2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) | 978-1-7281-9638-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/MLBDBI51377.2020.00100

# Intelligent detection and recognition system for mask wearing based on improved RetinaFace algorithm

Bin Xue, Jianpeng Hu, Pengming Zhang
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science
Shanghai, China

*Abstract*—**The COVID-19 can be transmitted through airborne droplets, aerosols and other carriers. In order to better reduce people's risk of infection, individuals need to wear masks to prevent the spread of the virus when going out to public places, seeking medical treatment, and taking public transportation. This paper is based on the improved RETINAFACE algorithm, which effectively realizes the detection of mask wearing, and on the basis of this algorithm, realizes the function of judging whether the mask is worn correctly. In the face recognition algorithm, this paper designs a face recognition algorithm with higher accuracy. The system combines a face mask wearing detection algorithm, a mask standard wearing detection algorithm and a face recognition algorithm. In addition, this article adds a voice prompt module to better assist the integrity of the system's functions. The test results of the final experiment show that the system can effectively achieve the purpose of face mask detection and recognition.**

***Keywords- Face mask detection; mask standard wearing detection; face recognition; RetinaFace algorithm***

## I. INTRODUCTION

Beginning in December 2019, the sudden new type of coronavirus pneumonia (COVID-19) quickly raged across the country and even the world [1]. As of July 15, 2020, more than 13.65 million confirmed cases have been reported in more than 220 countries and regions around the world, and more than 580,000 patients have died. At present, it is still continuing to spread on a large scale [2]. The new type of coronavirus is highly infectious. It can be spread through contact, droplets, aerosols and other carriers in the air, and it can survive for 5 days in a suitable environment [3]-[4]. The "Guidelines for the Prevention of New Coronavirus Infection Pneumonia" issued by the National Health Commission emphasized that when individuals go out to public places, seek medical treatment and take public transportation, they need to wear medical surgical masks or N95 masks to prevent the spread of the virus. Therefore, it is everyone's responsibility to wear masks in public places during the epidemic, but this requires not only the conscious compliance of the individual, but also the adoption of certain measures to supervise and manage.

At present, although there is no algorithm specifically applied to face mask wearing detection, with the development of deep learning in the field of computer vision [5-7], neural network-based target detection algorithms are used in pedestrian target detection, face detection, and remote sensing image targets. Detection, medical image detection and natural scene text detection are widely used in fields [8]-[11]. Face recognition algorithms rely on a high degree of recognition accuracy, and have huge application potential in classroom attendance, identity authentication, access control systems, login and unlocking, and social media platforms [12].

At present, face recognition devices on the market have relatively single functions and have relatively high requirements on faces. When the face is in a state of large-area occlusion, the recognition accuracy drops rapidly. Especially in the face of the current epidemic situation where all people wear masks, the capabilities of traditional face recognition systems appear to be stretched. Taking into account that we will try our best to resume production and work while ensuring people's safety, we have designed a smart detection and recognition system for mask wearing. The system is mainly composed of face mask detection algorithm and face recognition algorithm. The main functions of the system can be divided into three parts: face mask detection, face recognition, and voice prompts. When multiple pedestrians pass by the camera, the camera equipped with this algorithm will first detect the pedestrian's face mask. When the pedestrian wears the mask normally, it will not give a voice prompt. When a pedestrian wears a mask incorrectly, the voice will announce to remind him to wear the mask correctly. When a pedestrian is not wearing a mask, the system will trigger the face recognition module to speak his name and remind him to wear a mask. The system can be used in high-speed rail stations, subways, shopping malls and other crowded areas.

Through researching related target detection algorithms, it is found that the deep learning model used for face detection can be applied to the task of mask wearing detection. In this paper, the more accurate face detection algorithm RETINAFACE [13] is used as the basic algorithm for mask face detection, and on this basis, the network structure of the RETINAFACE algorithm is improved, and the attention mechanism is introduced to meet the needs of new functions; In this system, we calculate the mask and the key point positions of the face, and the confidence that the mask is worn on different faces is returned to determine whether the person wears the mask in a standard manner. The calculation is fast and accurate, and the algorithm is stable and efficient; for the current popular ones For the face recognition method, we use the DEEPFACE [14] algorithm. The algorithm divides the face recognition problem into several related sub-problems. Each sub-problem is completed by a machine learning algorithm, and a pipeline is constructed to recognize faces through the following four steps.

(1) Find all the faces in the picture

(2) Adjust the different postures of the face

(3) Coding the face

DOI 10.1109/MLBDBI51377.2020.00100

(4) Find the person's name from the code

Considering that the system needs to detect and recognize different face poses, various lighting and other actual scenes, after the collection and comparison of the mask face data set, this project finally chose the WIDER FACE [15] data set and MAFA [16] There are a total of 7959 images in the data set, including 6067 in the training set, 300 in the validation set, and 1592 in the test set. And done data preprocessing operations such as color reversal, random cropping, and horizontal flipping; in addition, the system has added a voice prompt function, and when it detects that it is not worn or worn out, it will broadcast information to personnel in time by calling win32API SPIVOICE speech synthesis service.

## II. Scheme demonstration and design

Since there is currently no algorithm specifically applied to face mask wearing detection, the traditional single-target recognition algorithm has a relatively single function, which requires relatively high requirements for the integrity of the face and the intensity of light, and the application scenarios are limited. For this reason, we have designed a set of intelligent detection and recognition system for multi-target mask wearing in the case of relatively dense human traffic. The main process of the system is shown in Figure 1:

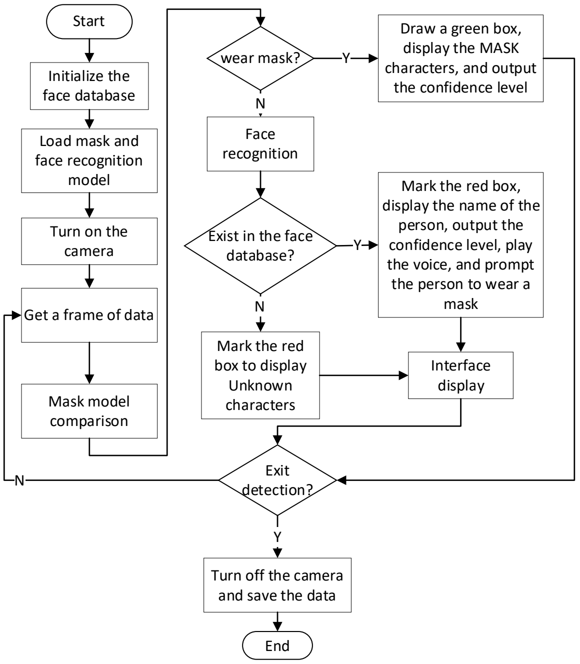

Start
Initialize the face database
Load mask and face recognition model
Turn on the camera
Get a frame of data
Mask model comparison
wear mask?
Y
Draw a green box, display the MASK characters, and output the confidence level
N
Face recognition
Exist in the face database?
Y
Mark the red box, display the name of the person, output the confidence level, play the voice, and prompt the person to wear a mask
N
Mark the red box to display Unknown characters
Interface display
Exit detection?
N
Y
Turn off the camera and save the data
End

Figure 1. System logic diagram

There are four main functional modules designed as follows:

(1) In terms of face mask detection, we have improved the commonly used RETINAFACE face algorithm, and added the mask detection function based on this algorithm, which can detect whether a person is wearing in real time, which is efficient and stable. At the same time, it also broadens the field of algorithm application. First obtain a frame of data through the camera, then process the image through the improved RETINAFACE algorithm, identify and locate the key points of the face through the face detection algorithm, and then locate the position of the mask through the target detection algorithm, if the mask covers the face including the nose and mouth The key point is that the personnel should wear it in a standard manner and mark the face with a green frame. If the mask only covers the key points of the face below the nose and mouth, it is considered that the mask is not properly worn, and the face is marked with an orange frame on the interface to show the confidence of wearing the mask.

(2) Regarding the detection of non-standard wearing of masks, we calculate the position of the key points of the mask and the face, return the confidence of wearing the mask, compare the threshold information, and judge whether the person wears the mask in a standard way.

(3) In terms of face recognition, we use the high-performance and widely used DEEPFACE face recognition library to encode and convert the face library images into 128-bit measurement values, and then do the same processing on the video frames, compare and find the processing After the data and the data in the face encoding library, return the measured value in the closest cognitive face library to obtain the person's name. After that, the person’s name and warning information are broadcast by voice, prompting the person to wear a mask.

(4) In terms of voice broadcast, first obtain the Com object of WINAPI, call the system's underlying voice synthesis service through WINAPI, and then pass in the information to be broadcast. Remind people who do not wear masks or who do not wear them properly to wear masks to reduce the risk of infection.

## III. Algorithm design

### *A. Principle of RETINAFACE algorithm*

Since the RETINAFACE algorithm can only perform face detection and does not meet the needs of the actual scene of this project, this article has been improved on the basis of the RETINAFACE algorithm. Section 3.2.1 introduces the principle of the RETINAFACE algorithm, the basic network of face mask detection, and section 3.2.2 introduces the improvements made to achieve specific functions based on the RETINAFACE algorithm.

RETINAFACE is a robust single-level face detection algorithm. The algorithm uses the advantages of multi-task joint additional supervised learning and self-supervised learning to perform pixel-level localization of faces of different scales. The algorithm incorporates excellent modeling ideas such as feature pyramid network, context network and task union.

#### *1) RETINAFACE feature extraction network*

In the feature extraction network of the RETINAFACE algorithm, five levels of feature pyramids from P2 to P6 are used, where P2 to P5 are connected by the corresponding residual network (Residual Network [18] output feature maps (C2 to C6) respectively Top-down and horizontal connection calculations. P6 is obtained by convolution sampling by C5 using a 3×3 convolution kernel with Stride=2. C1 to C5 use ResNet-512 on the ImageNet-11 dataset. The trained residual layer [19] uses the "XAVIEER" method [20] for the P6 layer to perform random initialization.

The RETINAFACE algorithm uses five independent context modules, corresponding to the five feature pyramid levels from P2 to P6, to increase the scope of the receptive field and enhance the robust contextual semantic segmentation capabilities. In addition, a Deformable Convolutional Network (DCN) [] is used to replace the 3×3 convolutional layer in the context module, which further strengthens the non-rigid context modeling ability.

*2) RETINAFACE multi-task loss function*

For a trained anchor box I, the multi-task joint loss function is defined as:

$$L{=}L_{\mathrm{cls}}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) \tag{1}$$

Among them, $L_{\mathrm{cls}}$ is the classification loss function, $p_i$ is the probability that the anchor box contains the predicted target, $p_i^* \in (0,1)$ means negative anchor box and positive anchor box. $L_{box}$ is the target detection frame regression loss function, $t_i = \{t_x, t_y, t_w, t_h\}$ represents the coordinate information of the prediction box related to the positive anchor box, for the same reason, $t_i = \{t_x^*, t_y^*, t_w^*, t_h^*\}$ represents the coordinate information of the prediction box related to the negative anchor box. $L_{pts}(l_i, l_i^*)$ is the facial landmark regression loss function, $l_i = \{l_{x1}, l_{y1}, \ldots l_{x5}, l_{y5}\}$ and $l_i = \{l_{x1}^*, l_{y1}^*, \ldots l_{x5}^*, l_{y5}^*\}$ are the five face landmark points predicted by the positive anchor point box and the five face landmark points labeled. $L_{pixel}$ represents the regression loss function of face dense points. $\lambda_1$, $\lambda_2$ and $\lambda_3$ represent the loss balance weight parameter, In the RETINAFACE algorithm, it is set to 0.25, 0.1 and 0.01 respectively, which means that the information of the detection frame and facial landmark points is added in the supervised learning.

## B. Improved RETINAFACE algorithm

This improved algorithm framework is divided into three parts: feature pyramid network, context network and multi-task joint loss. Among them, the backbone network in the feature pyramid is ResNet-512, which is used for feature extraction and introduces an attention mechanism module to enhance the expressive ability of feature maps. In the multi-task joint loss, irrelevant facial intensive point regression loss is discarded, which improves the training speed and efficiency of the algorithm model.

*1) Feature extraction network*

Use the pre-trained Resnet-512 as the backbone network of the feature pyramid network for feature extraction. Except for the 7×7 convolution in the first layer, the remaining 4 layers are composed of residual connection units. For the RES_N layer, it contains n A residual connection unit. Using residual connection can effectively solve the problem of gradient disappearance or gradient explosion when deep network training.

In the residual connection unit, for the input feature vector x and the output feature vector y, the calculation formula established by the residual connection is:

$$\mathrm{y} = \sigma(f(x, \{W_i\}) + x) \tag{2}$$

Among them, $\sigma$ represents the linear correction unit (Rectified Linear Unit, RELU) activation function, $W_i$ represents the weight parameter, $f(x, \{W_i\})$ represents the residual mapping that needs to be learned. For the three-layer residual connection unit in the figure, the calculation method is shown in formula (3). The addition operation is performed by shortcut connection and element-wise addition, and after the addition, the RELU activation function is used for nonlinearization again.

$$f(x, \{W_i\}) = W_{3\sigma}(W_{2\sigma}(W_1 x)) \tag{3}$$

*2) Improved self-attention mechanism*

This system introduces the attention mechanism module in the mask face detection algorithm. Its internal structure is shown in Figure 2, which mainly includes the Pyramid Attention Mechanism (PAM) [21] and the Self-Attention Mechanism (Self-Attention) , SA) [22]. Pyramid attention mechanism can enhance the expressive ability of feature maps, and self-attention can make better use of the above relationship of features and improve the descriptive ability of attention feature maps.

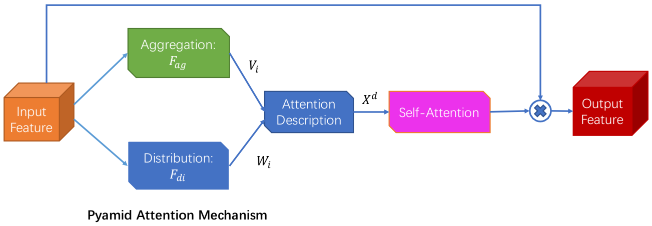


Figure 2. Attention mechanism network

*3) Multitask joint loss*

Based on the loss function of the face detection algorithm RETINAFACE, in order to improve the training speed and detection efficiency of the algorithm, this system only retains the relevant classification loss, the detection frame regression loss and the facial landmark regression loss, and optimizes them. Removed facial dense spots and return loss. The total loss function is expressed as:

$$L{=}L_{\mathrm{cls}}(p_i, p_i^*) + \lambda_1 p_i^* L_{box}(t_i, t_i^*) + \lambda_2 p_i^* L_{pts}(l_i, l_i^*) \tag{4}$$

Each variable is defined as formula (1), where the classification loss $L_{\mathrm{cls}}(p_i, p_i^*)$ is a two-class classification (complete face and masked face) made by the cross-entropy loss function. The detection frame regression loss $L_{box}(t_i, t_i^*)$ uses the Smooth-L1 loss function. The facial landmark regression $L_{pts}(l_i, l_i^*)$ uses the Smooth-L1 loss function to normalize the five detected facial landmark points. In addition, this paper sets the loss balance weight parameters $\lambda_1$ and $\lambda_2$ to 0.3 and 0.1 respectively.

## C. Detection principle of non-standard mask wearing

*1) Principle of face key point detection*

Face key point detection is a key step in the field of face recognition and analysis. It is the premise and breakthrough point for the correct detection of masks. Therefore, we conducted research on face key point detection for deep learning methods, and marked 68 key points on the face as shown in the figure. From Figure 3, we find that we use the eyes, nose and mouth as the key points to detect whether the mask is worn correctly. From the figure, it is found that the key points of the cheek and mouth and nose have shifted to varying degrees after wearing the mask. The distance between the center of the nose and the eyebrows, and the confidence level is output. Determine whether to wear masks as a standard.

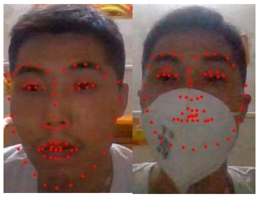


Figure 3. Changes of facial key point position information

*2) Implementation*

In order to reduce interference and accurately detect masks, we first use the FACE_LOCATION function to quickly locate the face, return key point data such as the face, nose, nose, and eyes, and then call the DETECT_MASK function to detect the mask feature information at the face position and calculate the confidence. Returns the label information of whether to wear a mask, the coordinates of the face position, etc. Because there may be multiple faces in the image, it is necessary to read out the face position information in a loop. In order to quickly process the mask classification detection, we store the information returned by FACE_LOCATION in the NOMASKDATA and MASKDATA arrays according to the returned label information.

### *D. Face recognition*

Face recognition generally consists of several parts: First, find all faces in a picture (faces are detected). For every face, no matter the light is bright or dark or facing away, it can still recognize the same person's face (different face poses). Then you can find out the unique features that can be used to distinguish others from each face, such as how big the eyes are, how long the face is, and so on. Finally, compare the characteristics of this face with all known faces to determine the person's name. These problems need to connect several machine learning algorithms together to build a pipeline so that the previous output can be used as the next input to start the pipeline operation.

(1) Face detection. Use the method of Histogram of Oriented Gradients [23], referred to as HOG. Find the face in this HOG image. Find the part of the image that looks most similar to some known HOG patterns.

(2) Different poses of the face. After separating the face in the picture, the next problem to be dealt with is the same face facing different directions, which is a different thing. The typical processing method is to use the algorithm of face landmark estimation. The basic idea is to find specific points (called landmarks) that are common on 68 people's faces-including the top of the chin, the outer contour of each eye, and the inner contour of each eyebrow. Then train a machine learning algorithm so that it can find these 68 key points on any face. Make some transformations to keep the picture relatively parallel, for example, do affine transformation to rotate and scale the picture.

(3) Face coding. Train a deep convolutional neural network and use it to generate 128 measurements for the face. The method of reducing complex original data (such as pictures) into a series of numbers that can be generated by a computer, and encoding the image of the face information database to generate 128-bit measurement values. The process of training a convolutional neural network to output the face embedding requires a lot of data and powerful computing power. After spending a certain amount of time, obtain the face coding library.

(4) Find out the person's name from the code, train a machine learning classification algorithm, use it to find the code value of the person closest to the measured value of the test image and the database and return the name.

The overall process of face recognition is shown in Figure 4:

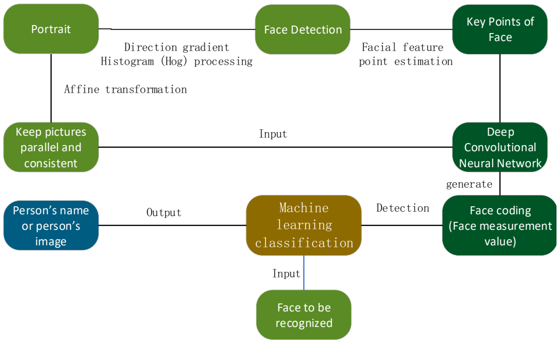

Portrait
Face Detection
Key Points of Face
Direction gradient Histogram (Hog) processing
Facial feature point estimation
Affine transformation
Keep pictures parallel and consistent
Input
Deep Convolutional Neural Network
generate
Person's name or person's image
Output
Machine learning classification
Detection
Face coding (Face measurement value)
Input
Face to be recognized

Figure 4. The overall process of face recognition

## IV. Experimental Results and Analysis

In order to verify the ability of the trained face mask detection model and face recognition model in complex scenes. This experiment is described by setting up some evaluation indicators for evaluation and using real scene detection and recognition videos.

### *A. Evaluation index*

This project will evaluate the face mask detection algorithm, and use the 1592 pictures set in section 4.1.1 for experimental test analysis. The evaluation indicators use the commonly used ROC curve (Receiver Operating Characteristic curve) [24], average precision AP (Average Precision), and mean average precision MAP (Mean Average Precision) commonly used in the field of target detection to evaluate the effect of the algorithm on face and mask wearing detection . Among them, the detection effect of a single target when the value of AP turns red is calculated as:

$$AP=\int_{0}^{1} p(r)dr \qquad (5)$$

Among them, is the mapping relationship between the true positive rate and the recall rate. The calculation method of the true positive rate and the recall rate is:

$$p = \frac{TP}{TP + FP}$$

$$r = \frac{TP}{TP + FN} \tag{6}$$

Among them, TP (True Positive, true number) represents the number of positive samples predicted as positive samples, FP (False Positive, false positive number) represents the number of negative samples predicted as positive samples, FN (False Negative, false Negative number) indicates the number of positive samples predicted as negative samples. In this paper, the ROC curve is drawn based on the relationship between the true rate and the false positive number to express the performance of target detection.

### *B. System overall function realization*

The system has completed the expected design and realized all the functions of face detection, mask wearing detection, wearing standard detection, face recognition, voice prompts, etc. After the system is turned on, within the detectable range of the camera, real-time multi-person detection, recognition and prompting in natural scenes can be realized.

### *C. Face mask detection experiment results*

In order to fully verify the test system, we have done two sets of experiments: different types of masks, two parts of multi-person testing, as shown in the figure below. As can be seen from the figure, no matter what kind of situation, the system can judge correctly. The system is stable and universal. Easy to apply.

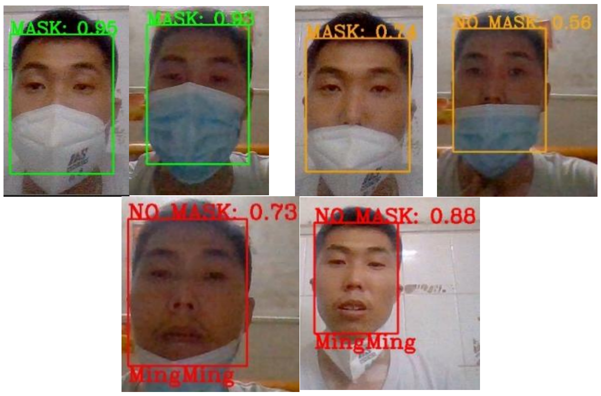
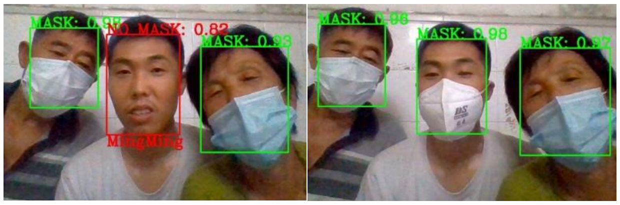


Figure 5. Test results under different types of masks



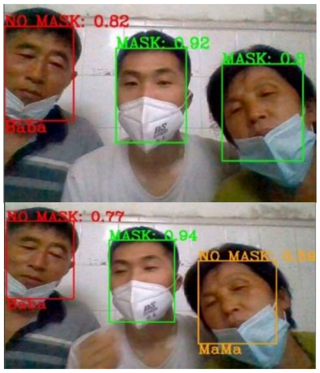


Figure 6. Multi-person detection effect

For model performance, we use the ROC curve to evaluate the mask recognition module. The ROC curve graph is a curve reflecting the relationship between sensitivity and specificity. The abscissa X axis is 1-specificity, also known as false positive rate (false positive rate), the closer the X axis is to zero, the higher the accuracy rate; the ordinate Y axis is called sensitivity, also known as true positive rate (sensitivity ), the larger the Y axis, the better the accuracy.

$$Sensitivity = \frac{TP}{TP + FN} \quad Specificity = \frac{TN}{TN + FP} \tag{8}$$

We compare with RETINAFACE and the results are as follows:

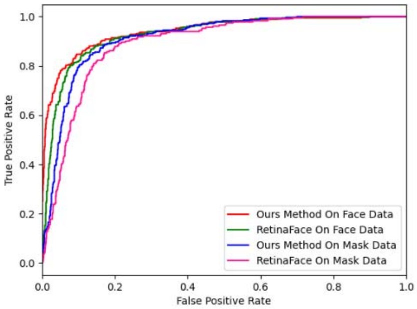


Figure 7. ROC curve characteristic diagram

### *D. Test results of irregular wearing of masks*

For this type of experiment, we designed three sets of experiments: covering the mouth and nose, exposing the nose, and having a mask under the mouth. Results as shown below. The mask irregularity detection partly depends on the returned confidence and the selection of the threshold. We select the threshold from 0.1 to 1.0 every 0.1 step. The experimental results are shown in Figure 8.

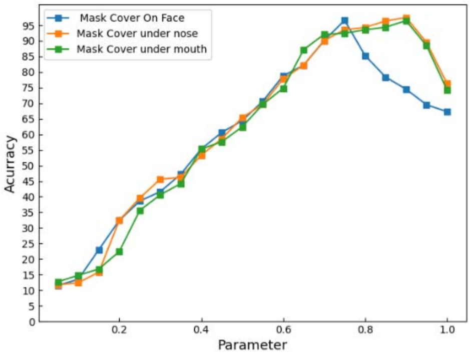


Figure 8. The experimental results of the irregular wearing of masks

### *E. Face recognition experiment results*

We test and train the face recognition model on the WILDER dataset, and we use the ROC curve to measure the performance of the model. It is observed that the ROC curve 0.6 increases sharply, and the AUC area reaches 0.94, indicating that the classification of the model is obvious and the effect is significant.

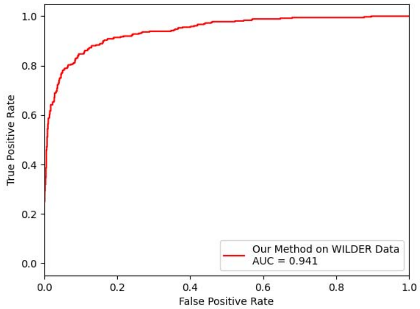


Figure 9. ROC curve of face recognition

## V. FUTURE WORK

In response to several problems encountered in the actual design of the system, we plan to improve the face mask intelligent detection and recognition system from the following aspects in the future:

(1) In order to increase the flexibility of the application area, the intelligent detection and recognition system for mask wearing can be transplanted to the mobile terminal in the future. By making a simple APP, the mobile terminal can be deployed in any area where detection and recognition are desired to achieve a more practical purpose.

(2) In order to enable the system to run on various machines with weak computing power, we plan to design a simpler network in the future, reduce the number of parameters in the model, reduce the cost of training, and improve the compatibility of the system with machines.

## REFERENCES

[1] Mehta P, McAuley D F, Brown M, et al. COVID-19: consider cytokine storm syndromes and immunosuppression[J]. Lancet (London, England), 2020, 395(10229): 1033.

[2] Fauci A S, Lane H C, Redfield R R. Covid-19—navigating the uncharted[J]. 2020.

[3] Bai Lang, Wang Ming, Tang Xiaoqiong, et al. Thinking on hot issues in the diagnosis and treatment of new coronavirus pneumonia[J]. West China Medicine, 2020, 35(2): 125-131.

[4] Chan J F, Yuan Shuofeng, Kok K H, et al.A familial clus-ter of pneumonia associated with the 2019 novel coro-navirus indicating person-to-person transmission: a study.of a family cluster[J].The Lancet, 2020, 395(10223): 514-523.

[5] Dekel R. Human prception in computer vision [J]. arXiv preprint arXiv: 1701. 04674, 2017.

[6] Sinha R K, Pandey R, Pattnaik R. Deep Learning for Computer Vision Tasks: A review [J]. arXiv preprint arXiv: l 804.03928, 2018.

[7] Bhandary A, Sudeepa K B, Chokkadi S, et al. A study on various state of the art of the art face recognition system using deep learning techniques[J].International Journal of Advanced Trends in Computer Science and Engineering, 2019, 8(4): 1590-1600.

[8] Liu Quan, Zhai Jianwei, Zhang Zongchang, et al. A review of deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 3-29.

[9] Milan C.A parallel fortran framework for neural networks and deep learning[J].ACM SIGPLAN Fortran Forum, 2019, 38(1): 4-21.

[10] Zhang S, Zhu X, Lei Z, et al. Detecting face with densely connected face proposal network[J].Neurocomputing, 2018, 284: 119-127.

[11] Diao W, Sun X, Zheng X, et al. Efficient saliency-based object detection in remote sensing images using deep belief networks[J].Geoscience and Remote Sensing Letters, 2016, 13(2): 137-141.

[12] Litjens G, Kooi T, Bejnordi B E, et al.A survey on deep learning in medical image analysis[J].Medical Image Analysis, 2017, 42: 60-88.

[13] Zhu C, Tao R, Luu K, et al. Seeing small faces from robust anchor's perspective[C]//Proceedings of the 2018 IEEE/ CVF Conference on Computer Vision and Pattern Recognition, 2018: 5127-5136.

[14] Zhang Cuiping, Su Guangda. Overview of Face Recognition Technology[J]. Chinese Journal of Image and Graphics: Series A, 2000 (11): 885-894.

[15] Deng J, Guo J, Zhou Y, et al. RetinaFace: single-stage dense face localisation in the wild [EB/OL]. [2020-02-10].https: // arxiv.org/abs/1905.00641.

[16] Taigman Y, Yang M, Ranzato M A, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 1701-1708.

[17] Yang S, Luo P, Loy C C, et al. Wider face: A face detection benchmark[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 5525-5533.

[18] Ge S, Li J, Ye Q, et al.Detecting masked faces in the wild with LLE-CNNs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017：426-434.

[19] Lawrence S, Giles C L, Tsoi A C, et al. Face recognition: A convolutional neural-network approach[J]. IEEE transactions on neural networks, 1997, 8(1): 98-113.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[21] Xavier G, Yoshua B.Understanding the difficulty of training deep feedforward neural networks[J].Journal of Machine Learning Research, 2010, 9: 249-256.

[22] Dai J, Qi H, Xiong Y, et al.Deformable convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 764-773.

[23] Li H, Xiong P, An J, et al. Pyramid attention network for semantic segmentation[J]. arXiv preprint arXiv:1805.10180, 2018.