A Mask Detection Method for Shoppers Under the Threat of COVID-19 Coronavirus

Wenxuan Han
The College of Software, Xinjiang University, Urumqi
830046, China
1950586584@qq.com

Zitong Huang
The College of Software, Xinjiang University, Urumqi
830046, China
342119334@qq.com

Alifu.kuerban*
The College of Software, Xinjiang University, Urumqi
830046, China
ghalipk@xju.edu.cn

Meng Yan
School of Computer Science; South China Normal University;
Guangzhou 511300
384884002@qq.com

Haitang Fu
School of Computer Science and Technology; Southwest
University for Nationalities; Chengdu 611200
hitom810@163.com

***ABSTRACT*—Object detection, which aims to automatically mark the coordinates of objects of interest in pictures or videos, is an extension of image classification. In recent years, it has been widely used in intelligent traffic management, intelligent monitoring systems, military object detection, and surgical instrument positioning in medical navigation surgery, etc. COVID-19, a novel coronavirus outbreak at the end of 2019, poses a serious threat to public health. Many countries require everyone to wear a mask in public to prevent the spread of coronavirus. To effectively prevent the spread of the coronavirus, we present an object detection method based on single-shot detector (SSD), which focuses on accurate and real-time face masks detection in the supermarket. We make contributions in the following three aspects: 1) presenting a lightweight backbone network for feature extraction, which based on SSD and spatial separable convolution, aiming to improve the detection speed and meet the requirements of real-time detection; 2) proposing a Feature Enhancement Module (FEM) to strengthen the deep features learned from CNN models, aiming to enhance the feature representation of the small objects; 3) constructing COVID-19-Mask, a large-scale dataset to detect whether shoppers are wearing masks, by collecting images in two supermarkets. The experiment results illustrate the high detection precision and real-time performance of the proposed algorithm.**

Keywords—Object detection; masks; feature fusion; COVID-19; spatial separable convolution.

I. INTRODUCTION

In December 2019, the World Health Organization (WHO) China Country Office was informed of cases of pneumonia of unknown aetiology in Wuhan City, Hubei Province, China [1]. So far, many confirmed cases have been confirmed in many countries, including medical staff. The Chinese government has taken timely public health measures including strengthening surveillance, conducting epidemiological surveys and limiting the inflow and outflow of population in Wuhan. This provides valuable experience for countries around the world to fight the coronavirus. Epidemiological investigations and genotyping have confirmed that COVID-19 is a highly infectious virus. To prevent the spread of the virus, scientists recommend that all people wear face masks in public.

Supermarket belongs to the personnel intensive place, be infected possibility is very high. Although there are inspectors at the door of the supermarket to check the masks and temperature of shoppers. However, in some supermarkets, there are still some people who do not wear masks, which poses a great threat to public safety. This, in other words, raises the possibility of one infected person passing the virus to another. Therefore, in this paper, we focus on real-time face masks detection, created a new dataset called COVID-19-Mask, which aims to automatically detect whether shoppers are wearing masks. Besides, we improved the SSD algorithm and designed a lightweight facemasks detection algorithm based on spatial separable convolution and Feature Enhancement Module (FEM).

This paper is organized as follows: Section 2 introduces the related works. Section 3 describes the dataset COVID-19-Mask. Our algorithm is given in Section 4. Section 5 is the experimental analysis and section 6 concludes the paper.

II. RELATED WORKS

With the rapid development of deep learning, especially deep convolutional neural networks (CNN), computer vision has been made significant advances in recent years on object recognition and detection [2]. The great majority of deep

DOI 10.1109/CVIDL51233.2020.00-54

learning methods for object detection have been designed for large objects but their performances on small-object detection are poor. Unfortunately, the objects in the created COVID-19-Mask dataset are smaller, generated from video captured by mobile devices at a distance.

Some efforts, in many areas, have been devoted to addressing small object detection problems [3 - 8]. The common method [3] [4] is to enhance the feature maps resolution of small objects by simply increase the scale of input images, which often results in heavy time consumption for training and testing. Some others [5 - 8] is centered on generating multi-scale representation which enhances high-level small-scale features by combining multiple lower-level features layers, which is simply increase the feature dimension. Next, we will introduce small object detection research in two areas.

A. Small object detection in remote sensing images

Small object detection in remote sensing images has been a popular problem in computer vision and various methods [9-13] have been proposed to address this challenging task. Traditional methods for this task include [9] [10]. In recent years, thanks to the development of deep learning, CNN-based approaches have been widely adopted in remote sensing small object detection due to their high accuracy. Zou et al. [11] designed a singular value decomposition network for ship detection in spaceborne optical images, which provides a simple yet efficient way to learn the features of remote sensing images. Cheng et al. [12] proposed a rotation-invariant CNN (RICNN) to detect multi object in high resolution optical remote sensing image. Ouyang et al. [13] proposed to combine CNN with the deformation model, which made the process of objection detection more sensitive through multiple models, multi-stage cascade, and other integrated approached.

B. Traffic sign detection

As everyone knows, for the unmanned vehicle to run safely, one of the most import factors is traffic sign detection and recognition. Sermanet et al. [14] proposed to feed multi-stage features to the classifier using connections that skip layers to boost traffic sign recognition. Zhu et al. [15] designed two CNNs for simultaneously localizing and classifying traffic signs. Jin et al. [16] proposed a hinge loss stochastic gradient descent method to train convolutional neural networks (CNNs), which provides better test accuracy and faster stable convergence.

III. COVID-19-Mask DATASET

To automatically detect whether shoppers are wearing masks in supermarket, we construct COVID-19-Mask, a new large-scale image dataset, by collecting images in two supermarkets. The new dataset is made up of the 2 types: wear a face mask, didn't wear a face mask. It should be noted that the images without masks were downloaded from the Internet. In addition, all the image labels were annotated with LabelImg and some samples are shown in Figure 1. Figure 2 shows the size distribution of the object to be detected in the COVID-19-Mask dataset. It can be found from Figure 2 that the sizes of most object are between 25^2 and 150^2 pixels in the COVID-19-Mask dataset. Table 1 shows the statistical details of the dataset.

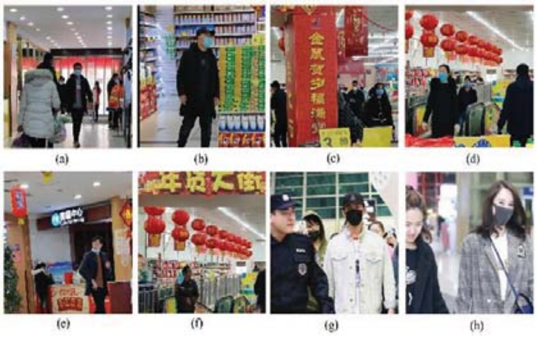

Figure 1. Some samples from the COVID-19-Mask dataset. (a)-(d) are examples of wearing masks taken in a supermarket. (e) and (f) are examples of non-standard wearing of masks taken in supermarkets, which fall under the category of not wearing masks. (g) and (h) are examples downloaded on the Internet.

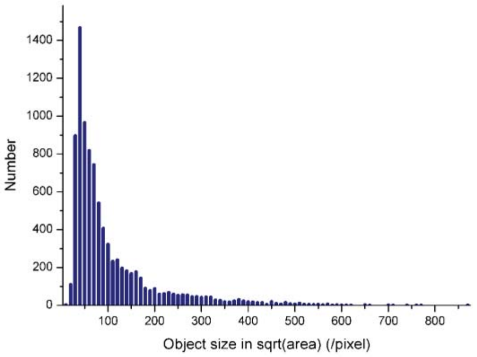

Figure 2. Histogram of object sizes in COVID-19-Mask dataset.

Table 1. The statistical details of the COVID-19-Mask.

Types	Number of Images	Pixels of Images	Number of Objects
wear a mask	4200	1024×1024	7214
didn't wear a mask	800	1024×1024	1658

IV. PROPOSED METHOD

Our goal is to leverage the SSD [17] to build novel detection network to detect whether shoppers are wearing masks in supermarket. However, a large number of experiments have proved that the original SSD has a high miss detection rate and false alarm rate for small object detection, so the SSD cannot be directly used for masks detection. Aiming at the problem of small object missing detection and low detection speed of the original SSD algorithm, several optimization strategies are proposed including a lightweight backbone network and Feature Enhancement Module (FEM). The overview framework of our proposed face masks detector is shown in Figure 3.

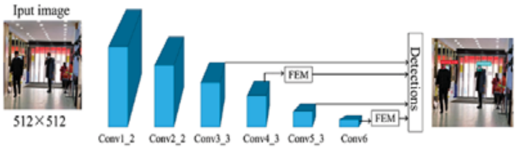

Figure 3. The overview of the proposed method. The Conv4_3 and Conv6 feature map is tailed by FEM.

A. *Lightweight backbone network*

The proposed lightweight backbone network for face masks detection is based on SSD and spatial separable convolution. Our method is based on the facts that: 1) feature maps from the shallow layer in VGG-16 contains more features about small objects [18], and 2) the computational cost of conventional convolutions and deep networks is large, which can lead to slower detection speeds.

Deep neural networks have great challenges in practical applications, because their CPU or GPU occupies a high amount, it is difficult to deploy on small devices, and their real-time performance is poor. In order to solve the problem of high CPU or GPU occupancy, many lightweight neural networks, such as Mobilenet [19] and EffNet [20], have been proposed. The core of EffNet is spatial separable convolution.

Different from conventional convolution, spatial separable convolution splits the convolution kernel into two smaller convolution kernels, and then performs convolution with two small convolution kernels respectively. The most common case is to split the 3 × 3 convolution kernel into 3 × 1 and 1 × 3 convolution kernels.

Assuming that the size of the convolution kernel is K × K, the size of the input image is L×W, and the number of channels is M, the calculation amount of conventional convolution is:

$$C = K^2 * L * W * M \qquad (1)$$

The computation amount of spatial separable convolution consists of two parts: 1×K convolution kernel, the computation amount is:

$$C_1 = K * L * W * M \qquad (2)$$

and the convolution kernel of K×1, the computation amount is:

$$C_2 = K * L * W * M \qquad (3)$$

The total computation is:

$$C = C_1 + C_2 = 2K * L * W * M \qquad (4)$$

We can see that the computation of the spatial separable convolution is only 2/K of the conventional convolution. The structure of spatial separable convolution is shown in Figure 4.

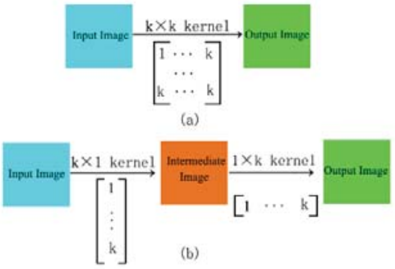

Figure 4. Comparison of spatial separable convolution(b) and conventional convolution(a).

B. *Feature Enhancement Module (FEM)*

Small objects detection is one of the rather challenging tasks in computer vision due to its limited resolution and information. In order to improve the detection accuracy of small objects and Inspired by the structure of Inception [21], we introduced the Feature Enhancement Module (FEM) to fuse the features generated by convolution layers with different kernel sizes, so as to enhance the representation capability of the network to the small objects. The Feature Enhancement Module (FEM) is shown in Figure 5.

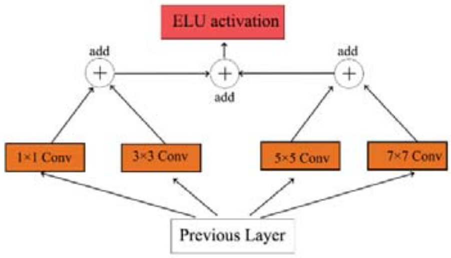

Figure 5. Structural diagram of Feature Enhancement Module (FEM).

C. *Constructing architecture*

As shown in Figure 3, we construct our architecture based on the SSD framework, and then design proper detection layers and default boxes settings, which are essential for high detection accuracy.

- Lightweight backbone network. The backbone network based on SSD. we reserve the convolutional layers from 'conv1_1' to 'conv6' and remove other layers because the too deeper convolutional layers behind is helpless for small objects detection but the computation is large. In order to obtain real-time detection effect, 'conv1_1' to 'conv6' were changed into spatial separable convolution. We select conv3_3, conv4_3, conv5_3 and conv6 as the detection convolution layers.
- Feature Enhancement Module (FEM). In order to enhance the representation capability of the network to the small objects, we introduced the Feature Enhancement Module (FEM). The detection layer of Conv4_3 and Conv6 is tailed by FEM.

- Default boxes parameters. To reduce the rate of missed detection, the scales in each detection layer should match as much as possible the scales of the objects to be detected. Because of the small objects in COVID-19-Mask dataset, we set up a series of small-scale default boxes. The parameters are shown in table 2.

Table 2. Default boxes parameters

Detection layers	Scales
conv3_3	0.02
conv4_3	0.1
conv5_3	0.2
conv6	0.4

D. Detection flow diagram

The task includes two modules, the training module and the detection module. In the training section, the COVID-19-Mask dataset was used to train the model to obtain a mask detector. In the detection stage, images are obtained in real-time from the surveillance video, and then use the trained detector to determine whether the shoppers in the pictures are wearing masks. A warning will be issued if a shopper is detected not wearing a mask. The overall flow diagram is shown in Figure 6.

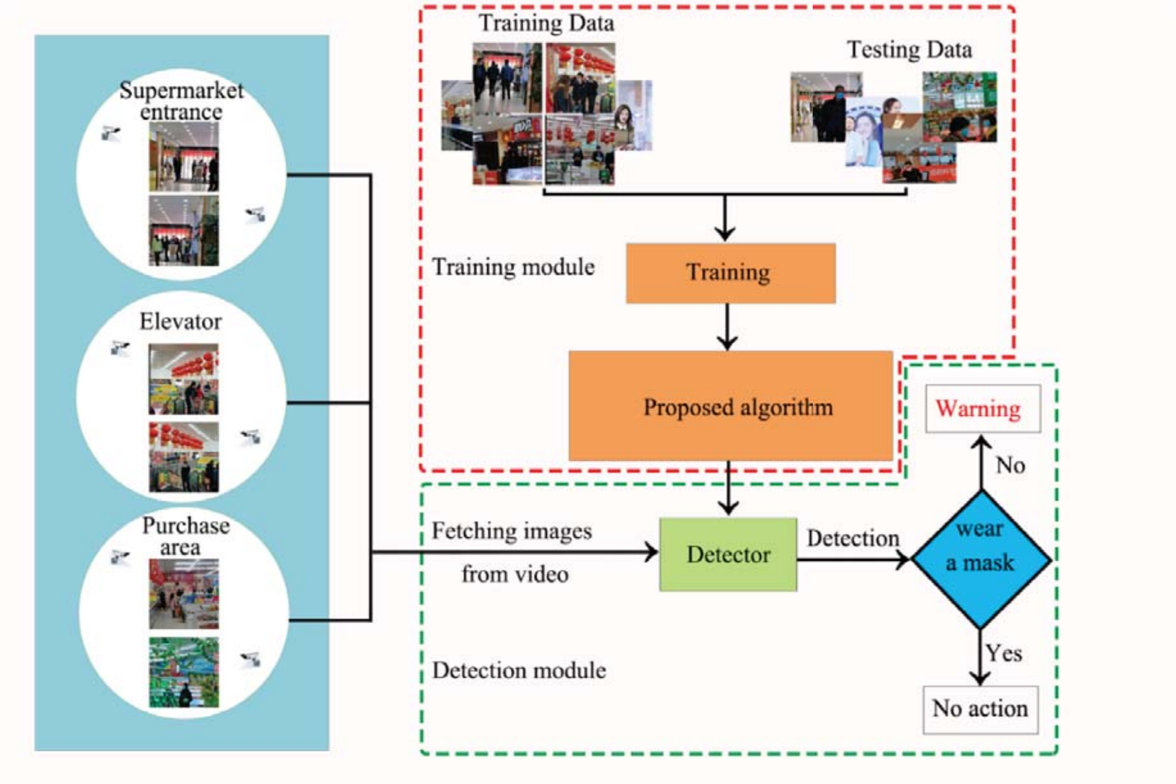

Figure 6. Flow diagram of face mask detection.

V. EXPERIMENTS

This section reports several experimental results based on the COVID-19-Mask data set. We first evaluate existing architectures [22], [17] and [23] on COVID-19-Mask dataset. Moreover, we evaluate proposed models' performance and analyze the effectiveness of backbone and FEM.

All experiments are conducted using NVIDIA GTX1080TI video card,11GB memory, CPU E5-2620.We trained and tested built on the deep learning framework, Keras. We adopt Adaptive Moment Estimation (Adam) as the optimization function to train our model and the learning rate starts with 0.001 and decrease to 0.0001 after 20k iterations. After 100k iterations, we stop training and the last model snapshot is used to evaluate the performance of object detection on the test set.

First, we conducted a series of experiments on other popular algorithms. The comparison between our experimental results and other models is shown in Table 3. As can be seen from Table 3, compared with other algorithms, the proposed method has excellent detection accuracy and real-time performance on the COVID-19-Mask dataset. Experimental results show that proposed method can achieve 90.9% accuracy which is 18% and 15.7% higher than SSD and YoloV3 respectively on COVID-19-Mask dataset. In terms of detection speed, the average detection time for proposed method processing a 512 × 512 pixels images are 0.12s, which is higher than SSD but lower than YoloV3.

Table3. Comparison of detection results on COVID-19-Mask dataset for different object detection algorithms. Run time means the average running time for detection small objects in an image of 512×512 pixels.

Method	mAP(%)	Wear a mask	Don't wear a mask	Run time(s)
Faster R-CNN [22]	74.4	70.5	78.3	0.21
SSD [17]	72.9	68.7	77.1	0.20
YoloV3 [23]	73.8	69.4	78.2	0.08
Our proposed	90.9	88.7	93.1	0.12

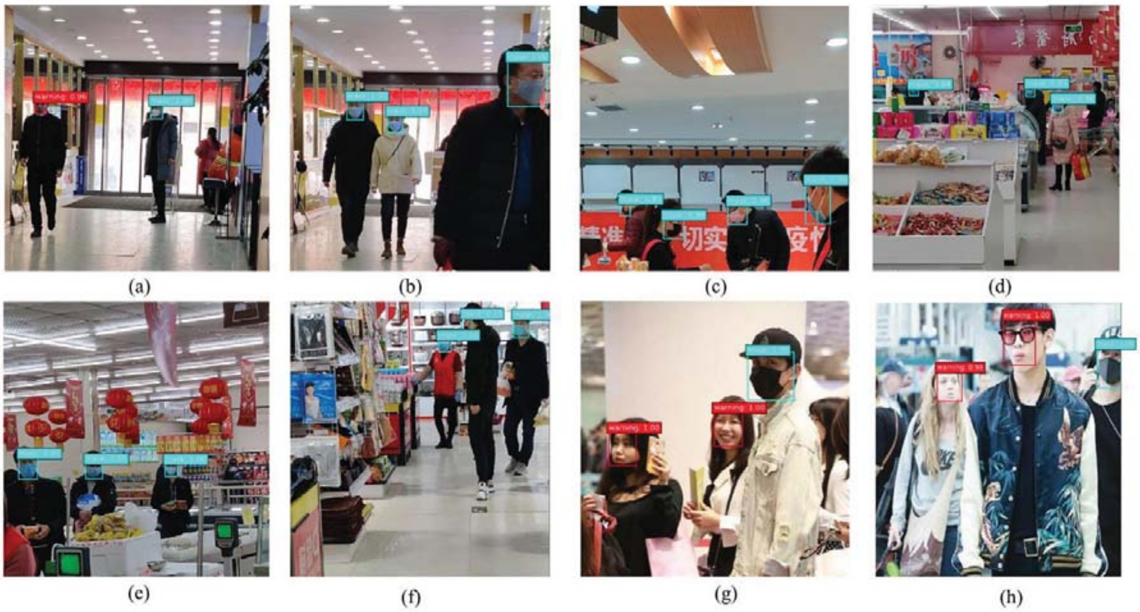

Figure 7. Detection results of proposed algorithm.

To evaluate the effectiveness of the proposed approach, we conduct the ablation experiments on the COVID-19-Mask datasets. The ablation experiments include four experiments and experimental results are shown in Table 4. Experiment 1 is the original SSD, and experiment 2 is the modified SSD with only 4 detection layers, which does not use spatial separable convolution and Feature Enhancement Module (FEM). In experiment 3, on the basis of experiment 2, the traditional convolution is changed into spatial separable convolution. Experiment 4 adds FFM on the basis of experiment 3.

Table4. Ablation experiments

Exp.np.	mAP(%)	Run time(s)
1	72.9	0.20
2	89.3	0.15
3	87.5	0.10
4	90.9	0.12

As can be seen from Table 4, compared with the original SSD, using the modified SSD with only 4 detection layers for training, the mAP of the algorithm improved by 16.4% to 89.3%. This is because compared with the original SSD, the default boxes scales of modified SSD design are more suitable for COVID-19-Mask datasets. Subsequently, the traditional convolution is changed to spatial separable convolution for training, the mAP is reduced by 1.8% but the running time is shortened. This shows that the spatial separable convolution inevitably loses a bit of information while reducing the parameters. Finally, we add FFM for training, and we can see that the mAP reached 90.9% and run time reached 0.12s. Experimental results show that the proposed method does help the real-time detection of face masks. Figure 7 shows the detection results of proposed method on COVID-19-Mask dataset.

VI. CONCLUSION

In this paper, we proposed a modified SSD method to detect whether shoppers are wearing masks in the supermarket. In order to detect whether shoppers are wearing masks, we created the COVID-19-Mask dataset, which can provide data for future studies. At the same time, in order to accurately detect masks in real time, we proposed a lightweight backbone network and Feature Enhancement Module (FEM), which improves the overall detection effect of the algorithm. We conducted a wide range of experiments and provided a comprehensive analysis of the performance of our model on the task of face mask detection. Experimental results show that the proposed method can effectively detect whether shoppers wear masks and can be applied to practice.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No.61562084.

REFERENCES

[1] Hui D S, I Azhar E, Madani T A, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—The latest 2019 novel coronavirus outbreak in Wuhan, China. International Journal of Infectious Diseases, 2020, 91: 264-266.

[2] Liu Y, Sun P, Highsmith M R, et al. Performance comparison of deep learning techniques for recognizing birds in aerial images//2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE, 2018: 317-324.

[3] Chen X, Kundu K, Zhu Y, et al. 3d object proposals for accurate object class detection//Advances in Neural Information Processing Systems. 2015: 424-432.

[4] [4] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector//European conference on computer vision. Springer, Cham, 2016: 21-37.

[5] Yang F, Choi W, Lin Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2016.

[6] Bell S, Lawrence Zitnick C, Bala K, et al. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2874-2883.

[7] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 5325-5334.

[8] Cao G, Xie X, Yang W, et al. Feature-fused SSD: Fast detection for small objects//Ninth International Conference on Graphic and Image Processing (ICGIP 2017). International Society for Optics and Photonics, 2018, 10615: 106151E.

[9] Li Y, Sun X, Wang H, et al. Automatic target detection in high-resolution remote sensing images using a contour-based spatial model. IEEE Geoscience and Remote Sensing Letters, 2012, 9(5): 886-890.

[10] Takacs G, Chandrasekhar V, Tsai S, et al. Unified real-time tracking and recognition with rotation-invariant fast features//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2010: 934-941.

[11] Zou Z, Shi Z. Ship detection in spaceborne optical image with SVD networks. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(10): 5832-5845.

[12] Cheng G, Zhou P, Han J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12): 7405-7415.

[13] Ouyang W, Wang X, Zeng X, et al. Deepid-net: Deformable deep convolutional neural networks for object detection//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2403-2412.

[14] Sermanet P, LeCun Y. Traffic sign recognition with multi-scale convolutional networks//The 2011 International Joint Conference on Neural Networks. IEEE, 2011: 2809-2813.

[15] Zhu Z, Liang D, Zhang S, et al. Traffic-sign detection and classification in the wild//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2110-2118.

[16] Jin J, Fu K, Zhang C. Traffic sign recognition with hinge loss trained convolutional neural networks. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(5): 1991-2000.

[17] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector//European conference on computer vision. Springer, Cham, 2016: 21-37.

[18] Zhang S, He G, Chen H B, et al. Scale Adaptive Proposal Network for Object Detection in Remote Sensing Images. IEEE Geoscience and Remote Sensing Letters, 2019, 16(6):864-868.

[19] Howard A G, Zhu M, Chen B, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. 2017.

[20] Freeman I, Roese-Koerner L, Kummert A. Effnet: An Efficient Structure for Convolutional Neural Networks// IEEE International Conference on Image Processing. IEEE, 2018.

[21] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning//Thirty-first AAAI conference on artificial intelligence. 2017.

[22] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7-12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 91-99.

[23] Redmon, Joseph, Farhadi, Ali. [IEEE 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - Honolulu, HI (2017.7.21-2017.7.26)] 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) - YOLO9000: Better, Faster, Stronger// :6517-6525.