




Face mask detection and classification via deep transfer learning

Xueping Su¹ · Meng Gao¹ · Jie Ren¹ · Yunhong Li¹ · Mian Dong¹ · Xi Liu² 

Received: 24 June 2021 / Revised: 23 October 2021 / Accepted: 25 November 2021 /
Published online: 9 December 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Wearing a mask is an important way of preventing COVID-19 transmission and infection. German researchers found that wearing masks can effectively reduce the infection rate of COVID-19 by 40%. However, the detection of face mask-wearing in the real world is affected by factors such as light, occlusion, and multi-object. The detection effect is poor, and the wearing of cotton masks, sponge masks, scarves and other items greatly reduces the personal protection effect. Therefore, this paper proposes a new algorithm for mask detection and classification that fuses transfer learning and deep learning. Firstly, this paper proposes a new algorithm for face mask detection that integrates transfer learning and Efficient-Yolov3, using EfficientNet as the backbone feature extraction network, and choosing CIoU as the loss function to reduce the number of network parameters and improve the accuracy of mask detection. Secondly, this paper divides the mask into two categories of qualified masks (N95 masks, disposable medical masks) and unqualified masks (cotton masks, sponge masks, scarves, etc.), creates a mask classification data set, and proposes a new mask classification algorithm that the combines transfer learning and MobileNet, enhances the generalization of the model and solves the problem of small data size and easy overfitting. Experiments on the public face mask detection data set show that the proposed algorithm has a better performance than existing algorithms. In addition, experiments are performed on the created mask classification data set. The mask classification accuracy of the proposed algorithm is 97.84%, which is better than other algorithms.

Keywords COVID-19 · Masked face detection · Masked face dataset · Mask classification

1 Introduction

In 2020, the rapid spreading of Corona Virus Disease (COVID-19) has forced the World Health Organization to declare COVID-19 as a global pandemic. Wearing a mask has an important effect on slowing the spread of the new epidemic. German researchers found that

✉ Xi Liu
liuxigdpu@163.com

¹ School of Electronics and Information, Xi'an Polytechnic University, 710048 Xi'an, China

² GuangDong Pharmaceutical University, 510000 Guangzhou, China

wearing masks effectively reduces the infection growth rate of COVID-19 by 40%. Moreover, a study published in the Proceedings of the National Academy of Sciences (PNAS) pointed out that the importance of wearing masks is far greater than social distancing and home isolation policies in preventing the spread and infection of COVID-19 [1]. COVID-19 is highly contagious, and the main transmission route is droplet and close contact [19]. Many countries emphasize that under the premise of sufficient masks, it is necessary to wear medical surgical masks or N95 masks for personal protection [22]. Wearing qualified masks can effectively resist the harm of viruses. Reference [26, 28] shows that the penetration rate of particles to cotton mask is close to 97%, and that of medical mask is 44%. The direct or indirect evidence in [29] shows that medical masks and N95 masks have better protective effects than cotton masks. It is everyone's responsibility to wear masks in public places to prevent the spread of the virus during the epidemic. This requires not only individual conscious compliance, but also certain means of supervision and management. Therefore, the study of mask wearing detection and mask type recognition has become very important for national management and public health.

At present, technologies based on deep learning and artificial intelligence have been widely applied to resist COVID-19, such as automatic CT image segmentation [34], automatic analysis [18], and detection of human respiration patterns [41]. However, there are relatively few researches on face mask detection and mask type recognition. Traditional face mask detection is applied in the security field to prevent terrorist attacks and reduce illegal and criminal behavior [4, 6]. After the outbreak of COVID-19, researchers conducted a variety of face mask occlusion studies. Wang et al [42] proposed the world's largest real-world masked face dataset and developed a multi-granularity masked face recognition model. Jiang et al [15] proposed a single-stage detector RetinaMask for mask-wearing detection. In [24], the author used Resnet50 for feature extraction and used decision trees, Support Vector Machines (SVM), and ensemble algorithm to classify face images wearing masks. The above research can effectively detect faces and faces wearing masks, and it is robust to other obstructions. However, in the detection of face masks in natural scenes, there are scenes such as poor lighting conditions, object occlusion, dense crowds, multiple objects, etc. In addition, there are case of wearing cotton masks, sponge masks, scarves, and other items, which greatly reduces the personal protection effect.

In summary, the new mask detection and classification algorithms based on transfer learning and deep learning are proposed in this paper. Firstly, the fusion of transfer learning and Efficient-Yolov3 algorithm is proposed for face mask-wearing detection in natural scenes. Secondly, the mask classification data set is created, and the mask classification algorithm fusing transfer learning and MobileNet [13] is proposed. The contributions of this paper are as follows:

1. The face mask detection dataset is small, and the training process is prone to overfitting, which reduces the robustness of the model. Besides, the loss function IOU (Intersection over Union) [45] is difficult to reflect the degree of coincidence between the predicted bounding box and ground-truth bounding box, resulting in poor regression effect. A new algorithm of face mask detection based on transfer learning and Efficient-Yolov3 is proposed to improve the detection accuracy, speed and model generalization ability, and improve the stability of Bounding Box (BB) regression.
2. In view of the problem that there is no public data set for mask classification. In this paper, the RMFD [42] data set and the MAFA [6] data set are segmented and normalized for face mask-wearing images. After preprocessing, the mask images are divided into

- qualified masks (N95 masks, disposable medical masks) and unqualified masks (cotton masks, sponge masks, scarves, etc.), and finally a mask classification data set is created.
3. Aiming at the problem of a large amount of calculation, many network parameters, and long training time in current classification algorithms based on convolutional neural networks under small sample data, a new mask classification algorithm fusing transfer learning and MobileNet [13] is proposed to improve the accuracy of mask classification and reduce training time.

The rest of the paper is organized as follows: In Section 2, the related research of objects detection and convolution neural networks are described. The two methods are proposed in Section 3, the fusion of transfer learning and Efficient-Yolov3 for face mask-wearing detection and the fusion of transfer learning and MobileNet for mask classification. The dataset, parameter settings, evaluation metrics, results and ablation experiments are shown in Section 4. Finally, the conclusion and future work prospects are described in Section 5.

2 Related work

As a typical task of object detection, the development process of face detection has experienced from traditional to deep learning. Among the typical algorithms for traditional object detection are: Viola and Jones [40] proposed an Adaboost face detection technology based on Haar features. The Adaboost algorithm is simple to implement, and is suitable for two-classification and multi-classification without over-fitting. However, the algorithm is susceptible to noise, has a long detection time, and is prone to missed detection. On the basis of Adaboost, Ma [25] et al used 4 types of Haar features to describe the face relationship, which reduced the detection time and the missed detection rate. However, the detection performance of this algorithm decreases in the case of occlusion and profile. In 2010, Pedro Felzenszwalb [5] proposed the Deformable Part-based Model(DPM) algorithm, which used the component model strategy of multi-component and pictorial structure to solve the multi-posture and multi-angle problem of the face. But this method requires artificially design the incentive template of the object, the workload is large, and it is not universal. In a word, the traditional object detection methods are limited to the effective description of image features, and can only rely on the experience to manually extract features, and then design classifiers for object detection based on the results and combined with the sliding window. and with complex steps, low accuracy and poor real-time performance.

With the great success of deep learning in image classification, many researchers have also applied this technology to face detection. Compared with traditional detection methods, face detection algorithms based on deep learning are more suitable for complex backgrounds and different face poses. Since the birth of the AlexNet [16] network in 2012, the convolutional neural network(CNN) has ushered in a historic breakthrough and has shown explosive development. From the continuously deepening VGG [35] network structure to the continuously widening Inception network [14, 36, 37] and Resnet [10, 12] network, and then to the lightweight MobileNet [13] network and EfficientNet [38] network. The focus of CNN research has also shifted from parameter optimization to designing network architectures, such as new architectures using attention-based information processing [39, 43]. The rapid development of CNN also provides new feature extraction strategies for object detection. Object detection based on deep learning is divided into two categories: two-stage detection and one-stage detection. The difference between the two is that the

former sets the detection process from coarse to fine, while the latter completes in one step [27, 46]. Two-stage detection algorithms mainly include RCNN [8], SSP-Net [11], Fast-RCNN [7], Faster-RCNN [33], etc. This type of algorithm improves the detection accuracy by constructing a Region Proposal Network (RPN), but the training time is longer and cannot meet real-time requirements. One-stage detection object detection algorithms pay more attention to the improvement of detection speed, such as YOLO(You Only Look Once) series of algorithms [2, 9, 30–32], SSD(Single shot multi-box detector) algorithm [23], and RetinaNet algorithms [21] etc. While generating detection box, one-stage detection algorithms classify and regress the category probability and position coordinates of object. Compared with two-stage detection algorithm, it greatly improves the detection speed. Therefore, the new face mask-wearing detection algorithm fusing transfer learning and Efficient-Yolov3 is proposed, and a new mask classification algorithm merging transfer learning and MobileNet is proposed in this paper.

3 Method

3.1 Fusion transfer learning and Efficient-Yolov3 for face mask detection

3.1.1 Efficient -Yolov3

The YOLOv3 [32] algorithm has undergone three generations of changes and evolutions. Compared with the previous two generations, it has the advantages of simple structure, high detection accuracy, and fast speed. The Yolov3 algorithm uses the DarkNet53 network as the backbone feature extraction network, which adopts the residual idea of ResNet [10], and uses multiple residual blocks for feature extraction, as shown in Fig. 1. The feature extraction steps of the DarkNet53 network are:

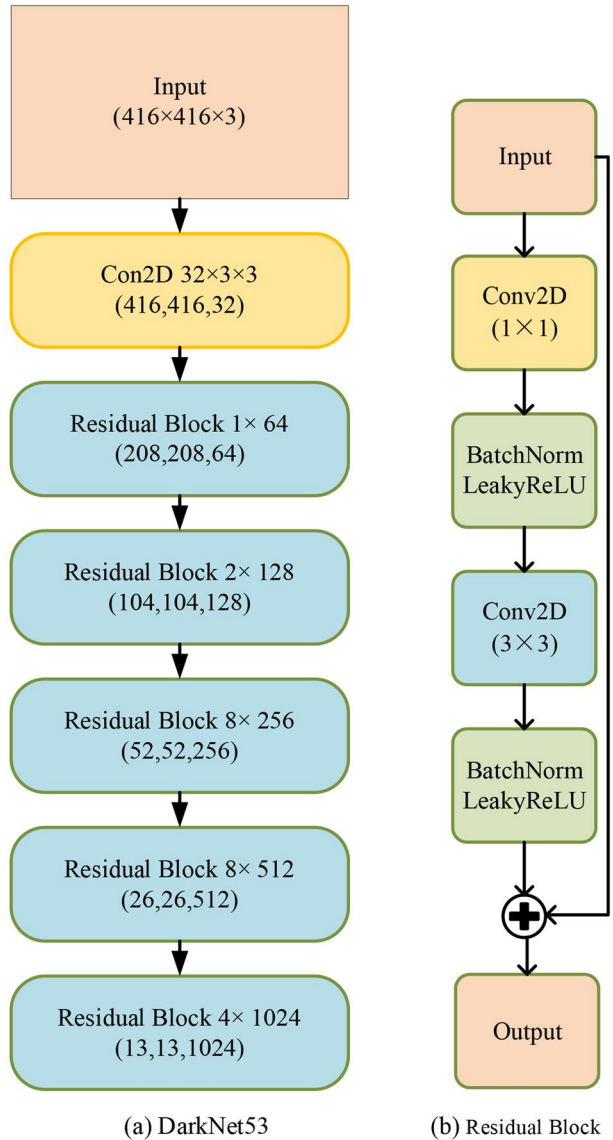
1. Residual Block first compresses the input size to 1/2 of the original size by a 1×1 convolution kernel, and then performs batch Normalization and Leaky ReLU. Ordinary ReLU sets all negative values to zero, while Leaky ReLU assigns a non-zero slope to all negative values. Its mathematical formula is expressed as:

$$y_i = \begin{cases} x_i & \text{if } x_i \geq 0 \\ \frac{x_i}{a_i} & \text{if } x_i < 0 \end{cases} \quad (1)$$

2. The input size is expanded to the original size through a 3×3 convolution kernel. After the convolution is completed, batch normalization and Leaky ReLU are also performed, and the result is input to the next stage. The DarkNet53 network is easy to optimize, and the accuracy is improved by increasing the depth of the network. The internal Residual block uses skip connections to alleviate the problems of gradient disappearance and gradient explosion.

However, the Darknet53 network classification performance is poor, and it adopts a multi-size and multi-level convolution kernel to increase the receptive field, increasing the amount of calculation while increasing the parameters. Although [17] used Yolov3 for face mask detection in natural scenes. On the one hand, the currently public face mask detection dataset is small, and over-fitting will occur during the training process, which reduces the

Fig. 1 Architecture of DarkNet53



robustness of the model. On the other hand, although the loss function IOU can reflect the distance between the predicted bounding box and ground-truth bounding box, it cannot reflect the degree of overlap between the two, which reduces the detection performance.

In response to the above problems, this paper proposes a new face mask detection algorithm that fuses transfer learning and Efficient-Yolov3. EfficientNet [38] was proposed by Google. The main advantages of this network are: 1) Using residual network to increase the depth of neural network, and extracting richer and more complex features of images through deep neural network; 2) Increasing the width of the network by increasing channel dimension, and capturing more fine-grained features; 3) Increasing

the resolution of the input image to make the network learn and express feature more abundant, and improving the accuracy of classification. For achieving higher accuracy and speed, EfficientNet uses compound coefficient φ to uniformly scale the baseline model. The model's scaling factor calculation formula are as follows:

$$d = \alpha^\varphi, w = \beta^\varphi, r = \gamma^\varphi \quad (2)$$

$$s.t. \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \quad (3)$$

$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \quad (4)$$

Where, α, β and γ are respectively the amplification ratio of depth, width, and resolution of the model. The flops of conventional convolution is proportional to d, w^2 and r^2 . But doubling the width of the network or the resolution will increase the flops by four times. In this paper, the constraint on φ is $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$. The overall network structure of EfficientNet is shown in Fig. 2(a). The network consists of 16 MBConvBlocks, 2 Conv2D layers, 1 Global Average Pooling2D layer, and 1 Dense layer, of which 16 MBConvBlocks are the main feature extraction network. The structure of MBConv is shown in Fig. 2(b). It is composed of DepthwiseConv2D and SENet. It uses Swish as the activation function. Swish has the characteristics of no upper bound and lower bound, smooth, and non-monotonic, and it is better than ReLU on the deep model. The formula is expressed as:

$$f(x) = x \cdot \text{sigmoid}(\beta x) \quad (5)$$

Where β is a constant or trainable parameter. The feature extraction process of MBConvBlock is as follows:

1. A 1×1 kernel is used to perform dimension upgrade operations on Inputs.
2. Through a Depthwise Convolution layer, this layer uses a 3×3 or 5×5 convolution kernel to perform convolution operations on the image. This process greatly reduces the amount of computation and ensures the accuracy of convolution by changing the multiplication operation to the addition operation.
3. A channel attention mechanism (SENet) is added as shown in Fig. 2(b). First, using a dimensionality reduction coefficient r to reduce the dimensionality of the FC layer, then activated with ReLU, and finally restore the FC layer to its original size. After the channel attention mechanism, a 1×1 convolution kernel is still used for dimensionality reduction.

The structure of the Efficient-Yolov3 algorithm in this paper is shown in Fig. 2. The steps of the algorithm are as follows:

1. After 3, 4, and 5 feature extractions of an image of 416×416 size by EfficientNet network, the size becomes $52 \times 52 \times 48$, $26 \times 26 \times 120$, and $13 \times 13 \times 352$.
2. The Feature Pyramid Networks(FPN) [20] structure of YOLOv3 adopts a 13×13 scale feature map that has been down-sampled 32 times as the deepest detection layer.
3. The $13 \times 13 \times 352$ scale feature map is twice up-sampling and then fused with the 26×26 scale feature map after 16 times down-sampling. The fusion result is used as an intermediate detection layer.

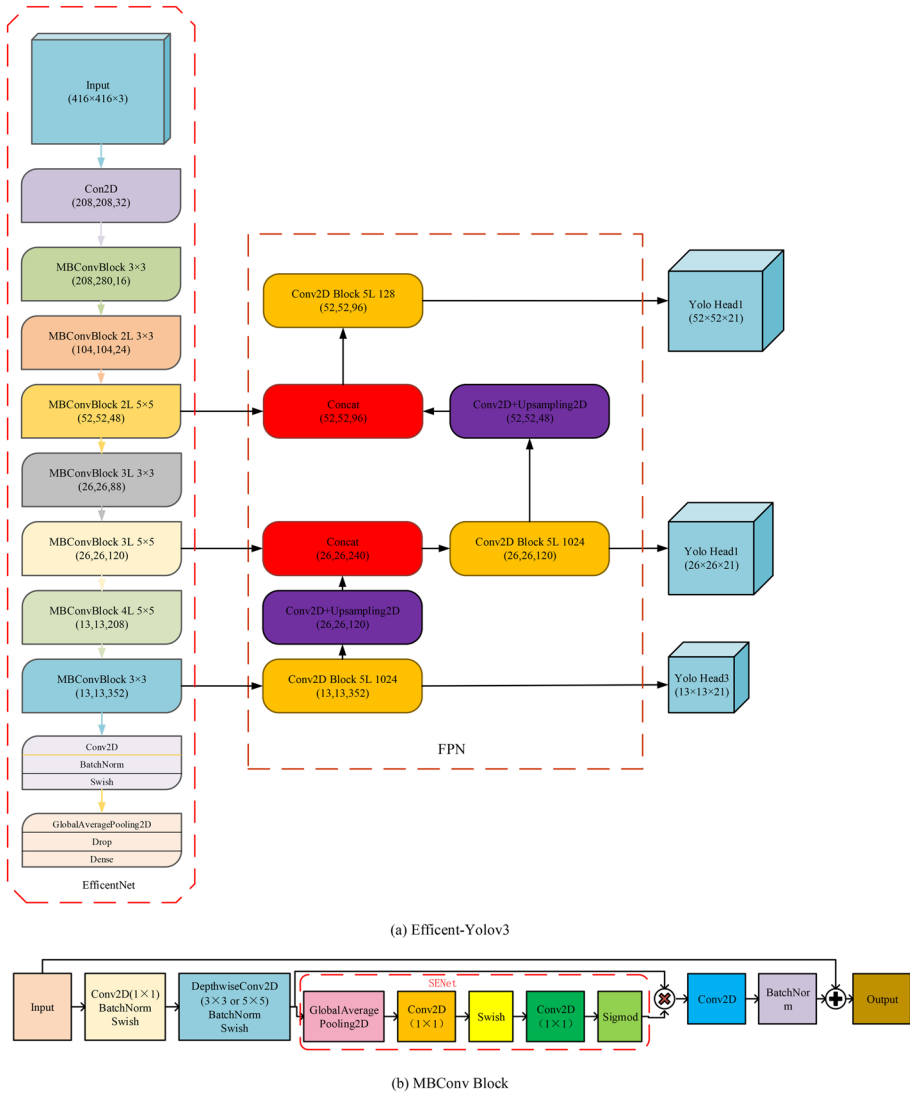


Fig. 2 Architecture of Efficient-Yolov3

4. The $26 \times 26 \times 120$ scale feature map is twice up-sampling and then merged with the 52×52 scale feature map after 8 times down-sampling. The fusion result is used as the shallow detection layer. The high-level semantic information transmitted from the shallow fusion deep layer is more conducive to detection. The constructed feature pyramid establishes high-level semantic information on multi-scale features, which can better detect objects of different scales and improve detection performance.

3.1.2 Deep transfer learning

In this paper, the method of deep transfer learning is introduced to apply the similarity of EfficientNet’s underlying elements in the ImageNet dataset to face mask detection. The process of transfer learning is shown in Fig. 3.

First, the weights of all levels of the ImageNet network model are loaded, and then the levels of extracting features described in EfficientNet such as edges, lines, colors, patterns, etc., are frozen. Only the level parameters for downstream classification task are adjusted. The purpose is to reduce the training parameters in a small amount of dataset and correctly train and adjust the model. The trained model not only inherits the hierarchical weight of the original model, but also possesses the training weight of the model itself.

3.1.3 Loss function

The loss function IOU is the most commonly used indicator in object detection. It has the characteristics scale invariance and can reflect the distance between the predict box and ground-truth. Its formula is expressed as:

$$IOU = \frac{|P \cap G|}{|P \cup G|} \tag{6}$$

$$L_{IOU} = 1 - IOU \tag{7}$$

Where P represents the predict box, and G represents the ground-truth. It can be seen from formula 6, when P and G do not intersect, $IOU = 0$, and the loss function $L_{IOU} = 1$, so that L_{IOU} is always equal to 1. In addition, as shown in Fig. 4, the predict box and ground-truth

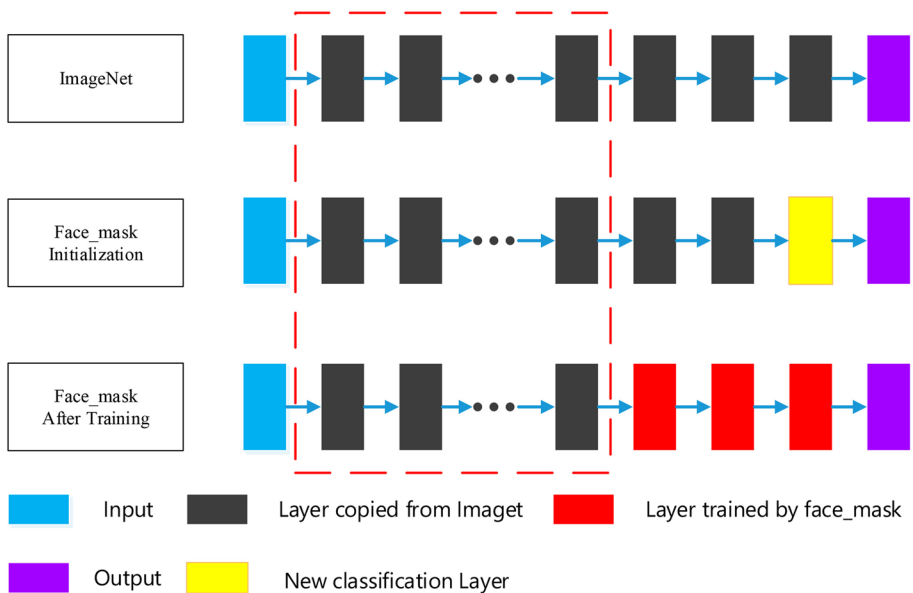
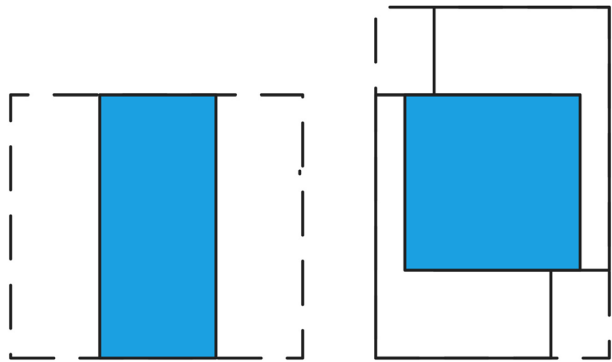


Fig. 3 Transfer learning framework

Fig. 4 Two ways of overlapping two rectangles with exactly the same IOU value



at different positions have the same IOU value, which cannot accurately reflect the degree of overlap between the two, resulting in poor regression

In response to the above problems, this paper selects CIOU (Complete-IOU) as the loss function. Different from IOU, CIOU includes the distance between the object and the anchor, the overlap area, scale, and penalty factor, so that the bounding box regression becomes more stable, and there will be no divergence during the training process. And the penalty factor takes into account the aspect ratio of the predicted box to fit the object box. The specific calculation formula is expressed as:

$$CIOU = IOU - \frac{\rho^2(p, g)}{c^2} - \alpha v \tag{8}$$

Where $\rho^2(p, g)$ is the Euclidean distance between the center points of the predicted box and the ground-truth, and c is the diagonal distance of the smallest closed area that can contain both the predict box and ground-truth. And α is a positive trade-off parameter, and v is the parameter used to measure the consistency of the aspect ratio. The formulas are as follows:

$$\alpha = \frac{v}{1 - IOU + v} \tag{9}$$

$$v = \frac{\pi^2}{4} \left(\arctan \frac{w^g}{h^g} - \arctan \frac{w^p}{h^p} \right) \tag{10}$$

The complete CIOU loss function formula is expressed as:

$$L_{CIOU} = 1 - IOU + \frac{\rho^2(p, g)}{c^2} + \alpha v \tag{11}$$

3.2 Fusion transfer learning and MobileNet of mask classification

3.2.1 MobileNet network model

The MobileNet model is a lightweight deep neural network proposed by Google for embedded devices such as mobile phones. Compared with other network models, MobileNet has fewer parameters and fast calculation speed. The core idea of MobileNet is deep separable convolution, and its structure is divided into deep convolution and point-wise convolution.

It is assumed that the input feature map is $DH \times DW \times M$, and the output feature map is $DH \times DW \times N$, the size of the convolution kernel is $DK \times DK$, the process is shown in Fig. 5.

For standard convolution, the parameters are:

$$DK \times DK \times M \times N \tag{12}$$

The calculation amount is:

$$DK \times DK \times M \times DH \times DW \times N \tag{13}$$

Using depth separable convolution, the parameters are:

$$DK \times DK \times M + M \times N \tag{14}$$

The calculation amount is:

$$DK \times DK \times M \times DH \times DW + DH \times DW \times N \times M \tag{15}$$

As shown in Fig. 5, the size of the input feature map is $12 \times 12 \times 3$, and the size of the output feature map is $12 \times 12 \times 4$. After applying the depth separable convolution, three 5×5 size convolution kernels are used to traverse the data of three channels respectively, and three feature maps are obtained. Before the fusion operation, four 1×1 size convolution kernels are used to traverse three feature maps, and the required calculation amount is $5 \times 5 \times 3 \times 12 \times 12 + 12 \times 12 \times 4 \times 3 = 12528$. In fact, the standard convolution calculation amount is $5 \times 5 \times 3 \times 12 \times 12 \times 4 = 43200$. It can be seen that the depth separable convolution can greatly reduce the calculation amount of the model.

This paper proposes a new method of mask classification that combines transfer learning and MobileNet model. The overall framework is shown in Fig. 6. First, loading the weights trained by ImageNet, then freezing the feature extraction layer of MobileNet, training the Global Average Pooling layer and the Softmax layer, and finally reinitializing the model, and fine-tuning the parameters of each layer to achieve the best classification effect.

4 Experiment result

4.1 Dataset

4.1.1 Face mask wearing detection dataset

The Face Mask Dataset [3] in this paper contains a total of 7959 images, including 6120 images in the training set, 3006 images from the MAFA [6] dataset (basically all images with masks), and 3114 images from the WIDER Face [44] dataset (basically all images

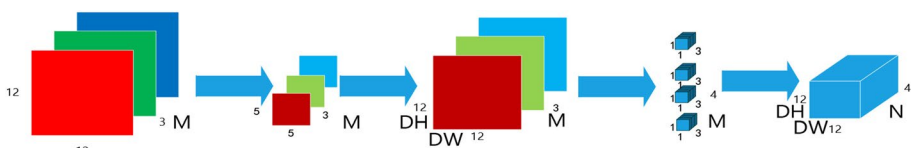


Fig. 5 Depth separable convolution

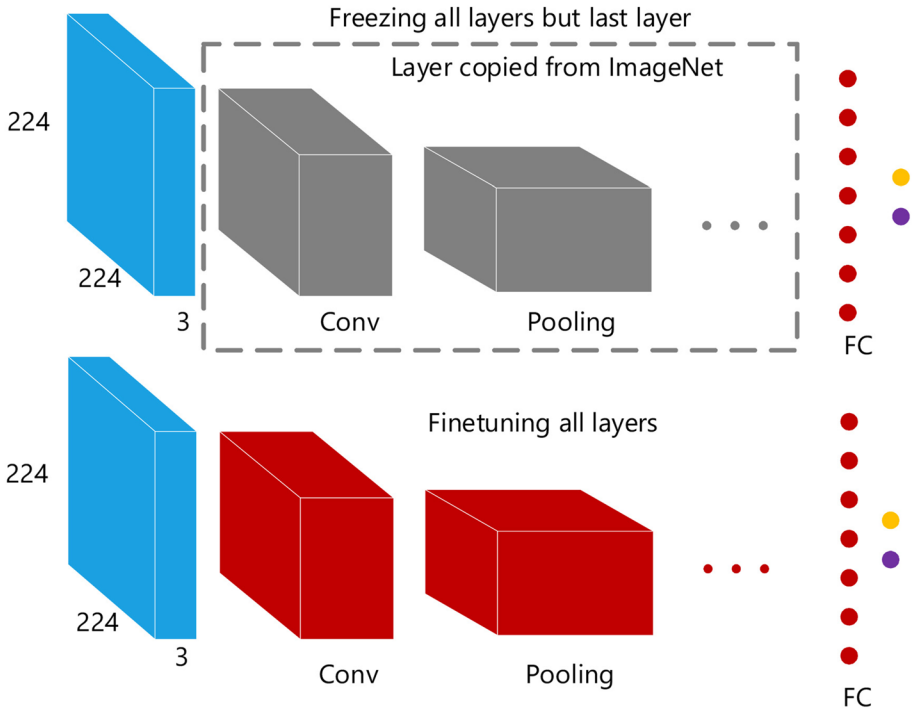


Fig. 6 Transfer learning-mobilenet

without a mask). The test set has a total of 1839 images, including 1059 images from the MAFA and 780 images from the WIDER Face. In this paper, the dataset is uniformly labeled. The labeled information mainly includes the upper left corner of the object box, the lower right corner coordinates, and the category. An example is shown in Fig. 7, and the corresponding labeled data is shown in Table 1.

4.1.2 Face mask classification dataset

In this paper, the face mask classification dataset cuts and filters the faces wearing masks in the RMFD [42] dataset and MAFA [6] dataset, and the size is normalized to 224×224 . Mask categories are divided into qualified masks (OK-mask) and unqualified masks (NG-mask). Qualified masks contain 1361 images, mainly N95 masks and disposable medical masks. Unqualified masks contain 1880 images, mainly including sponge masks, cloth masks, and scarves, etc. The mask classification dataset contains a total of 3241 images for mask classification, some of which are shown in Fig. 8. The mask classification dataset of

Table 1 Sample image annotation data

Object	Box
face	381,207,618,561
face_mask	156,141,393,418

Fig. 7 Dataset annotation example diagram



Fig. 8 Mask classification dataset

this article is available at the following website: <https://github.com/Kyrie-leon/Face-Mask-Classification-Dataset>.

4.2 Parameter settings

This paper trains the model on NVIDIA GeForce RTX 1080 Ti. The dataset is divided into training set, validation set, and test set with ratios of 0.7, 0.1, and 0.2. The algorithm is developed using the Tensorflow deep learning framework and implemented based on Python.

In the face mask-wearing detection experiment, this paper is divided into two stages of training, both using Adam optimizer. The first stage freezes the first 378 layers of EfficientNet for rough optimization. The learning rate is $\alpha=0.001$, and 25 epochs are trained. The stochastic gradient descent algorithm is used as the optimization algorithm. The second stage is to unfreeze the first 378 layers of EfficientNet for more detailed network learning,

with a learning rate $\alpha=0.0001$, and training for 25 epochs. For the EfficientNet network, the input image size is 416×416 , and the batch size is 2.

In the mask classification experiment, this paper is divided into two stages of training, both using Adam optimizer. The first stage freezes the first 81 layers of the MobileNet network to optimize the Softmax layer, the learning rate is $\alpha=0.001$, and the training is performed for 25 epochs. The gradient descent algorithm is used as an optimization algorithm. In the second stage, the 81-layer network before thawing is optimized for all networks, the learning rate is $\alpha=0.0001$, and the training is 25 epochs.

4.3 Evaluation metrics

Object detection indicators use average precision AP (Average Precision), mAP (Mean Average Precision), and frame rate per second (FPS). The AP value reflects the detection effect of a single object, and mAP is the mean value of the average accuracy of all categories in the dataset. The calculation formulas are as follows:

$$AP = \int_0^1 p(r)dr \quad (16)$$

$$mAP = \frac{\sum_{i=1}^N AP_i}{N} \quad (17)$$

Where $p(r)$ represents the mapping relationship between precision and recall. The calculation formulas for precision and recall are expressed as:

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

Where TP , FP and FN denoted the true positive, false positive and false negative, respectively. Image classification uses *Accuracy*, *Precision*, *Recall*, *F1 – measure* as the evaluation criteria, and the calculation formulas are as follows:

$$Accuracy = \frac{TP}{TP + FP} \quad (20)$$

$$F1 - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (21)$$

4.4 Result and analysis

4.4.1 The result of wearing face mask

This paper verifies the effectiveness and superiority of the proposed algorithm through two sets of experiments. The first group selects the Yolov3 algorithm as a comparison, and the second group compares with other related research algorithms.

The first set of experimental results are shown in Table 2, and some of the experimental results are shown in Fig. 9. As shown in Table 2. In this paper, the face AP value on the Yolov3 object detection network is 93.88%, and the value in the face mask wearing detection is as high as 98.18%, while the AP value of Yolov3 object detection algorithm in face and face mask wearing detection is 90.82% and 95.26% respectively. Compared with the Yolov3 algorithm, the AP values of the face and face_mask of the fusion transfer learning and Efficient-Yolov3 algorithm proposed in this paper are increased by 3.06% and 2.92%, mAP increased by 2.99, and FPS increased by 1.71. In general, the mAP of this method is as high as 96.03%, and has achieved the best performance in both face and face mask-wearing detection. In terms of detection efficiency, the algorithm designed in this paper can reach 14.62 frames per second, the main reason for the fast speed of our method is that the amount of model parameters is reduced by about 4 times. It can be seen from Fig. 9 that the algorithm in this paper has high robustness for face occlusion and good performance for multi-object detection in complex scenes.

The second set of experimental results are shown in Table 3. It can be seen from Table 3. that the proposed algorithm has a significant improvement of AP and mAP values than other methods. Baseline's model parameters are only 1.01 million, and the FPS is 23.12. Although the speed can reach real-time detection, the small number of convolution

Table 2 Detection average precision of Efficient-Yolov3

Methods	face(%)	face_mask(%)	mAP(%)	FPS($F \cdot S^{-1}$)	Params(M)
Yolov3	90.82	95.26	93.04	12.91	61.58
Ours	93.88	98.18	96.03	14.62	15.91



Fig. 9 Part of detection results

Table 3 Comparison with related works

Methods	face(%)	face_mask(%)	mAP(%)	FPS($F \cdot S^{-1}$)	params(M)
Baseline [3]	89.60	91.90	90.75	23.12	1.01
Yolov3 [32]	90.82	95.26	93.04	12.91	61.58
Yolov4 [2]	91.22	96.59	93.91	11.32	64.01
Yolov5 [9]	93.03	98.14	95.59	18.45	7.07
Ours	93.88	98.18	96.03	14.62	15.91

kernels results in the low accuracy and the mAP is only 90.75%. Compared with Yolov3 and Yolov4, our parameters are reduced by 4 times. In addition, our parameters are only increased by 2 times than Yolov5, but our method achieves higher accuracy. In summary, the proposed algorithm in this paper has obvious advantages over other algorithms.

In addition, as shown in Table 3. The experimental results show that the AP value of the face is lower than the AP value of the face_mask. The main reason is that the feature point information of face is rich and the structure is complex, and the face information between different people is different, such as facial features, skin color, gender, age and so on. In addition, faces are also disturbed by expressions (happy, angry, sad, fear, disgust, anger, etc.) and occlusions (hands, paper, towels, etc.), which result in the failure of the classifier to learn robust and accurate features at the same time. However, due to the large area covered by masks and most people wear the same type of masks, the mask information is regular and easy to distinguish and detect, resulting in a high AP value, as shown in Fig. 9.

4.4.2 The result of mask classification

Since the mask classification research is proposed for the first time in this paper, there is currently no relevant research to compare. The experiment selects the commonly used classification network to compare with the algorithm proposed in this paper to verify the superiority and efficiency. For the classification of masks, this paper pays more attention to the people who wear unqualified masks, namely NG-mask. This paper uses NG-mask as a positive example in mask classification. Figure 10 shows the results of different network models on the test set. For the test set *Precision*, VGG16 is only 92.49%, while the other three models are all higher than 95%, and the precision value of InceptionV3 is 98.40%, lower than our method 98.40%, with a difference of 0.04%. For *Recall*, *F1 – measure*, and *Accuracy*, the models in this paper have achieved the best performance, respectively 97.87%, 98.13%, and 97.84%, while InceptionV3 and VGG16 have poor classification results. It can be seen from the various values of the test set that the performance of our method is

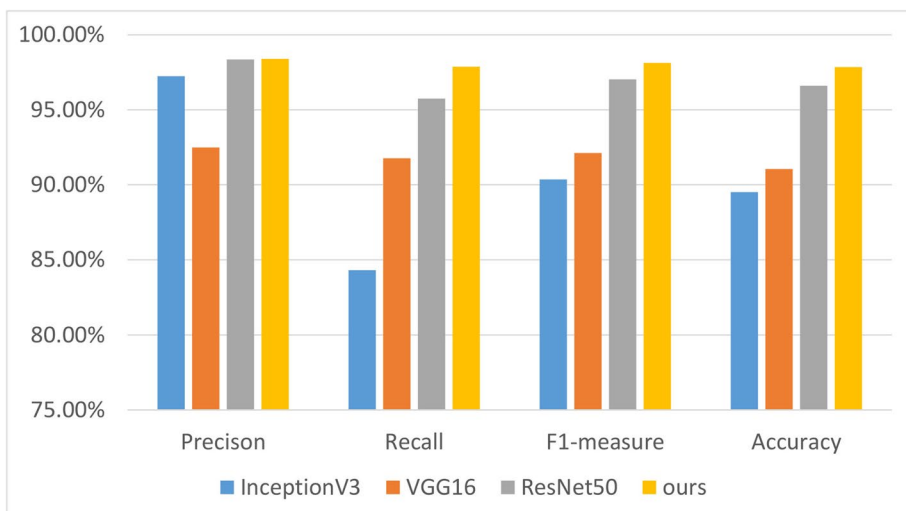


Fig. 10 Classification results of different models in the test set

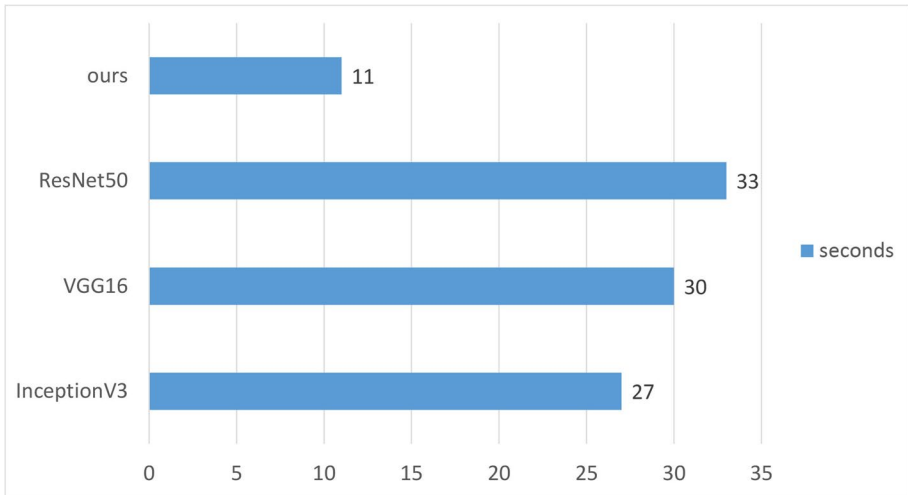


Fig. 11 The consumption of each epoch training time

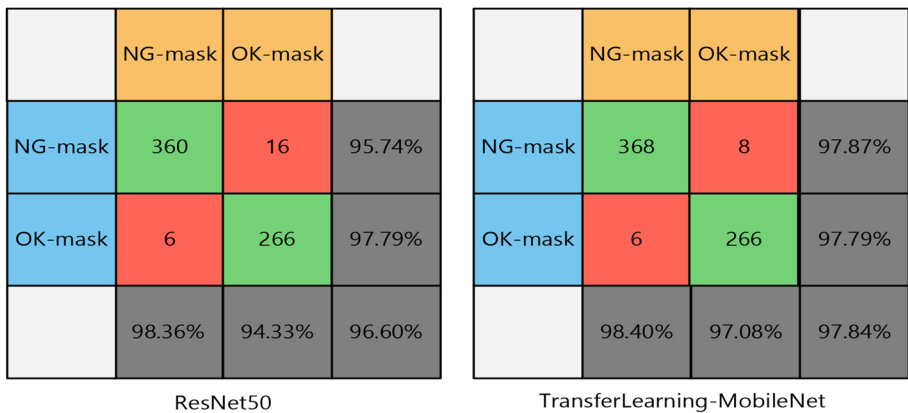


Fig. 12 The confusion matrix of ResNet50 and TransferLearning-MobileNet on test set

similar to that of ResNet50 on the precision of unqualified masks, but it is better than other networks on *Recall*, *F1Score*, and *Accuracy*.

Figure 11 shows the time consumed in each epoch of training for different models on the dataset. Since the training time depends on the dataset size and batch size, Fig. 11 shows the training consumption time obtained in this paper under the condition that the dataset size is 2982 and the batch size is 16. It can be seen from the Fig. 11 that the training time of ResNet50 is the longest 33s, while the training time of our method is the shortest 11s. It can be seen that the introduction of transfer learning greatly reduces the time consumed for model training.

Figure 12 shows the confusion matrix of different models on the test set. It can be seen from Fig. 12 that InceptionV3 and VGG16 have a poor classification effect. ResNet50 and our algorithm have similar performance in the classification of unqualified masks. The details of the model classification effect can be seen intuitively through the confusion

matrix. After comparing and classifying all the test set data and the model prediction results, the confusion matrix data table is obtained to further prove the superiority of the proposed method. As shown in Fig. 12, the rows of the matrix represent the *Recall* and the columns represent the *Precision*. For unqualified masks(NG-mask), the Precision of TransferLearning-MobileNet is 98.04%, and the recall rate is 97.84%. For qualified masks(OK-mask), the *Precision* of TransferLearning-MobileNet is 97.08% and the *Recall* is 97.79%. The *Accuracy* of the model proposed in this paper is 97.84%. Compared with ResNet50, the two models have the same *Recall* on qualified masks, and the indicators of this model are higher than the ResNet50. Therefore, the following conclusions can be drawn:

1. The overall performance of the TansferLearning-MobileNet network for mask classification is better than other networks.
2. The TansferLearning-MobileNet network consumes less time on the training dataset.

4.5 Ablation study

In order to further verify the superiority of the algorithm proposed in this paper, this paper conducts ablation experiments on face mask detection and mask classification.

4.5.1 Face mask detection ablation study

This paper conducted ablation experiments on Backbone, Transfer Learning and *CIoU* respectively, and the experimental results are shown in Table 4.

Backbone: Replacing Darknet53 with EfficientNetB2-Transer Learning, it can be seen that the accuracy of face and face_mask increased by 1.97%, and the FPS has increased by 0.17. The reason is that the deeper network structure of EfficientNetB2 can extract more features. At the same time, the attention mechanism can be introduced to gather features on faces and masks, and the model has fewer parameters, thereby improves the accuracy and speed of detection. If EfficientNetB4 is used as a Backbone, the increase in network parameters causes a decrease in speed, and the AP value of face and face_mask also decreased. The reason is that the amount of data is small and the phenomenon of overfitting has occurred.

Transfer Learning: Transfer learning uses the training weights of ImageNet dataset as the weights of our dataset, which make neural network get better performance without using a large amount of data. Due to the small amount of training sample data, the network cannot learn more robust and stronger features, and the performance decline is

Table 4 Ablation study of Yolov3

Backbone	face	face_mask	mAP	FPS
Darknet53-CIoU	90.82%	95.26%	93.04%	12.91
EfficientnetB2	75.64%	89.96%	82.80%	12.99
EfficientnetB2-Transer Learning	92.79%	97.23%	95.01%	13.08
EfficientnetB2-Transer Learning- CIoU	93.88%	98.18%	96.03%	14.62
EfficientnetB4-Transer Learning- CIoU	88.62%	94.86%	91.74%	11.87

serious. The experimental results show that face and face_maks are reduced by 17.15% and 7.27% respectively without transfer learning.

CIOU: Compared with *IOU* as a loss function, *CIOU* can better reflect the distance, overlap rate, and scale information between the prediction box and ground-truth, which improves the overall performance of the network. It can be seen from Table 4 that the *mAP* is improved by 1.02%.

4.5.2 Mask classification ablation study

The ablation results of MobileNet in this paper are shown in Table 5. Without transfer learning, the *Accuracy* of MobileNet in the testset decreased by 15.43%, *Precision* and *Recall* decreased by 16.46% and 21.02 respectively for OK-mask data, and *Precision* and *Recall* decreased by 14.81% and 15.46% respectively for NG-mask data. Experimental results show that transfer learning can enhance the classification accuracy of the model.

5 Conclusion

In this paper, a fusion transfer learning and Efficient-Yolov3 of face mask-wearing detection algorithm is proposed to effectively solve the problems of poor lighting conditions, multiple objects, and occlusion in natural scenes. The algorithm uses EfficientNet as the backbone network of feature extraction and chooses *CIOU* as the loss function to improve the speed and accuracy of model detection. At the same time, transfer learning is introduced to improve the training speed and enhance the generalization ability of the model. The final *mAP* is 96.03%, and the FPS is 15. In addition, a dataset of mask classification is created to divide masks into qualified masks (N95 masks, disposable medical masks) and unqualified masks (cloth masks, sponge masks, scarves, etc.). A fusion transfer learning and MobileNet of mask, classification algorithm is proposed. This algorithm can effectively distinguish the types of masks and make detailed classification of the types of masks, with an accuracy of 97.84%. Experimentally verified that the algorithm proposed in this paper can effectively detect the wearing of face masks and classify the type of masks, which helps protect public health and plays a positive role in promoting the epidemic.

Table 5 Ablation study of MobileNet

model	OK-mask		NG-mask		Accuracy
	Precision	Recall	Precision	Recall	
MobileNet	80.62%	76.47%	83.59%	86.70%	82.41%
MobileNet- Transer Learning	97.08%	97.79%	98.40%	97.87%	97.84%

Acknowledgements This work is supported by National Natural Science Foundation of China under Grant 61902301, Shaanxi natural science basic research project under Grant 2021JQ692, the Scientific Research Program funded by Shaanxi Provincial Education Department, under Grant 19JK0364 and 20JK0647, Science and Technology Project of Xi'an Science and Technology Bureau (grant no. 21XJZZ0020).

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

- Betsch C, Korn L, Sprengholz P, Felgendreff L, Eitze S, Schmid P, Böhm R (2020) Social and behavioral consequences of mask policies during the covid-19 pandemic. *Proceedings of the National Academy of Sciences* 117(36):21851–21853
- Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection
- Chiang D (2020) Detect faces and determine whether people are wearing mask. <https://github.com/AIZOOTech/FaceMaskDetection>
- Ejaz MS, Islam MR, Sifatullah M, Sarker A (2019) Implementation of principal component analysis on masked and non-masked face recognition. In: 2019 1st International conference on advances in science, engineering and robotics technology (ICASERT). IEEE, pp 1–5
- Felzenszwalb PF, Mcallester DA, Ramanan D (2008) A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE Conference on computer vision and pattern recognition
- Ge S, Li J, Ye Q, Luo Z (2017) Detecting masked faces in the wild with lle-cnns. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2682–2690
- Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
- Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 580–587
- GitHub (2021) Yolov5-master. <https://github.com/ultralytics/yolov5/git/>
- He K, Zhang X, Ren S, Jian S (2016) Deep residual learning for image recognition. In: IEEE Conference on computer vision & pattern recognition
- He K, Zhang X, Ren S, Sun J (2015) Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37(9):1904–1916
- He K, Zhang X, Ren S, Sun J (2016) Identity mappings in deep residual networks
- Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. [arXiv:1704.04861](https://arxiv.org/abs/1704.04861)
- Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift
- Jiang M, Fan X (2020) Retinamask: A face mask detector. [arXiv:2005.03950](https://arxiv.org/abs/2005.03950)
- Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. *Communications of the Acm*
- Li C, Wang R, Li J, Fei L (2020) Face detection based on YOLOv3
- Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q et al (2020) Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*
- Li Y, Guo F, Cao Y, Li L, Guo Y (2020) Insight into covid-2019 for pediatricians. *Pediatric Pulmonology* 55(5):E1–E4
- Lin TY, Dollar P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: 2017 IEEE Conference on computer vision and pattern recognition (CVPR)
- Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: 2017 IEEE International conference on computer vision (ICCV)
- Liu C, Diab R, Naveed H, Leung V (2020) Universal public mask wear during covid-19 pandemic: Rationale, design and acceptability. *Respirology (Carlton, Vic.)* 25(8):895
- Liu W, Anguelov D, Erhan D, Szegedy C, Berg AC (2016) Ssd: Single shot multibox detector. In: European conference on computer vision
- Loey M, Manogaran G, Taha MHN, Khalifa NEM (2020) A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. *Measurement* 167:108288

25. Ma S, Lu B (2016) A face detection algorithm based on adaboost and new haar-like feature. In: IEEE International conference on software engineering & service science
26. MacIntyre CR, Seale H, Dung TC, Hien NT, Nga PT, Chughtai AA, Rahman B, Dwyer DE, Wang Q (2015) A cluster randomised trial of cloth masks compared with medical masks in healthcare workers. *BMJ open* 5(4):e006577
27. Ning C, Menglu L, Hao Y, Xueping S, Yunhong L (2020) Survey of pedestrian detection with occlusion. *Complex Intell Syst*:1–11
28. Qaseem A, Etzeandia-Ikobaltzeta I, Yost J, Miller MC, Abraham GM, Obley AJ, Forcica MA, Jokela JA, Humphrey LL (2020) Use of n95, surgical, and cloth masks to prevent covid-19 in health care and community settings: living practice points from the american college of physicians (version 1). *Annals of internal medicine* 173(8):642–649
29. Raina MacIntyre C, Jay Hasanain S (2020) Community universal face mask use during the covid 19 pandemic-from households to travellers and public spaces. *Journal of Travel Medicine* 27(3):taaa056
30. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: *Computer vision & pattern recognition*
31. Redmon J, Farhadi A (2017) Yolo9000: Better, faster, stronger. In: *IEEE Conference on computer vision & pattern recognition*
32. Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. *arXiv e-prints*
33. Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137–1149
34. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M, Xue Z, Shi Y (2020) Lung infection quantification of covid-19 in ct images with deep learning. [arXiv:2003.04655](https://arxiv.org/abs/2003.04655)
35. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition
36. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: *2016 IEEE Conference on computer vision and pattern recognition (CVPR)*
37. Szegedy C, Wei L, Jia Y, Sermanet P, Rabinovich A (2015) Going deeper with convolutions
38. Tan M, Le QV (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. [arXiv:1905.11946](https://arxiv.org/abs/1905.11946)
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30:5998–6008
40. Viola P (2001) Rapid object detection using a boosted cascade of simple features. In: *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*
41. Wang Y, Hu M, Li Q, Zhang XP, Zhai G, Yao N (2020) Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with covid-19 in an accurate and unobtrusive manner. [arXiv:2002.05534](https://arxiv.org/abs/2002.05534)
42. Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, Yi P, Jiang K, Wang N, Pei Y et al (2020) Masked face recognition dataset and application. [arXiv:2003.09093](https://arxiv.org/abs/2003.09093)
43. Woo S, Park J, Lee JY, So Kweon I (2018) Cbam: Convolutional block attention module. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 3–19
44. Yang S, Luo P, Loy CC, Tang X (2016) Wider face: A face detection benchmark. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 5525–5533
45. Zheng Z, Wang P, Liu W, Li J, Ren D (2020) Distance-iou loss: Faster and better learning for bounding box regression. In: *AAAI Conference on artificial intelligence*
46. Zou Z, Shi Z, Guo Y, Ye J (2019) Object detection in 20 years: A survey. [arXiv:1905.05055](https://arxiv.org/abs/1905.05055)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.