performance of the network, so that the model has a better feature extraction effect on dense targets.

2.2 Single Stage Target Detection Methods

The single stage target detection method omits the target-like window generation step and directly detects the predicted target. Therefore, compared with the two-stage target detection method, its prediction speed is faster. The main representatives of single-stage target detection methods include YOLO (You Only Look Once) [13], SSD (Single Shot Multi-Box Detector) [14] and RetinaNet. YOLO only uses an end-to-end neural network to complete the target detection. Although the process is simple and the structure is simplified, the detection error is large. SSD integrates the excellent design of two-stage detection method, uses multi-scale convolution feature map to predict the results and sets a prior box, which improves the accuracy, speed and other performance to a certain extent. For the low effect of setting anchor box, the proposed method adopts an adaptive anchor box generation method based on clustering to effectively reduce the poor effect caused by the size mismatch between common anchor frame and target on the detection results of dense targets.

2.3 Existing Small Target Detection Methods

Many scholars have proposed different methods for small target visual tasks. Hu et al. [15] proposed an efficient small-scale face detection model by exploring the scale, image resolution and context information. Zhang et al. [16] proposed a scale-invariant one-stage face detection model. Wang et al. [17] improved SSD and proposed a small target detection method based on mobile GPU platform. But dense target detection is more difficult than small target detection. In dense small target scenes, there is always the occlusion between targets, which causes the interference between each other [18]. For the dense target detection, Lin et al. [19] proposed the Focal Loss. Focal Loss is a specific cost sensitive method. By adding new regular terms to the cross-entropy Loss function, it changes the weight contributions of foreground, background, hard-to-classify samples and classified samples to the Loss function, and it plays an important role in solving the imbalance problem. Meanwhile, Focal Loss also has a good effect on the detection of dense small targets. Gonthier et al. [20] proposed a small target detection method combining with image semantic segmentation. This method extracted the features of small targets based on full convolutional neural network with semantic segmentation. Finally, support vector machine was used to classify small target regions [21]. Song et al. [22] proposed an improved small target detection method based on Faster R-CNN. Based on the candidate box selection model, the method used super-resolution technology to clarify the fuzzy small target and then carry out classification and regression, which improved the performance of small target detection to a certain extent. Dong et al. [23] proposed a perception generative adversarial network, which improved the detection effect of small targets by mapping the features of small targets to similar features of large targets to reduce the differences. However, the current small target detection methods do not consider the scale difference of different targets in the actual scene, nor consider the unique attributes of dense targets, so they cannot

be directly applied to the dense target detection problem [24]. In this paper, according to the characteristics of dense target and the complex environmental factors, the shortcomings of traditional dense targets detection are improved. This paper combines multi-scale training with Inception network structure to extract feature information effectively. Multi-scale training is used to make the model more robust to small targets. By using the convolution kernel of different sizes to extract the feature information of different scales, all the features are connected in series, which effectively solves the problem of losing the important feature information of the dense targets.

3 Proposed Dense Targets Detection Model

In order to solve the challenges caused by the difference of shape, scale and perspective of dense targets, this paper proposes a novel multi-scale network based on data-driven and visual perception to solve the problem of dense target detection. As shown in Fig. 2, the feature expression model mainly includes three parts: multi-scale training model, anchor point window generation module based on shape priori, and feature fusion module based on multi-size convolution kernel.

At present, target detection algorithms based on deep learning generally only use one scale for training. The model trained from a single scale image can effectively deal with the general detection tasks. However, due to the large difference in target scale, dense target and high noise interference in complex environment, it is often difficult to effectively deal with these problems by using only a single scale. Therefore, this paper proposes a multi-scale model training method. In the training phase, a two-stage operation is used to process the image as shown in Fig. 3.

3.1 Anchor Point Window Generation Module Based on Shape Priori

Different from directly detecting the position of the interested target in the image, some current target detection algorithms first generate a series of anchor point windows before classification detection, so as to facilitate the neural network to carry out the subsequent classification and recognition tasks. The size of the anchor window reflects the size of the detected target to some extent. Taking the candidate box

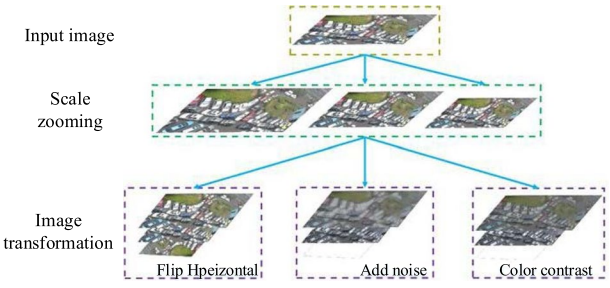

Fig. 3 Multi-scale training model

generated by Region Proposal Network as an example, the anchor window in Faster R-CNN generally contains three ratios, 1:1, 1:2 and 2:1, and has three scales: 128, 256 and 512. Nine reference anchor boxes are set at each position of the feature spectrum, which can cover the target with a side length of 70–768. Thus, the objects with some small size or side length less than 70 will not be detected. Therefore, the adjustment of anchor window is very important for dense target detection.

The current anchor-free target detector marks all features as positive that fall in the predetermined center area of the boundary box of the ground truth label in space. This approach will cause Label noise during training, as some of these features with positive labels may be on the background or occlusions, or are not discriminative features at all. In this paper, we optimized the detector to reduce label noise in the anchor-free target detector. Specifically aggregating predictions that stem from all features into one prediction allows the model to reduce the contribution of non-discriminatory features in training.

In this paper, the width-height ratio and size of the anchor frame suitable for dense target detection are firstly obtained as priors by c [25], so as to reduce the optimization difficulty in positioning and improve the detection rate of the candidate box. In the traditional k-means clustering method, the measurement standard of similarity between data objects is distance, that is, the smaller distance denotes the higher similarity. So they are more likely to belong to the same cluster. According to the analysis, if the Euclidean distance is used to calculate the distance in this paper, the clustering results may have error, that is, the larger the size of the anchor window is, the larger the error will be. Since the purpose of clustering is to determine more accurate initial anchor window parameters, that is, to improve the Intersection-over-Union (IoU), the error is independent of the size of anchor window. Therefore, this paper conducts clustering according to IoU of anchor point window and real annotation box, and defines the following distance function through IoU:

$$d(B, C) = 1 - IoU(B, C) \tag{1}$$

where B represents the labeled target window. C represents the clustering center. Since the convolutional neural network has the characteristics of translational invariant and the position of the anchor point window is fixed by each grid, the cluster center point C in Eq. (1) is represented by width and height. According to Eq. (1), the distance to the clustering center is closer, the result is better, that is, the greater IoU denotes the better result.

In this paper, according to the shape priori of the detection target, especially the scale characteristics, the corresponding size of the anchor point is designed to achieve a better effect. According to cross validation, the number of clustering centers is set $K=9$ in this paper. Figure 4 shows the clustering results of the labeled window in all the training samples.

3.2 Feature Fusion Module Based on Multi-Size Convolution Kernel

In the target detection task, the context information of the target plays an extremely important role. The context reflects the correlation between the target

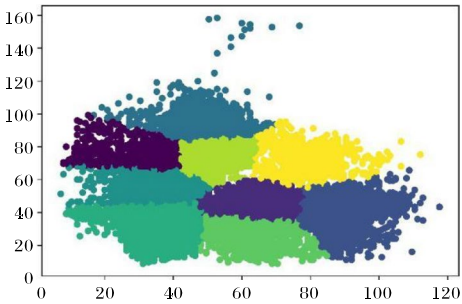

Fig. 4 Clustering results of target windows with different scales

and some local areas or backgrounds, and the use of these information can provide useful clues for target detection. However, the existing convolution layer cannot effectively capture the context information of the target. In view of the above shortcomings, this paper uses the feature fusion module based on multi-scale convolution kernel to extract the context information of the target. Since convolution kernels of different scales have different receptive fields, the smaller convolution kernel can extract the finer feature information. Therefore, the smaller convolution kernel is more suitable for the extraction of detailed information and other features, while the larger convolution kernel is more suitable for the extraction of high-level abstract features. However, if only a single convolution kernel is used, it is easy to cause incomplete feature information.

In this paper, a feature fusion module based on multi-scale convolution kernel is proposed to improve the traditional convolution layer in the network. As shown in Fig. 5, the feature fusion module based on the multi-scale convolution kernel contains five parallel branching structures, namely 1×1 convolution kernel, 3×3 pooling, 3×3 convolution kernel, 5×5 convolution kernel and 7×7 convolution kernel. Finally, the features in the five channels are fused in series. In order to obtain features of the same size before feature fusion, the padding values of 3×3 pooling, 3×3 convolution kernel, 5×5 convolution kernel and 7×7 convolution kernel are set as 1, 1, 2, 3 respectively for expansion. Different convolution kernel sizes have different sizes of receptive fields. Multiple convolution kernels are used to extract the feature information of different scales in the image, and finally the feature information of different scales is fused to make full use of the

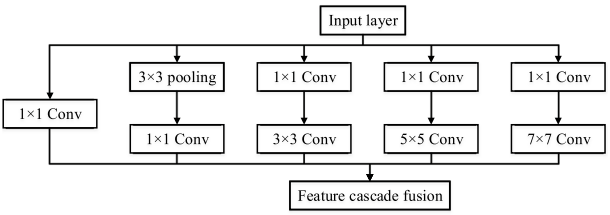

Fig. 5 Feature fusion module based on multi-size convolution kernel

multi-scale feature information. The information of each layer is utilized to obtain a better representation of the image features.

For feature extraction of large-size objects, if the size of the convolution kernel is large enough, the extracted semantic information will be relatively complete; if the size of the convolution kernel is too small, the extracted semantic information will be incomplete and scattered in the features extracted by different convolution kernels.

For feature extraction of small objects, the large-size convolution kernel is easy to introduce noise when extracting semantic information, resulting in difficulty in detecting small objects, while the small-size convolution kernel can extract more complete semantic information of small objects. When detecting dense targets in complex scenes, the network structure is extracted by using convolutional kernel features with different sizes. The multi-scale object feature information is obtained by the convolution kernel of different receptive fields, and the performance and effect of dense target detection can be improved by making full use of the feature information of the multi-receptive fields.

4 Experiment Results and Analysis

4.1 Data Set

Considering that dense targets widely exist in UAV aerial photography scenes and vehicles frequently appear in daily life, in order to verify the effectiveness and reliability of the proposed method in this paper, the CARPK data set is used to evaluate the new method. CARPK is currently the largest vehicle data set taken by UAV in the parking lot scene, which is mainly used to evaluate vehicle detection and counting tasks. The dataset collects 1448 images from four different parking lots and labels about 90,000 vehicle targets. The resolution of image is 720×1280 pixels. In this data set, 989 images are selected as the training set, and the remaining 459 images are as the testing set. The maximum number of targets in a single image reaches 188 vehicles. In the experiment, 80% training images are training sets and 20% are validation sets. The setting of all the hyperparameters is performed on randomly separated training sets, so this data set is very challenging.

4.2 Evaluation Indexes

In this paper, the Average Precision (AP) is used as the evaluation index to measure the accuracy of the proposed algorithm. The measurement comprehensively considers the positioning accuracy and classification accuracy, in which the box matching threshold IoU is set as 0.5. Firstly, the prediction results of the model are arranged in descending order according to the confidence value of each predicted value, and the prediction is judged by combining the box matching threshold IoU. Then, the recall rate and accuracy are calculated according to the prediction results.

For each recall rate, the accuracy greater than or equal to the recall rate is selected, and the average accuracy is defined as the average of accuracy under these recall rates. The running time is defined as the average time of testing ten images.

4.3 Experimental Environment and Setting

The experiment environment of this paper is: Tensor Flow, Intel Xeon E5-2620 v4@2.10 GHZ, Nvidia Titan X1060. The proposed method in this paper is based on the classic Faster R-CNN method, and the basic network uses Resnet50 pre-trained on ImageNet. During the training phase, the size of the input network is set to 1280×720. For data augmentation, each training sample is obtained from the original image by multi-level multi-resolution module. During the testing phase, the initial confidence threshold is set to 0. The stochastic gradient descent method is used to optimize the objective function, and the basic learning rate is set as $5e^{-4}$.

4.4 Comparison Experiments

This section mainly evaluates the performance of the presented method and the existing methods including Faster R-CNN, SSD, YOLOv2, YOLOv3, SA + CF + CRT (scale-adaptive + Circular Flow + Counting regularization Term) [17], GANet (Guided Attention Network) [25]. In addition, due to the large number of dense targets, the ROI pooling layer is used in the experiment to fine-tune the final detection results. The window number of network output is adjusted from 100 to 300.

Table 1 shows the results with different methods on the CARPK dataset. As can be seen from Table 1, the presented method in this paper has obvious advantages, achieving the best performance of 96.5%. The average accuracy of proposed method improves by 9.5%, 15.7%, 11.3%, 27.8%, 6.8% and 6.5% than that of Faster R-CNN, SSD, YOLOv2, YOLOv3, SA + CF + CRT and GANet respectively. Compared with the general target detection algorithms, this proposed method uses the multi-scale feature fusion method to effectively solve the problem of missing the important feature information in the complex scene according to the basic features of the dense target. At the same time, the adaptive strategy of generating anchor box is adopted to effectively reduce the negative effect of the errors generated by using the traditional anchor frame on the dense target detection results.

Table 1 The comparison with different methods

Method	AP/%	Time/s
Faster RCNN	87.1	3.6
YOLOv2	80.9	3.4
YOLOv3	85.3	2.9
SSD	68.8	2.5
SA + CF + CRT	89.8	2.3
GANet	90.1	1.8
Proposed	96.6	1.2

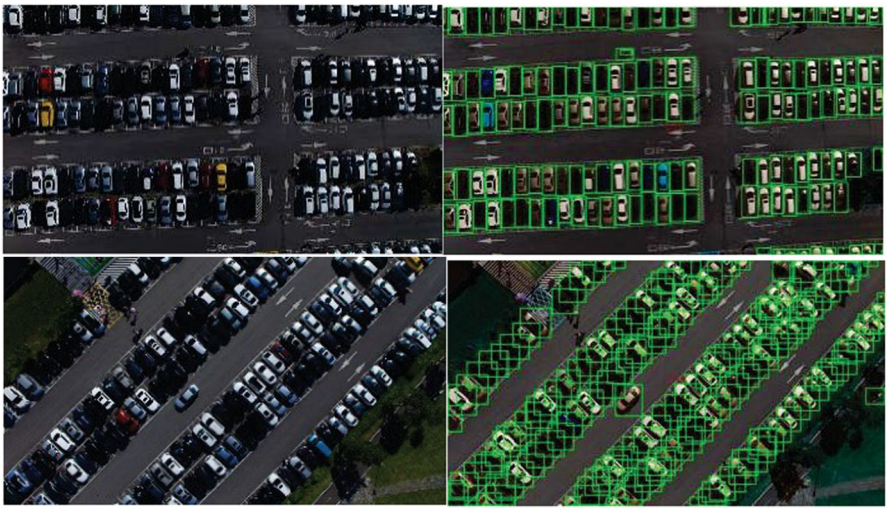

Fig. 6 Detection result with proposed method

As can be seen from the subjective effect shown in Fig. 6, the proposed method can accurately detect vehicle targets with dense distribution, which is feasible in practical application.

4.5 Ablation Experiments

In this paper, a deep network model based on multi-scale fusion is proposed. Resnet50 feature extraction network is used to replace the VGG16 structure in the original Faster R-CNN network. Resnet structure can accelerate the training of neural network, and the accuracy of the model can also be greatly improved. At the same time, the proposed three modules in this paper, namely multi-scale training module, anchor point window generation module based on shape priori and feature fusion module based on multi-size convolution kernel, are used to detect dense targets. In order to verify the effectiveness of the proposed method for dense target detection in complex scenes, the proposed algorithm is trained on CARPK data set with the same strategy, and each improved module is selected for ablation experiment. It

Table 2 The results with different modules

Method	AP/%
Resnet50	92.6
Resnet50 + adaptive anchor point	95.0
Resnet50 + adaptive anchor point + multi-scale training	96.1
Resnet50 + adaptive anchor point + multi-scale training + feature fusion	96.9

includes adaptive anchor points, multi-scale training and feature fusion. The results are shown in Table 2.

From Table 2, it can be summarized as follows:

1. After the use of the adaptive anchor frame generation module, AP is improved by 2.4% on the basis of the original model, which indicates that the adaptive anchor frame generation method based on clustering can help to obtain a better anchor frame window, thus improving the accuracy of dense target detection.
2. On the basis of using the adaptive anchor frame generation module and then using the multi-scale training method, the AP increases by 1.1% compared with only using the adaptive anchor frame generation module. This experimental result verifies the effectiveness of the multi-scale training method.
3. The first two experiments verify the effectiveness of the cluster-based adaptive anchor module and multi-scale training respectively. We further analyze the impact of feature fusion module on performance. It can be seen from Table 2 that the feature fusion module based on different convolution kernels can further improve the performance of dense target detection. In this paper, the new method combining the three modules achieves the best performance, and the AP increases by 4.3% than that of the original model, indicating that the combination of the three modules in this paper is effective for detecting dense vehicle targets.

5 Conclusion

In order to improve the accuracy of dense vehicle target detection, a dense target detection method based on multi-scale fusion network is proposed in this paper to solve the shortcomings of existing target detection technologies. The proposed multi-scale network training strategy can effectively capture the missing details of target density. This paper not only analyzes the shape variation of dense targets at different scales, but also considers the large appearance difference of dense distributed targets. By using the convolution kernel of different sizes to extract the feature information of different scales, the problem of target information loss in the existing detection model is effectively solved, and the automatic recognition of dense targets is realized. The validity of the proposed method is verified on public data sets. In the future work, the problem of multi-class dense target detection will be further studied.

References

1. Helmy, H., Kamaluddin, M. T., & Iskandar, I. (2022). Investigating spatial patterns of pulmonary tuberculosis and main related factors in Bandar Lampung, Indonesia using geographically weighted poisson regression. *Tropical Medicine and Infectious Disease, 7*(9), 212.
2. Yin, S., Li, H. & Teng, L. (2020). Airport detection based on improved faster RCNN in large scale remote sensing images. *Sensing and Imaging, 21*, 2020. https://doi.org/10.1007/s11220-020-00314-2
3. Yin, S., & Li, H. (2020). Hot region selection based on selective search and modified fuzzy C-means in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 5862–5871. https://doi.org/10.1109/JSTARS.2020.3025582

4. Wang, K., Du, S., Liu, C., et al. (2022). Interior attention-aware network for infrared small target detection. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–13.
5. Ren, S., He, K., Girshick, R., et al. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 39*(6), 1137–1149.
6. Chen, F. (2022). Modal decomposition of an incoherent combined laser beam based on the combination of residual networks and a stochastic parse gradient descent algorithm. *Applied Optics, 14*, 61.
7. Liang, J., Xu, F., & Yu, S. (2022). A multi-scale semantic attention representation for multi-label image recognition with graph networks. *Neurocomputing, 491*, 14–23.
8. Wang, Y., & Derr, T. (2021). Tree decomposed graph neural network. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 2040–2049.
9. Liu, J., Wang, X., & Tai, X. C. (2022). Deep convolutional neural networks with spatial regularization, volume and star-shape priors for image segmentation. *Journal of Mathematical Imaging and Vision, 64*(6), 625–645.
10. Tseng, K. K., Lin, J., Chen, C. M., et al. (2021). A fast instance segmentation with one-stage multi-task deep neural network for autonomous driving. *Computers & Electrical Engineering, 93*, 107194.
11. Dewi, C., Chen, R. C., Yu, H., et al. (2021). Robust detection method for improving small traffic sign recognition based on spatial pyramid pooling. *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18.
12. Girshick, R. (2015). Fast R-CNN. In *2015 IEEE international conference on computer vision (ICCV)*, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169.
13. Redmon, J., Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.
14. Liu, W. et al. (2016). SSD: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, M. Welling (Eds.) *Computer vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9905*. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2
15. Hu, P., & Ramanan, D. (2017). Finding tiny faces. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1522–1530, doi: https://doi.org/10.1109/CVPR.2017.166.
16. Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., Li, S. Z. (2017). S3FD: Single shot scale-invariant face detector. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 192–201, https://doi.org/10.1109/ICCV.2017.30.
17. Wang, Y., Hou, J., Hou, X., et al. (2021). A self-training approach for point-supervised object detection and counting in crowds. *IEEE Transactions on Image Processing, 30*, 2876–2887.
18. Wang, X., Yi, S., Liu, D., et al. (2020). Accurate playground localisation based on multi-feature extraction and cascade classifier in optical remote sensing images. *International Journal of Image and Data Fusion, 11*(3), 233–250.
19. Lin, T., Goyal, P., Girshick, R., He, K., Dollár, P. (2020). Focal loss for dense object detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42*, no. 2, pp. 318–327, https://doi.org/10.1109/TPAMI.2018.2858826.
20. Gonthier, N., Ladjal, S., & Gousseau, Y. (2022). Multiple instance learning on deep features for weakly supervised object detection with extreme domain shifts. *Computer Vision and Image Understanding, 214*, 103299.
21. Krishna, H., & Jawahar, C. V. (2017). Improving small object detection. In *2017 4th IAPR Asian conference on pattern recognition (ACPR)*, pp. 340–345, https://doi.org/10.1109/ACPR.2017.149.
22. Song, R., Huang, Y., Xu, K., et al. (2021). Electromagnetic inverse scattering with perceptual generative adversarial networks. *IEEE Transactions on Computational Imaging, 7*, 689–699.
23. Dong, X., Qin, Y., Gao, Y., et al. (2022). Attention-based multi-level feature fusion for object detection in remote sensing images. *Remote Sensing, 14*(15), 3735.
24. Yin, X., Sasaki, Y., Wang, W., et al. (2020). YOLO and K-means based 3D object detection method on image and point cloud. arXiv:2004.11465
25. Cai, Y., Du, D., Zhang, L., et al. (2019). Guided attention network for object detection and counting on drones. arXiv:1909.11307