Face Mask Wearing Detection Based on YOLOv5

Yunshan Xie

School of Computer Science and Engineering

Xi'an Technological University

Xi'an, 710021, Shaanxi, China

E-mail: 2398576240@qq.com

Zhiyi Hu

Engineering Design Institute

Army Research Loboratory

Beijing, 100042, China

E-mail: 763757335@qq.com

Jun Yu

School of Computer Science and Engineering

Xi'an Technological University

Xi'an, 710021, Shaanxi, China

E-mail: yujun@xatu.edu.cn

Abstract—**In recent years, COVID-19 has swept the world, and people in crowded public places are usually large. In order to reduce the risk of virus transmission, stop the spread of the epidemic and reduce cross-infection, wearing masks correctly has become an important measure to prevent the virus. Aiming at the time-consuming and laborious situation of wearing masks manually, this paper proposes a mask wearing detection method based on yolov5. The input layer is mainly used for mosaic data enhancement, that is, adaptive anchor box and adaptive image scaling technology; Yolov5 in backbone mainly adopts focus and CSP (cross stage partial) structure; The neck layer adopts spp (spatial pyramid pooling) module and FPN (feature pyramid networks) + pan (pixel aggregation network) structure; The output mainly adopts ciou for the bounding box loss function_ Loss is the average index of NMS (non maximum suppression). This method uses 8000 preprocessed images as the data set and trains 200 epochs to get the final model. The algorithm visually displays the training and test results through tensor board, and inputs the pictures captured by the camera into the model to detect whether the face wears a mask. The accuracy, recall jond mean accuracy (map) of the algorithm on the test set are 94.8%, 89.0% and 93.5% respectively, which are higher than the detection results of yolov3 and yolov4 algorithms.**

Keywords-Mask Detection; Yolov5; CIOU Loss Function

I. INTRODUCTION

By April, 2022, more than 500million cases of COVID-19 had been confirmed in the world, and more than 6.2 million cases had died. Under the normal epidemic situation, wearing masks in public places is an important means of effective epidemic prevention. Therefore, mask wearing detection has become a core work of epidemic prevention. At present, the main detection method is manual detection, which is not only time-consuming and laborious, but also increases the risk of virus infection. Therefore, it will be of great practical significance to build a mask detection and monitoring system to realize the automation and intelligence of epidemic prevention and control.

At present, some scholars have done research on mask wearing detection. Ren Yu [1] and others proposed a fast r-cnn mask wearing detection algorithm based on deep learning,

which accelerated the convergence of the model by migrating the weight of the pre training model on the Imagenet data set, and the accuracy of the final test set reached 89.41%. Dong Yanhua [2] and others proposed a SSD network based on residual structure. By adding residual structure before the positioning and classification of SSD network, the feature extraction network and classification and positioning layer are separated, which effectively solves the dual task of learning local information and high-level information at the same time. Finally, the average detection accuracy reaches 92.3%. Cao chengshuo [3] and others proposed a YOLO mask algorithm, which is based on YOLOv3, introduces the attention mechanism into the feature extraction network, and uses the feature pyramid and path aggregation strategy for feature fusion. The average accuracy of the algorithm is 93.33%. Based on YOLOv4, Cheng Tinghao [4] and others redesigned the anchor frame size with K-means clustering algorithm and improved the network structure. The recall rate and accuracy rate on the test set reached 88.20% and 92.30% respectively.

This paper adopts the YOLOv5 target detection network model. Compared with YOLOv3 and YOLOv4, the innovation of YOLOv5 is that it uses the focus structure in the backbone, which reduces flops and improves the training speed. Compared with the ordinary convolution operation in YOLOv4neck structure, YOLOv5 uses CSP_X structure, which strengthens the fusion ability of network features and improves the detection accuracy. Overall, the performance of YOLOv5 is better than that of YOLOv4 and YOLOv3.

II. YOLOV5 NETWORK MODEL
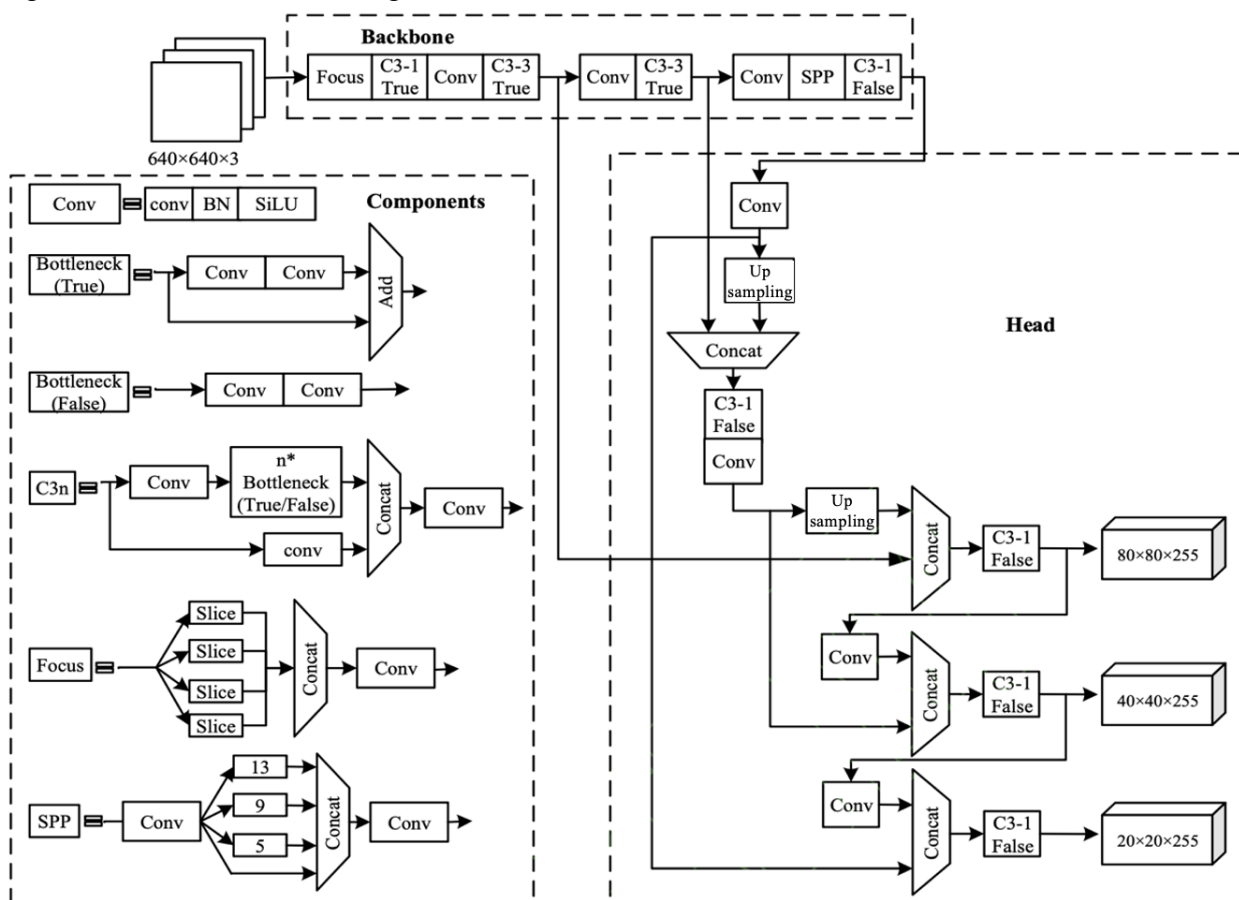
A. YOLOv5 algorithm

Figure 1.   YOLOv5s network structure diagram

Yolov5 algorithm is developed on the basis of yolov4 and yolov3. Compared with yolov4, the architecture of yolov5 is about 90% smaller [9]. In terms of accuracy, the performance of yolov5 is better than that of yolov3 and yolov4 algorithms in the current market.YOLOv5 is a single-stage target detection algorithm. YOLOv5 [10] series can be divided into YOLOv5s [11], YOLOv5m, YOLOv5l [12] and YOLOv5x [13]. Among them, YOLOv5s has the smallest network. Although the accuracy is slightly worse than the other three, it has the fastest speed. Because of the speed requirement of mask detection algorithm, this paper selects YOLOv5s network model.

The network structure diagram of YOLOv5s is shown in Figure 1. It can be seen that the model is mainly divided into four parts: input, backbone, neck and prediction. Mosaic data enhancement is used at the input end to improve the detection effect of small targets [14]. In addition, using the adaptive anchor frame calculation method [15], different data sets will have anchor frames with initial length and width, which can get a larger intersection and union ratio, and greatly improve the efficiency of training and prediction [16]; The backbone part is a convolutional neural network that aggregates and forms image features on different image granularity [17], mainly including Focus structure and CSP structure. The neck part is between the backbone and prediction. It is a network layer of a series of mixed and combined image features, which transmits the image features to the prediction part [18]. The prediction part is the final detection part, which mainly predicts the image features, generates the boundary box and predicts the type of target [19].

B. Input

The input terminal includes mosaic data enhancement [20], image size processing and adaptive anchor box calculation. Masaic data enhancement adopts 4 pictures to be spliced in the way of random scaling, random clipping and random arrangement. Using mosaic data enhancement can greatly enrich the data set and add many small targets. YOLO algorithms need to change the input image size into a fixed size. The standard size of the image in this paper is $640 \times$ six hundred and forty $\times 3$。 It is necessary to set the initial knowledge anchor box before network training. The initial knowledge anchor box of YOLOv5 is [10, 13, 16, 30, 33, 23], [30, 61, 62, 45, 59119], [116, 90156198373326].

Adaptive image scaling refers to the unified scaling of images with different lengths and widths to a standard size, and then used as a data set for detection and processing. After the picture is scaled, if there are many black edges filled on both sides of the picture, there will be information redundancy and affect the speed. In this regard, yolov5 has been improved and the letterbox function has been modified, so that the original picture can adaptively reduce information redundancy, that is, reduce black edges. The steps of yolov5 algorithm for image adaptive scaling are to calculate the scaling scale, calculate the size after scaling, and calculate the black edge filling value. The calculation formula is as follows (1):

$$\begin{cases} l_3 = l_2 + np \cdot mod\left(\left(l_1 - w_1\right)\right) \cdot min\left\{\frac{l_2}{l_1}, \frac{w_2}{w_1}\right\}, 32\right) \\ w_3 = w_1 \cdot min\left\{\frac{l_2}{l_1}, \frac{w_2}{w_1}\right\} + np \cdot mod\left(\left(l_1 - w_1\right) \cdot min\left\{\frac{l_2}{l_1}, \frac{w_2}{w_1}\right\}, 32\right) \end{cases} \quad (1)$$

Where: l_1, w_1 is the length and width of the original image; l_2, w_2 is the length and width of the original scaled size; l_3, w_3 is the length and width of adaptive scaling size; $\frac{l_2}{l_1}, \frac{w_2}{w_1}$ is the scaling factor. Select a small scaling factor and multiply it by the length and width of the original image size to obtain l_2 and w_2; $(l_1 - w_1) \cdot min\left\{\frac{l_2}{l_1}, \frac{w_2}{w_1}\right\}$ is the height to be filled, and then in numpy use np.mod to remove the remainder, get 8 pixels, and divide 8 by 2 to get the values that need to be filled in at both ends of the image adaptation.

Adaptive anchor box calculation, setting anchor boxes with initial length and width for different data sets. In the network training of YOLOv5, the predicted anchor frame is output on the set initial anchor frame, the error between the predicted anchor frame and the ground truth of the actual calculation anchor frame is calculated, and the data is updated by backpropagation, and the network parameters are optimized through multiple iterations. The anchor box is calculated. In network training, the network outputs the prediction frame on the basis of the initial anchor frame, then compares it with the real frame, calculates the gap between the two, and then updates it in reverse to iterate the network parameters.

C. Backbone

YOLOv5 uses the focus and CSP structures in backbone [21]. The CSP[22] structure is shown in the network structure diagram in Figure 1, The Focus structure is the operation of the YOLOv5 algorithm that is different from YOLOv3 and YOLOv4 ,and the key step of focus structure is slicing operation, as shown in Figure 2. For example, the image of 608 * 608 * 3 is input into the focus structure, and the slicing operation is adopted. First, it becomes the feature map of 304 * 304 * 12, and then after a convolution operation of 32 convolution cores, it

finally becomes the feature map of 304 * 304 * 32. The original intention of the design of the Cross Stage Partial structure is to reduce redundant computation and enhance gradients. The difference between YOLOv5 and YOLOv4 is that only the backbone network in YOLOv4 uses CSP structure. Two CSP structures are designed in YOLOv5. Take YOLOv5s network as an example, CSP1_ X structure is applied to backbone network, another CSP2_ X structure is applied to neck.
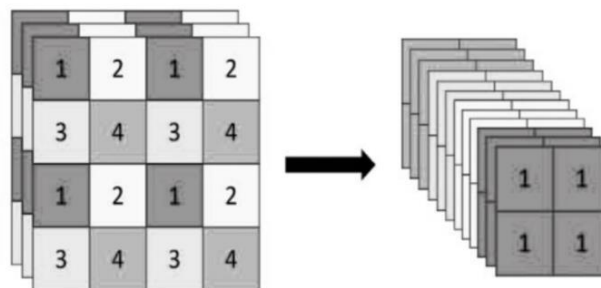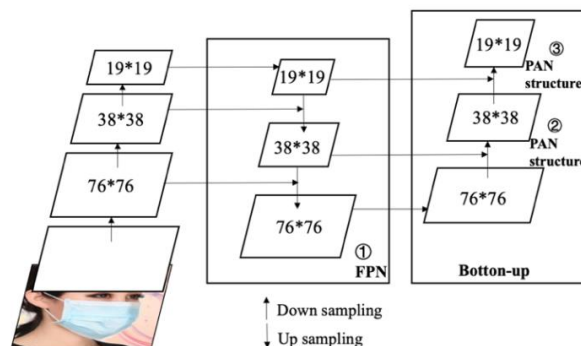

Figure 2.   Slicing operation

D. Neck

Figure 3.   FPN + PAN structure

The network structure design of neck adopts the structure of FPN + PAN. FPN [23] uses a top-down side connection to build a high-level semantic feature map on all scales and construct the classic structure of the feature pyramid. PAN [24] uses a bottom-up feature pyramid. After passing through a multi-layer network in the middle of FPN, the target information at the bottom has been very blurred. Therefore, pan is added to make up for and strengthen the

positioning information. The specific structure is shown in Figure 3.

E. Prediction

Prediction includes bounding box loss function and non maximum suppression (NMS). YOLOv5 uses CIOU_Loss [25], which effectively solves the problem of overlapping bounding boxes. In the post-processing process of target detection, for the screening of many target frames, weighted NMS operation is adopted to obtain the optimal target frame.

Target detection algorithms often output multiple overlapping prediction frames for the same target, resulting in false detection. Therefore, non maximum suppression algorithm [26] is generally used as post-processing technology to suppress redundant prediction frames in order to obtain the final detection results. NMS [27] and soft NMS are commonly used non maximum suppression algorithms [28]. It mainly reduces the redundant prediction frames for the same target [29], reduces the number of false positive prediction frames and improves the accuracy of target detection by suppressing the confidence of the prediction frame with non maximum confidence [30] (hereinafter referred to as the non maximum box). However, NMS and soft NMS only use IOU as the judgment standard [31] to suppress the prediction box that highly overlaps with the confidence maximum box. This kind of algorithm can successfully suppress redundant prediction frames, but it will also suppress adjacent targets [32]. Therefore, this paper uses weighted NMS operation [33]. Compared with the traditional non maximum suppression, weighted NMS does not directly eliminate those frames with the same category as the current rectangular frame whose IOU is greater than the threshold in the process of rectangular frame elimination, but weights them according to the confidence of network prediction to obtain a new rectangular frame, Take the rectangular box as

the rectangular box of the final prediction, and then eliminate those boxes [34].

III. EXPERIMENT AND RESULT ANALYSIS

A. Experimental preparation

There are 8000 pictures in this experimental data set, including two categories: mask and no mask. Mask indicates that the personnel have worn masks, and no mask indicates that the personnel have not worn masks. 80% of the images in the data set are used as training sets and 20% as test sets. Some dataset images are shown in Figure 4.

Figure 4. Dataset partial image

The experiment uses pytoch to build the network framework, the initial learning rate is set to 0.01, and 200 epochs are trained iteratively. Firstly, the obtained images are normalized and preprocessed, and then the processed images are input into the YOLOv5 network model training to obtain the best weight data, and then the images are tested and analyzed. The experimental process is shown in Figure 5.
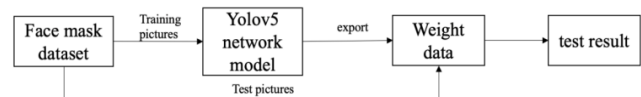

Figure 5. Experimental flow chart

B. Mosaic data enhancement

Mosaic splices four pictures by randomly selecting them and randomly cutting, arranging and scaling them. The effect picture is shown in

Figure 6. This method enriches the background and small targets of the detection object, and improves the robustness of the network to a certain extent. In addition, after mosaic data enhancement, it is equivalent to processing four pictures at a time, and the batch size increases implicitly, which makes the initially set batch size value do not need to be large, and a good model can be obtained, reducing the performance requirements of GPU.

Figure 6.   Mosaic data enhancement

C. Experimental environment

The configuration of algorithm experiment platform is shown in Table 1.

TABLE I.   EXPERIMENTAL ENVIRONMENT
CONFIGURATION

Parameter	Configuration
CPU	Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz
GPU	RTX 3090
Memory	60G
Display memory	24G
System environment	Ubuntu18.04
Experimental platform	PyTorch1.6.0、Python3.8
Accelerated environment	CUDA10.1

D. YOLOv5 network training

In this paper, the training epoch is 200 and the batch size is 64. The official pre training weight of YOLOv5 is used to accelerate the convergence of the model. In the experiment, the learning rates of BN layer, weight layer and Bais

layer are lr0, lr1 and lr2 respectively, in which lr0 and lr1 change the same. The learning change rate is shown in Figure 7.

In this paper, the loss function adopts CIOU_ Loss, as shown in formula (2). Where v is used to measure the consistency of the relative proportion of two rectangular boxes, α Is the weight coefficient, from α It can be seen from the definition of parameters that the loss function will be more inclined to optimize in the direction of increasing overlapping areas, especially when the IOU is zero. CIOU_Loss comprehensively considers the overlapping area, center distance and aspect ratio, and further considers the relative proportion of rectangular frame, which makes the detection effect further.

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2\left(\mathbf{b},\mathbf{b}^{gt}\right)}{c^2} + \alpha v \qquad (2)$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2$$

$$\alpha = \frac{v}{\left(1 - IoU\right) + v}$$

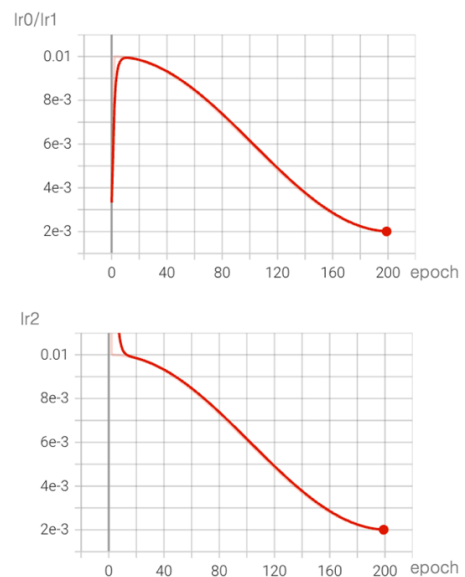

Figure 7.   Learning rate curve

E. Experiment and result analysis

The results of this experiment will be from mAP@0.5, mAP@0.5: 0.95, Precision and Recall. The calculation formula is as follows.

$$Precision ?= \frac{TP}{TP + FP} \qquad (3)$$

$$Recall ?= \frac{TP}{TP + FN} \qquad (4)$$

$$AP = \int_0^1 PdR \qquad (5)$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \qquad (6)$$

TP, FP and FN in (3) and (4) above refer to the number of correct inspection frames, false inspection frames and missed inspection frames respectively. AP value is the area of P-R curve, and N in formula (6) represents the total number of detection categories, which is 2 in this paper. mAP@0.5 it refers to the average AP of all categories when IOU is set to 0.5, mAP@0.5: 0.95 refers to the average mAP on different IOU thresholds. The IOU value ranges from 0.5 to 0.95 in steps of 0.05.

While outputting the prediction box, this paper will also output the classification confidence score belonging to the box. Generally speaking, for a certain model, by adjusting the confidence threshold, the positive or negative of the predicted value can be changed, and the Precision and Recall also change accordingly. By observing the changes in the confidence thresholds of Precision and Recall followers, the quality of the model can be evaluated to a certain extent. If a model maintains a stable Precision at a high level while Recall grows, it proves that the model has better performance. If a model needs to lose a lot of Precision in exchange for the improvement of Rccall, the performance of the model is poor. Typically, researchers use the Precision-Recall curve to measure the model's trade-off between Precision and Recall. The mean precision of each category is calculated from the area under its corresponding Precision-Recall curve. Generally, the higher the AP of each category, the better. The average precision of the evaluation indicators commonly used in target detection is the average of each category of AP.

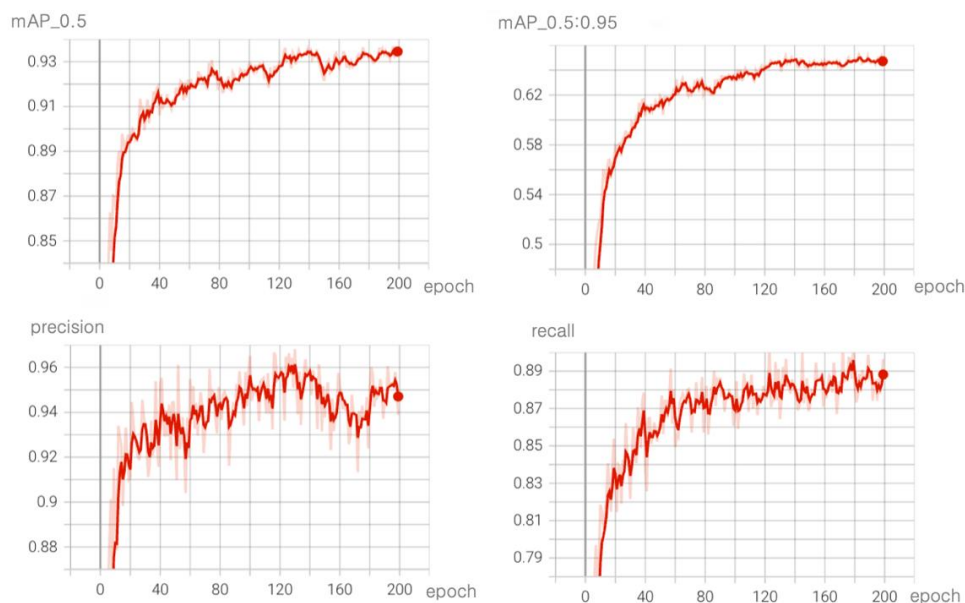The experimental process and experimental test results are shown in Figure 8 and Table 2 respectively.

Figure 8.   Experimental training process

Model	Precision/%	Recall/%	mAP@0.5/%
YOLOv5s	94.8	89	93.5
YOLOv4	76.2	85.4	51.2
YOLOv3	73.6	82.3	48.9

It can be seen from table 2 that the Precision, Recall and mAP of YOLOv5s on the test set are 94.8%, 89% and 93.5% respectively. These three indicators are significantly higher than YOLOv3 and YOLOv4 algorithms.

After the algorithm training in this paper, input the data of the test set into the network model to get the test results, as shown in Figure 9. If the mask is worn, the value on the target box represents the confidence of classification. If the mask is not worn, no mask will be displayed on the target box. According to the performance analysis of various data, the training effect of YOLOv5 network model is ideal.

Figure 9.   Experimental test effect

IV. CONCLUSION

To sum up, in order to achieve the accuracy and real-time of detection and liberate the manual from the complex detection work, this paper proposes a personnel mask wearing detection method based on yolov5, which can effectively solve the problem of time-consuming and laborious manual detection of mask wearing. If the mask is not worn in the public place, it will be detected in the mask confidence box. If the mask is not worn in the public place, it will be detected and displayed. Experiments show that the accuracy, recall and average accuracy (map) on the test set reach 94.8%, 89.0% and 93.5% respectively. The detection effect is better than that of yolov4 and yolov3, and the overall effect is ideal. Further improvements can be made in the follow-up, such as how to detect the non-standard wearing of personnel masks, how to lighten the model, and how to verify and improve the network model in practical application.

REFERENCES

[1] Ren Yu, Liu Quanjin, Huang Zhong, Hu langtao, Liu Guoming. Mask wearing detection algorithm based on fast r-cnn and transfer learning [J] Journal of Anqing Normal University (NATURAL SCIENCE EDITION), 2021, 27(04): 25-30. DOI: 10.13757/j.cnki.cn34-1328/n. 2021. 04. 005.

[2] Dong Yanhua, Zhang Shumei, Zhao Junli.SSD mask detection based on residual structure [J] Computer technology and development, 2021, 31(12):67-72.

[3] Cao chengshuo, Yuan Jie. Mask wearing detection method based on Yolo mask algorithm [J] Progress in laser and Optoelectronics, 2021, 58(08):211-218.

[4] Cheng Tinghao, Cui Yuchao, Wu Xinmiao. Research on mask wearing detection in public places based on yolov4 [J] Modern computer, 2021(16):134-140.

[5] https://zhuanlan.zhihu.com/p/172121380, 2021-04-2

[6] Wang hengtao, Zhang Shang, Zhang Chaoyang, Liu Zhanwei Lightweight PCB defect detection based on yolov5 [J / OL] Radio Engineering: 1-9 [2022-05-10] http://h-p.kns.cnki.net.lib-ycfw.xatu.edu.cn/kcms/detail/13.1097.TN.20220509.1015.004.html

[7] Zhang Xuan, Yan Mingzhong, Zhu Daqi, Guan Yang. Marine ship detection and classification based on YOLOv5 model [J]. Journal of Physics: Conference Series, 2022, 2181(1).

[8] Li Chao, Cao Yining, Peng Yakun. Research on Automatic Driving Target Detection Based on YOLOv5s [J]. Journal of Physics: Conference Series, 2022, 2171(1).

[9] Guo Xiaotong, Zuo Min, Yan Wenjing, Zhang Qingchuan, Xie Sijun, Zhong Iker. Behavior monitoring model of kitchen staff based on YOLOv5l and DeepSort techniques [J]. MATEC Web of Conferences, 2022, 355.

[10] Vadym Slyusar, Mykhailo Protsenko, Anton Chernukha, Vasyl Melkin, Oleh Biloborodov, Mykola Samoilenko, Olena Kravchenko, Halyna Kalynychenko, Anton Rohovyi, Mykhaylo Soloshchuk. Improving the model of object detection on aerial photographs and video in unmanned aerial systems [J]. Eastern-European Journal of Enterprise Technologies, 2022,1(9).

[11] He Yi, Li Han Dong Mask wearing recognition in complex scenes based on improved yolov5 model [J] Microprocessor, 2022, 43 (02): 42-46

[12] Chen Jiping, Chen Yongping, Xie Yi, Zhu Jianqing, Zeng huanqiang Ghost Yolo: lightweight mask face detection algorithm [J / OL] Signal processing: 1-13 [2022-05-10] http://h-p.kns.cnki.net.lib-ycfw.xatu.edu.cn/kcms/detail/11.2406.TN.20220322.1023.002.html

[13] Wang Xinran, Tian Qichuan, Zhang Dong Overview of research on wearing detection of face masks [J / OL] Computer engineering and application: 1-15 [2022-05-10] http://h-p.kns.cnki.net.lib-ycfw.xatu.edu.cn/kcms/detail/11.2127.TP.20220124.1558.014.html

[14] Guo Lei, Wang Qiulong, Xue Wei, Guo Ji Mask wearing detection in dim light based on attention mechanism [J] Journal of University of Electronic Science and technology, 2022, 51 (01): 123-129

[15] Yu Shuo, Li Hui, GUI Fangjun, Yang Yanqi, LV Chenyang Research on real-time detection algorithm of mask wearing based on yolov5 in complex scenes [J] Computer measurement and control, 2021,29 (12): 188-194 DOI:10.16526/j.cnki. 11-4762/tp. 2021.12.035.

[16] Zhu Xinpeng, Li Dan Detection system for wearing masks in railway station based on yolo5face [J] Electronic testing, 2021 (24): 50-52 DOI:10.16520/j.cnki. 1000-8519.2021.24.017.

[17] Chen Zhaojun, Chu Jun, Zeng lunjie. Multi category mask wearing detection based on dynamic weighted category balance loss [J / OL] Journal of graphics:1-10[2022-04-26].

[18] Zhang Qian.Research on acceleration of SAR image target detection algorithm based on deep learning [D] Beijing Jiaotong University, 2021.DOI:10.26944/d.cnki.gbfju.2021.000566.

[19] Li Bao. Design and implementation of pedestrian detection algorithm based on deep learning [D] Beijing University of Posts and Telecommunications, 2021.DOI:10.26969/d.cnki.gbydu.2021.000482.

[20] Niu Ruixin, Zhao Zhengjian, Zhang Shiyuan, Dang Jie, Du Lu.Ship detection in optical remote sensing image based on FPN feature extraction [J] Radio Engineering, 2021,51(11):1296-1302.

[21] Yu Wenqing. Research and application of natural scene text detection algorithm based on deep learning [D] Hebei University of Geosciences, 2022.DOI:10.27752/d.cnki.gsjzj. 2022.000068.

[22] Li Yongshang, Ma Ronggui, Zhang Meiyue.Improve the monitoring video traffic flow statistics of yolov5s + deepsort [J] Computer engineering and Application,2022,58(05):271-279.

[23] Zhang Zhiyan, Li Dan Face detection with mask on campus based on yolo5face [J] Electronic test, 2021 (23): 40-42 + 99 DOI:10.16520/j.cnki. 1000-8519.2021.23.012.

[24] Fan qinrui, Li Dan Mask face detection based on yolo5face [J] Electronic production, 2021 (19): 61-63 + 8 DOI:10.16589/j.cnki. cn11-3571/tn.2021.19.019.

[25] Ye Xingyu Research on mask wearing detection algorithm based on deep learning [J] Information and computer (theoretical Edition), 2021, 33 (18): 72-76

[26] Zhang Luyao, Han Hua Face mask detection based on yolov5s [J] Intelligent computer and application, 2021, 11 (09): 196-199

[27] Yu Fangxu, Wang Shuai, Liu Shichao, Liang Peng, Yu Xi Campus mask wearing detection system based on convolutional neural network [J] Electronic components and information technology, 2021, 5 (08): 19-20 + 24 DOI: 10.19772/j.cnki. 2096-4455.2021.8.009.

[28] Peng Cheng, Zhang Qiaohong, Tang Zhaohui, GUI Weihua Research on mask wearing detection method based on yolov5 enhancement model [J] Computer engineering, 2022, 48 (04): 39-49 DOI: 10.19678/j.issn. 1000-3428.0061502.

[29] Liu Qiyuan Research on single-stage complex face detection method [D] People's Public Security University of China, 2021 DOI: 10.27634/d.cnki. gzrgu. 2021.000271.

[30] Tan Shilei, BIE xiongbo, Lu Gonglin, Tan Xiaohu Real time detection of personnel wearing masks based on yolov5 network model [J] Laser journal, 2021,42 (02): 147-150 DOI:10.14016/j.cnki. jgzz. 2021.02.147.

[31] Wang Feng Improved yolov5 artificial intelligence detection and recognition algorithm for wearing masks and helmets [J] Architecture and budget, 2020 (11): 67-69 DOI:10.13993/j.cnki. jzyys. 2020.11.021.