Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computer Methods and Programs in Biomedicine 221 (2022) 106888

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

SMD-YOLO: An efficient and lightweight detection method for mask wearing status during the COVID-19 pandemic

Zhenggong Han[a], Haisong Huang[a,d,*], Qingsong Fan[a], Yiting Li[b], Yuqin Li[c], Xingran Chen[a]

[a] Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang 550025, Guizhou, China
[b] College of Big Data Statistics, GuiZhou University of Finance and Economics, Guiyang 550025, Guizhou, China
[c] Stomotological Hospital of Guizhou Medical University, Guiyang 550004, Guizhou, China
[d] Chongqing Vocational and Technical University of Mechatronics, Chongqing 400036, China

ARTICLE INFO

Article history:
Received 13 February 2022
Revised 30 April 2022
Accepted 11 May 2022

Keywords:
COVID-19
Computer vision
Face mask detection
YOLO
Object detection

ABSTRACT

Background and Objective: At present, the COVID-19 epidemic is still spreading worldwide and wearing a mask in public areas is an effective way to prevent the spread of the respiratory virus. Although there are many deep learning methods used for detecting the face masks, there are few lightweight detectors having a good effect on small or medium-size face masks detection in the complicated environments.
Methods: In this work we propose an efficient and lightweight detection method based on YOLOv4-tiny, and a face mask detection and monitoring system for mask wearing status. Two feasible improvement strategies, network structure optimization and K-means++ clustering algorithm, are utilized for improving the detection accuracy on the premise of ensuring the real-time face masks recognition. Particularly, the improved residual module and cross fusion module are designed to aim at extracting the features of small or medium-size targets effectively. Moreover, the enhanced dual attention mechanism and the improved spatial pyramid pooling module are employed for merging sufficiently the deep and shallow semantic information and expanding the receptive field. Afterwards, the detection accuracy is compensated through the combination of activation functions. Finally, the depthwise separable convolution module is used to reduce the quantity of parameters and improve the detection efficiency. Our proposed detector is evaluated on a public face mask dataset, and an ablation experiment is also provided to verify the effectiveness of our proposed model, which is compared with the state-of-the-art (SOTA) models as well.
Results: Our proposed detector increases the AP (average precision) values in each category of the public face mask dataset compared with the original YOLOv4-tiny. The mAP (mean average precision) is improved by 4.56% and the speed reaches 92.81 FPS. Meanwhile, the quantity of parameters and the FLOPs (floating-point operations) are reduced by 1/3, 16.48%, respectively.
Conclusions: The proposed detector achieves better overall detection performance compared with other SOTA detectors for real-time mask detection, demonstrated the superiority with both theoretical value and practical significance. The developed system also brings greater flexibility to the application of face mask detection in hospitals, campuses, communities, etc.

1. Introduction

At present, there are more than 200 million cases of infection with the Corona Virus Disease 2019 (COVID-19) worldwide. Because respiratory infection viruses, toxic and harmful gasses, and droplets suspended et al. in the air can enter the lungs of humans to cause pneumonia, toxic reactions and even nerve damage [1]. Responding to the rapid spread of the brutal virus, Governments have started implementing new rules forcing people to wear face masks in public places [2]. During the epidemic, wearing disposable medical masks or type N95 respirators is a very important means of protecting against other respiratory diseases not just for the COVID-19 [3]. It can provide good protection against the virus for people going out and also an extremely effective and economical means of prevention and control for society [4,5]. Therefore, it is of great practical significance to realize the detection for mask wearing status in public places (such as hospitals, campuses etc.).

In recent years, a large number of studies have used deep learning to complete object detection and are widely used in biomedicine [6–8], lesions detection [9–11], face detection [12–14] and other fields [15–18]. The existing machine learning and deep learning methods have achieved some results in the task of face mask detection [19—22], however, there are still limitations in

* Corresponding author.

https://doi.org/10.1016/j.cmpb.2022.106888
0169-2607/

the complicated environments. For example, the face mask detector for real-time has low recall, detection accuracy and speed; The overlap of the face masks and the shadow of the background result in missing detection of some small or medium-size targets; The larger detection models consume a lot of computer resources. Until the advent of You Only Look Once (YOLO) series (v1 [23],v2 [24],v3 [25],v4 [26],v5 [27],X [28] et al.),as the advanced single-stage object detectors, they broke the dominance of the two-stage method in object detection. Among them, YOLOv4 is another milestone object detector. And it directly classifies and predicts the object at each position in the entire original image. Contrasted to previous version, it modifies the backbone network, activation function, loss function, data enhancement, etc. While maintaining a good accuracy, it solved the pain point, the speed problem, in the detection. Meanwhile, some lightweight detectors appeared, which greatly shortened the amount of model parameters and training time, such as MobileNet [29–31], EfficientNet [32], EfficientDet [33], GhostNet [34] and YOLOv3-tiny. Whereas YOLOv4-tiny is released by authors of YOLOv4. As a lightweight version, its parameters are merely about 6 million, which is equivalent to 1/10 of the original. The network structure is simpler, the detection speed is faster. It also uses fewer computation resources, however, its detection accuracy is reduced.

To solve the above exiting problems to balance the detection accuracy and speed, in this paper we utilize the variants of YOLO and propose a new variant of YOLOv4-tiny for small or medium-size face masks detection, named SMD-YOLO. It is mainly used for masks detections of small or medium faces in size. It improves the detection accuracy of the model on the premise of meeting the real-time performance. This detection method we proposed is a more lightweight model for detecting faces with masks, faces without masks, and masks wearing incorrectly during the COVID-19 pandemic. Meanwhile, we also develop a face mask detection and monitoring system using the detector. The main contributions and innovations will bring greater flexibility to the application of face mask detection in the hospitals, campuses etc.

The following are the primary contributions of this work:

1. A face mask detection and monitoring system is developed for mask wearing status by using an improved variant of YOLOv4-tiny network with two feasible improvement strategies: network structure optimization and K-means++ clustering algorithm.
2. Our proposed detector improves the detection accuracy with less weight parameters and calculation amount while meeting the real-time requirements by employing improved residual module, cross fusion module, enhanced dual attention mechanism, improved spatial pyramid pooling module, modified activation function and depthwise separable convolution module.
3. We evaluate the performance of the proposed detector adopting an ablation experiment based on YOLOv4-tiny. The mAP relative to the original detector is improved by 4.56% and the speed reaches 92.81 FPS. Meanwhile, the quantity of parameters and the FLOPs are reduced by 1/3, 16.48%, respectively.
4. The proposed detector is compared with a few state-of-the-art detectors to demonstrate the superiority of our proposed model balancing the accuracy and speed through comprehensive evaluation of various indicators.

The rest of this article is organized as follows: Section 2 reviews previous related work. Section 3 describes our face mask detection and monitoring system and proposed detection method in detail. Section 4 introduces the experimental details and results, including the face mask dataset, the experimental environment, and the results of the experiments. Section 5 discusses and analyzes the experimental results. And Section 6 presents the conclusion and future prospects of this paper.

2. Related research

In recent years, face detection has always been a research hotspot in object detection. The task of face masks recognition in public areas can be achieved by deploying an efficient object recognition algorithm on surveillance devices. The different shapes and postures of people wearing masks, the influence of ambient light and occlusion interference, etc. make the detection task more difficult and more challenging [35]. Novel proposals by improving the existing masks detectors, face detectors or other object detectors can achieve enthralling results and escalate the research in this domain.

Lin et al. [36] proposed an modified LeNet (MLeNet) network for detecting masked faces in the wild through video surveillance and violence video retrieval. The authors modified the number of units in output layer and increased the number of feature maps with smaller filter size to manually design a learn-based feature and classifier training paradigm. The proposed work increased the training dataset by horizontal reflection and learned MLeNet via combining both pre-training and fine-tuning. The final accuracy rate of 71% was achieved on a real-world masked face detection dataset. Ge et al. [37] proposed a novel dataset of Masked Faces (MAFA) with 30,811 Internet images and 35,806 masked faces, and a model based on a convolutional neural network called LLE-CNNs to find the normal and masked face in the wild. The proposed model was composed of proposal, embedding and verification modules. On the MAFA dataset, the AP value of their LLE-CNNs model reached up to 76.4%. Hussain et al. [38] presented an Internet of Things-based Smart Screening and Disinfection Walk-through Gate (SSDWG) for all public areas entrance. The authors implemented a self-designed real-time deep learning models classified individuals who wear the face mask properly and without a face mask on a combination dataset of MAFA, Masked Face-Net and Bing. Moreover, using a transfer learning approach, they achieved a detection accuracy rate of 99.81%. Loey et al. [39] proposed a hybrid model using deep and classical machine learning for face mask detection. Resnet50 was used as the feature extraction network (FEN), meanwhile decision trees, Support Vector Machine (SVM), and ensemble algorithm were employed for the classification process of face masks. On the Real-World Masked Face Dataset (RMFD), the Simulated Masked Face Dataset (SMFD) and the Labeled Faces in the Wild (LFW) dataset, the proposed technique achieved more than 99% detection accuracy, respectively. At the same year, Loey et al. [40] also proposed a blended face mask detection method which combined ResNet-50 as the feature extraction component with YOLOv2 as the detection component. Mean Intersection over Union (IOU) was used to estimate the best number of anchor boxes as well. The authors obtained the final 81% AP using the Adam optimizer and transfer learning on the two public Medical Masks Dataset (MMD) and Face Mask Dataset (FMD).

Singh et al. [41] accomplished the task of face masks detection using YOLOv3 and Faster R-CNN [42] on the custom dataset which is composed of MAFA, WIDER FACE and many manually prepared images. By taking IOU = 0.5, the authors achieved the AP value of 55% and 62%, respectively. Wu et al. [43] proposed a novel face mask detection framework (FMD-YOLO), which deep residual network was combined with Res2Net module in the FEN and an enhanced path aggregation network was used for feature fusion. Moreover, to improve the detection efficiency and accuracy, localization loss was designed and adopted in model training phase and Matrix NMS method was employed in the inference stage. Finally, on the two public dataset MD-2 and MD-3, the authors achieved the mAP value of 66.4% and 57.5% at IOU = 0.5 level, respectively. Su et al. [44] proposed a modified YOLOv3 for detecting faces and determine whether people are wearing mask, and fused transfer learning and MobileNet for mask classification.

The authors employed EfficientNet as the backbone network and chose CIOU as the loss function. In the dataset combined MAFA and WIDER Face, the mAP value of 96.03% and the FPS value of 14.62 are achieved. Cao et al. [45] proposed a novel object detector namely MaskHunter for the real-time mask detection. The authors modified CSPDarknet19 in the backbone on the basis of YOLOv4, and introduced SPP modules, the feature pyramid network (FPN) modules and path aggregation network (PAN) modules in the neck. In addition, a novel improved Mosaic data augmentation method and a novel mask-guided module were proposed to enhance the discrimination ability of face mask especially in the night environment. The Average Precision (AP) value of 94% and the speed value of 74 FPS were achieved in the end. Jiang et al. [46] proposed a mask detector Squeeze and Excitation (SE)-YOLOv3 with relatively balanced effectiveness and efficiency. The authors integrated the attention mechanism in the backbone network of YOLOv3, employed GIoU as regression loss. The mAP of 71.9% was obtained on their proposed Properly Wearing Masked Face Detection Dataset (PWMFD), which is 8.6% higher than the original network, and the speed is almost identical. The detection time of an image reached 43.2 ms. Yu et al. [47] proposed a face mask recognition and standard wear detection algorithm based on the improved YOLOv4. The authors adopted adaptive image scaling algorithms, introduced modified CSPDarkNet53 and PANet network structures in the feature layer. The results tested on the dataset screening from the published RMFD and MaskedFace-Net and showed that the mAP and speed can reach 98.3% and 54.57 FPS, respectively.

Nagrath et al. [48] proposed a face masks classification model by combining Single Shot Detector (SSD) as a detector with MobileNetv2 [49] as a classifier to realize the real-time detection for people wearing or not wearing face masks. It has a classification accuracy of 92.64% and a speed of 15.71 FPS on the custom dataset. Kumar et al. [50] proposed a novel face masks detection dataset consisting of 52,635 images for four different categories namely, with masks, without masks, masks incorrectly, and mask area. Further, the authors tested with eight variants of the YOLO, and among all tiny variants YOLOv4-tiny achieved a mAP value of 57.71%. Finally, new architectures modifications were proposed in the FEN so that mAP was improved by 2.54% for YOLOv4-tiny. In the same year, these authors [51] proposed a novel face mask vision system that is based on an improved YOLOv4-tiny object detector with spatial pyramid pooling (SPP) [52] module and additional YOLO detection layer as well. By using K-means++ clustering to extract the best priors for anchor boxes, the proposed improved network achieved a mAP value of 64.31% on their published dataset [50] which was 6.6% higher than the original. Roy et al. [53] proposed the Moxa3K Benchmark Dataset (MOXA) consisting of 3000 images for persons with masks and without masks. To meet the monitoring platform with limited computing power, the authors used a few popular detectors such as YOLOv3, YOLOv3-tiny, SSD and Faster R-CNN, and evaluated on their MOXA dataset. The mAP value of 63.99%, 56.27%, 46.52% and 60.05% were attained, respectively. Moreover, YOLOv3-tiny and SSD acquired the 138 FPS, 67.1 FPS speed, respectively. The research progress of related work is shown in Table 1.

To sum up, there is a lacking exploration of a few lightweight detectors suitable for real-time surveillance applications, but it can be achieved by enhancing and improving the exiting state-of-the-art object detection model. To avoid threatening the health and safety of public areas, at present, the detection task of incorrect wearing of masks can be more of a concern because it is more likely to cause misjudgment. Therefore, we propose an improved face mask detector integrated into a face mask detection and monitoring system on the basis of YOLOv4-tiny to achieve better performance.

3. Methods

In this section, to prevent people from removing masks or not wearing them correctly in public places, we introduced a face mask detection and monitoring system to use for ensuring the proportion of people wearing masks correctly in areas. And to achieve better overall performance, a novel efficient and lightweight detection method SMD-YOLO based on YOLOv4-tiny for small or medium-size masks wearing status are described in detail.

3.1. Face mask detection and monitoring system

The face mask detection and monitoring system is developed for this work including four major components, namely the monitoring system (scene monitor), detecting control system (face mask detector, statistics and controller), alarm system (acousto-optic alarm and voice prompt) and restricted flow system (such as gate). To be specific, firstly the monitoring system is used to obtain the videos and images of the environment to be detected and upload it to the detecting control system after collection by wireless or wired transmission. Then, in the detecting control system, the face mask detector detects the face mask wearing status of people in this area for transmitting to the statistics module. Moreover, the proportion of wearing masks properly will be calculated and the

Table 1
Research progress of the related work.

Work	Year	Method	Dataset	Public	Task	Result
Lin et al. [36]	2016	MLeNet	Custom	No	Detection	AP = 71%
Ge et al. [37]	2017	LLE-CNNs	MAFA	Yes	Detection	AP = 76.1%
Hussain et al. [38]	2021	CNN	Custom	No	Classification	AC = 99.81%
Loey et al. [39]	2021	ResNet50+SVM	RMFD SMFD LFW	Yes	Classification	AC = 99.64%
Loey et al. [40]	2021	YOLOv2+ResNet	Custom	No	Detection	AP = 81%
Singh et al. [41]	2021	YOLOv3	Custom	Yes	Detection	AP = 55%
		Faster R-CNN				AP = 62%
Wu et al. [43]	2022	FMD-YOLO	MD-2	Yes	Detection	mAP = 66.4%
			MD-3	Yes		mAP = 57.5%
Su et al. [44]	2022	Efficient-YOLOv3	FMD	Yes	Detection	mAP = 96.03%, Speed = 14.62 FPS
Cao et al. [45]	2020	MaskHunter	Custom	No	Detection	AP = 94%, Speed = 74 FPS
Jiang et al. [46]	2021	(SE)-YOLOv3	PWMFD	Yes	Detection	mAP = 71.9%, Speed = 23 FPS
Yu et al. [47]	2021	Improved YOLOv4	Custom	No	Detection	mAP = 98.3%, Speed = 54.57 FPS
Nagrath et al. [48]	2021	SSD+MobileNetv2	Custom	Yes	Classification	AC = 92.64%, Speed = 15FPS
Kumar et al. [51]	2021	YOLOv4-tiny-SPP	Custom	Yes	Detection	mAP = 64.31%
Roy et al. [53]	2020	YOLOv3	MOXA	Yes	Detection	mAP = 63.99%
		YOLOv3-tiny				mAP = 56.27%, Speed = 138 FPS
		SSD				mAP = 46.52%, Speed = 67.1FPS
		Faster R-CNN				mAP = 60.05%

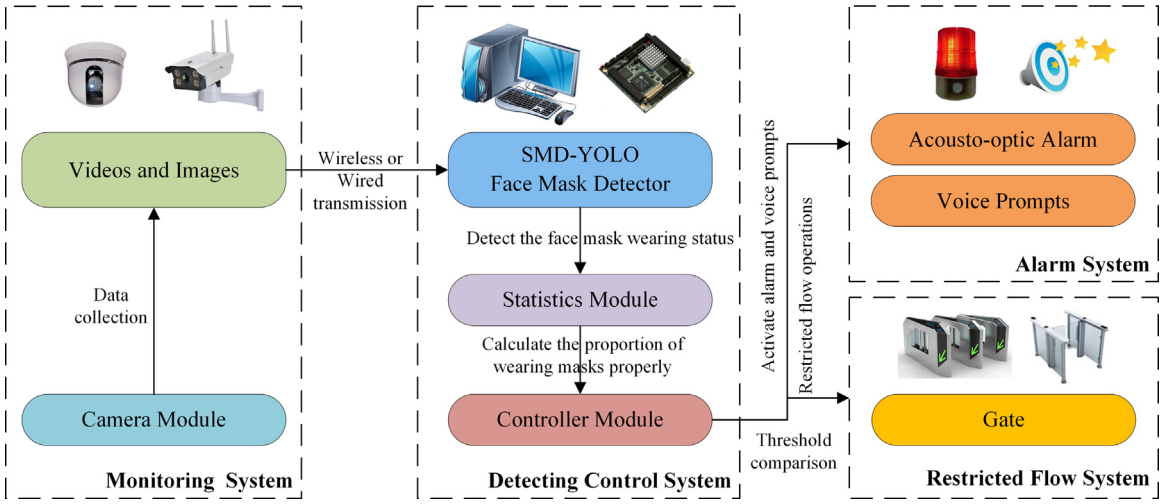

Fig. 1. Workflow of the face mask detection and monitoring system.

calculated result can be delivered to the controller. Finally, the controller triggers to perform corresponding operations by comparing the set thresholds. If the proportion of wearing masks properly is less than the set threshold, it will be fed back to alarm system for activating the acousto-optic alarm and voice prompts and restricted flow equipment for turning off the entrance of the gate. On the contrary, turn off the regional current limiting to maintain normal access. The detailed workflow of the proposed face mask detection and monitoring system is shown in Fig. 1.

The face mask detector of the above detecting control system is the most important part of the proposed the face mask detection and monitoring system, which can be on a computer system or also belong to different computing systems from the statistics and controller modules. Since the accuracy of face mask detector directly affects the subsequent operation of the alarm and restricted flow devices, it is indispensable to develop a lightweight face mask detection method with high detection accuracy during the COVID-19 pandemic.

3.2. Proposed SMD-YOLO detector

YOLOv4-tiny is a lightweight version based on YOLOv4, which utilizes the CSPDarknet53-tiny network contained 28 layers as the backbone network. Three CSPBlock modules are adopted to replace the ResBlock modules in the original residual network. Two feature layers (13×13 and 26×26) are used for object classification, and the FPN is used to merge the effective feature layers to improve the detection accuracy. YOLOv4-tiny also employs the CSPnet structure and performs channel segmentation in the feature extraction network.

In this part, we fine-tune the YOLOv4-tiny network. The optimized YOLOv4-tiny network address 3 shortcomings in the original model. 1) Poor detection ability and relatively low recall for small and medium-sized face masks. 2) Interference on face mask detection from background environments such as light, building and decoration. 3) The larger model is difficult to deploy on small and real-time computing equipment. For the former, it is generally because of the shallow depth of the network that the feature maps for small and medium-sized objects are not expressed enough. For the middle, it lacks the means to quickly filter out high-value information from a large number of features. For the latter, currently object detection needs to seek further improvements in precision and less storage consumption. The structure scheme diagram of the proposed novel SMD-YOLO detector is shown in Fig. 2.

The SMD-YOLO detector adequately fuses the feature information of deep and shallow layers to benefit for small targets detection through combining the improved residual module with cross fusion module. Then, by means of the modified SPP module, attention mechanism, activation function and depthwise separable convolution, the small or medium-size targets of the images are attached more importance to, meanwhile, redundant information is removed to reduce the influence of background. Finally, the detector can improve the detection accuracy with less weight parameters and calculation amount while meeting the real-time requirements.

3.2.1. Improved residual module

In different scenarios, the feature extraction is very important to classify for face mask detection. The convolution layer of mask feature extraction network can effectively analyze the features of masks. According to the principle of convolution operation [54], the number of convolution parameters is related to the size of the convolution kernel, the number of input and output feature map channels. Let $(K, K, C_{\text{int}}, C_{out})$ represent a standard convolution operation. Where C_{int}, C_{out} denote the number of input channels and output channels of a $K \times K$ convolutional layer, respectively. In the case of ignoring bias, the quantity of the convolution parameters is $K \times K \times C_{\text{int}} \times C_{out}$. Thus, the convolutional parameter number has a significant impact on the training time, operation speed and lightness of neural networks. In addition, the number of convolutional layers also affects the number of convolutional parameters. Adding a convolutional layer inside the residual module can decompose the problems that need to be learned hierarchically by means of deepening the network. It is more effective for the network to learn local features, and helpful to improve the classification accuracy of each target.

As is exhibited in Fig. 3, to enhance the feature extraction capability of network, a residual module, divided into enhanced and lightweight modules, is improved to obtain more features. It is used for replacing the 3 original CSPBlock modules to achieve multi-scale changes in channels. The enhanced module, shown in Fig. 3(**b**), is to fully integrate and extract the feature map via using a 1×1 point convolutional layer and a 3×3 convolutional layer, before the first 3×3 convolutional layer splitting in the CSPBlock. While the lightweight module, shown in Fig. 3(**c**), replaces the first 3×3 convolutional layer of CSPBlock with 1×1 and 3×3 convolutional layers to reduce the quantity of parameters for network's convolutional operations.

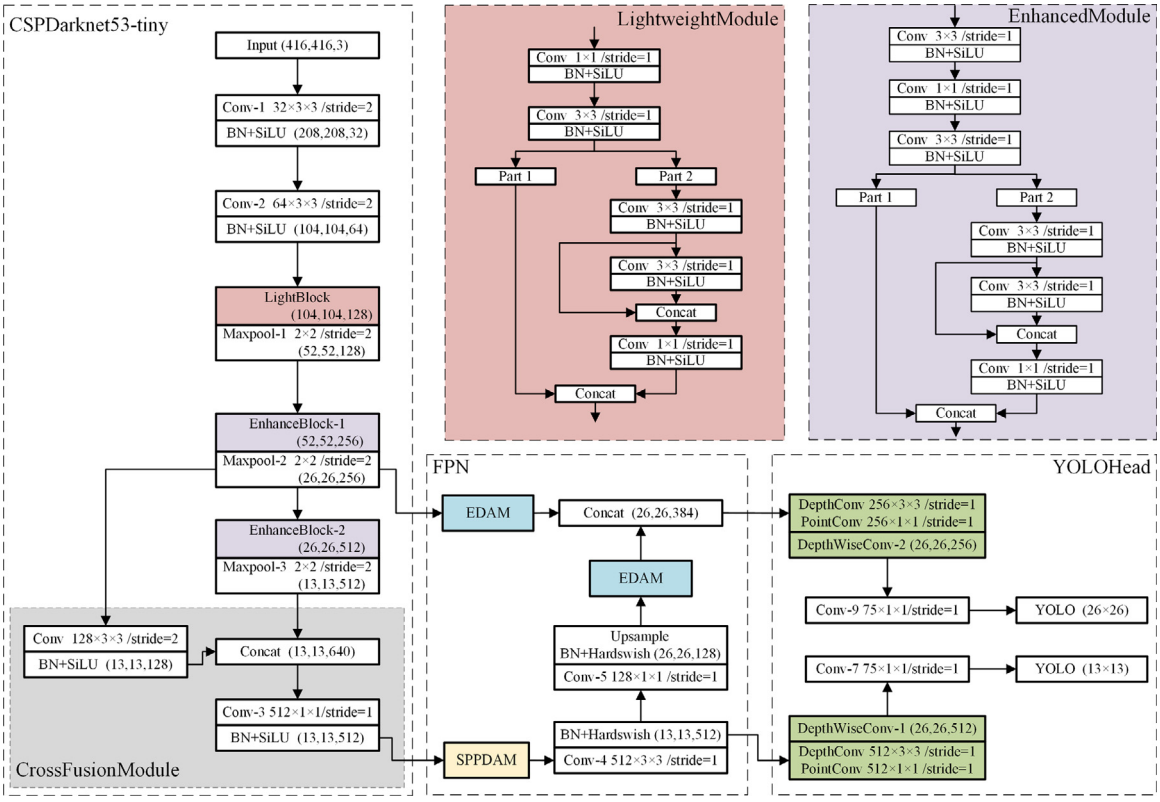

Fig. 2. Structure scheme diagram of SMD-YOLO detector.

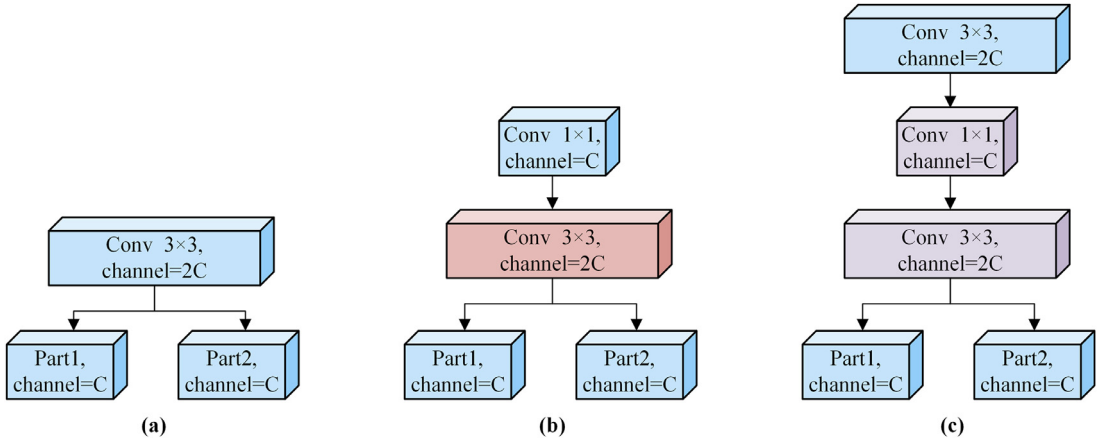

Fig. 3. Structure diagrams of partial residual module. (a) The first 3 × 3 convolution layer of original CSPBlock. (b) Design of an enhanced module based on (a) in SMD-YOLO. (c) Design of a lightweight module based on (a) in SMD-YOLO.

Taking the first CSPBlock (image size 104 × 104) as an example, the quantity of parameters in Fig. 3**(a), (b), (c)** are 68 608, 89 088, and 52 224, respectively. Hence, the improved residual module can be optimized as an enhanced module with stronger feature extraction capacity, or also as a lightweight module with fewer calculation parameters. Considering not increasing too many parameters and calculation meanwhile, we place a lightweight module in the front and 2 enhanced modules in back to improve detection performance for small or medium objects.

3.2.2. Cross fusion module

Deep convolutional networks tend to lose key location information of small objects when extracting feature map information. Generally, the deepest feature map of the network contains only a single layer of semantic information, resulting in less semantic information obtained by the shallow feature map through the reverse path, which is not conducive to the detection of small targets. Whereas the feature maps of different scales include different feature information and are more adaptable to objects of different sizes. To enhance the expressive power of deep features, we design a lightweight deep feature cross fusion module increased on the basis of the original backbone network. The global and local contextual information of deep multi-scale feature maps is extracted using parallel paths, and the semantic information of the deepest feature maps is fused. It helps to determine the exact location of different objects and better resolve the local ambiguity problem. The design scheme of a cross fusion module is demonstrated in Fig. 4.

Down-sampling the feature map with 26 × 26 resolution after Maxpool-2 to convert to 13 × 13 resolution by means of a 3 × 3

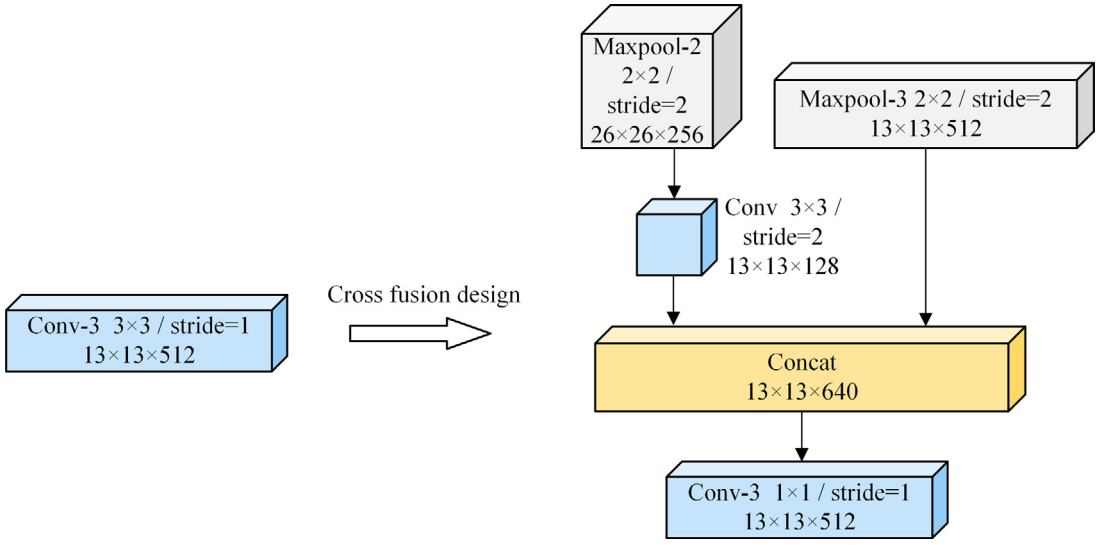

Fig. 4. Design scheme of a cross fusion module.

convolutional layer with stride 2. Subsequently, it is concatenated with the 13 × 13 feature map after Maxpool-3 to form a feature map possessed 13 × 13 resolution and 640 channels. Then, integrating the features through a convolutional layer with kernel size 1 and stride 1, the number of channels is restored to the original 512. The above operations, a substitute of the Conv-3 which is taken up the most calculations and parameters in the original backbone, would be used to formulate a cross fusion module. The purpose of our work is to reduce the entire quantity of convolution parameters, and to improve the lightweight and multi-resolution fusion capability of the network. Thereby, the detection ability of the smaller objects can be optimized.

In terms of the parameter number, the cross fusion module reduces drastically to 622,592, which is 26.39% of the original 3 × 3 standard convolution 2359,296. Moreover, in the backbone, it fuses the 13 × 13 and 26 × 26 feature maps so that the input of the subsequent FPN has richer semantic information, which is beneficial to improve the utilization of shallow features.

3.2.3. Depthwise separable convolution module

In the YOLO Head, for object detection the detector extracts 13 × 13 and 26 × 26 feature layers, which implements classification forecasting via two conventional convolutions in terms of structure. Howard et al. [29] have confirmed in their research that depthwise separable convolution (DSC) can effectively decrease the amount of parameters and calculation compared with standard convolution.

Different from the standard convolutional operation, the depthwise separable convolution operation is divided into depthwise convolution (DWConv) and pointwise convolution (PWConv). Although it greatly reduces the parameters of the model, it is important that the feature extraction capability of the convolutional layer is basically unaffected. At the same time, it also expands the activation range of neurons and effectively improves the accuracy of model recognition. Assuming that a DSC layer with $K \times K$ depthwise convolution has the same quantity of the input and output channel as the standard $K \times K$ convolutional layer. Thus, the quantity of parameters in this layer can be calculated as $K \times K \times C_{\text{int}} + C_{\text{int}} \times C_{out}$. We replace the 3 × 3 conventional convolution of detection head with a depthwise separable convolution to output the position and category confidence information of the mask. The structure and parameters of the improved convolutional layer are respectively shown in Fig. 5 and Table 2.

Table 2
Quantity of convolutional layer parameters comparison.

Scale	Layer	$K \times K$	C_{int}	C_{out}	Parameters
26 × 26	standard	3 × 3	384	256	884,736
	DSC				101,760
13 × 13	standard	3 × 3	512	512	2359,296
	DSC				266,752

Therefore, the parameters of a DSC layer are less than 1/8 of that of a standard 3 × 3 convolutional layer, saving a lot of parameters. It means the improved can speed up the training and test processes. Using the DSC in the YOLO Head, it greatly filters non-target information namely background and environment to improves target detection accuracy.

3.2.4. Enhanced dual attention mechanism

With the primary objective of intensive feature expression ability, an attention mechanism is introduced into the model through training fewer parameters so that the important region of an input image is focused [55–57]. Quite a few attention modules, such as squeeze and exception network (SENet) [58], convolutional block attention module (CBAM) [59], efficient channel attention (ECA-Net) [60], etc., have been proposed successively and been proved the effectiveness of them. Compared with SENet, which only pays attention to channel features, CBAM is an attention module that combines the spatial and the channel to effectively help the transmission of information in the network. It can enhance useful features in feature maps, suppress useless features, and achieve better results in practical applications. While ECA is based on SENet improvement. It only uses a small number of parameters using a method of adaptively selecting the size of one-dimensional convolution kernels to achieve performance improvement. In this paper, we propose a fresh effective dual attention module (EDAM) for DCNN combined CBAM with ECA.

EDAM is still composed of a channel and spatial attention module, enhancing the important spatial and channel features in the feature map so that the network can grasp the "key" learning of the target features during the training process, as shown in Fig. 6. Firstly, the input feature maps are processed through the channel attention module. Global average pooling F^c_{avg} and global max-pooling F^c_{max} are used to extract richer high-level features. A fast one-dimensional (1D) convolution under adaptive selection of k is

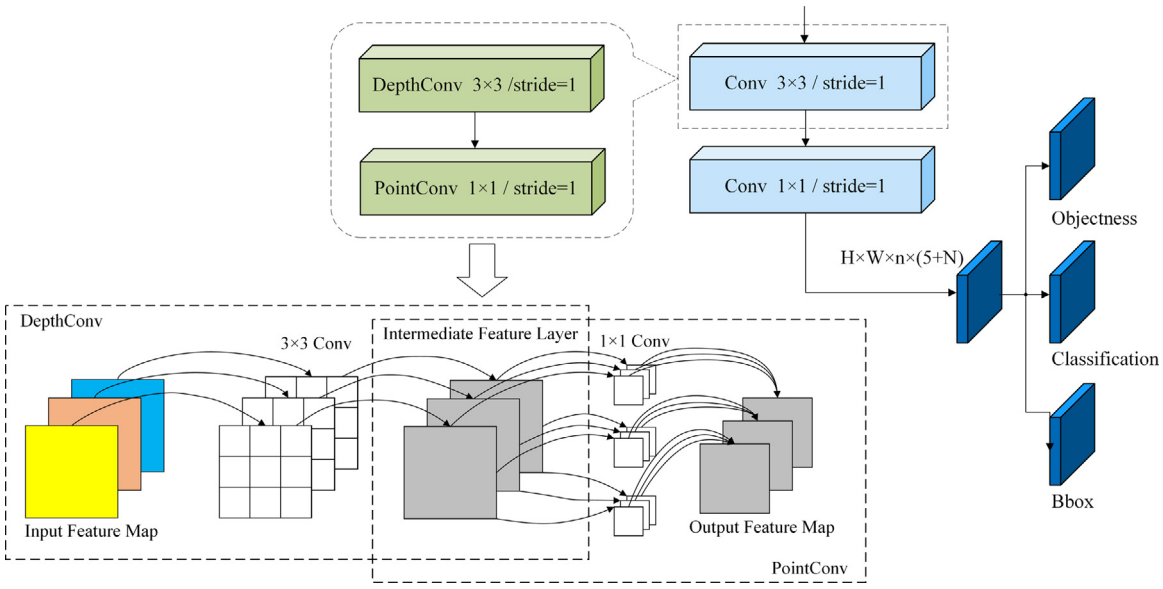

Fig. 5. Structure diagram of the improved prediction network.

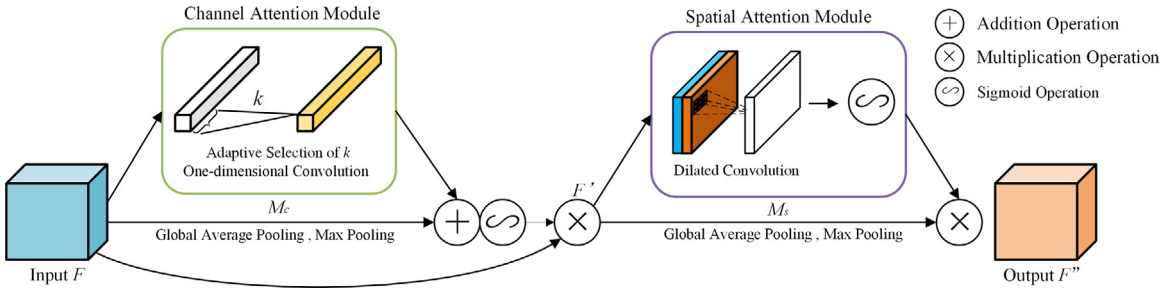

Fig. 6. Schematic diagram of an EDAM module.

adopted to aggregate the features in the k neighborhood channels. Then, the features generate the final channel attention module output via a sigmoid activation, and the corresponding elements of the output are multiplied. In short, the channel attention is computed as equations (1):

$$\begin{aligned} M_c(F) &= \sigma\left(Conv_{1d}^{k\times k}(AvgPool(F)) + Conv_{1d}^{k\times k}(MaxPool(F))\right) \\ &= \sigma\left(Conv_{1d}^{k\times k}\left(F_{avg}^{c}\right) + Conv_{1d}^{k\times k}(F_{max}^{c})\right) \end{aligned} \tag{1}$$

where σ denotes the sigmoid function and $Conv_{1d}^{k\times k}$ represents 1D convolution with kernel size k adaptively determined. The formula used for computing k is given by equations (2):

$$k = \left|\frac{\log_2 C + b}{\gamma}\right|_{odd}, \gamma = 2, \quad b = 1 \tag{2}$$

where C denotes channel dimension, and $|^*|_{odd}$ indicates the nearest odd number of *. In this paper, γ and b are both hyperparameters and set to 2 and 1 throughout subsequent experiments, respectively.

The spatial attention module mainly explores the internal relationship of feature maps at the spatial level, that is, the importance of salient regions, and complements the channel attention module. The output of the channel attention module is used as the input required by the spatial attention module. Then, the feature map of the spatial attention module is obtained after a dilated convolutional layer with 7 × 7 kernel size. Dilation is the introduction of a new parameter called the dilation rate into the standard convolution. The dilation rate is used to control the spacing of each value when the convolution kernel processes the data to realize the increase of the convolution layer's receptive field under the condition of the same amount of computation. In brief, the spatial attention is computed as Equations (3):

$$\begin{aligned} M_s(F') &= \sigma\left(f_{dilation}^{7\times 7}\left(AvgPool(F')\right); f_{dilation}^{7\times 7}\left(MaxPool(F')\right)\right) \\ &= \sigma\left(f_{dilation}^{7\times 7}\left(F'^{s}_{avg}; F'^{s}_{\max}\right)\right) \end{aligned} \tag{3}$$

where σ denotes the sigmoid function and $f_{dilation}^{7\times 7}$ represents a dilated convolutional operation with the filter size of 7 × 7. The dilation parameter of convolutional operation in our work is set to 2.

EDAM combines the advantages of both CBAM and ECA, which enhances ability of the model to extract features. It not only captures information across channels, but also senses the size and the position features of the face masks, which enables the model to more accurately identify targets and lock onto target locations.

3.2.5. Spatial pyramid pooling dual attention module

Spatial pyramid pooling in deep convolutional networks (SPP) [52] is a structure in which the feature maps are merged into a fixed-length feature vector through cross aggregation operations. Where the Max-pool layer of SPP enlarges the receptive field while maintaining the translation invariance of the feature map. The SPP module obtains the local receptive field and near-global receptive field information of the feature map by using the Max-pool layer of different kernel sizes, and performs feature fusion. The operation of fusing the receptive fields of different scales can effectively

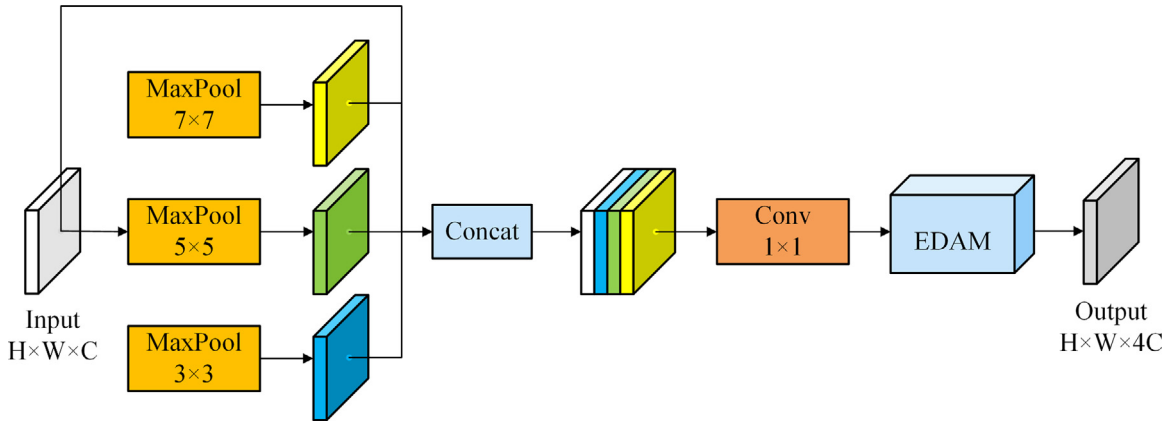

Fig. 7. Structure diagram of a SPPDAM module.

enrich the expressive ability of the feature map and enhance the acceptance range of the output features of the backbone network. Moreover, it also can separate important contextual information, and effectively improve the detection performance of the model. The problems about information loss and non-uniform scale could be solved. For strengthening the ability to detect small or medium-size face masks of different scales, we propose an improved spatial pyramid pooling dual attention module (SPPDAM) based on SPP to enhance the receptive field of deep semantic features. The structure of a SPPDAM module is exhibited in Fig. 7.

Considering that there are relatively many face mask objects with dense crowds, the max-pooling kernel size of SPP should be matched to the size of feature map that needs to be pooled as much as possible. Therefore, we modify respectively the original max-pooling kernel size to 3, 5, and 7. In this way, the most significant features could be retained at diverse scales, and the receptive field of feature maps for local region would be heightened. In the SPPDAM module, the max-pooling operations are executed for an input feature map (H, W, C) with three different kernel sizes of 3×3, 5×5 and 7×7 under stride 1. Where padding around the feature map is set to 1,2 and 3 to ensure that the input and output dimensions of the feature map remain unchanged. Concatenating the input to the feature maps equipped with cross local receptive field after that, output result (H, W, 4C) for the mask location is obtained through feeding into EDAM. Hence, more abundant local feature information could be acquired, and more mask features would be captured by the network to enhance the expression ability of feature map feature information and achieve better detection effect.

3.2.6. Modified activation function

For faster detection, the LeakyReLU activation function is employed for YOLOv4-tiny. Although LeakyReLU solves the problem of dead neuron, it cannot provide a consistent relationship prediction for positive and negative inputs so that the accuracy will inevitably decrease. To prevent the accuracy from falling too seriously, the activation function of the network is modified to compensate the accuracy.

Due to the maximum's unlimitedness, smoothness and non-monotonicity of SiLU, the saturation can be avoided, the nonlinearity can be increased meanwhile. When the input value is greater than 0, the SiLU and ReLU activation functions are roughly equal. When the input value is less than 0, the activation function approaches 0. More importantly, SiLU has better stability. It has been proved in Elfwing's experiments [61] that the global minimum can act as a "soft bottom" to suppress the updating of the weight value and avoid gradient explosion in the case that the derivative is 0. Thus, we use SiLU instead of LeakyReLU to become the activation function of backbone, the shallow part, of the network.

Meanwhile, in the deep of network, the Swish function can provide better accuracy than the ReLU without sacrificing too much detection speed [62,63]. In a lightweight model, computing the sigmoid function is expensive. In order to reduce the computational cost, a Hardswish function with comparable Swish performance is used [64]. While the Hardswish, proposed by Howard [31] in the MobileNetv3 architecture, was only used at the second half of their model. Because it is confirmed that most of the benefits swish are realized by using them only in the deeper layers by those authors. Thus, we merely utilize the Hardswish function behind the backbone in our network. The formulas used for computing SiLU and Hardswish is given by Eqs. (4) and (5):

$$SiLU(x) = x \cdot sigmoid(x) \tag{4}$$

$$Hardswish(x) = \begin{cases} 0 & if\ x \le -3, \\ x & if\ x \ge -3, \\ x \cdot (x+3)/6 & otherwise \end{cases} \tag{5}$$

3.3. Anchor boxes dimension clustering

In the object detection task, selecting an appropriate anchor can significantly improve the detection speed and accuracy. The size of anchor boxes in the original YOLO v4-tiny is obtained using K-means clustering algorithm. According to the K-means algorithm [65], the problem with its initialization is that final clusters might produce clusters incorrectly partitioned. To solve this drawback, K-means++ [66], an improved version based on K-means, is utilized for the face mask data set, which selects the initial centroids from data points that are far away from one another to build better clusters. The main algorithm for implementing K-means++ is presented in Algorithm 1. The input of the algorithm not only includes category information but also the location and size information of annotation frame relative to the original image, namely, the text file of $(x_i,y_i,w_i,h_i),i \in \{1, 2, \ldots, N\}$.Where (x_i,y_i) is the center coordinate of the annotation box, (w_i,h_i) is the actual width and height, and N is the total number of annotation boxes. The output is the width and height of k groups of anchor boxes$(w_i,h_i),i \in \{1, 2, \ldots, k\}$.

For a selected position in the face mask data set, the algorithm first defines an initial centroid $C_{[w_1,h_1]}$ by choosing a box at random (lines 01 to 04 in the algorithm). Then, the shortest distance d_{IOU} in Eq. (6) between the existing cluster center and the rest of anchor boxes is calculated. Calculate the probability P in Eq. (7) that each box is selected as the next cluster center as well. And add up above

Algorithm 1
K-means++ clustering algorithm.

```
01:  Initialization {
02:  define label path to obtain (x_i,y_i,w_i,h_i),i ∈ {1, 2, ..., N}
03:  define k, N, d_IOU, and grid size}
04:  C_[w_1,h_1] ← getRandomPoint (N)
05:  Main loop
06:  while (k ≤ 6)
07:    for each (w_i,h_i),i ∈ {1, 2, ..., N}
08:      calculate D[i] = d_IOU(x_[w_i,h_i], C_[w_1,h_1]),P[i] // the shortest distance measured by IOU
09:      accumulate Sum(D[i])
10:    end for
11:    for each (w_j,h_j),j ∈ {2, ..., N}
12:      r ← getRandomValue (Sum(D[j]))
13:      calculate r − =D[j]
14:      if r ≤ 0
15:        C_[w_j,h_j] is the next cluster center // update a new cluster center
16:    end for
17:  k = k + 1
18:  end while
19:  print(C_[w,h]) // obtain all the initial centers of clustering
20:  run the standard K-means algorithm utilizing above k initial clustering centers
```

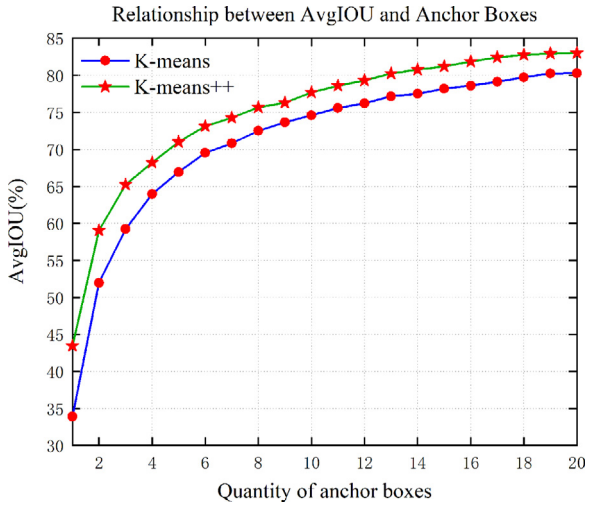

Fig. 8. Relationship of AvgIOU and the quantity of anchor boxes.

distances to get $Sum(d)$ (lines 07 to 10 in the algorithm)

$$d_{IOU} = 1 - IOU(box, centroid) \tag{6}$$

$$P = \frac{d_{IOU}^2}{\sum_{i=1}^{n} d_{IOU_i}^2} \tag{7}$$

when selecting a new cluster center, first take the random value r that falls between 0 and$Sum(d)$. If there is r satisfying $r \le D[j]$, $j = \{2, 3..., N\}$, then $C_{[w_j,h_j]}$ is the next cluster center (lines 11 to 16 in the algorithm). Repeat the above steps until k cluster centers are screened out. Finally, run the standard K-means algorithm utilizing above k initial clustering centers to obtain the width and height of anchor boxes.

The K-means++ algorithm optimizes the center selection method of the initial clustering and greatly reduces the dependence of clustering results on the k value to obtain the better clustering effect. As the number of anchor boxes increases, the calculation amount of the model will also increase, shown in Fig. 8. To ensure the accuracy of the prediction results and avoid selecting too many anchor boxes brings a huge amount of computation, therefore, our work selects 6 anchor boxes consistent with the number of the original detector to achieve a good balance between the complexity and recall rate of the model.

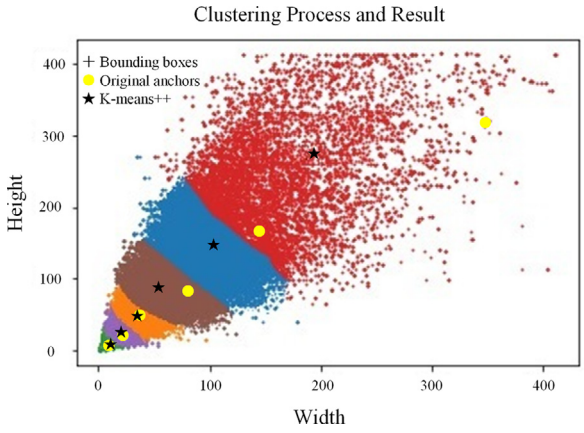

Fig. 9. Distribution of bounding boxes and anchor boxes.

The size of Anchor box is set relative to feature map. When the quantity of anchor boxes is set to 6, the corresponding AvgIOU is 73.12%. Compared with the K-means algorithm, the AvgIOU of the anchor boxes clustered by K-means++ algorithm is improved by 3.58%, that is, the degree of overlap between the anchor box and the bounding box is improved. Then, the K-means++ clustering experiment is repeated for 20 times, and the size of anchor box corresponding to the highest average IOU value in these 20 experiments is finally selected. Therefore, the more suitable width and height of the anchor boxes for detecting face masks are obtained as (11, 16), (21, 33), (35, 58), (56, 98), (105, 155) and (192, 275). The scheme diagram of clustering process and result is shown in Fig. 9. It expresses the distribution of the width and height of bounding boxes. Meanwhile, the difference between the anchor boxes before and after clustering can be seen from the figure. Where the axis of coordinates represents the width and height of the face mask targets' bounding boxes in the images respectively.

4. Results

4.1. Data set

To verify the effectiveness of our proposed detector, we adopt the public face masks detection dataset released by Kumar et al. [50] in 2021. The reason for choosing this dataset is that the releasers themselves have used it and published the research achievement [51] in public. In the meanwhile, their work has the same application scenario as us, aiming to meet the challenge of

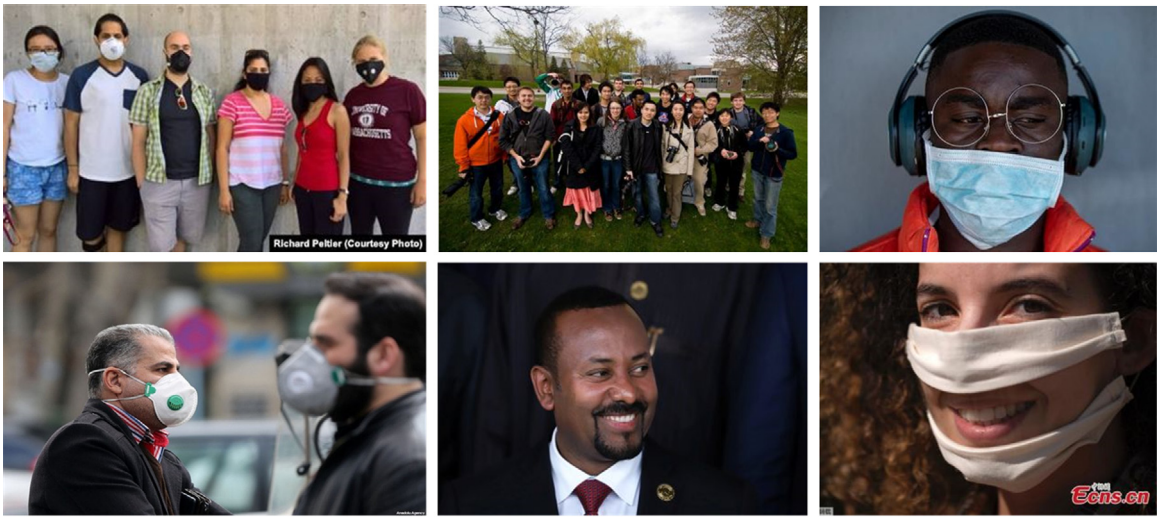

Fig. 10. Image samples of data set for different class labels. (a) With mask (WM). (b) Without mask (WOM). (c) Mask worn incorrectly (WMI) and Mask area (MA).

low accuracy, slow speed and high false detection rate of face mask detection.

The original approximately 11,000 images of this dataset were created by crawling images from the internet using Google and Bing APIs, and resized to a size of 416 × 416 to meet the input size requirement of the YOLO network. Factors such as mask type and color are fully considered to meet the richness of the data set. The dataset contains images of people wearing face masks, medical masks and people not wearing face masks in four classes with labels with mask (WM), without mask (WOM), mask worn incorrectly (WMI) and mask area (MA). It should be noted that an image is not only one of the four classes but may contain two, three or all classes at the same time. Approximately 50,000 bounding boxes were applied over 11,000 images to gather rich and precise information for each class of the dataset. Moreover, to enhance the size of the dataset, the strategy of data augmentation, such as rotation, shearing, flipping, and shift, was employed, increasing the dataset from 11,000 images to 52,635 images. The dataset is divided into training, validation and test set in a ratio of 8:1:1 respectively to obtain the training set with 42,115 images, test set with 5260 images, and validation set with 5260 images. The three sets are independent of each other and marked as a text document in YOLO format. The image samples corresponding to different class labels are exhibited in Fig. 10.

As shown in Fig. 11, the visualization results are obtained by analyzing the dataset. Fig. 11**(a)** demonstrates the categories of face mask objects, and the number of bounding boxes for each class. And Fig. 11**(b)** reveals that the normalized distribution of bounding boxes' center points. The darker the color in the picture is, the more concentrated the center point of the target is. Combined with Fig. 9, it can be seen that the distribution of objects' center in the data set is relatively uniform, and the proportion of small or medium-size mask objects is larger. However, there is a phenomenon that the quantity of objects is not balanced in categories, and there is occlusion between objects, which is in line with daily practical application scenarios.

4.2. Experiment environment and settings

4.2.1. Experimental environment and hyperparameter

In our experimental environment, the performance of SMD-YOLO proposed is evaluated on seven NVIDIA GeForce RTX 2080 Ti GPUs with 11 GB of RAM. The CUDA version is 10.1 and the cuDNN version is 7.6. The details of the hardware and software environment are shown in Table 3.

Table 3
Hardware and software environment.

Device	Configuration
Operating System	Ubuntu 18.04.5 LTS
Processor	Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz
GPU Accelerator	CUDA 10.1, cuDNN 7.6
GPU	GeForce RTX 2080Ti (× 7), 11G
Framework	PyTorch
Compilers	Spyder, Anaconda
Scripting Language	Python 3.6

We utilize the PyTorch deep learning framework and deploy the aforementioned tunings. During the training phase, a few hyperparameters need to be initialized. By virtue of transfer learning [67], the network adopts the strategy of freeze training and unfreeze training, which can speed up the training efficiency and prevent the weight from being destroyed. The hyperparameter initialization of SMD-YOLO is shown in Table 4.

4.2.2. Evaluation indicators

To better evaluate and compare the novel detector properly, we primarily adopt the following five indicators: precision rate (P), recall rate (R), mean value of average precision (mAP), F1-score (F1) and frames per second (FPS). These indicators have been widely used for classification and detection visual tasks. The formulas used for computing these indicators is given by Eqs. (8)–(12):

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

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

$$F1 - score = \frac{2 \times P \times R}{P + R} \tag{10}$$

$$AP_i = \int_0^1 P_i(R_i) dR_i \tag{11}$$

$$mAP = \frac{1}{n} \sum_{i=1}^{n} AP_i \tag{12}$$

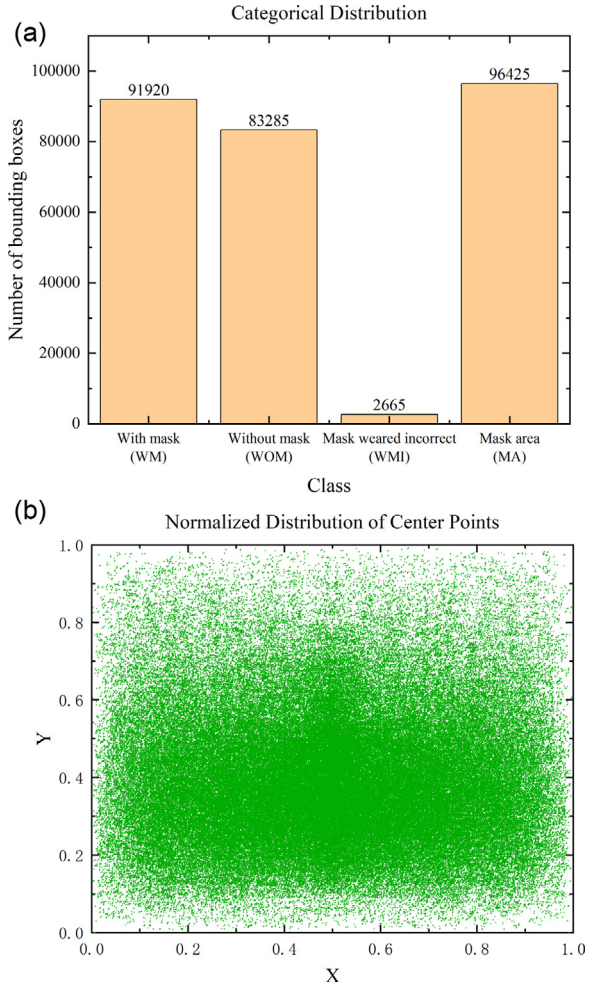

Fig. 11. Visualization analysis diagrams of the face mask detection dataset. (a) Categorical distribution. (b) Normalized distribution of center points.

Where TP, FP, and FN represent the quantity of true positive samples, false positive samples, and false negative samples, respectively. The precision rate denotes the proportion of the real target predicted by the detector, and the recall rate indicates the proportion of all real targets detected. The F1-score is generally in the form of the harmonic mean of precision rate and recall rate. The higher the F1-score is, the better the detection effect of the detector is, normally. The AP represents the area under the Precision-Recall (P-R) curve, and is a metric used for evaluating the overall performance of each category on the test model. While the mAP calculates the mean of all APs for each category to determine the whole detection ability of a detector. Using the mAP, as the primary key indicator, can justify a detector that performed best overall to detect face mask objects specifically. FPS denotes the frames rate per second, that is, the number of images that can be processed per second, and it is used to evaluate the speed of object detection. The larger the value is, the faster the detection speed is.

While evaluating the test performance of the detector through mAP and FPS, we consider the quantity of parameters (Params) and the floating-point operations (FLOPs) of the detector as well. FLOPs can be understood as the amount of calculation, used to measure the complexity of the model. It is generally a very large number, and this article uses BFLOPs (billion float operations) as the unit. Params represents the quantity of parameters of the model, which directly determines the weight size of the model, and affects the memory usage during inference. For the convolutional layer, the formula used for computing FLOPs is given by Eqs. (13):

$$FLOPs = 2 \times H \times W \times \left(C_{in}K^2 + 1\right) \times C_{out} \tag{13}$$

Where W, H respectively represent the width and height of an input feature map, C_{in},C_{out} respectively denote the number of input and output channels, and K is the size of the convolution kernel.

4.2.3. Experimental settings

Single-factor and combined-factor ablation experiments are set up to verify the effectiveness of the proposed improvements. To further demonstrate the superiority of SMD-YOLO, it has been compared with other SOTA object detectors in the practical application of face mask detection. These SOTA detectors for comparison contain not only YOLOv4, EfficientDet-D1, YOLOX_s, YOLOv3-tiny and YOLOv4-tiny, but also a few detectors combined with lightweight module, such as EfficientNetv2-Yolov3, MobileNetv3-YOLOV4, GhostNet-YOLOV4 and MobileNetv2-SSD.All the experiments are pre-trained on the PASCAL VOC2007 and 2012 data set in advance. After obtaining the pre-training weights, transfer learning is used for training and comparison tests on the data set. Mosaic data augmentation, label smoothing and cosine annealing learning rate are utilized in all experiments. In other words, these comparison detectors and SMD-YOLO have the same hardware configuration and software environment.

4.3. Ablation experiment results

We divide the aforementioned improvements into five-part factors. Where KC represents the modified anchor boxes with K-means++ clustering. LEM denotes the improved backbone network adopting lightweight module, enhanced module and cross fusion module. SPEA refers to the enhanced feature extraction network with SPPDAM and EDAM. SIHW indicates the modified activation functions with SiLU and Hardswish. DW represents the improved detection head with depthwise separable convolution. The AP_{WMI}, AP_{WM}, AP_{WOM} and AP_{MA} denote the AP values of WMI, WM, WOM and MA, respectively.

The results of different experimental schemes on model performance are shown in Table 5 at an IOU value of 0.5. The No. 1 is the original YOLOv4-tiny detector reproduced, and it is the baseline. And the No. 2∼6 are the experiments adding a certain improvement factor to the original, that is, the single-factor experiment. The purpose of setting up the single-factor experiment is to prove

Table 4
Hyperparameter initialization of SMD-YOLO.

Hyperparameter	Value	Hyperparameter	Value
Size of Input Image	416 × 416	Freeze Training Mode	True
Batch Size	32	Freeze Epoch	50
Momentum	0.9	Freeze Learning Rate	0.001
Mosaic	True	Training Epoch	500
Cosine Annealing Learning Rate	True	Unfreeze Learning Rate	0.0001
Label Smoothing	0.005	Optimizer	Adam

Table 5
Results of different experimental schemes on model performance.

N	KC	LEM	SPEA	SIHW	DW	AP_{WMI}/%	AP_{WM}/%	AP_{WOM}/%	AP_{MA}/%	mAP/%	FPS/(f·s^{-1})
1	×	×	×	×	×	3 1.91	81.56	60.81	75.52	62.45	127.91
2	√	×	×	×	×	37.55	81.36	63.13	76.24	64.57	130.22
3	×	√	×	×	×	40.61	80.09	61.84	77.66	65.05	113.08
4	×	×	√	×	×	34.96	82.95	65.88	75.93	64.93	104.94
5	×	×	×	√	×	34.58	82.32	63.36	75.74	64.00	121.39
6	×	×	×	×	√	33.84	81.35	66.75	75.61	64.39	132.59
7	√	√	√	×	×	39.31	83.06	64.95	76.03	65.84	94.35
8	√	√	√	×	√	41.72	82.54	67.11	75.36	66.57	97.27
9	√	√	√	√	×	42.30	82.25	64.84	77.25	66.66	88.78
10	√	√	√	√	√	41.91	82.11	68.12	75.88	67.01	92.81

the contribution degree for mAP value under a single factor condition. As seen from No. 2∼6 experiments in the table, each improvement has an increase for mAP value of the algorithm. Where there is the greatest impact on the mAP value from 62.45% to 65.05% through using the enhanced, lightweight and cross fusion modules in the backbone of the network, increased by 2.6%. Then, adopting a combination of SPPDAM and EDAM reaches 64.93%, 2.48% higher than the baseline. Meanwhile, choosing suitable anchor boxes and depthwise separable convolution module can improve the prediction accuracy by about 2% while basically ensuring the detection speed. In addition, the modified activation functions have a slight increase in overall accuracy. However, it is unavoidable for decreasing the detection speed because of the addition of new modules increasing the model complexity.

Hui [68] and Liu [69] have proved in their experiments that the multi-factor combination strategy of selecting the appropriate anchor boxes, modifying the backbone network and increasing the attention mechanism can improve the value of mAP accordingly. Therefore, we also implement a combination of these three improvements simultaneously, as No. 7 experiment. It achieves the mAP value of 65.84%, 3.39% higher than the baseline. The No. 8–9 comparison experiments are to test an impact on mAP adding the deep separable convolution and activation function respectively based on the No. 7 combined strategy. The results prove that the mAP is risen by 0.73% and 0.82% on the basis of No. 7 experiment, respectively reaching 66.57% and 66.66%. Finally, the No. 10 experiment is our proposed model, which achieves the mAP value of 67.01%. This shows that the mAP is risen by at least 0.35% over the No. 7–9 experiments and 4.56% higher than the No. 1 experiment. Meanwhile, the detection speed reaches the FPS value of 92.81.

The P-R curves of the four categories are demonstrated in Fig. 12 in the No. 1 and No. 10 experiments. Meanwhile, the area under the PR curve and enclosed by the coordinate axis of each category is the AP value of the corresponding category. Where for each category the blue curve is the P-R curve of the original YOLOv4-tiny, while the red curve is the P-R curve of proposed SMD-YOLO detector in our work. The higher the curve is to the upper right, the larger the area enclosed by the curve and the coordinate axis is, the higher the corresponding AP value is and the better the performance is. As shown in the figure, our proposed detector achieves good detection performance.

In the whole, contrasted with the original detector, the AP values of our proposed model increases by 10.00% (WMI), 0.55% (WM), 7.31% (WOM) and 0.36% (MA) for each category, respectively. In terms of AP values for each category, the improvement of WMI and WOM category is more conspicuous. Table 6 indicates the comparison of the performance indicators values for each category between the baseline and our proposed work. It is seen from the comprehensive indicators shown by the F1 value, the WMI, WM, WOM and MA categories reach 0.57, 0.80, 0.68 and 0.77, respectively, greatly improving by 16%, 2%, 6% and 1%.

In order to better understand the effect of the improved network, the heat maps visualization of the baseline and our proposed SMD-YOLO are demonstrated in Fig. 13, which is the visualization of the prediction results. The following images are all from the test set of the face mask dataset, and the size of the face mask target decreases from left to right. Where the darker the color, the more attention the model pays to it. It can be seen that our proposed enhanced dual attention mechanism strengthens the learning of these local features. And it can also prove that the method of improving the performance of the model by integrating and improving attention mechanism is feasible and efficient.

4.4. SOTA comparison results

The recent related work about the face mask detection has basically been on variants of YOLO3, YOLOV4, SSD and Faster-RCNN, incorporating a few single-stage lightweight models as the backbone of network. Therefore, we chose a few typical variant combinations for comparison. The comparisons of detection results are shown in Tables 7 and 8 in the same experimental conditions. These detectors can describe the features of face mask objects accurately and the detection accuracy is high.

The backbone of network and input size of the image, and the detection results for AP indicator of different detectors on the test set is shown in the Table 7. It can be seen from the table, there are a good performance on our proposed SMD-YOLO for the WMI and WOM categories, almost higher than the other lightweight detectors with an input image size of 416 × 416.

However, for the actual applications and occasions such as hospitals, campuses, communities and so on, it is very important for the face mask detection schemes to have a good real-time performance and a relative high accuracy. To further compare the detection effects of our proposed SMD-YOLO with other detectors, we also evaluate the various indicators shown in Table 8.

As seen from the table, the mAP value of our proposed detector is above the medium level. However, our detector is more compet-

Table 6
Comparison of the performance indicators values for each category.

Models	WMI			WM			WOM			MA		
	P/%	R/%	F1	P/%	R/%	F1	P/%	R/%	F1	P/%	R/%	F1
YOLOv4-tiny (Baseline)	59.88	31.29	0.41	83.62	72.58	0.78	67.31	57.67	0.62	80.28	71.32	0.76
Proposed Work	64.34	50.65	0.57	83.78	77.36	0.80	70.36	65.20	0.68	79.66	74.03	0.77

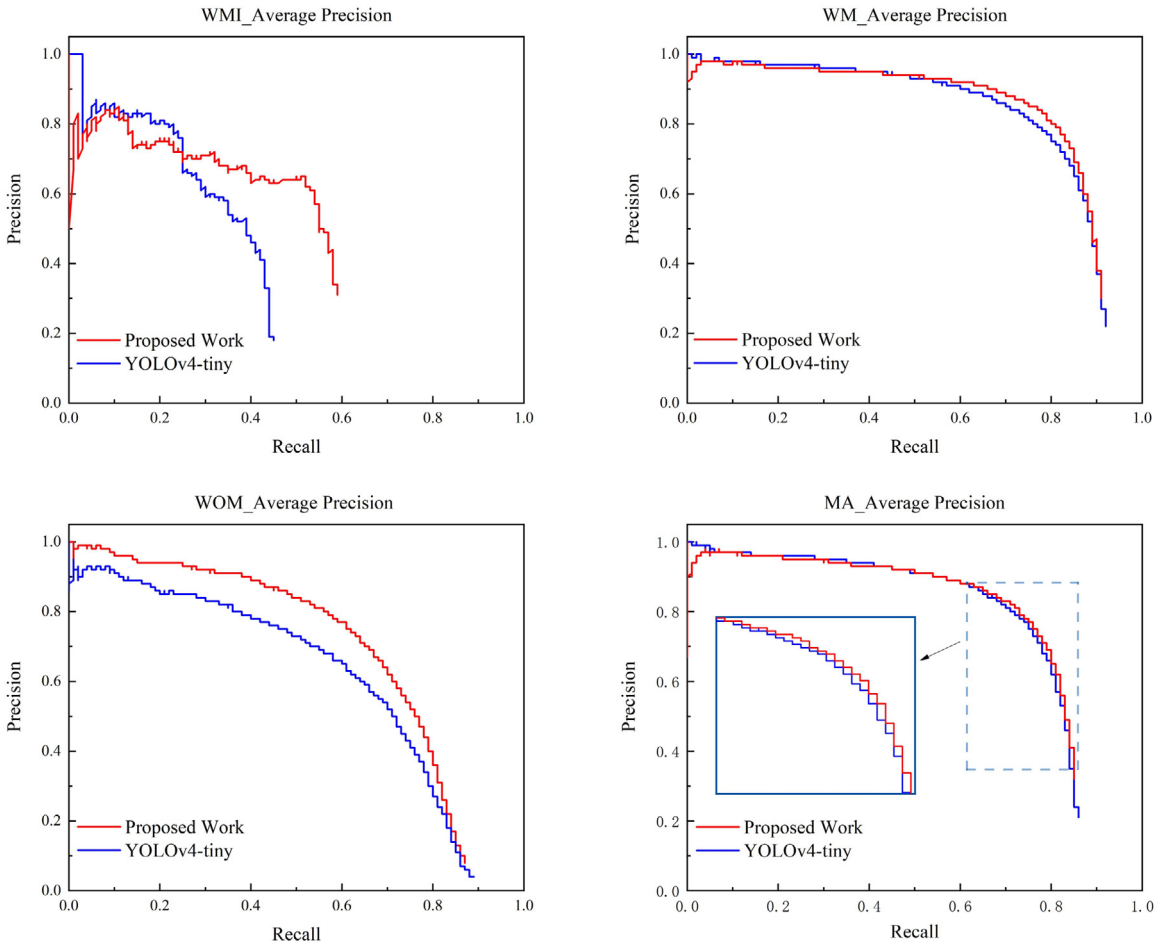

Fig. 12. Comparison of P-R curves of different categories.

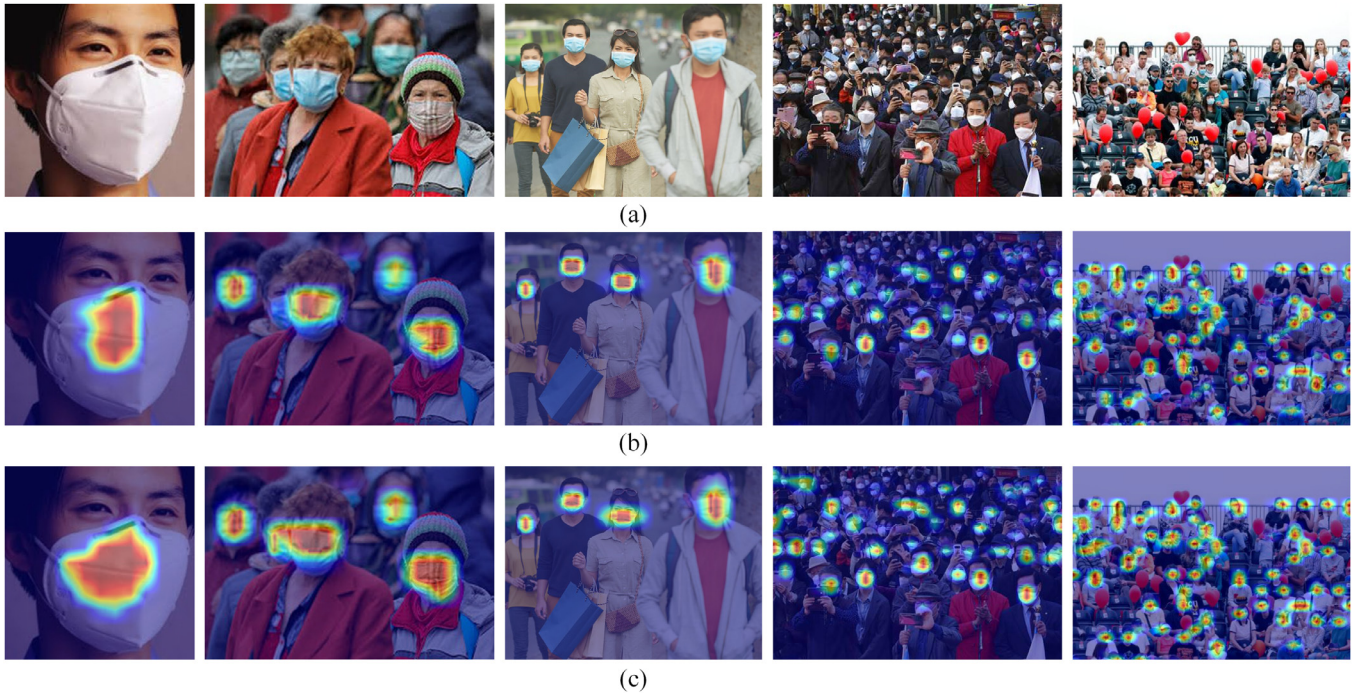

Fig. 13. Comparison of two detectors for visualization of the heat maps. (a)Original images of test set. (b)Baseline. (c)Our proposed SMD-YOLO.

Table 7
Detection results for AP indicator of different detectors on the test set.

Models	Backbone	Input Size	AP_{WMI}/%	AP_{WM}/%	AP_{WOM}/%	AP_{MA}/%
YOLOv4	CSPDarknet53	416 × 416	49.60	86.95	71.26	81.82
YOLOX_s	CSPDarknet	640 × 640	40.47	87.04	70.35	78.20
EfficientDet-D1	EfficientNet-b1	640 × 640	47.23	88.62	64.87	76.69
MobileNetv3-YOLOv4	MobileNet-v3	416 × 416	38.74	84.04	64.88	79.46
GhostNet-YOLOv4	GhostNet	416 × 416	40.25	85.18	65.88	77.68
EfficientNetv2-YOLOv3	EfficientNet-b2	416 × 416	30.90	85.37	68.71	80.46
MobileNetv2-SSD	MobileNet-v2	300 × 300	33.19	84.85	69.25	76.56
YOLOv4-tiny	CSPDarknet53-tiny	416 × 416	31.91	81.56	60.81	75.52
YOLOv3-tiny	Darknet53-tiny	416 × 416	11.24	72.48	48.69	73.86
Proposed Work	Improved CSPDarknet53-tiny	416 × 416	41.91	82.11	68.12	75.88

Table 8
Results for various indicators of different detectors on the test set.

Models	*mAP*/%	FPS/(f·s^{-1})	Params/M	BFLOPs/s	Inference Time /ms	Weight Size/Mb
YOLOv4	72.41	28.75	63.9485	47.82	34.78	244.3
YOLOX_s	69.02	47.01	8.9385	42.62	21.27	34.3
EfficientDet-D1	69.35	36.16	6.5561	8.95	27.65	25.4
MobileNetv3-YOLOv4	66.78	35.54	11.3144	11.26	28.14	53.7
GhostNet-YOLOv4	67.25	29.48	11.0139	10.41	33.92	42.4
EfficientNetv2-YOLOv3	66.36	31.49	8.4601	7.33	31.76	60.0
MobileNetv2-SSD	65.96	113.33	3.9413	2.25	8.82	15.8
YOLOv4-tiny	62.45	127.91	5.8787	10.92	7.82	22.5
YOLOv3-tiny	51.57	116.29	8.7423	9.45	8.60	33.1
Proposed Work	67.01	92.81	3.7573	9.12	10.77	14.4

itive in terms of the overall performance. Compared with the original YOLOv4-tiny, the quantity of parameters in our proposed SMD-YOLO is merely about 3.7573E+06, which is approximately 2/3 of the original quantity, and the least among these detectors. Simultaneously, the FLOPs value reduces from 10.92B to 9.12B, a decrease of about 16.48%. And the size of the weight reaches 14.4 Mb decreases by 36%. The overall performance of our proposed detector is improved, which shows the effectiveness of the novel backbone, neck and head structure proposed in this work.

To visualize the model's performance between these evaluation indicators more intuitively, we adopt a normalized histogram to represent, shown in Fig. 14. Where the symbol (-) in the figure represents its opposite number, and the numerical values are obtained via the Min-Max standard normalization process. The closer the value of the indicator is to 1, the better the performance reflects.

It can be seen from Table 7, 8 and Fig. 14 that the overall detection performance of SMD-YOLO marked red star is the best for real-time detection, considered accuracy and speed. Among the above lightweight detectors, YOLOX_s, EfficientDet-D1 and GhostNet-YOLOv4 perform better than our proposed detector in accuracy, but far inferior in terms of speed, at least a gap of 45FPS even more. MobileNetv2-SSD and YOLOv3-tiny detect faster than our proposed model, however, their accuracy is not as high as ours. Although the mAP value of EfficientNetv2-YOLOv3 is almost same as our detector, the detection speed which have the FPS of 31.49 is slower. Moreover, YOLOv4 achieves the mAP value of 72.41%, 5.4% higher than our model. The reason is that YOLOv4 has three detection heads, one of which is dedicated to detecting small targets. Nevertheless, the YOLOv4-tiny, same as YOLOv3-tiny, just has two detection head removing one detection path for faster detection. In general, our SMD-YOLO has shown a greater lead in the comprehensive evaluation of various indicators.

The detection results of the proposed SMD-YOLO detector in a few applications with real-world face masks wearing status are illustrated in Fig. 15. These figures show the detection performance of the proposed detector under the varied environments in the images. On the whole, the proposed model always finishes the detection properly. Especially during the COVID-19 pandemic, it can be used as a useful and beneficial tool at hospitals, schools, communities, etc. for detecting people wearing or not wearing a face mask and detecting the mask on any region of the face. Furthermore, it can effectively reduce the workload of the staff who need to maintain the order, and promptly remind those people not wearing face masks or wearing masks incorrectly to take precautions.

Firstly, Fig. 15**(a)–(c)** illustrate the results about the medium individual faces with some distinctive and characteristic face masks, and all the face masks were successfully identified. Next, Fig. 15**(d)–(g)** are applied in the hospital scenario, including queuing to verify the health code and itinerary code at the entrance of the hospitals (a specific kind of epidemic prevention measures in China), and queuing or registering of the small scale in a hospital department. Where **(d)–(f)** gave an almost correct estimate about the face mask targets and the region of area covered by a mask. Then, the detection results show that face masks detection for small size targets seems to perform well and there are merely few missed instances in the indoor and outdoor public areas of dense crowdedness in Fig. 15**(g)–(i)**. Where the crowded scenarios are more complicated scenes that are clear near and blurred in the distance. The reasons for the missed detection are mainly due to the overlapping between masks, the occlusion of masks and body parts (such as head, back, and hands). Besides, the low resolution of the input detection image may cause the mask in small targets to be similar to the background. Afterwards, the detection effect in the nighttime environment is exhibited in Fig. 15**(j)** and **(k)**. The detection results indicate that the proposed detector, which possesses good adaptability to environmental changes, can almost accurately detect the masks wearing status of people. It can deal with the influence of external light source and environment on the detection model. Finally, Fig. 15**(l)** and **(m)** are the applications of detection inside the hospital corridor through the camera of monitor. Its detection effect is directly affected by the pixels of the camera.

Moreover, to fully demonstrate the superiority of our proposed SMD-YOLO, we compared with a few recent related works in the task of face masks detection. Table 9 shows the performance comparison of our proposed detector with the previous works. As can

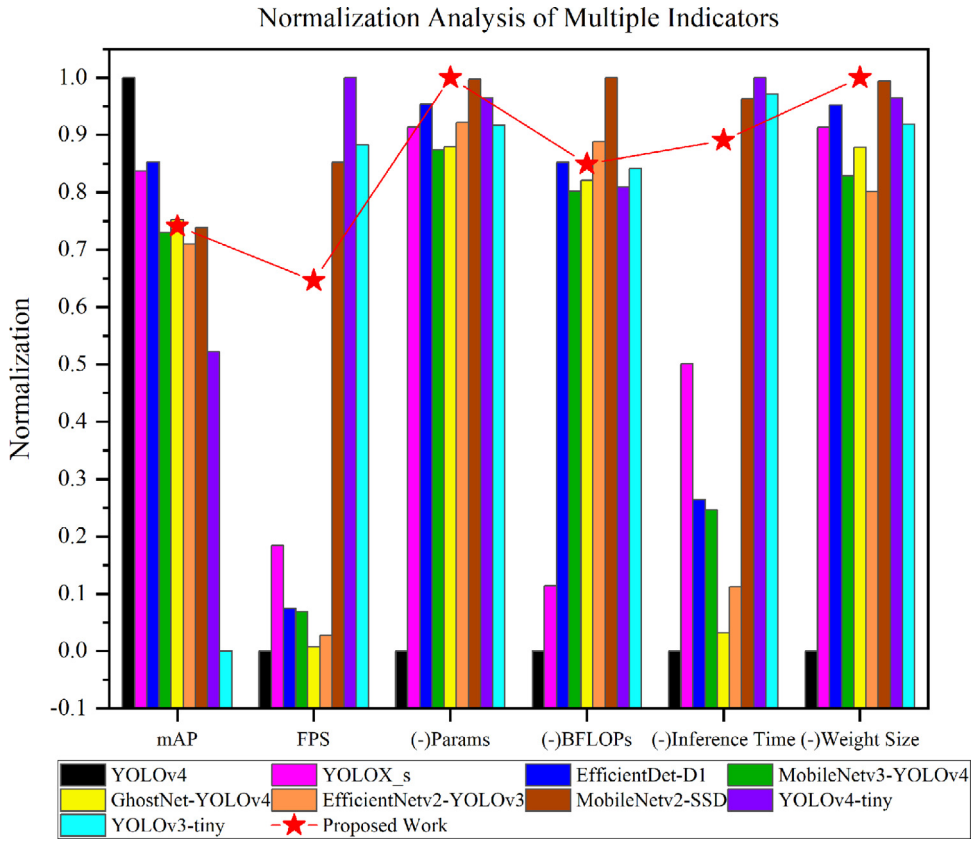

Fig. 14. Normalization analysis of multiple indicators.

Table 9
Performance comparison of the different detectors.

Dataset	Work	Detectors	AP_{WMI}/%	AP_{WM}/%	AP_{WOM}/%	AP_{MA}/%	mAP/%
Face Mask Dataset [50]	Nagrath et al. [48]	SSDMNV2	34.19	85.85	70.25	77.56	66.96
	Kumar et al. [51]	YOLOv4-tiny-SPP	27.64	86.31	58.86	84.42	64.31
	Proposed Work	SMD-YOLO	41.91	82.11	68.12	75.88	67.01
MOXA [53]	Roy et al. [53]	MobileNetv2-SSD	/	/	/	/	46.52
		Inceptionv2-F-RCNN	/	/	/	/	60.05
		YOLOv3-tiny	/	71.48	41.06	/	56.27
	Proposed Work	SMD-YOLO	/	80.94	41.86	/	61.40

be seen from the table, in the dataset [50] employed in this paper, some AP indicators of our proposed detector are obviously better than Nagrath et al. [48] and Kumar et al. [51]. And in the MOXA dataset, the values of evaluation indicators are also higher than Roy et al. [53]. In the whole, our detector performs better than other schemes for the corresponding data set in terms of mAP value. It indicates that our proposed detector is a significant improvement in the face mask detection based on visual images.

5. Discussion

In this work, the effectiveness of our proposed SMD-YOLO model is proved from both single-factor and multi-factor in an ablation study. From the overall experimental results of ablation experiment, these improved models based on multi-factor combined strategy would generate a decrease in FPS values. However, they have little impact on real-time performance in actual use and still meet a real-time requirement, which can take into account the detection accuracy and speed as well.

It needs to be noted that compared with the baseline model of the original detector in Table 6, the AP values of the WM and MA categories are merely increased by 0.55% and 0.36% respectively, which cannot improve too much. However, the recall rate value of the WM category is increased by 4.78%, while its value of the MA category is added by 2.71% sacrificing a little bit of the precision rate. It means that the missed detection rate for area with a mask on the people's face is reduced. From a management and control point of view, it is very important for managers or governmental officers to increase the recall rate value in era of COVID-19 pandemic, especially in areas with stricter supervision. For the detection of people wearing a mask or not and wearing masks incorrectly, all the evaluation metrics values almost have been significantly improved.

Moreover, it can be found from Table 7 that the AP values for each category of the tiny model are nearly lower than other lightweight combination models. The reason for the great difference in the AP values of tiny detectors is that the lightweight network reduces the feature layer and detection layer to varying degrees. Owing to the relatively simple structure, it is difficult to detect small objects in a complex background and crowds of people. Overall, the performance of our proposed detector outperforms the benchmark.

Fig. 15. The detection results of SMD-YOLO detector with real-world face masks wearing status under the various environments. (a)(b)(c) On medium individual faces with a distinctive mask. (d)(e) At the entrance of the hospitals. (f)(g) When small queuing or registering in a small scale at the hospitals. (h)(i) In the densely crowded public areas indoors and outdoors. (j)(k) In the nighttime environment. (l)(m) Inside the hospital corridor from the camera of monitor.

In the SOTA comparison, with the increase in the complexity of the network structure after our improvement, the detection speed is reduced to a certain extent compared with the original detector. However, the FPS value of our detector is still better than most lightweight detector on the test set. In other words, the extremely fast detection of YOLOv4-tiny makes it feasible to sacrifice partial speed to improve the accuracy. In term of the training and testing time, our proposed detector decreases dramatically the time-consuming at least a half, compared to the other detectors in the same experimental environment. Moreover, our proposed detector facilitates deployment on devices with low computing power due to the less weight parameters and calculation amount.

6. Conclusion

In this paper, an efficient and lightweight detection method based on YOLOv4-tiny, and a face mask detection and monitoring system are proposed for mask wearing status, aiming at solving the shortcomings in the small or medium-size masks detection. To improve the detection accuracy on the premise of ensuring the real-time face masks recognition, two feasible improvement strategies are proposed: 1) K-means++ clustering algorithm is to generate anchor boxes suitable for the face mask dataset, making the network easier to train and the parameters to converge more readily. 2) Network structure optimization is to balance the accuracy of detection and speed. Firstly, the improved residual module and cross fusion modules are to extract the features of small or medium-size targets effectively. Next, the enhanced dual attention mechanism heightens the mask features expression and focusing ability of detection model to mask areas. Then, the improved spatial pyramid pooling module strengthen the receptive field of deep semantic features. Besides, the combination of activation functions is benefit for effective and smooth transmission of parameters to compensate the detection accuracy. Finally, the depthwise separable convolution reduces the quantity of parameters to improve the detection speed. The experiment results show that the final mAP value of our proposed model increased from 62.45% to 67.01%, the Params value dropped from 5.8787E+06 to 3.7573E+06, and the FLOPs value dropped from 10.92B to 9.12B compared with the original. With the FPS value of 92.81, the model can significantly improve the detection capacity of small target masks in public places where crowds gather under the premise of ensuring real-time performance.

In future work, the impact of unbalanced data sets will be considered in the preprocessing part. In addition, we will keep working on theoretical and practical application research. In theory, we will study on the lightweight detection of small targets equipped with more efficient feature matching mechanisms. In practical application, we will attempt to use the proposed model in various fields related to the COVID-19 pandemic, such as telemedicine for skin diseases or sore throat, and further do research on the robustness of the model.

Statements of ethical approval

As the data come from the public data set, the informed consent is exempted.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (52165063), the National Natural Science Foundation of China (51865004), the Natural Science Foundation of Guizhou Province (Qiankehe support normal [2021] No. 445), the Natural Science Foundation of Guizhou Province (Qiankehe support normal [2021] No. 172), the Natural Science Foundation of Guizhou Province (Qiankehe support normal [2021] No. 397), theNatural Science Foundation of Guizhou Province (Qiankehe support normal [2022] No. 165), the Natural Science Foundation of Guizhou Province (Qiankehe support normal [2022] No. 272), the Natural Science Foundation of Guizhou Province (Qiankehe basis [2020] No. 1Y232).

References

[1] J.T. Brooks, J.C. Butler, Effectiveness of mask wearing to control community spread of SARS-CoV-2, JAMA- J. Am. Med. Assoc. 325 (2021) 998–999.
[2] G. Cao, L. Shen, R. Evans, Z. Zhang, Q. Bi, W. Huang, R. Yao, W. Zhang, Analysis of social media data for public emotion on the Wuhan lockdown event during the COVID-19 pandemic, Comput. Methods Programs Biomed. 212 (2021) 106468.
[3] V.C.-C. Cheng, S.-.C. Wong, V.W.-M. Chuang, S.Y.-C. So, J.H.-K. Chen, S. Sridhar, K.K.-W. To, J.F.-W. Chan, I.F.-N. Hung, P.-.L. Ho, K.-.Y. Yuen, The role of community-wide wearing of face mask for control of coronavirus disease 2019 (COVID-19) epidemic due to SARS-CoV-2, J. Infect. 81 (2020) 107–114.
[4] C. Sun, Z. Zhai, The efficacy of social distance and ventilation effectiveness in preventing COVID-19 transmission, Sustainable cities and society, 62 (2020) 102390.
[5] C.Y. Liu, J. Berlin, M.C. Kiti, E. Del Fava, A. Grow, E. Zagheni, A. Melegaro, S.M. Jenness, S.B. Omer, B. Lopman, K. Nelson, Rapid review of social contact patterns during the COVID-19 pandemic, Epidemiology 32 (2021) 781–791.
[6] M. Elgendy, C. Sik-Lanyi, A. Kelemen, A novel marker detection system for people with visual impairment using the improved tiny-YOLOv3 model, Comput. Methods Programs Biomed. (2021) 205.
[7] C. Munoz-Lopez, C. Ramirez-Cornejo, M.A. Marchetti, S.S. Han, P. Del Barrio–Diaz, A. Jaque, P. Uribe, D. Majerson, M. Curi, C. Del Puerto, F. Reyes-Baraona, R. Meza-Romero, J. Parra-Cares, P. Araneda-Ortega, M. Guzman, R. Millan-Apablaza, M. Nunez-Mora, K. Liopyris, C. Vera-Kellet, C. Navarrete-Dechent, Performance of a deep neural network in teledermatology: A single-centre prospective diagnostic study, J. Eur. Acad. Dermatol. Venereol. 35 (2021) 546–553.
[8] Z. Liu, L. Wang, Y. Meng, T. He, S. He, Y. Yang, L. Wang, J. Tian, D. Li, P. Yan, M. Gong, Q. Liu, Q. Xiao, All-fiber high-speed image detection enabled by deep learning, Nat. Commun. (2022) 13.
[9] T.K. Yoo, J.Y. Choi, H.K. Kim, I.H. Ryu, J.K. Kim, Adopting low-shot deep learning for the detection of conjunctival melanoma using ocular surface images, Comput. Methods Programs Biomed. 205 (2021) 106086.
[10] Y. Yang, F. Shang, B. Wu, D. Yang, L. Wang, Y. Xu, W. Zhang, T. Zhang, Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image, IEEE Trans. Cybern. (2021).
[11] T.K. Yoo, J.Y. Choi, Y. Jang, E. Oh, I.H. Ryu, Toward automated severe pharyngitis detection with smartphone camera using deep learning networks, Comput. Biol. Med. (2020) 125.
[12] H. Ge, Z. Zhu, Y. Dai, B. Wang, X. Wu, Facial expression recognition based on deep learning, Comput. Methods Programs Biomed. (2022) 215.
[13] K.H. Lin, H.M. Zhao, J.J. Lv, C.Y. Li, X.Y. Liu, R.J. Chen, R.Y. Zhao, Face detection and segmentation based on improved mask R-CNN, Discrete Dyn. Nat. Soc. (2020) 2020.
[14] A.S. Ahmed, H.A. Ibrahim, B.B. Sundaram, P. Karthika, IEEE, small scale targeted face detection using deep convolutional neural network, in: 5th International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC)Electr Network, 2021, pp. 889–893.
[15] H. Li, L.B. Deng, C. Yang, J.B. Liu, Z.Q. Gu, Enhanced YOLO v3 tiny network for real-time ship detection from visual image, IEEE Access 9 (2021) 16692–16706.
[16] Q.S. Fan, H.S. Huang, Y.T. Li, Z.G. Han, Y. Hu, D. Huang, Beetle antenna strategy based grey wolf optimization, Expert Syst Appl (2021) 165.
[17] Q.S. Fan, H.S. Huang, Q.P. Chen, L.G. Yao, K. Yang, D. Huang, A modified self--adaptive marine predators algorithm: framework and engineering applications, Eng. Comput. (2021).
[18] B. Han, Y. Wang, Z. Yang, X. Gao, Small-scale pedestrian detection based on deep neural network, IEEE Trans. Intell. Transp. Syst. 21 (2020) 3046–3055.
[19] G.K.J. Hussain, R. Priya, S. Rajarajeswari, The face mask detection technology for image analysis in the Covid-19 surveillance system, International Conference on Computing, Communication, Electrical and Biomedical Systems (ICCCEBS) 1916, 2021.
[20] M. Besnassi, N. Neggaz, A. Benyettou, Face detection based on evolutionary Haar filter, Pattern Anal. Appl. 23 (2020) 309–330.
[21] Z. Zakaria, S.A. Suandi, J. Mohamad-Saleh, Hierarchical Skin-AdaBoost-Neural Network (H-SKANN) for multi-face detection, Appl. Soft. Comput. 68 (2018) 172–190.
[22] B. Pushyami, C.N. Sujatha, B. Sanjana, N. Karthik, Real-time face mask detection using machine learning algorithm, in: 2nd International Conference on Advances in Computer Engineering and Communication Systems (ICACECS)Electr Network, 2021, pp. 347–357.
[23] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[24] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.
[25] J. Redmon, A. Farhadi, Yolov3: an incremental improvement, arXiv preprint, (2018).
[26] A. Bochkovskiy, C.-.Y. Wang, H.-Y.M. Liao, Yolov4: optimal speed and accuracy of object detection, arXiv preprint, (2020).
[27] Ultralytics, YOLOv5: Open source neural networks in python, Available online: https://github.com/ultralytics/yolov5/. (2020) Accessed 9 June 2020.
[28] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: exceeding yolo series in 2021, arXiv preprint, (2021).
[29] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv preprint, (2017).
[30] A. Howard, A. Zhmoginov, L.-.C. Chen, M. Sandler, M. Zhu, Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation, (2018).
[31] A. Howard, M. Sandler, G. Chu, L.-.C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Searching for mobilenetv3, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1314–1324.
[32] M. Tan, Q.V. Le, EfficientNet: rethinking model scaling for convolutional neural networks, 36th International Conference on Machine Learning (ICML)Long Beach, CA, 2019.
[33] M. Tan, R. Pang, Q.V. Le, Efficientdet: scalable and efficient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10781–10790.
[34] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, C. Xu, Ghostnet: more features from cheap operations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1580–1589.
[35] G. Wang, H. Ding, Z. Yang, B. Li, Y. Wang, L. Bao, TRC-YOLO: a real-time detection method for lightweight targets based on mobile devices, IET Computer Vision 16 (2022) 126–142.
[36] S. Lin, L. Cai, X. Lin, R. Ji, Masked face detection via a modified LeNet, Neurocomputing 218 (2016) 197–202.
[37] S. Ge, J. Li, Q. Ye, Z. Luo, IEEE, detecting masked faces in the wild with LLE-CNNs, in: 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)Honolulu, HI, 2017, pp. 426–434.
[38] S. Hussain, Y. Yu, M. Ayoub, A. Khan, R. Rehman, J.A. Wahid, W. Hou, IoT and deep learning based approach for rapid screening and face mask detection for infection spread control of COVID-19, Appl. Sci.-Basel 11 (2021).
[39] M. Loey, G. Manogaran, M.H.N. Taha, N.E.M. Khalifa, A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic, Measurement 167 (2021).
[40] M. Loey, G. Manogaran, M.H.N. Taha, N.E.M. Khalifa, Fighting against COVID-19: a novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection, Sustain. Cities Soc. (2021) 65.
[41] S. Singh, U. Ahuja, M. Kumar, K. Kumar, M. Sachdeva, Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment, Multimed. Tools Appl. 80 (2021) 19753–19768.
[42] S.Q. Ren, K.M. He, R. Girshick, J. Sun, R.C.N.N. Faster, Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149.
[43] P. Wu, H. Li, N. Zeng, F. Li, FMD-Yolo: an efficient face mask detection method for COVID-19 prevention and control in public, Image Vis. Comput. 117 (2022) 104341 104341.
[44] X. Su, M. Gao, J. Ren, Y. Li, M. Dong, X. Liu, Face mask detection and classification via deep transfer learning, Multimed. Tools Appl. 81 (2022) 4475–4494.
[45] Z. Cao, M. Shao, L. Xu, S. Mu, H. Qu, MaskHunter: real-time object detection of face masks during the COVID-19 pandemic, IET Image Process. 14 (2020) 4359–4367.

[46] X. Jiang, T. Gao, Z. Zhu, Y. Zhao, Real-time face mask detection method based on YOLOv3, Electronics 10 (2021).
[47] J. Yu, W. Zhang, Face mask wearing detection algorithm based on improved YOLO-v4, Sensors 21 (2021).
[48] P. Nagrath, R. Jain, A. Madan, R. Arora, P. Kataria, J. Hemanth, SSDMNV2: a real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2, Sustainable Cities and Society (2021) 66.
[49] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-.C. Chen, Ieee, MobileNetV2: inverted residuals and linear bottlenecks, in: 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)Salt Lake City, UT, 2018, pp. 4510–4520.
[50] A. Kumar, A. Kalia, K. Verma, A. Sharma, M. Kaushal, Scaling up face masks detection with YOLO on a novel dataset, Optik 239 (2021).
[51] A. Kumar, A. Kalia, A. Sharma, M. Kaushal, A hybrid tiny YOLO v4-SPP module based improved face mask detection vision system, J. Ambient Intell. Humaniz. Comput. (2021).
[52] K.M. He, X.Y. Zhang, S.Q. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 1904–1916.
[53] B. Roy, S. Nandy, D. Ghosh, D. Dutta, P. Biswas, T. Das, MOXA: A deep learning based unmanned approach for real-time monitoring of people wearing medical masks, Trans. Indian Natl. Acad. Eng. 5 (2020) 509–518.
[54] R.P. Martinez, I. Schiopu, B. Cornelis, A. Munteanu, Real-time instance segmentation of traffic videos for embedded devices, Sensors 21 (2021).
[55] S. Chen, R. Zhan, W. Wang, J. Zhang, Learning slimming SAR ship object detector through network pruning and knowledge distillation, IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens. 14 (2021) 1267–1282.
[56] R. Cheng, X. He, Z. Zheng, Z. Wang, Multi-scale safety helmet detection based on SAS-YOLOv3-tiny, Appl. Sci. Basel 11 (2021).
[57] Y. Lin, R. Cai, P. Lin, S. Cheng, A detection approach for bundled log ends using K-median clustering and improved YOLOv4-Tiny network, Comp. Electron. Agric. (2022) 194.
[58] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[59] S. Woo, J. Park, J.-.Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.
[60] Q. Wang, B. Wu, P. Zhu, P. Li, Q. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[61] S. Elfwing, E. Uchibe, K. Doya, Sigmoid-weighted linear units for neural network function approximation in reinforcement learning, Neural. Netw. 107 (2018) 3–11.
[62] Z.C. Ouyang, J.H. Cui, X.Y. Dong, Y.Q. Li, J.W. Niu, SaccadeFork: A lightweight multi-sensor fusion-based target detector, Inform. Fusion 77 (2022) 172–183.
[63] D.H. Kang, Y.J. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, Struct. Health Monit. Int. J. (2022).
[64] A.M. Roy, R. Bose, J. Bhaduri, A fast accurate fine-grain object detection model based on YOLOv4 deep neural network, Neural. Comput. Appl. 34 (2022) 3895–3921.
[65] A. Vouros, E. Vasilaki, A semi-supervised sparse K-Means algorithm, Pattern Recognit. Lett. 142 (2021) 65–71.
[66] H. Huang, X. Tang, F. Wen, X. Jin, Small object detection method with shallow feature fusion network for chip surface defect detection, Sci. Rep. 12 (2022) 3914.
[67] X.P. Su, M. Gao, J. Ren, Y.H. Li, M. Dong, X. Liu, Face mask detection and classification via deep transfer learning, Multimed. Tools Appl. 81 (2022) 4475–4494.
[68] T. Hui, Y.L. Xu, R. Jarhinbek, Detail texture detection based on Yolov4-tiny combined with attention mechanism and bicubic interpolation, IET Image Process. 15 (2021) 2736–2748.
[69] C.Y. Liu, Y.Q. Wu, J.J. Liu, J.M. Han, MTI-YOLO: A light-weight and real-time deep neural network for insulator detection in complex aerial images, Energies 14 (2021).