Applied Intelligence (2021) 51:3012–3025
https://doi.org/10.1007/s10489-020-02100-9

Cropping and attention based approach for masked face recognition

Yande Li[1] · **Kun Guo**[1] · **Yonggang Lu**[1] · **Li Liu**[2]

Accepted: 26 November 2020 / Published online: 1 February 2021

Abstract

The global epidemic of COVID-19 makes people realize that wearing a mask is one of the most effective ways to protect ourselves from virus infections, which poses serious challenges for the existing face recognition system. To tackle the difficulties, a new method for masked face recognition is proposed by integrating a cropping-based approach with the Convolutional Block Attention Module (CBAM). The optimal cropping is explored for each case, while the CBAM module is adopted to focus on the regions around eyes. Two special application scenarios, using faces without mask for training to recognize masked faces, and using masked faces for training to recognize faces without mask, have also been studied. Comprehensive experiments on SMFRD, CISIA-Webface, AR and Extend Yela B datasets show that the proposed approach can significantly improve the performance of masked face recognition compared with other state-of-the-art approaches.

Keywords Masked face recognition · Cropping-based approach · Attentional mechanism

1 Introduction

The epidemic of COVID-19 has shocked the world seriously and is threatening the lives and health of people all over the world. In order to protect us from the viruses, we must wear masks when going out, especially in places with many people, which poses a huge challenge for face recognition. However, in personal identification scenarios of airports and stations, and in authentication scenarios of communities, schools, companies, etc., we have to take off our mask for face recognition. When the mask is removed, it is easily infected by viruses in the air, and contact with the mask frequently also increases the chance of exposure to the virus. Therefore, it is very important to explore solutions for masked face recognition, which will also provide support for the prevention and control of infectious diseases that may occur in the future.

Face **R**ecognition (FR) has always been popular since Turk et. al [1] proposed historical eigenface approach in the early 1990s. Then local-feature approaches and shallow feature learning have been the focus of research for a long time. Until the emergence of DeepFace [2] and DeepID [3] in 2014, which shifted the research focus to deep learning-based approach. Current state-of-the-art methods, such as ArcFace [4] and CosFace [5], have achieved over 99.5% accuracy on LFW dataset. Deep learning has achieved trustworthy success in general face recognition, but the deep features still cannot solve the uncontrollable variations like illumination, pose and occlusion, etc., especially occlusion. A review on deep Learning-based face recognition can be found in [6].

Facial occlusion is still one of the most challenging problems in FR. In real life, there are different types of occlusion objects and different occlusion areas. The detection of occlusions is a major problem in conventional occlusion FR. Ge et.al [7] proposed LLE-CNNs for masked face detection. Wang et al. [8] proposed Face Attention Network (FAN) to

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

✉ Yonggang Lu
ylu@lzu.edu.cn

✉ Li Liu
dcsliuli@cqu.edu.cn

Yande Li
liyd19@lzu.edu.cn

Kun Guo
guok19@lzu.edu.cn

[1] Lanzhou University, Lanzhou, China

[2] Chongqing University, Chongqing, China

improve face detection performance in the occluded cases, an anchor-level attention was introduced to highlight the face region features. Occlusion face detection requires complex network computing to achieve satisfied results. Then how to deal with the area damaged by occlusion is another important issue. The core problem is to reduce the negative impact of occlusion in FR. Du et al. [9] adopted a dictionary containing occluded images to construct a novel reconstruction model for occlusion FR. Song et al. [10] utilized the feature-level difference between occluded and occlusion-free face pairs to eliminate the influence of damaged features. In [11], the non-occluded images synthesized by Generative Adversarial Networks (GAN) were used for refined face recognition. A review on occlusion FR can be found in [12].

Masked **F**ace **R**ecognition (MFR) is a special case in occlusion FR. Different from routine occlusion FR, there are three main difficulties in MFR. Firstly, there lacks of large face dataset with masks. Secondly, the features of mouth and nose are severely damaged, and the effective features are greatly reduced. Finally, the face wearing a mask is hard to detect. At the same time, the existing deep learning methods are difficult to solve two special cases: using masked data for training and using non-mask data for testing; using non-mask data for training and using masked data for testing. However, the two cases are crucial in some special situations. For example, conventional FR systems cannot recognize faces wearing masks under the COVID-19 epidemic; The police hold a large number of occlusion images of the suspects, but lack of clean face images.

To tackle the above difficulties, we subdivide MFR into three masked cases and one normal case, as shown in Fig. 1. Simulated Masked Face Recognition Dataset (SMFRD) [13] is adopted to train our MFR model. MTCNN [14] is introduced to process original images, including face area detection, face key points detection and face aligned. To reduce the negative impact of damaged regions, we propose two solutions, which are attention-based approach and cropping-based approach. In attention-based approach, the masked features are given a lower weight, and features around eyes are given a higher weight. In cropping-based approach, the face images after removing the masked regions are used for model training. And the optimal cropping for each case is explored. Lastly, the attention-based approach is integrated with cropping-based approach to maximize the advantages of both approaches.

Overall, the main contribution in this paper lies in threefolds.

- We propose a cropping-based approach and explore the optimal cropping for each MFR case.
- We tackle the difficulties of MFR in two special application scenarios: using masked faces for training to recognize faces without mask; using faces without mask for training to recognize masked faces.
- The integration of CBAM and cropping-based approach shows superior performance over other state-of-the-art approaches on MFR.

The rest of the paper is organized as follows. In Section 2, we summarize the work related to occlusion FR, attention mechanism and MFR for COVID-19. Section 3 introduces CBAM, cropping method, integration approach and network structure in detail. Experiment settings, datasets, results and discussion are introduced in Section 4. Lastly, we draw conclusions around the experimental results in Section 5.

2 Related work

In this section, we will highlight the works relevant to occlusion FR, attention mechanism and MFR for COVID-19.

2.1 Occlusion face recognition

Occlusion FR is one of the most challenging tasks and has attracted the attention of many researchers. In general, the methods of occlusion FR can be divided into three

Fig. 1 Cases for MFR. Case 2 and case 3 are two special cases and case 4 is conventional face recognition

categories: image reconstruction, occlusion discarding and deep learning-based approaches.

Image Reconstruction Image reconstruction means that the probe face image is re-expressed by a linear combination of gallery images. A pioneering work is Sparse Representation based Classification (SRC) on robust occlusion FR [15]. Then several extended version were proposed for specific problems in FR, such as Extended SRC (ESRC) for under-sampled FR [16], Group Sparse Coding (GSC) for increasing discriminative power [17]. These methods focused on linear feature space and did not take occlusion into consideration. Yuan et al. [18] proposed support vector discrimination dictionary and Gabor occlusion dictionary based SRC (SVGSRC) for occlusion FR. Sparse representation and particle filtering are combined in [19] for tracking target under partial occlusion and illumination variation. In [20], a real FR system that could solve illumination variation, image misalignment and partial occlusion was proposed. Cen et al. [21] proposed a classification scheme based on depth dictionary representation for robust occlusion FR, which used a convolutional neural network as a feature extractor, and then used the dictionary to linearly encode the extracted depth features. To utilize 2D structure of error images, Yang et al. [22] presented a 2D image-matrix-based error model named Nuclear norm based Matrix Regression (NMR) for occlusion FR, which showed superiority over previous regression-based methods. Chen et al. [23] proposed a sparse regularized NMR method by introducing l_1-norm constraint instead of l_2-norm on the representation of the NMR framework, which was beneficial to recover low-rank error images under occlusion and illumination changes. Du et al. [9] proposed Nuclear Norm based Adapted Occlusion Dictionary Learning (NNAODL) for occlusion FR, where a dictionary containing occluded images was adopted to construct a novel reconstruction model.

The image reconstruction approaches have made great process in occlusion FR, but there are three main limitations. (1) These methods need overcomplete dictionary, but the gallery images of each subject are usually inadequate in practical FR scenarios. (2) A large increase in gallery images will lead to a sharp increase in the complexity of the sparse representation solution. (3) These approaches cannot be generalized well because that they require probe images have identical subjects with the gallery images.

Occlusion Discarding Occlusion discarding approach aims to discard features corresponding to the occluded part in FR, which is based on the fact that the features damaged by the occluded part have a negative effect on FR. Generally, this approach often takes two steps. First, the occlusion part is detected from the face image, and then the remaining clean part is used for the recognition process. SVM was often used as a binary classifier to detect the occluded area in images, and then the non-occluded part is used for face recognition [24, 25]. Andres et al. [26] calculated the difference between occluded and non-occluded face images of the same subject to detect the occluded regions and discard them at recognition phase. However, the shallow features used in above methods have limited the discriminative ability. Song et al. [10] developed Pairwise Differential Siamese Network (PDSN) to generate mask dictionary using the differences between the top conv features of occluded and occlusion-free face pairs, which indicated the correspondence between occluded facial areas and damaged feature elements. This approach aims to eliminate the negative effects of damaged areas from depth features. However, the requirements of paired pictures are difficult to satisfy in practical applications. In addition, the above approaches spend a lot of efforts to detect and discard the occlusion regions, the cropping-based approach proposed in this paper aim to simplify these steps by finding out the optimal cropping.

Deep Learning In recent years, deep learning has achieved great success in FR. Deep features has shown superior performance over shallow features and is popular in occlusion FR. In [27], Dynamic Feature Matching (DFM) approach, combining FCN and SRC, was proposed to recognize partial faces of arbitrary size, where the deep features after the last pooling layer were linearly represented by gallery feature maps. Similar as the SRC related work, an overcomplete dictionary is necessary in DFM. In [10], the differences between the top convolution features of occluded and clean face pairs were used to establish mask dictionary, then the damaged features were discarded by querying the mask dictionary. But this approach requires paired images that are not easily satisfied. Duan et al. [11] proposed an end-to-end BoostGAN model for occlusion but profile FR, in which the occluded image was first used to synthesize non-occluded image, then the non-occluded image was used for refined face recognition. However, GAN-based methods are hard to reproduce the details of the key points on the face, especially for large area occlusion like facemask.

These approaches show the latest progress of deep learning technology in occlusion FR. However, considering that the key discriminative features of the nose and mouse are completely damaged by the facemask, most of current approaches are not suitable for MFR in practical life. In this paper, we introduce attention mechanism in ResNet50 network for MFR, we hope it can focus on more expressive and discriminative features like eyes and assign a very low weight to the masked area.

2.2 Attention mechanism

Attention Mechanism (AM), inspired by human attention mechanisms, is first known for its excellent performance in natural language processing [28]. It has become an important component of neural networks, and has been widely used in various fields of deep learning, such as image processing [29, 30], speech recognition [31, 32] and natural language processing [33, 34].

In image classification task, attention has been applied successfully. Wang et al. proposed a Residual Attention Network for image classification by stacking multiple attention module [35], where it encoded top-down attention mechanism into bottom-up top-down feedforward convolutional structure in each Attention Module. However, its computational overhead is too large and it is not convenient to be integrated with other pre-existing CNN architectures. Hu et al. proposed Squeeze-and-Excitation (SE) block [36] to recalibrate channel-wise features by squeezing global spatial information into a channel descriptor, which could merge with other state-of-the-art CNN architectures to produce more expressive features. In SE, the author utilized global average pooling after convolutional layer to generate channel-wise statistic, without considering other pooling methods. Park et al. proposed Bottleneck Attention Module (BAM) [37], in which spatial attention and temporal attention are added to form an attention map. The same to SE, they did not use global max pooling. In [38], Woo et al. proposed Convolutional Block Attention Module (CBAM) that contained channel attention module and sparse attention module. Compared with SE, in addition to increasing the spatial attention module, they argued that global max pooling was complementary to global average pooling. Also, they discussed the arrangement of spatial and channel attention and found that the channel-first order performs better. We adopt CBAM in our MFR network and it shows a superior performance over other existing state-of-the-art methods.

In more specific FR task, there is also some excellent attention-based works. Shao et al. proposed an adaptive attention learning module to extract refined feature map [39] for face alignment and facial action unit detection. To recognize video face recognition, Rao et al. proposed an Attention-aware Deep Reinforcement Learning (ADRL) method to find key area of face video frames [40]. Zhang et al. [41] combined spatial attention mechanism and GAN framework to edit face attribute. Wang et al. [42] proposed Region Attention Networks for facial expression recognition with occlusion and variant pose, results showed the performance of facial expression recognition with attention module was improved significantly. Face Attention Network (FAN) [8], was propsed to improve face detection performance in the occluded cases, an anchor-level attention was introduced to highlight the face region features. In this paper, we introduce attention module to emphasize the features around eyes and despise the features of damaged area.

2.3 MFR for COVID-19

In this section, we highlight the works relevant to MFR for COVID-19.

Masked face recognition The published literature rarely studies face recognition with masks, and indeed some AI companies have developed related products. For example, the MFR schemes of Sensetime technology and Hanvon technology achieved an accuracy rate of 85%, and the MFR accuracy of Minivision technology in a community scene reached 90% [43]. They do not open their datasets and experimental details, and some methods are suitable for a small range of actual scenarios, but unfortunately the accuracy is not very high.

Face mask recognition Face mask recognition refers to identify whether a person is wearing a mask. Different from MFR, face mask recognition is a two-class recognition problem, while MFR is a multi-class recognition problem. Therefore, face mask recognition is much more easier than MFR. Tencent youtu achieved an accuracy of over 99% for mask recognition [44], Baidu also disclosed their mask recognition solution ("https://ai.baidu.com/tech/face/mouth-mask?track=cp:ainsem|pf:pc|pp:chanpin-renlianshibie|pu:renlianshibie-kouzhaojianceshibie|ci:|kw:10011448").

Masked face Dataset Ge et.al introduced a masked face dataset named MAFA that contains 35806 masked faces and 30811 internet images, and proposed LLE-CNNs for masked face detection [7]. Wang et al. proposed three datasets related to masked face detection and recognition, namely Masked Face Detection Dataset (MFDD), Real-world Masked Face Recognition Dataset (RMFRD) and Simulated Masked Face Recognition Dataset (SMFRD) [13]. The RMFRD dataset contains many images of celebrities wearing masks and corresponding images without masks, but there are too few images of wearing masks. The images in the SMFRD are synthetic images from Webface and LWF dataset. There are about 500,000 images from 10,000 subjects.

3 Proposed methods

In this section, we firstly make a investigation on CBAM. Secondly, cropping method on masked face for finding out the optimal cropping is explored. Then we describes how

to integrate cropping-based approach with CBAM module. Lastly, we introduce network framework in detail.

3.1 Preliminary Investigation in CBAM

Convolutional Block Attention Module (CBAM) [38] contains Channel Attention Module and Spatial Attention Module (see Fig. 2), which can be widely used in varies CNNs to get refined features. CBAM aims to focus on more influential parts of feature maps from channel and spatial dimensions. Given an input feature map $\mathbf{G} \in \mathbb{R}^{C\times H\times W}$, let $\mathbf{M_c} \in \mathbb{R}^{C\times 1\times 1}$ denotes channel attention map and $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$ denotes spatial attention map. Then the feature refining process can be indicated as follows:

$$\begin{aligned} \mathbf{G}' &= \mathbf{M_c} \odot \mathbf{G}, \\ \mathbf{G}'' &= \mathbf{M_s} \odot \mathbf{G}' \end{aligned} \tag{1}$$

Where $\odot$ means element-wise multiplication.

Channel attention module. The channel attention module focus on the inter-channel relationship of features. In CBAM, both global average-pooling and max-pooling operations are used to aggregate spatial information of a feature map, generating two spatial information descriptors, namely $\mathbf{G^c_{avg}} \in \mathbb{R}^{c\times 1\times 1}$ and $\mathbf{G^c_{max}} \in \mathbb{R}^{c\times 1\times 1}$. Then the two descriptors are fed into a shared multi-layer perceptron (MLP) with one hidden layer to generate the final channel attention map: $\mathbf{M_c} \in \mathbb{R}^{c\times 1\times 1}$.

Overall, the compute process of channel attention map is as follows:

$$\begin{aligned} \mathbf{M_c} &= \sigma(MLP(AvgPool(\mathbf{G})) + MLP(MaxPool(\mathbf{G}))) \\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{G^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{G^c_{max}}))); \end{aligned} \tag{2}$$

Where σ represents sigmoid function, $\mathbf{W_0} \in \mathbb{R}^{c/r\times c}$, $\mathbf{W_1} \in \mathbb{R}^{c\times c/r}$.

Spatial attention module The spatial attention module focus on the inter-sparse relationship of features, which aims to explain 'where' is more influential among the whole feature map. The same to channel attention module as mentioned before, global average-pooling and max-pooling operations are all adopted to generate information descriptors. However, the two operations are along with the channel axis, generating two 2D maps: $\mathbf{G^s_{avg}} \in \mathbb{R}^{1\times H\times W}$ and $\mathbf{G^s_{max}} \in \mathbb{R}^{1\times H\times W}$. Then the two maps are concatenated together and a convolution layer is adopted to generate the final spatial attention map: $\mathbf{M_s} \in \mathbb{R}^{1\times H\times W}$. The compute process of spatial attention map is as follows:

$$\begin{aligned} \mathbf{M_s} &= \sigma(f \otimes ([AvgPool(\mathbf{G}); MaxPool(\mathbf{G})])) \\ &= \sigma(f \otimes ([\mathbf{G^s_{avg}}; \mathbf{G^s_{max}}])); \end{aligned} \tag{3}$$

Where σ represents sigmoid function, $\otimes$ denotes convolution process, f is a convolution kernel.

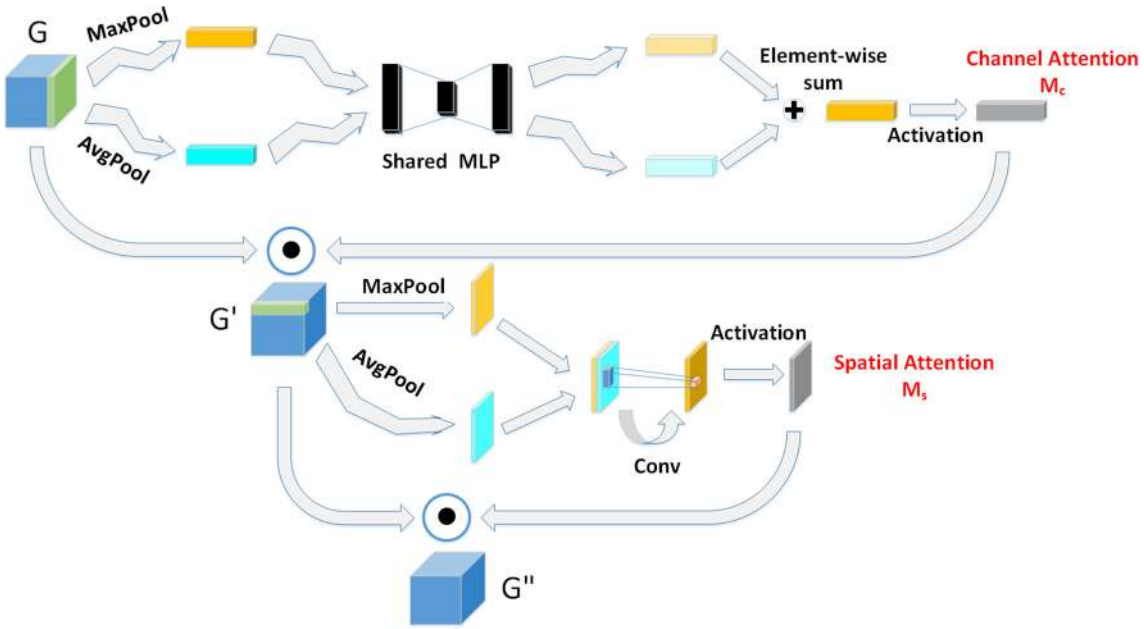

Fig. 2 Diagram of Convolutional Block Attention Module

3.2 Cropping method in masked face

MFR is seriously affected by mask occlusion, and the effective facial features are greatly reduced. Different from the routine occlusion on face, the relative position of facemask is fixed. So there are some masked face-specific solutions. No matter which method is used, the mask has permanently damaged the corresponding area of face image. Inspired by this fact, we consider completely removing the damaged area or giving a lower weight to the masked area. This is the core idea of our solutions for MFR. It is worth emphasizing that the cropping-based method can support the two aforementioned special application scenarios, which cannot be supported by other deep learning methods. However, where is the optimal cropping?

Given a masked face image, the key points of face are generated by using MTCNN [14]. Assuming the coordinate of the key point of the left eye is $E_l(x_1, y_1)$, the right eye is $E_r(x_2, y_2)$. Then the midpoint position of the two eyes can be denoted as $E_m((x_1 + x_2)/2, (y_1 + y_2)/2)$. We define the Euclidean distance between the key points of eyes as the reference distance $L = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$. The x-axis coordinate of the cropping point is $(x_1 + x_2)/2$, the y-axis is $(y_1 + y_2)/2 + l$, indicated by $C((x_1 + x_2)/2, (y_1 + y_2)/2 + l)$, where $l \subseteq [0.4L, 1.2L]$. Finally, crop the images through the cropping points parallel to the x-axis.

The image examples at different cropping proportions are shown in Fig. 3.

3.3 The integration of attention-based and cropping-based approaches

The attention-based approach is adopted to focus on the regions around eyes, which assigns a higher weight to discriminative area and a lower weight to masked features. The cropping-based approach can not only simplify the process of occlusion detection and discarding, but also support the two aforementioned special application scenarios in case 2 and case 3. To maximize the advantages of both approaches, we integrate cropping-based approach with attention-based approach. Firstly, we crop face images at different cropping proportions to find out the optimal cropping, and the face images after optimal cropping are used for model learning. Then the CBAM module is embedded in every convolution block of resnet50 to refine feature maps. In other words, this approach can be viewed as a joint effect of two kinds of attention mechanisms, one is cropping-based approach that focus on the remaining parts after cropping, the other one is attention-based approach that focus on the regions around eyes.

3.4 Network architecture

The refined ResNet50 network [45] is employed as our baseline module, and the attention ResNet framework for MFR is presented in Fig. 4. CBAM attention module is embedded in every convolution block of ResNet50 to refine feature maps. The model of aembedding CBAM in the

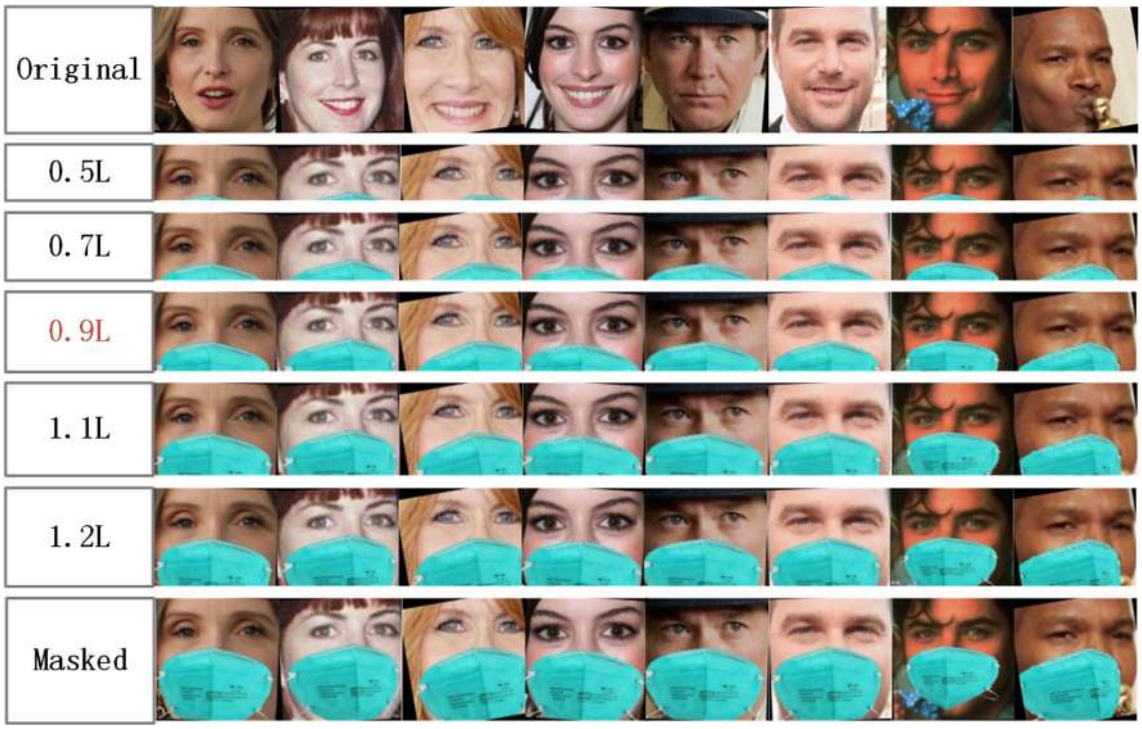

Fig. 3 Examples of cropped images at different cropping proportions

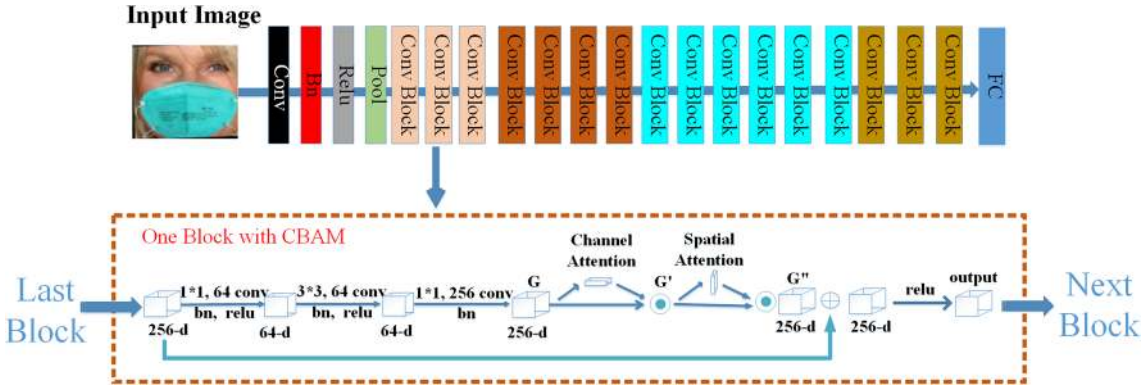

Fig. 4 Framework of proposed attention ResNet in this paper. The CBAM module is adopted in every convolution block in ResNet50

second convolution block of ResNet50 is described in detail. G represents the feature maps after convolution operations, which are 256-d feature maps in the second convolution block. Then the attention from channel-wise and spatial-wise is multiplied to G, generating refined feature maps G'', which are also 256-d feature maps that are same to G. Lastly, the output of this block is obtained by adding input feature maps and G'', which is also the input of next convolution block.

CBAM is a lightweight general-purpose module, the amount of computation added by its embedding is almost negligible. It is suitable for varies of CNN networks and can be seamlessly integrated into ResNet50 architecture, and end-to-end training can be carried out together with ResNet itself. The attention ResNet not only inherits the advantages of ResNet, which avoids the problem of gradient disappearance in deep neural network, but also the attention mechanism brought by CBAM can refine feature maps and enhance the performance of ResNet.

4 Experiment and discussion

In this section, we firstly introduce the experimental settings and datasets. Then the main content of the experiment in this paper can be divided into three steps: Attention-based approach for MFR, Cropping-based approach for MFR and the integration of the two above approaches. After that, we visualize the network with CAM to explain the results.

4.1 Settings and datasets

Settings The standard MTCNN is used to detect the face of all images and the key points of two eyes. After performing similarity transformation accordingly, we obtain the aligned face images. We crop each image in different proportions according to the key points of eyes. Assuming that the pixel distance between two eyes is L, the face datasets with masks are cropped at the proportions form $0.4L$ to $1.2L$ lower than the midpoint of the two eyes.

150 classes are randomly selected from CASIA-WebFace dataset, and the corresponding 150 classes are selected from SMFRD dataset. In each category, we divide the training set and test set according to the ratio of 4:1.

The initial learning rate is 0.1. Every 30 epochs, the learning rate will become one tenth of the original. Each training requires 120 epochs. Batchsize is 36 and momentum is 0.9. Our experiment is conducted under below environment : Ubuntu 16.04, GTX 1080 Ti, PyTorch 0.4.1, Python 3.6.

Datasets Four benchmark datasets for FR are used in this paper, namely SMFRD, Webface, AR and Extend Yela B. The SMFRD and Webface are used for finding out the optimal cropping and selecting attention module. And AR and Extend Yela B datasets are used for verifying the effect of attention mechanism on FR.

SMFRD [13] SMFRD dataset, designed for masked face recognition, contains about 500,000 images from 10,000 people. The images can be divided into two categories: Masked-Webface and Masked-LFW dataset, which are generated by wearing masks for the images in Webface and LFW datasets using GAN network.

Webface [46] 500,000 images of 10,000 people are collected from the IMBb website. Similarity clustering is used to remove part of the noise on the images. This dataset is often used as training set in FR tasks.

AR [47] AR dataset is a benchmark dataset that is widely used in varies of occlusion FR tasks. The images in this dataset are occluded by scarf and sunglass, and totally about 4,000 images from 126 subjects are included.

Extend Yela B [48] This dataset is widely used in face recognition for its different poses and illumination conditions. Totally 16128 images from 28 people under 9 poses and 64 illumination conditions are contained.

4.2 Attention-based approach for MFR

In this subsection, we conduct experiment to explore the effectiveness of CBAM attention module on MFR. We use the selected 150 categories from WebFace dataset and Masked-WebFace dataset, as well as AR dataset, Extend Yela B dataset and Masked-LFW dataset. ResNet50 without attention mechanism is used as baseline model and several popular attention modules for CNNs, including SE and BAM, are adopted into ResNet50 as contrast architectures.

Performance of different attention modules on Masked-Webface Dataset In this part, we test the performance of the aforementioned attention modules on the four MFR cases. For each attention module, the model trained on masked face images is shared by case 1 and case 2, the model trained on clean face images is shared by case 3 and case 4. The FR performance of different attention modules on webface dataset is listed in Table 1.

We have three important findings in Table 1. Firstly, it is easy to notice that the recognition accuracy in case 2 and case 3 is very low. Therefore, traditional deep learning is hard to solve the two special cases: using masked faces for training to recognize faces without mask; using faces without mask for training to recognize masked faces. Then we find that the attention models have advantages over the baseline model in case 1 and case 2, but there is no significant improvements in case 3 and case 4. Considering the different training sets between case 1, case 2 and case 3, case 4, we believe that when the model is trained using clean images, the ResNet network is able to learn enough discriminative features and emphasize the key regions of faces. However, when using masked images to train the model, the model with attention mechanism is less affected by the facemask than the original ResNet. In addition, the CBAM model shows superior performance over the other attention modules and this is the reason why we adopt CBAM into our MFR network. However, we also notice that CBAM can only improve the performance of CNNs, but it cannot fundamentally solve the problems in MFR.

Table 1 The comparison of different attention modules on Masked-Webface dataset

Cases	Baseline(%)	BAM(%)	SE(%)	CBAM(%)
Case1	85.925	87.725	85.526	**88.5**
Case2	41.925	45.292	44.122	**46.828**
Case3	43.329	43.721	43.444	43.689
Case4	94.211	94.327	94.376	94.411

Bold entries are the necessary findings

Performance of different attention modules on AR Dataset and Extend Yela B Dataset AR dataset contains sunglasses and scarf occlusion, which belongs to case 1 of occlusion FR. Extend Yela B dataset is a clean face dataset, which belongs to case 4. This part is used to verify the results on Webface. The performance of different attention modules on AR dataset and Extend Yela B dataset is listed in Table 2.

In Table 2, we find that CBAM still preforms best among the other modules on AR dataset, which is consistent with the result above. However, none of the attention modules show significantly better performance than the baseline model on Extend Yela B dataset. This results are consistent with case 4 in Table 1, that is, the attention mechanism has a positive effect on occlusion FR, but has a limited effect on non-occlusion FR.

Performance of different test categories on Maksed-LFW Dataset Maksed-LFW is used for verifying the model generated on Maksed-Webface dataset, which determines whether two images belong to the same subject by calculating feature similarity. The more the categories, the less inter-class distance, and the more challenging the task. In this part, we adopt the model in case 1 as the test model. Because the model on Maksed-Webface is trained on the randomly selected 150 categories, the recognition performance on Masked-LFW is seriously influenced by the number of categories in testset.

The results of different test categories on Masked-LFW dataset are listed in Table 3. With the increase of test categories, the recognition accuracy is decreasing, and finally reaches an accuracy rate of 82.8648% in all categories. When the number of test categories is less than 700, the accuracy of MFR is higher than 90%. The results further verify that the attention-based approach is able to realize MFR with a high accuracy.

4.3 Cropping-based approach for MFR

Though the CBAM module has shown some improvements on MFR, but it has not achieved a qualitative improvement when faced with case 2 and case 3. Therefore, we proposes

Table 2 The comparison of different attention modules on AR dataset and Extend Yela B Dataset

	Baseline(%)	BAM(%)	SE(%)	CBAM(%)
AR	97.600	97.800	97.600	98.400
Yela B	99.474	99.342	99.342	99.474

Table 3 Test results of different test categories on Masked-LFW dataset

Categories	300	400	500	600	700	800	900	1000	3000
CBAM(%)	96.8153	94.958	91.8644	91.0364	90.3981	89.4422	88.1295	87.1753	82.8648

a cropping-based approach to discard the masked regions from face images.

The croping-based approach not only saves the computing resources consumed due to occlusion detection, but also fundamentally solves the problems of case 2 and case 3. But where to crop is the most important factor affecting the performance of this method. In this part, we focus on finding the optimal cropping on the Masked-Webface dataset for case 1, case 2 and case 3.

The recognition performance of the three MFR cases at different cropping proportions is tested using CBAM, and the results are shown in Table 4. The optimal cropping in case 1 appears at $0.9L$, where the recognition accuracy is 91.529%, which increases the accuracy by 3.029% compared to uncut case. In case 2 and case 3, the optimal cropping proportions is $0.7L$. Compared with uncut, the accuracy rate in case 2 is increased by 40.025%, and the accuracy rate in case 3 is increased by 38.844%. Therefore, an important conclusion is that the cropping-based approach ia able to support the special application scenarios in case 2 and case 3 and significantly improves recognition performance. In addition, we find that the accuracy rate in case 1 is higher than that in case 2, and the accuracy rate in case 2 is higher than that in case 3. We think this is because the models of case 1 and case 2 are learned from occlusion data, and the model of case 3 is learned from non-occlusion data. So the remaining occlusion part affects the different accuracy rates in the three cases.

The performance at different cropping proportions without attention is also tested. In Table 5, we can draw some conclusions similar to Table 4. The optimal cropping proportion is still $0.9L$ in case 1 and $0.7L$ in case 2 and case 3. Compared with the uncut image, the accuracy rate is increased by 1.138% in case 1, 41.196% in case 2, 38.092% in case 3. The above results verify the superior performance of the cropping-based method on the special problems in case 2 and case 3. In addition, the accuracy rates in Table 5 are generally lower than those in Table 4, which further shows that the positive effect of attention mechanism on occlusion FR is universal.

The line charts of the croping performance at different proportions with CBAM module and without attention are shown in Fig. 5. Intuitively, it can be found that the accuracy in three cases increases first and then decreases. With the increase of cropping proportion, the remaining masked part is also increasing, which brings more useless features.

Table 4 The MFR performance at different cropping proportions with CBAM

Cropping proportion(*L)	Case1 Acc(%)	Case2 Acc(%)	Case3 Acc(%)
0.4	80.066	80.122	74.929
0.5	82.152	82.796	76.962
0.55	83.371	83.906	79.12
0.6	84.293	85.507	81.25
0.65	85.263	85.976	82.01
0.7	85.461	**86.853**	**82.533**
0.75	87.292	86.211	81.439
0.8	89.163	85.843	81.359
0.85	89.31	85.162	80.101
0.9	**91.529**	84.497	76.348
0.95	91.01	82.969	75.912
1	90.95	80.01	74.606
1.1	90.212	77.212	71.421
1.2	89.492	73.012	66.437
Uncut	88.5	46.828	43.689

Bold entries are the necessary findings

Table 5 The MFR performance at different cropping proportions without attention

Cropping proportion(*L)	Case1 Acc(%)	Case2 Acc(%)	Case3 Acc(%)
0.4	79.423	77.426	74.162
0.5	81.547	79.206	75.124
0.55	82.169	79.461	77.426
0.6	83.962	81.427	80.012
0.65	83.991	82.529	80.602
0.7	84.101	**83.121**	**81.421**
0.75	85.264	82.112	80.042
0.8	85.921	82.011	79.429
0.85	86.796	81.864	78.926
0.9	**87.063**	81.519	75.42
0.95	86.977	80.974	74.194
1	86.539	80.112	73.12
1.1	86.32	76.421	69.521
1.2	86.101	73.112	64.421
Uncut	85.925	41.925	43.329

Bold entries are the necessary findings

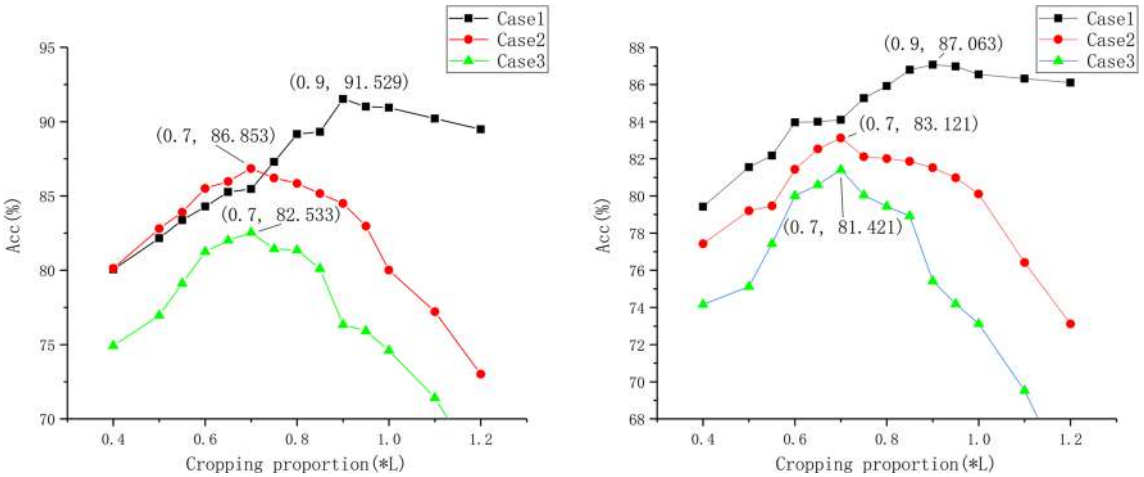

Fig. 5 The cropping performance at different proportions with CBAM module (left) and without attention(right) on Masked-Webface dataset

However, before the accuracy decreases, the increase of useful features (such as the exposure of information on both sides of the cheek, facial contour, etc.) plays a major role. This can also be intuitively explained from Fig. 3.

4.4 The integration of attention-based and cropping-based approaches

CBAM has shown its excellent ability to focus on discriminative areas, and the croping-based approach has also been proven to greatly improve the accuracy of MFR. The cropping-based approach is integrated with CBAM module to combine advantages of the two approaches. The recognition performance of different models on Masked-Webface dataset at the optimal cropping proportion is listed in Table 6. Our approach shows superior performance over the other approaches in all three MFR cases. In particular, compared with the state-of-the-art occlusion face recognition method PDSN [10], our approach can improve the recognition accuracy by at least 0.104% in case 1, 17.427% in case 2 and 18.507% in case 3. It is worth emphasizing that the PDSN method requires paired data, and it takes a lot of time and computing resources for occlusion detection and mask learning. However, for routine FR tasks, the recognition accuracy is decreased by 2.2% and 2.149% compared with Arcface [4] and Cosface [5] methods, respectively. We consider that the cropping-based method will greatly reduce the effective features of clean face images, resulting in a decrease in accuracy.

As shown is Fig. 6, our approach not only improves the performance in case 1, but also support the special problems in case 2 and case 3. In case 3 and case 4, compared with the baseline model, several attention models show a very limited positive effect because these models are trained with clean face images, which is consistent with the previous conclusion. Overall, the integration of the optimal cropping and CBAM module can realize the best recognition accuracy for MFR.

4.5 Network Visualization with CAM

In order to qualitatively analyze and interpret the experimental results, Class Activation Mapping (CAM) [49] technique is adopted to localize the discriminative ares. Since the filters in CNNs play a role of object detector, each generated feature map corresponds to a region of the original image. Based on this fact, CAM adopts global average pooling on feature maps before fully connected layer to find out the most discriminative feature map in classification task, then it highlights the corresponding discriminative areas on

Table 6 Performance comparison between our approach and state-of-the-art approaches

Cases	Baseline(%)	BAM(%)	SE(%)	CBAM(%)	Cosface(%)	Arcface(%)	PDSN(%)	Ours(%)
Case1	85.925	87.725	85.526	88. 5	87.906	88.01	91.421	**91.525**
Case2	41.925	45.292	44.122	46.828	45.372	45.494	69.426	**86.853**
Case3	43.329	43.721	43.444	43.689	43.551	44.012	62.914	**81.421**
Case4	94.211	94.327	94.376	94.411	94.761	**94.812**	94.794	92.612

Bold entries are the necessary findings

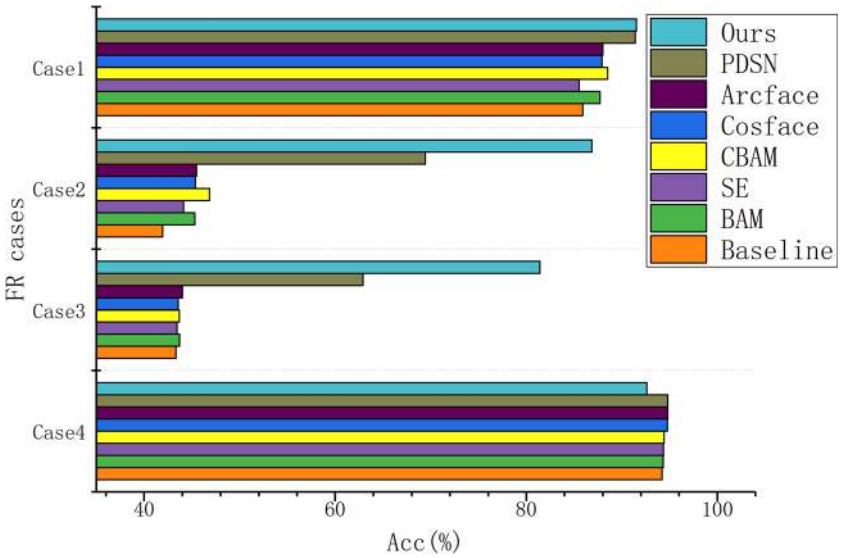

Fig. 6 Performance comparison of MFR with different approaches on Masked-Webface dataset

the original image by deconvolution. The generated class activation maps of different methods are shown in Fig. 7, the CAM of PDSN is not completed because its special mask learning process.

In Fig. 7, it is easy to find that the CAMs of different approaches vary greatly. Among them, Baseline and SE have focused a lot of attention on the mask, while BAM and CBAM have focused more on the area around eyes. Especially the CBAM method, which assigns a very low weight to the mask part. Arcface and Cosface, the state-of-the-art FR approaches, can also focus on the area around eyes. More importantly, as can be seen from the CAM maps

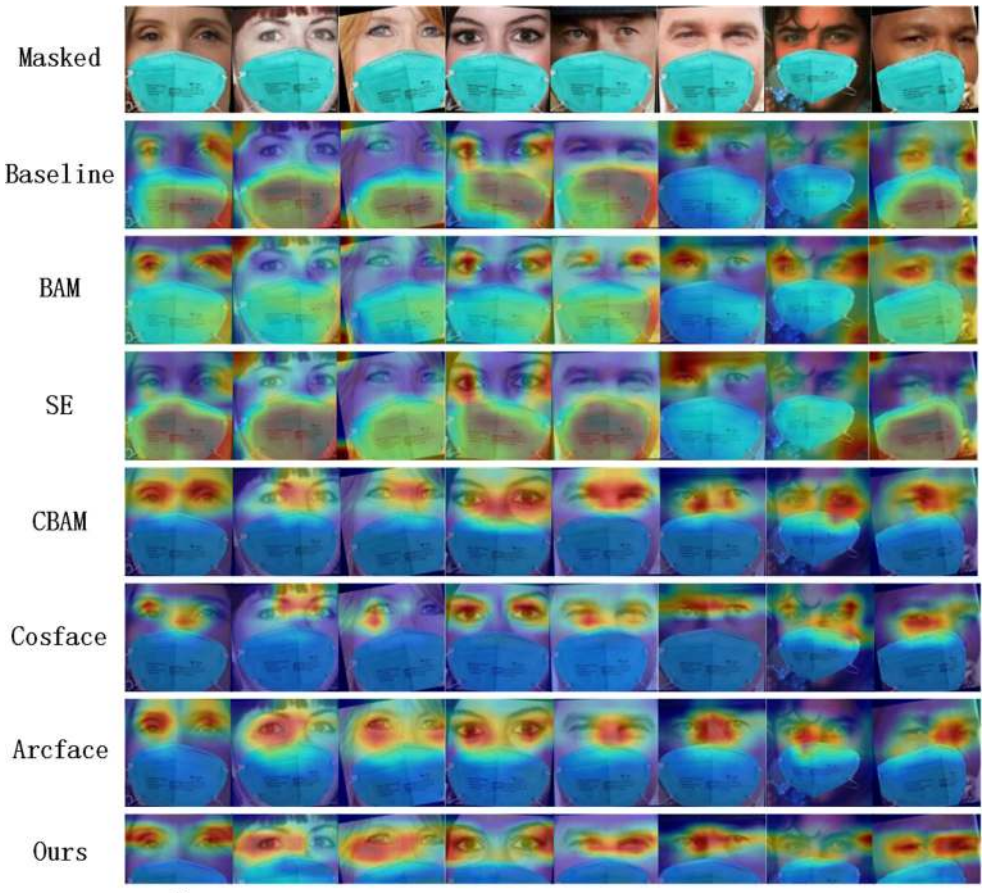

Fig. 7 Examples of class activation maps of different approaches

of our approach, the most discriminative areas are not all areas above the mask, but the regions around two eyes.

5 Conclusion

The epidemic of the COVID-19 forces people to wear masks when going out, but the existing face recognition systems cannot work when faced with masks. In this paper, we propose two approaches to solve the above difficulties, which are attention-based approach and cropping-based approach. In attention-based approach, CBAM attention module is adopted to focus on the the area around eyes, which shows superior performance on masked face recognition over the other attention modules. In cropping-based approach, we explore the optimal cropping for each case in masked face recognition. Results shows that the optimal cropping proportion is about $0.9L$ in case 1 and $0.7L$ in case 2 and case 3. This approach shows its excellent ability to support two special cases: using masked faces for training to recognize faces without mask; using faces without mask for training to recognize masked faces. Lastly, we integrate the above approaches to optimize masked face recognition performance. Results show that our approach can increase the recognition accuracy by 0.104% in case 1, 17.427% in case 2 and 18.507% in case 3.

Overall, the findings of this paper can be summarized as follows:

1. In occlusion face recognition, such as AR, Masked-Webface datasets, the attention mechanism can significantly improve the recognition performance. However, in routine face recognition, the attention mechanism shows limited performance.
2. The cropping-based approach can effectively support the two special application scenarios: using masked faces for training to recognize faces without mask; using faces without mask for training to recognize masked faces.
3. The optimal cropping proportion is around $0.9L$ in case 1 and $0.7L$ in case 2 and case 3. The integration of the optimal cropping and CBAM module achieves the best recognition accuracy for MFR.
4. Compared with the state-of-the-art approach, our approach realizes the optimal masked face recognition performance.

Acknowledgements This work was supported by grants from the National Major Science and Technology Projects of China (grant nos. 2018AAA0100703), the National Natural Science Foundation of China (grant nos. 61977012, 61977054), the Central Universities in China (grant nos. 2019CDJGFDSJ001, XDJK2019B023), the Chongqing Provincial Human Resource and Social Security Department (grant no. cx2017092), the National Key R&D Program of China (Grants No. 2017YFE0111900, 2018YFB1003205).

Compliance with Ethical Standards

Conflict of interests The authors declare that they have no conflict of interest.

References

1. Turk M (1991) Eigenfaces for recognition. J Cogn Neurosci:3
2. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification, pp 1701–1708
3. Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification, pp 1988–1996
4. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4690–4699
5. Wang H, Wang Y, Zhou Z, Ji X, Gong D, Zhou J, Li Z, Liu W (2018) Cosface: Large margin cosine loss for deep face recognition, pp 5265–5274
6. Wang M, Deng W (2018) Deep face recognition: A survey
7. 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
8. Wang J, Yuan Y, Yu G (2017) Face attention network: An effective face detector for the occluded faces. arXiv:1711.07246
9. Du L, Hu H (2019) Nuclear norm based adapted occlusion dictionary learning for face recognition with occlusion and illumination changes. Neurocomputing 340:133–144
10. Song L, Gong D, Li Z, Liu C, Liu W (2019) Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In: Proceedings of the IEEE international conference on computer vision, pp 773–782
11. Duan Q, Zhang L (2020) Look more into occlusion: Realistic face frontalization and recognition with boostgan. IEEE Transactions on Neural Networks:1–15
12. Lahasan B, Lutfi SL, San-Segundo R (2019) A survey on techniques to handle face recognition challenges: occlusion, single sample per subject and expression. Artif Intell Rev 52(2):949–979
13. 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
14. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
15. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2008) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31(2):210–227
16. Deng W, Hu J, Guo J (2012) Extended src: Undersampled face recognition via intraclass variant dictionary. IEEE Trans Pattern Anal Mach Intell 34(9):1864–1870
17. Huang J, Nie F, Huang H, Ding C (2013) Supervised and projected sparse coding for image classification. Twenty-Seventh AAAI Conference on Artificial Intelligence
18. Yuan L, Li F (2016) Face recognition with occlusion via support vector discrimination dictionary and occlusion dictionary based sparse representation classification. In: 2016 31st Youth Academic annual conference of chinese association of automation (YAC). IEEE, pp 110–115
19. Li G, Liu Z, Li H-B, Ren P (2016) Target tracking based on biological-like vision identity via improved sparse representation and particle filtering. Cogn Comput 8(5):910–923

20. Wagner A, Wright J, Ganesh A, Zhou Z, Mobahi H, Ma Y (2011) Toward a practical face recognition system: Robust alignment and illumination by sparse representation. IEEE Trans Pattern Anal Mach Intell 34(2):372–386
21. Cen F, Wang G (2019) Dictionary representation of deep features for occlusion-robust face recognition. IEEE Access 7:26595–26605
22. Yang J, Luo L, Qian J, Tai Y, Zhang F, Xu Y (2016) Nuclear norm based matrix regression with applications to face recognition with occlusion and illumination changes. IEEE Trans Pattern Anal Mach Intell 39(1):156–171
23. Chen Z, Wu X-J, Kittler J (2019) A sparse regularized nuclear norm based matrix regression for face recognition with contiguous occlusion. Pattern Recogn Lett 125:494–499
24. Min R, Hadid A, Dugelay J-L (2014) Efficient detection of occlusion prior to robust face recognition. The Scientific World Journal 2014
25. Priya GN, Banu RW (2014) Occlusion invariant face recognition using mean based weight matrix and support vector machine. Sadhana 39(2):303–315
26. Andrés AM, Padovani S, Tepper M, Jacobo-Berlles J (2014) Face recognition on partially occluded images using compressed sensing. Pattern Recogn Lett 36:235–242
27. He L, Li H, Zhang Q, Sun Z (2018) Dynamic feature learning for partial face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7054–7063
28. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. Computer Science
29. Fu J, Zheng H, Mei T (2017) Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4438–4446
30. Zheng H, Fu J, Mei T, Luo J (2017) Learning multi-attention convolutional neural network for fine-grained image recognition. In: Proceedings of the IEEE international conference on computer vision, pp 5209–5217
31. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. In: Advances in neural information processing systems, pp 577–585
32. Bahdanau D, Chorowski J, Serdyuk D, Brakel P, Bengio Y (2016) End-to-end attention-based large vocabulary speech recognition. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4945–4949
33. Parikh A, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference. In: Conference on Empirical Methods in Natural Language Processing
34. Zhou X, Wan X, Xiao J (2016) Attention-based lstm network for cross-lingual sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing
35. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual attention network for image classification
36. Hu J, Shen L, Albanie S, Sun G, Wu E (2017) Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
37. Park J, Woo S, Lee JY, Kweon IS (2018) Bam: Bottleneck attention module
38. Woo S, Park J, Lee JY, Kweon IS (2018) Cbam: Convolutional block attention module
39. Shao Z, Liu Z, Cai J, Ma L (2018) Deep adaptive attention for joint facial action unit detection and face alignment. In: Proceedings of the European conference on computer vision (ECCV), pp 705–720
40. Rao Y, Lu J, Zhou J (2017) Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE international conference on computer vision, pp 3931–3940
41. Zhang G, Kan M, Shan S, Chen X (2018) Generative adversarial network with spatial attention for face attribute editing. In: Proceedings of the European conference on computer vision (ECCV), pp 417–432
42. Wang K, Peng X, Yang J, Meng D, Qiao Y (2019) Region attention networks for pose and occlusion robust facial expression recognition. arXiv: Computer Vision and Pattern Recognition
43. Face masks are effective for epidemic prevention and control? http://ai.cps.com.cn/article/202002/937650.html
44. Tencent youtu overcomes the problem of mask recognition, the accuracy rate of mask wearing recognition exceeds 99°. https://www.jiqizhixin.com/articles/2020-02-23
45. He K, Zhang X, Ren S, Jian S (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision & pattern recognition
46. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv:1411.7923
47. Martinez AM, Benavente R Ar face database, 2000
48. Georghiades AS, Belhumeur PN, Kriegman DJ (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23(6):643–660
49. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929

Yande Li received his Bachelor degree and Master degree in school of information science and engineering, Lanzhou University, China in 2016 and 2019, respectively. He is currently a PhD student in computer application technology, Lanzhou University, China. His research focuses on occlusion face recognition, facial expression recognition and action recognition.

Kun Guo is a master student of Computer Science in Information Science and Engineering College of Lanzhou University. He received his bachelor degree in Computer Science degree in Lanzhou University, Lanzhou, China. He has research experience in applying machine learning in human motion detection and object recognition.

Yonggang Lu is now working as a professor in the School of Information Science and Engineering, Lanzhou University, Lanzhou, China. He is a member of Chinese Computer Federation, IEEE and ACM. He received both the B.S. and M.S. Degrees in Physics from Lanzhou University, Lanzhou, China in 1996 and 1999 respectively. Later he received the M.S. and Ph.D. Degrees in Computer Science from New Mexico State University, Las Cruces, NM, USA in 2004 and 2007 respectively. He finished some of the Ph.D. work at Los Alamos National Lab, NM, USA. His main research interests include artificial Intelligence, machine learning, pattern recognition, image processing and bioinformatics. He has presided over a Chinese Nation Science Foundation project and other projects and has published over 50 research papers.

Li Liu is an associate professor at Chongqing University. He is also serving as a Senior Research Fellow of School of Computing at the National University of Singapore. Li received his Ph.D. in Computer Science from the Université Paris-sud XI in 2008. He had served as an associate professor at Lanzhou University in China. His research interests are in pattern recognition, data analysis, and their applications on human behaviors. He aims to contribute in interdisciplinary research of computer science and human related disciplines. Li has published widely in conferences and journals with more than 50 peer-reviewed publications. Li has been the Principal Investigator of several funded projects from government and industry.