ISSN: 2664-9640

Mask wearing detection in public places algorithm based on improved SSD

Sicheng Li[1, 2], Shunyong Zhou[1, 2, *] and Yalan Zeng[1, 2]

[1] Sichuan University of Science & Engineering, Zigong 643000, China;

[2] Sichuan Provincial Key Laboratory of Artificial Intelligence, Sichuan.

* Corresponding Author

Abstract

How to quickly detect the wearing of masks in public is a hot issue and one of the most useful methods to block coronavirus spread. Aiming at the problems of complex structure of mask detection algorithm, different detection target scales, insufficient feature extraction and difficult training, a mask wearing detection model based on improved SSD (single shot multibox detector) algorithm is proposed. Compared with the ordinary SSD network architecture, this model will be smaller and easier to be applied in the terminal system with low configuration. This algorithm improves the detection accuracy of masks by predicting on a variety of scale feature maps, using default frames with different aspect ratios, and Non-maximum suppression at the end of the classifier. Experimental results show that the average accuracy of the algorithm reaches 91.75%, and it still has a good effect on small target detection, which basically meets the actual needs of use.

Keywords

SSD; mask wearing detection; target detection; multi-scale learning.

1. Overview

The COVID-19 has spread to the world. This virus is highly infectious and has a long incubation period. It will spread through droplets and air. It is estimated that one person will die from the new crown every 8 seconds. The health and social order have caused a certain impact. Although the number of people vaccinated in my country is increasing, COVID-19 and its variant strains still pose a threat to people's health, and we need to pay attention to the protection of this virus. Leung et al. [1] proposed that wearing a mask can effectively suppress the spread of COVID-19. At present, in most public places, such as supermarkets, hospitals, etc., manual reminders to wear masks are required. This method greatly wastes human resources. With the development of deep learning, the use of artificial intelligence detection technology can quickly determine whether people have the conditions to enter and exit the occasion.

Therefore, mask wearing detection technology is one of the key prevention methods for the supervision of disease control departments, and its related algorithm research is particularly important. At present, some scholars have researched mask wearing detection algorithms. Literature [2] uses traditional target detection algorithms such as HSV+HOG to detect mask wearing behavior. Although a high detection accuracy is achieved, traditional detection methods have high time complexity and poor robustness; literature [3] combines deep learning method designed a mask detection algorithm. Literature [4] improved the YOLO algorithm and designed a new mask detection algorithm. Both algorithms have achieved good detection results, but these two algorithms only work on masks. Whether it is worn or not has been tested, and the incorrect wearing of the mask has not been fully considered, and the demand for mask

ISSN: 2664-9640

detection under real conditions cannot be fully met. Mask detection requires accurate recognition of human faces, and it is necessary to judge whether the mask is worn and whether it is worn correctly. In the case of wearing a mask, most of the features of the face are covered, which brings a certain amount of interference to the detection of wearing a mask. Mask detection has higher requirements for the algorithm's detailed feature learning and processing capabilities. In addition, there are fewer public data sets related to mask wearing and the data is not complete, and a new data set needs to be re-established.

Target detection technology is a key technology to realize the task of mask wearing detection, and it is also a research hotspot in the direction of computer vision, which is widely used in various fields [5-7]. In recent years, fast R-CNN [8], Faster R-CNN [9], YOLO [10-12], SSD [13] and other excellent deep learning target detection algorithms have been born. Among them, the SSD algorithm has been studied and applied by many scholars because of its excellent performance in detection speed and accuracy. For example, literature [14] detects urban outdoor advertising panels based on this algorithm; literature [15] detects foreign objects on the surface of coal mine belt conveyors based on this algorithm; literature [16] detects wild panda videos based on this algorithm; Literature [17] detects items in subway security images based on this algorithm. These works have made corresponding improvements to the SSD algorithm for specific scenarios, and have achieved very good detection results. SSD is based on the VGG-16 network, uses a pyramid feature structure, makes predictions on different feature map mappings, and uses default boxes with different scales of aspect ratios in the feature map, which solves the problem of feature loss of small targets and improves small targets the detection accuracy. Based on the SSD model, this paper realizes the detection task of masks.

2. SSD detection algorithm

2.1. SSD detection algorithm model

SSD is one of the classic algorithms in the field of target detection, and it is still the mainstream target detection algorithm. The network structure of SSD is shown in Figure 1. The SSD network is based on a fully convolutional network structure. It replaces the fully connected layer of the basic network VGG16[18]with a convolutional layer and adds several auxiliary convolutional layers at the end of the VGG16 network to gradually reduce the size of the feature map. Used to extract feature maps of different scales. SSD uses the output of conv4_3, fc6, fc7, conv6_2, conv7_2, conv8_2, and conv9_2 layers as feature maps of different scales for detection, and the corresponding feature maps are 38*38, 19*19, 10*10, 5*5, 3*3 and 1*1, respectively, to deal with objects of different sizes, which improves the prediction speed and accuracy.

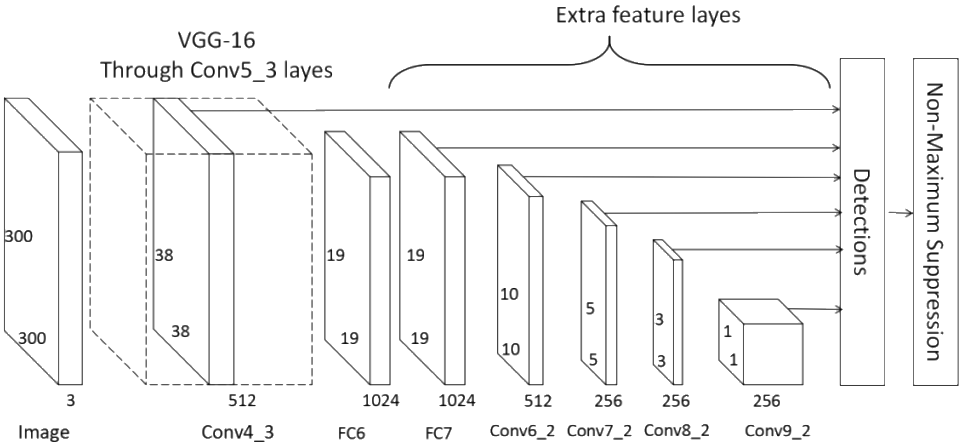

Figure 1: SSD algorithm network structure

ISSN: 2664-9640

2.2. Feature layer candidate frame mechanism

SSD adopts a multi-scale feature map method, and area selection boxes of different sizes and aspect ratios are set on different scale feature maps. The definition of the area candidate box is calculated as follows:

$$S_k = S_{min} + \frac{S_{max} - S_{min}}{m - 1}(k - 1), k \in [1, m] \quad (1)$$

Among them: m is the number of characteristic layers; S_{min}=0.2 is the minimum characteristic layer scale; S_{max}=0.9 is the highest characteristic layer scale; the middle characteristic layer scales are uniformly distributed.

Area candidate boxes have different aspect ratios $a_r = \in \left\{1,2,3,\frac{1}{2},\frac{1}{3}\right\}$.The width and height of the area candidate frame are $\omega_k^a = S_k\sqrt{a_r}$, $h_k^a = S_k/\sqrt{a_r}$.At the same time, add a scale $S_k' = \sqrt{S_k S_{k+1}}$ to the area candidate frame with an aspect ratio of 1.The center coordinates of each area candidate box are $\left(\frac{i+0.5}{\omega_{fk}}, \frac{j+0.5}{h_{fk}}\right)$ ω_{fk} is the width of the k-th feature map, and h_{fk} is the height of the k-th feature map,$i \in [0, \omega_{fk}]$,$j \in [0, h_{fk}]$.

Therefore, for each feature layer grid, there are a total of 6 types of default frames. This default frame has different scales in different feature layers, and different aspect ratios in the same feature layer, basically, it can cover targets of various shapes and sizes in the input image. For the first layer of features, given S_k and S_{k+1}, the interval between the remaining five layers is step. The step specifies a calculation intermediate number corresponding to the original image from the feature. min_ratio and max _ratio are the size ratios of the original image. It can be understood that (for a 300*300 SSD) the smallest default frame area is 0.2*300*300, and the maximum is 0.9*300*300. The default frame parameters on each feature layer are shown in Table 1.

Table 1: Default frame parameters on each feature layer

Feature layers	w*h	k	1:1	1:2	2:1	1:3	3:1
conv4_3	38*38	4	30*30	21*42	42*21	Null	Null
fc7	19*19	6	60*60	42*84	84*42	35*105	105*35
conv6_2	10*10	6	111*111	79*158	158*79	65*195	195*65
conv7_2	5*5	6	162*162	116*232	232*116	95*285	285*95
conv8_2	3*3	4	213*213	154*308	308*154	Null	Null
conv9_2	1*1	4	264*264	191*382	382*191	Null	Null

2.3. Loss function

The loss function of SSD has the same principle as the loss function in Faster R-CNN, and it consists of classification and regression. During the SSD training process, the position and target category are regressed. The target loss function is the sum of the positioning loss (loc) and the confidence loss (conf), and its expression is as follows:

$$L(x, c, l, g) = \frac{1}{N}(L_{conf}(x, c) + \alpha L_{loc}(x, l, g)) \quad (2)$$

Among them, N is the number of candidate frames that match the labeled frame; if N=0, set Loss=0; x is the matching result of the regional candidate frame and the real frame of different categories, if it matches x=1, otherwise x=0; c is The confidence of the predicted object category;

ISSN: 2664-9640

l is the position offset information of the prediction frame; g is the offset between the real frame and the regional candidate frame; α is the position loss weight parameter is usually set to 1.

The training objective of SSD training is derived from the objective function of MultiBox, but the article expands it so that it can handle multiple target categories. The specific process is to let each candidate frame calculate the similarity with the real frame through the Jaccard coefficient, and only those with a threshold greater than 0.5 can be included in the candidate list; assuming that N frames with a matching degree higher than 50% are selected, let i Represents the i-th default box, j represents the j-th real box, and p represents the p-th class.Then x_{ij}^{p} represents the Jaccard coefficient that the i-th candidate box matches the j-th labeled box of category p. If it does not match, x_{ij}^{p}=0. The total objective loss function is the weighted sum of localization loss (loc) and confidence loss (conf).

3. Improved SSD algorithm

3.1. SSD area candidate frame settings

The conventional SSD algorithm is based on VGG16, after which fc6, fc7 and Extra Feature Layers are added. Extra Feature Layers consists of four blocks: conv6_2, conv7_2, conv8_2, and conv9_2. VGG16 is developed on the basis of AlexNet, its structure is straight and consists of five groups of convolutional layers and three fully connected layers. When the mask detection task is implemented according to the conventional SSD algorithm, the model will consist of nine sets of convolutional layers and three fully connected layers, and the model is relatively large. If you want to apply it to the terminal system of the actual production and life scene, you need to reduce the model and then embed it into the actual scene.

Make the following changes to the SSD: ①Keep only the prediction layers conv4_3, fc7, conv6_2, conv7_2, conv8_2, and delete all subsequent convolutional layers. ②conv4_3 sets 4 area candidate frames, fc7, conv6_2, conv7_2, conv8_2 sets 5 area candidate frames respectively.

The size of the input layer of the SSD algorithm before the improvement is 300*300. This article has been reduced to 260*260. The fc6 and conv9_2 convolutional layers are removed, and the remaining convolutional layers are reduced accordingly. See Table 2 for the improved SSD prediction layer area candidate frames.

Table 2: Improved SSD candidate frame for each prediction layer area

Feature layers	Feature layer size	anchor min_size	anchor max_size	aspect_ratio
conv4_3	30*30	0.04	0.056	1,0.62,0.42
fc7	17*17	0.08	0.11	1,0.62,0.42
conv6_2	9*9	0.16	0.22	1,0.62,0.42
conv7_2	5*5	0.32	0.45	1,0.62,0.42
conv8_2	3*3	0.64	0.72	1,0.62,0.42

The improved model backbone network has only 8 convolutional layers, plus the positioning and classification layers, and there are only 24 layers in total, so the model is very small, with only 1.015 million parameters. The actual convolution process of the improved SSD network structure is shown in Figure 2.

ISSN: 2664-9640

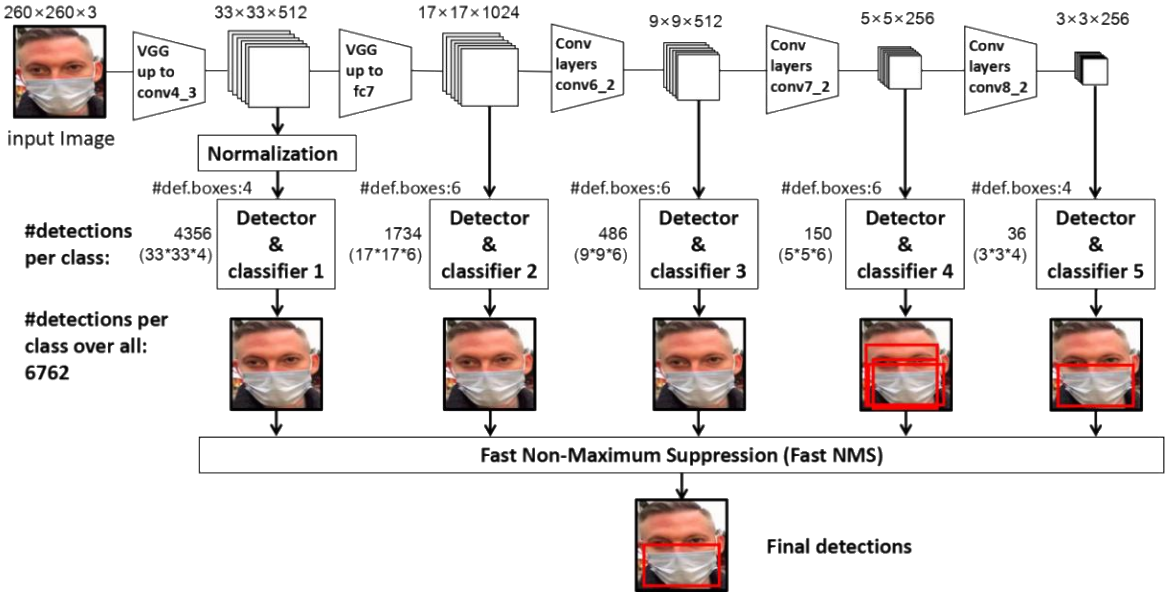

Figure 2: Improved SSD algorithm network structure

3.2. Rejection loss

Aiming at the problem of SSD's poor detection effect on overlapping targets, a repulsion loss was adding on the original SSD loss function [19], the final SSD loss function is as follows:

$$\mathrm{L} = \mathrm{L}(\mathrm{x}, \mathrm{c}, \mathrm{l}, \mathrm{g}) + \gamma \mathrm{L}_{\mathrm{RepGT}} \tag{3}$$

Let $P_+ = \{P\}$ denote a collection of area candidate frames that match at least one real frame ($I_O U$>0.5), and $G_+ = \{G\}$ denote a collection of all real frames. For a given candidate frame $P \in P_+$, assign a real frame with the largest $I_O U$ value as its designated target, as follows:

$$G_{attr}^{P} = argmax I_O U(G, P)\ , G \in G_+ \tag{4}$$

Because the rejection loss is to make the regional candidate frame repel the adjacent real frame except its designated target, for $P \in P_+$, its repelling target is the real target with the largest $I_O U$ value besides its designated target.

$$G_{ReP}^{P} = argmax I_O U(G, P)\ , G \in G_+ \{G_{attr}^{P}\} \tag{5}$$

Let B^P be the predicted frame returned from the candidate frame P. The $I_O G$ between B^P and G_{ReP}^{P} is calculated as follows:

$$I_O G(B^P, G_{ReP}^{P}) = \frac{area(B^P \cap G_{ReP}^{P})}{area(G_{ReP}^{P})} \tag{6}$$

Then the rejection loss is calculated as follows:

$$\mathrm{L}_{\mathrm{RepGT}} = \frac{I_O G(B^P, G_{ReP}^{P})}{|P_+|} \tag{7}$$

ISSN: 2664-9640

4. Experiment and result analysis

4.1. Experimental data set

The experiment involves two data sets, namely the data set with and without masks respectively. The two parts are mainly derived from the data sets publicly available on the Internet. Use the Labelme tool to mark the target and generate the corresponding json file with the category labels as 'have_mask' and 'no_mask'.

(a) no_mask

(b)have_mask

Figure 3: Labeled data set

A total of 2402 pictures of the mask wearing detection data set, of which 1796 are not wearing a mask and 606 are wearing a mask.

Table 3: Number of labels by category

Category	Number
have_mask	606
no_mask	1796
Total category	2402

Aiming at the problem that the amount of data in the data set is not high enough, the experiment uses random data enhancement methods such as cropping, rotation, and color transformation

ISSN: 2664-9640

to expand the training data. While increasing the amount of data and the number of samples of incorrectly wearing masks, it can also alleviate the overfitting phenomenon during the training process and improve the performance and robustness of the algorithm. The information of the mask wearing data set after data enhancement is shown in Table 4.

Table 4: Mask wearing dataset

Type of data	Original quantity	Current quantity
Training	1680	2450
Test	480	700
Validation	240	350
Total type of data	2400	3500

4.2. Experimental environment

The experiment in this article is based on the Pytorch framework, the programming language is python3.6, the operating system is Windows10, the integrated development environment is PyCharm, the network input size is 260*260, and the initial learning rate is set to 0.001.

4.3. Experimental results and analysis

This article mainly uses Mean Average Precision (mAP) and Average Precision (AP) to evaluate the model. Among them, mAP and AP consider both precision (Precision, P) and recall (Recall, R), as shown below:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

Average Precision (AP) is used to evaluate the performance of the model on the test set, as shown in equation (10). Multi-category detection results are usually measured by Mean Average Precision (mAP). The mAP in this paper is shown in formula (11):

$$AP = \int P(R)dR \tag{10}$$

$$mAP = \frac{AP_{have\_mask} + AP_{no\_mask}}{2} \tag{11}$$

The P-R curve of this experiment is shown in Figure 4.

The picture detection result is shown in Figure 5. Among them, 'no_mask' represents not wearing a mask, and 'have_mask' represents wearing a mask. The value on the target box represents the confidence level of each category label. The test results show that the situation of wearing a mask is well distinguished during the test. When the number of targets in the picture is small, the confidence level can reach more than 90%. For the detection of dense crowds, most of the targets can be identified, but the confidence level is relatively low.

ISSN: 2664-9640

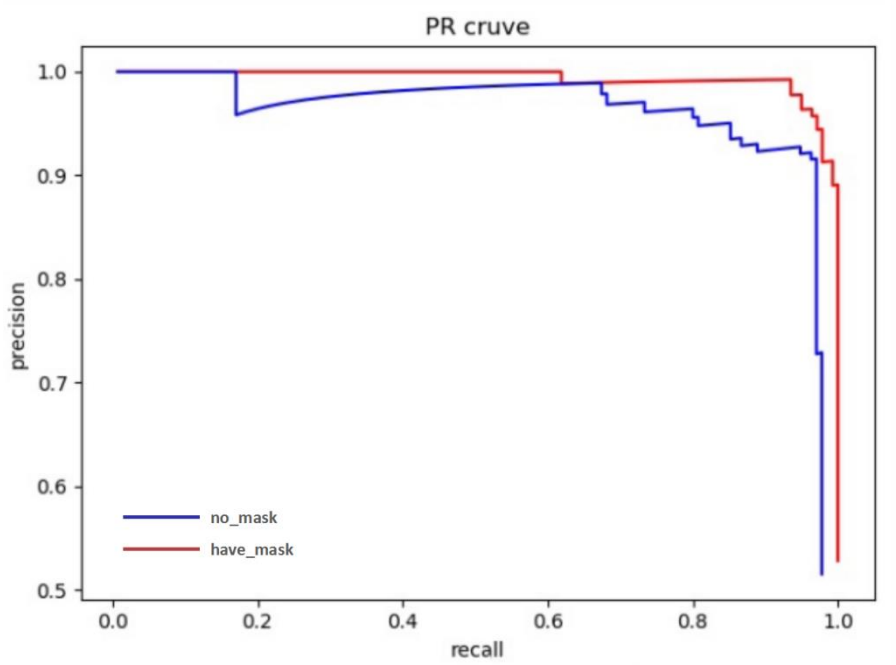

Figure 4: P-R curve of various targets

Figure 5: Test results

The real-time video detection results are shown in Figure 6. This experiment tried to wear the mask correctly, put the mask on the chin, take the mask off to the throat, and did not wear the mask. Experimental results prove that the detection algorithm still has a good detection effect when wearing a mask incorrectly.

In terms of the detection effect of the comprehensive experiment, the improved SSD network model has a high degree of confidence for faces and masks. At the same time, it also has a faster detection speed. It has more advantages in low-end terminals and has a more lightweight architecture. Can meet the needs of actual use.

ISSN: 2664-9640

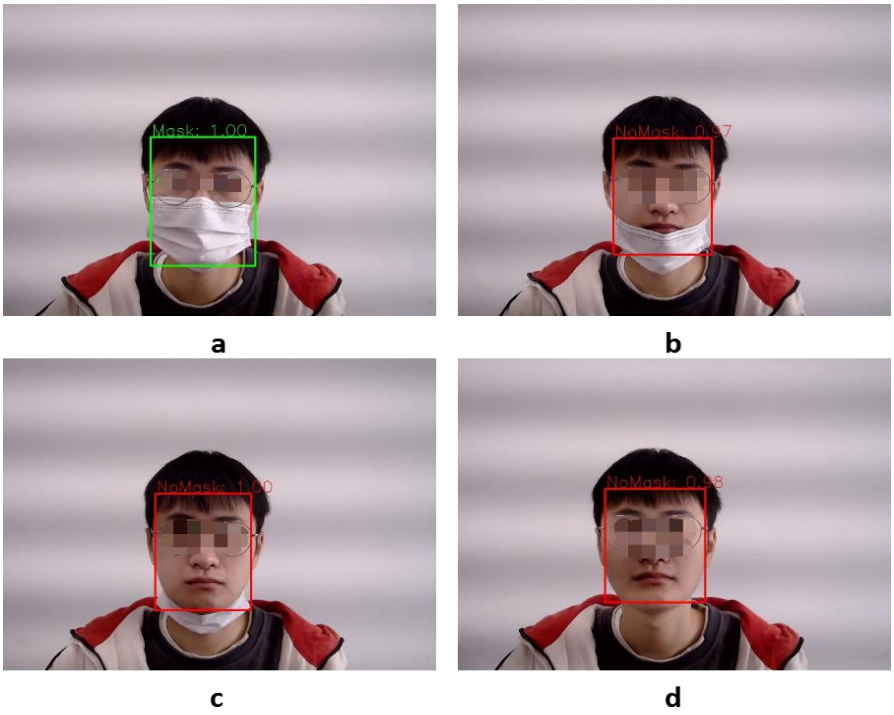

Figure 6: The real-time video detection results

5. Conclusion

In summary, this article uses an improved SSD algorithm to achieve real-time detection of masks. The detection ability of small targets is enhanced through multi-scale feature maps, and the confidence level reaches 91.75%, which meets the needs of daily use. However, there are still certain shortcomings. In the face of overly complex scenes, it is easy to miss the detection. The main problems are that the low-level feature map has fewer default frames and insufficient feature extraction for small targets. Later improvements will be made based on this problem, and the model will be promoted after verification.

References

[1] Liu Y, Gayle A A, Annelies W S, et al. The reproductive number of COVID-19 is higher compared to SARS coronavirus [R]. Journal of Travel Medicine, 2020.

[2] HE Yumin, WANG Chaohui, GUO siyu, et al. Research on face mask detection algorithm based on HSV+hog feature and SVM [J/OL]. Journal of Measurement Science and Instrumentation. http://kns.cnki.net/kcms/detail/14. 1357. TH. 20210315. 0832. 002. html.

[3] Jiang M, Fan X, Yan H. Retina facemask: A face mask detector [J]. arXiv preprint arXiv: 2005. 03950, 2020, 2.

[4] Abbasi S, Abdi H, Ahmadi A. A face-mask detection approach based on YOLO applied for a new collected dataset [C]//2021 26th International Computer Conference,Computer Society of Iran(CSICC). IEEE, 2021: 1-6.

[5] LI Wenbin, HE ran. Aircraft target detection in remote sensing image based on depth neural network [J]. Computer Engineering, 2020, 46(7): 268-276.

[6] Liu G, Nouaze J C, Touko Mbouembe P L, et al. YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3 [J]. Sensors, 2020, 20(7): 2145. DOI: 10.3390/s20072145.

[7] CHENG Shuhong, ZHOU Bin. Recognition of characters in aluminum wheel back cavity based on improved Convolution Neural Network [J]. Computer Engineering, 2019, 45(5): 182-186.

[8] GIRSHICK R. Fast R-CNN [C]//Proceedings of IEEE International Conference on Computer Vision. Santiago, Chile: IEEE Press, 2015: 1440-1448.

ISSN: 2664-9640

[9] 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.

[10] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection [C].//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE Press, 2016: 779-788.

[11] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE Press, 2017: 6517-6525.

[12] REDMON J, FARHADI A. YOLOv3: an incremental improvement [EB/OL]. [2019-11-02]. https://arxiv.org/abs/1804. 02767.

[13] LIU W, ANGUELOVD, ERHAN D, et al. SSD: single shot multibox detector [C]//Proceedings of the 2016 European Conference on Computer Vision. Amsterdam: EC-CV, 2016: 21-37.

[14] Morera Á, Sánchez Á, Moreno A B, et al. SSD vs. YOLO for detection of outdoor urban advertising panels under multiple variabilities [J]. Sensors, 2020, 20(16): 4587. DOI: 10. 3390/s20164587.

[15] Wang Y, Wang Y, Dang L. Video detection of foreign objects on the surface of belt conveyor underground coal mine based on improved SSD [J]. Journal of Ambient Intelligence and Humanized Computing, 2020: 1-10.

[16] Fang J, Yang H, Chen P, et al. A detection algorithm of giant panda in wild video image based on wavelet-SSD network [C]//2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2020: 3655-3660.

[17] ZHANG Zhen, LI Mengzhou, LI Haofang, MA Junqiang. Improved SSD algorithm and its application in subway security detection [J]. Computer Engineering, 2021, 47(7): 314-320.

[18] Simonyan, K. Zisserman, A.: Very deep convolutional networks for largescale image recognition. International Conference on Learning Representations [C]. USA: IEEE, 2015. 714-723.

[19] WANG X, XIAO T, JIANG Y, et al. Repulsion Loss: Detecting Pedestrians in a Crowd[J].arXiv preprint arXiv: 1711. 07752, 2017.