FCOSMask: Fully Convolutional One-Stage Face Mask Wearing Detection Based on MobileNetV3

Yang Yu
East China Normal University, Shanghai 200062, China
MoEEngineeringResearchCenterforSoftware/HardwareCo-designTechnology51205902120@stu.ecnu.edu.cn

Jie Lu
Software Engineering Institute, East China Normal University, Shanghai 200062 China
71194501135@stu.ecnu.edu.cn

Chao Huang
East China Normal University, Shanghai 200062 China
MoEEngineeringResearchCenterforSoftware/HardwareCo-designTechnology51194501130@stu.ecnu.edu.cn

Bo Xiao*
East China Normal University, Shanghai 200062 China
MoEEngineeringResearchCenterforSoftware/HardwareCo-designTechnologybxiao@sei.ecnu.edu.cn

ABSTRACT

Wearing masks correctly in public is one major self-prevention method against the worldwide Coronavirus disease 2019 (COVID-19). This paper proposes FCOSMask, a fully convolutional one-stage face mask wearing detector based on the lightweight network, for emergency epidemic control and long-term epidemic prevention work. MobileNetV3 is applied as the backbone network to reduce computational overhead. Thus, complex calculation related to anchor boxes is avoided in the anchor-free method, and Complete Intersection over Union (CIoU) loss is selected as the bounding box regression loss function to speed up model convergence. Experiments show that compared to other anchor-based methods, detection speed of FCOSMask is improved around 3 to 4 times on self-established datasets and mean average precision (mAP) achieves 92.4%, which meets the accuracy and real-time requirements of the face mask wearing detection task in most public areas. Finally, a Web-based face mask wearing system is developed that can support public epidemic prevention and control management.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Object detection**;

KEYWORDS

Face mask wearing detection, Lightweight network, Anchor free, Multi-level prediction

ACM Reference Format:
Yang Yu, Jie Lu, Chao Huang, and Bo Xiao*. 2021. FCOSMask: Fully Convolutional One-Stage Face Mask Wearing Detection Based on MobileNetV3. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3487075.3487078

CSAE 2021, October 19–21, 2021, Sanya, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487078

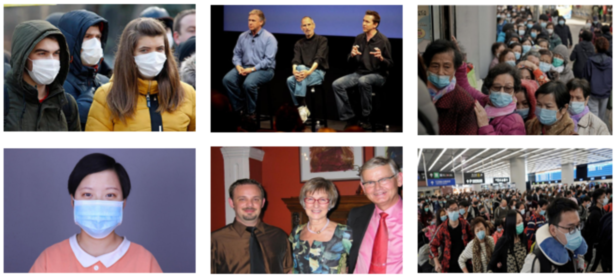

Figure 1: Examples of Face Mask Wearing Dataset.

1 INTRODUCTION

Since the end of 2019, COVID-19 has affected the world. Until July 2021, the cumulative number of confirmed cases worldwide exceeded 100 million, and the death toll reached 3.97 million. Wearing masks can greatly reduce the probability of coronavirus infection through droplet spread. Therefore, in many public areas, people are required to wear masks. However, manually checking the wearing of masks consumes a great deal of human resources. And when the flow of people is large, it is easy to miss the detection. Therefore, the realization of face mask wearing detection algorithm becomes significantly important.

Compared with face detection [1, 2], face mask wearing detection not only requires identifying the face target, but also judging whether the target wears a mask. The criterion of wearing masks is if the mask covers the nose and mouth of a human face. For some targets that wear a mask without covering the nose and mouth, we regard them as not wearing a mask. So far there have been few public datasets related to face masks and the data is incomplete. Therefore, we need to re-integrate and process them. Some images of our dataset can be seen in figure 1. In some complex scenes, the detection target that is small or partially covered become the tricky bit of our detection. In addition, due to the rapid flow of people in public places, the face mask wearing detection algorithm has high requirements to real-time performance.

In this paper, we propose FCOSMask, a one-stage face mask wearing detection method in a per-pixel prediction fashion to achieve an anchor free and proposal free solution. It avoids the disadvantages of using anchor box object detection, i.e., causing extremely unbalanced positive and negative samples, introducing too many hyperparameters related to the anchor box and resulting in increased computational complexity. In order to meet the real-time performance of the face mask detection algorithm in public places, the speed of the face mask detectors is further improved. We choose the lightweight network MobileNetV3 [3] as our backbone network for feature extraction which is improved on the basis of MobileNetV1 [4] and MobileNetV2 [5]. MoileNetV3 not only has a small amount of model parameters, but also reduces computational overhead and improves detection speed. Moreover, we choose CIoU as the bounding box regression loss function. CIoU comprehensively involves the center point distance, overlap rate and the aspect ratio information between the ground-truth and the bounding box to speed up the model convergence [6].

The main contributions of this paper are as follows:

- We proposed a fully convolutional anchor-free face mask wearing detection algorithm based on lightweight network—FCOSMask. Compared with anchor-based detectors, the detection speed is increased by 3 to 4 times, and mAP as high as 92.4%.
- A face mask wearing detection system has been designed and implemented, which can help the epidemic prevention and control management in public areas.
- Multiple datasets from github, kaggle and other websites were collected for integration. In the end, the dataset contains about 20,000 images and annotations.

2 RELATED WORK

2.1 Anchor-based Face Mask Wearing Detectors

Currently, most of the detectors for face mask wearing are based on anchor boxes. Among them, SSDMNV2 [7] proposes the SSD-based face mask detection algorithm which uses the Single Shot MultiBox Detector (SSD) [8] algorithm to detect the face and crop it. It inputs the cropped face into the MobileNetV2 network to detect whether there is a mask. The face mask detection algorithm RETINAFACE-MASK [9] based on RetinaNet [10], which uses focal loss as the loss function. There are also face mask detection algorithm based on the You Only Look Once (YOLO) series. The research work in [11] uses YOLOv2 [12] for medical mask detection. And the work in [13] compares YOLOv3 [14] and Fast R-CNN [15] for face mask recognition.

2.2 Anchor-free Detectors

While the shortcomings of anchor-based detectors were recognized, in recent years, anchor-free object detection algorithms emerged gradually. The earliest exploration of the anchor free method was DenseBox [16], which proved that a single fully convolutional network can detect objects with severe coverage and different scales. It laid the foundation for the Fully Convotional One-Stage Object Detection (FCOS) [17] algorithm. The later exploration of the anchor-free model can be roughly divided into two categories: key point detection and dense prediction. Among them, the anchor-free detection algorithm based on key points is represented by CornerNet [18], which predicts the position of the bounding box through the upper left corner and the lower right corner. ExtremeNet [19], through a given heat map, detects all peaks to extract key points, and groups all key points according to the geometric structure to generate a bounding box. The anchor-free algorithm based on dense prediction includes FCOS and Feature Selective Anchor-Free (FSAF) [20]. In FCOS, Feature Pyramid Network (FPN) [21] is applied to make multi-level predictions. At the same time, similar to Fully Convolutional Networks (FCN) for semantic segmentation [22], FCOS proposes to perform object detection in a per-pixel prediction fashion. FSAF allows each instance to select the best feature layer to optimize the network, so there is no need for anchor boxes to limit the selection of features.

3 DESIGNED OF FCOSMASK

3.1 FCOSMask Approach

Here we propose the FCOSMask approach, a fully convolutional one-stage face mask wearing detector based on the lightweight network, for emergency epidemic control and long-term epidemic prevention work. It is optimized towards mask wearing detection tasks on top of FOCS that directly views locations as training samples instead of anchor boxes as in anchor-based detectors. Therefore FCOSMask can use as many foreground samples as possible to train the regressor. In contrast, anchor-based detectors can only regard anchor boxes and ground-truth boxes with sufficiently high Intersection over Union (IoU) as positive samples. After feature extraction, FCOSMask applies multi-level prediction to detect targets of different sizes at various levels of feature layers. Multi-level prediction can solve the ambiguity and low recall caused by overlapping of real bounding boxes and improve the FCN-based detector to the same level as the anchor-based detector. At the end of each feature level, four convolutional layers were added to obtain the classification branch and the regression branch. Meanwhile, a center-ness layer paralleled with the classification branch is included, which can suppress some low-quality bounding boxes generated by the points far from the center. The overall architecture of FCOSMask is shown in figure 2

3.2 Backbone based on MobileNetV3

Face mask wearing detector has high requirements for real-time performance due to the rapid flow of people in public places. Therefore, we choose lightweight network MobileNetV3 as our backbone network for feature extraction, which has a smaller volume and less calculation amount. MobileNetV3 inherits the depth-wise separable convolutions of MobileNetV1 to reduce the amount of parameter calculation, and the inverse residual structure with a linear bottleneck of MobileNetV2 to extract more feature information. Here the network design improve the swish activation function into h-swish activation function to increase the speed while maintaining accuracy, and use the h-swish activation function in the deeper channels. Other improvements includes the introduction of the lightweight attention module, which makes the model focus more on the identifiable features related to face mask wearing, effectively alleviating the interference of other surrounding features on the detection of face mask wearing. Moreover, MobileNetV3 is divided into MobileNetV3-Large and MobileNetv3-Small. In our architecture design, we use the MobileNetV3-Small structure in order to improve the detection speed, and then apply other tricks

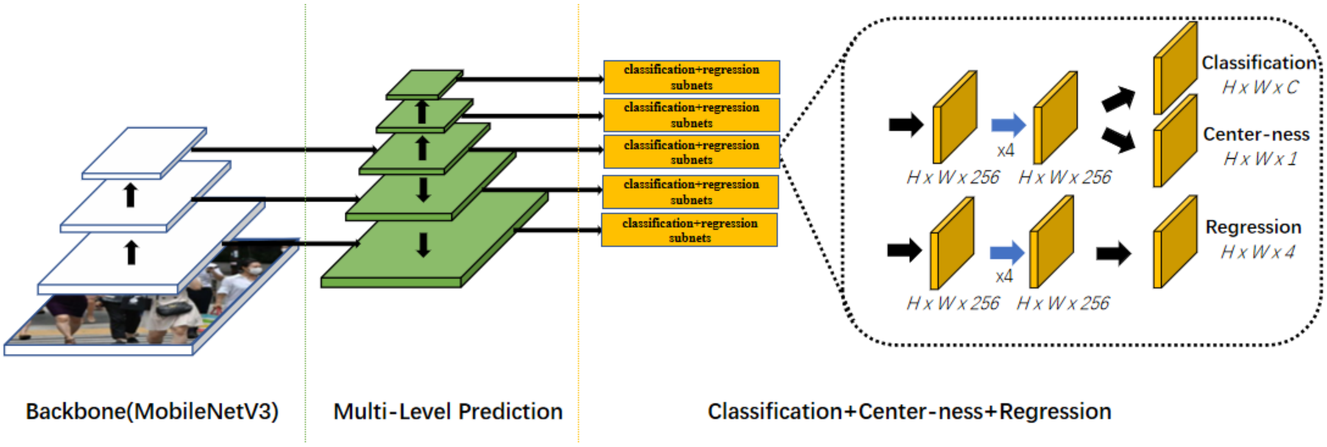
classification+regression subnets
classification+regression subnets
classification+regression subnets
classification+regression subnets
classification+regression subnets
H x W x 256
x4
H x W x 256
Classification
H x W x C
Center-ness
H x W x 1
H x W x 256
x4
H x W x 256
Regression
H x W x 4
Backbone(MobileNetV3)
Multi-Level Prediction
Classification+Center-ness+Regression

Figure 2: Overall Architecture of FCOSMask.

Table 1: MobileNetV3 Structure in Our Network

Input	Operator	#out	SE	NL	s
800 x 1024 x 3	Conv2d, 3 x 3	16	-	HS	2
400 x 512 x 16	Bottleneck,3 x 3	16	√	RE	2
200 x 256 x 16	Bottleneck,3 x 3	24	-	RE	2
100 x 128 x 24	Bottleneck,3 x 3	24	-	RE	1
100 x 128 x 24	Bottleneck,5 x 5	40	√	RE	2
50 x 64 x 40	Bottleneck,5 x 5	40	√	HS	1
50 x 64 x 40	Bottleneck,5 x 5	40	√	HS	1
50 x 64 x 40	Bottleneck,5 x 5	48	√	HS	1
50 x 64 x 48	Bottleneck,5 x 5	48	√	HS	1
50 x 64 x 48	Bottleneck,5 x 5	96	√	HS	2
25 x 32 x 96	Bottleneck,5 x 5	96	√	HS	1
25 x 32 x 96	Bottleneck,5 x 5	96	√	HS	1

to increase the detection accuracy. In order to connect with the following multi-level prediction, we omitted some of operations at the end of MobileNetV3-Small. The MobileNetV3-Small structure in our network is shown in Table 1. Our image input size is 800*1024. Where SE represents whether joining the Squeeze-And-Excite module, and NL denotes the type of nonlinearity used. Here HS stands for h-swish and RE stands for ReLU.

3.3 Loss Function

In FCOSMask, the loss function contains three classes, namely the classification loss $\mathbf{L_{cls}}$, the center-ness loss $\mathbf{L_{cnt}}$ and the regression loss $\mathbf{L_{reg}}$. Where $\mathbf{L_{cls}}$ is focal loss, which is mainly to solve the problem of the imbalance of the positive and negative sample ratio. $\mathbf{L_{cnt}}$ applies the cross-entropy loss function for training. $\mathbf{L_{reg}}$ is the IoU loss as in UnitBox [23]. However, we use CIoU loss for regression instead of IoU loss. Compared with the IoU loss function, CIoU not only normalizes the coordinate scale and solves the situation where IoU is zero, but also takes the overlapping area of the bounding box, the distance between the center points, and the aspect ratio between the anchor box and the target box all into account, thus can make the model converge faster. CIoU loss calculation formula is as follows

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{1}$$

Where $\mathbf{b}$, $\mathbf{b}^{\text{gt}}$ represent the center point of the bounding box and ground-truth box respectively, $\rho^2(\mathbf{b}, \mathbf{b}^{gt})$ means the Euclidean distance between the center point of the bounding box and ground-truth box, and c^2 means the diagonal distance of the smallest closure area that can contain both the bounding box and ground-truth box. α is a trade-off parameter, and v reflects the consistency of the aspect ratio. The calculation formulas of α and v are defined as,

$$\alpha = \frac{v}{(1 - IoU) + v} \tag{2}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{3}$$

4 EXPERIMENTS AND RESULTS

The core system configuration in our experiments is NVIDIA GeForce RTX 3080, and the operating system is Ubuntu 20.14.

Table 2: Ablation Study of FCOSMask

Data Augmentation	CIoU	APwith	APwithout	mAP
×	×	86.6	89.8	88.2
√	×	89.5	93.7	91.6
×	√	87.9	90.3	89.1
√	√	90.1	94.6	92.4

4.1 Datasets

The current datasets on face mask wearing are few and incomplete. We collect several datasets from github, kaggle and other websites for integration, in order to further remove duplicates and low-quality samples, relabel them, divide the samples into with mask and without mask, and generate the VOC data format. The final dataset contains 20,000 images. Training set contains 15,385 images and validation set contains 4615 images. In training set, there are 11,397 with mask target samples and 19,770 without mask target samples. Part of the image is shown in figure 1

4.2 Ablation Study

With MobileNet-Small as our backbone and FCOS as the detector, we added the data augmentation and loss function CIoU for ablation experiments, as shown in table 2. The results are analyzed as follows:

4.2.1 CIoU+DIoU NMS.. We replace the original bounding box regression loss function IoU with CIoU. There the overlap area, center point distance and aspect ratio between different targets are fully considered. At the same time, we replace the Non-Maximum Suppression (NMS) [24] in post-processing with Distance-IoU (DIoU) NMS in YOLOv4 [25] to match it. The results of the experiment show that mAP increases by nearly 1% point.

4.2.2 Data augmentation. After data analysis, we found that among the 15,385 images in the training set, 60% were single-sample images. This makes it easy to miss the detection of pictures containing multiple samples in complex scenes. Therefore, we need data augmentation. We use mosaic in YOLOv4 to augment for 50% of the images. The principle of mosaic is to stitch 4 pictures by randomly zooming, cutting, and arranging them. In this way, the background information of the backbone is enriched, and the information extraction capability of the backbone is enhanced. For the other 50% of images, we use a "fill-in" method for image stitching. We intercept the bounding boxes of all images in the training set, and for each image input to the network, randomly select multiple intercepted bounding boxes for stitching, thereby simulating complex scenes. Experimental results show that through this method, the detection performance is significantly enhanced about 2.5%. Figure 3,4 shows the effect after mosaic and "fill-in" method respectively.

4.3 Comparison with Other Detectors

On the self-established dataset, we compared FCOSMask with state-of-the-art anchor-based detectors SSD, RetinaNet, and YOLO series on the detection accuracy and detection speed. We take mAP as accuracy evaluation index, and frames per second (FPS) as speed evaluation index. We found that SSD, RetinaNet performed slightly worse both in accuracy and speed. Our method outperforms the anchor-based counterpart SSD by 5.5% and RetinaNet by 3.7%. The speed is increased by 3 to 4 times. For the YOLO series, we selected the most lightweight network YOLOv5s, which only accounts for 14M weight. By comparison, we found that although we have a slightly lower detection speed, but the detection accuracy is improved by 0.8%. The specific comparison information can be found in the table 3. Our final detection results are illustrated in figure 5

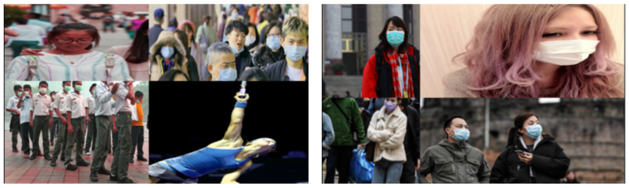

Figure 3: Data Augmentation with Mosaic.

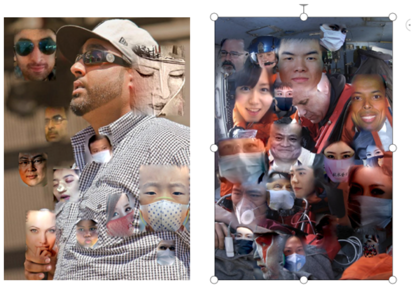

Figure 4: Data Augmentation with "fill-in" Method.

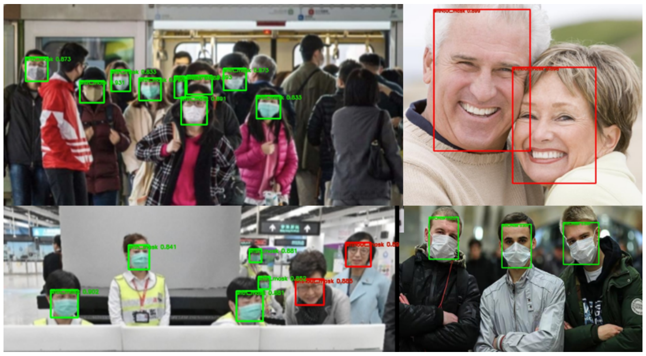

Figure 5: Some Detection Results.

Table 3: Comparison with Anchor-based Detectors

Method	APwith	APwithout	mAP	FPS
SSD	84.7	89.1	86.9	12
RetinaNet	86.8	90.6	88.7	15
YOLOv5s	89.5	93.7	91.6	54
Ours	90.1	94.6	92.4	50

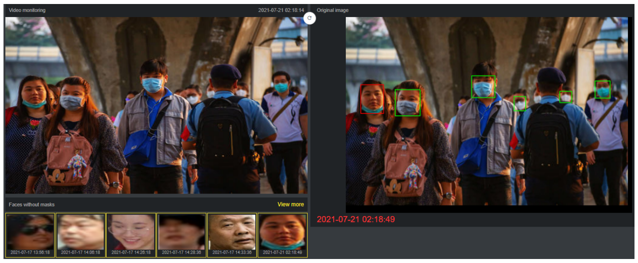

Figure 6: Web Page Front-End of the Detection System.

5 FACE MASK WEARING DETECTION SYSTEM

We design a face mask wearing detection system, which can help detect face mask wearing in public areas. This system realizes the video stream data captured by the terminal camera after it is connected to the camera. The back-end detects images in the video in real time. If someone is not wearing a mask, it will be recorded and a voice prompt will be given to the manager. The front-end and the back-end of our system are implemented by Spring Boot framework and Vue.js framework respectively. Storing images through the MySQL database. Moreover, RabbitMQ middleware is used to communicate with the camera and the back-end. The camera collects the video, and sends the video stream to the back-end through RabbitMQ. If a frame is pushed to the back-end program and the algorithm detects that there is someone not wearing a mask, the back-end program will push this frame to the front-end page for display through the WebSocket technology. As we can see from figure 6, the Web page is composed of three modules, which are video monitoring, faces without masks and original image. The upper left corner is the real-time video monitoring captured by the camera. Once algorithm detects that someone without a mask in a certain frame of the video stream, the original image will be displayed on the upper right corner. Then the face of the person who is not wearing a mask is cut out in the original image and displayed on the lower left part of the page.

6 CONCLUSION

In this paper, we propose FCOSMask, a fully convolutional one-stage face mask wearing detector based on lightweight network. Unlike anchor-based detectors, we apply a per-pixel prediction method to avoid the complex computations associated with anchor boxes. We collected and relabeled more than 20,000 images for training. On the self-established dataset, FCOSMask performs better than anchor-based detectors both in the aspect of speed and accuracy. The mAP can reach 92.4%, and speed up to 50 fps. The experiment results meet the accuracy and real-time requirements for the detection of face mask wearing in public places. In addition, we designed and implemented a Web-based epidemic prevention and control management system, which can apply the FCOSMask algorithm in a variety of practical scenarios. In the future, under the premise of ensuring accuracy, we will use methods such as pruning to streamline the network structure to further increase the detection speed.

ACKNOWLEDGMENTS

This work was supported by ECNU project - research on key technologies of WSN (No: 48102840)

REFERENCES

[1] Xu T , Du D K , He Z , *et al* (2018). PyramidBox: A Context-assisted Single Shot Face Detector[J].
[2] Huang X , Deng W , Shen H , *et al* (2020). PropagationNet: Propagate Points to Curve to Learn Structure Information[J]. IEEE.
[3] Howard A , Sandler M , Chen B , *et al* (2020). Searching for MobileNetV3[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.
[4] Howard A G , Zhu M , Chen B , *et al* (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications[J] .
[5] Sandler M , Howard A , Zhu M , *et al* (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[6] Zheng Z , Wang P , Liu W , *et al* (2020). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[C]// AAAI Conference on Artificial Intelligence.
[7] Pn A , Rj A , Am A , *et al.* SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2 - ScienceDirect[J]. Sustainable Cities and Society, 66.
[8] Liu W , Anguelov D , Erhan D , *et al* (2016). SSD: Single Shot MultiBox Detector[J].Springer, Cham.
[9] Jiang M , Fan X (2020). RetinaMask: A Face Mask detector[J].
[10] Lin T Y , Goyal P , Girshick R , *et al*(2017). Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, PP(99):2999-3007.
[11] Ml A , Gmb C , Mhnt D , *et al* (2020). Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection - ScienceDirect[J]. Sustainable Cities and Society,.
[12] Redmon J , Farhadi A (2017). YOLO9000: Better, Faster, Stronger[J]. IEEE Conference on Computer Vision & Pattern Recognition, 6517-6525.
[13] Singh S, Ahuja U , Kumar M , *et al* (2021). Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment[J]. Multimedia Tools and Applications, 1-16.
[14] Redmon J , Farhadi A (2018) . YOLOv3: An Incremental Improvement[J]. arXiv e-prints.
[15] Girshick R (2015). Fast R-CNN[J]. arXiv e-prints.
[16] Huang , Yang Y , Deng Y , *et al* (2015). DenseBox: Unifying Landmark Localization with End to End Object Detection[J]. Computer Science.
[17] Tian Z , Shen C , Chen H , *et al* (2020). FCOS: Fully Convolutional One-Stage Object Detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE.
[18] Law H, Deng J (2020) . CornerNet: Detecting Objects as Paired Keypoints[J]. International Journal of Computer Vision, 128(3):642-656.
[19] X Zhou, J Zhuo, Krhenbühl, Philipp (2019). Bottom-up Object Detection by Grouping Extreme and Center Points[J]. .
[20] Zhu C , He Y , Savvides M (2019). Feature Selective Anchor-Free Module for Single-Shot Object Detection[C]// 2019 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR). IEEE.

[21] Lin T Y, Dollar P , Girshick R , *et al* (2017). Feature Pyramid Networks for Object Detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society,.

[22] Shelhamer E , Long J , Darrell T (2016). Fully Convolutional Networks for Semantic Segmentation[J].

[23] Yu J , Jiang Y , Wang Z , *et al* (2016). UnitBox: An Advanced Object Detection Network[M]. ACM.

[24] Neubeck A , Gool L (2006) . Efficient Non-Maximum Suppression[C]// International Conference on Pattern Recognition. IEEE Computer Society.

[25] Bochkovskiy A , Wang C Y , Liao H (2020) . YOLOv4: Optimal Speed and Accuracy of Object Detection[J].