Masked face recognition with BF-FaceNet and multi-view features

Xueping Su (yifeichongtian1201@163.com)
Xi'an Polytechnic University
Dandan Sun
Xi'an Polytechnic University
Yunhong Li
Xi'an Polytechnic University
Lina Yao
Xi'an Polytechnic University
Matthias Ratsch
Reutlingen University

Research Article

Keywords: masked face recognition, BoTNet, face attention augmentation model, multi-view features

Posted Date: July 4th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3127760/v1

Additional Declarations: No competing interests reported.

Masked face recognition with BF-FaceNet and multi-view features

Xueping Su[1*], Dandan Sun[1†], Yunhong Li[1†], Lina Yao[1†] and Matthias R$\ddot{a}$tsch[2†]

[1*]School of Electronics and Information, Xi'an Polytechnic University, Xi'an, 710048, China.
[2]Department of Engineering, Reutlingen University, Reutlingen, 72762, Germany.

*Corresponding author(s). E-mail(s): yifeichongtian1201@163.com;
Contributing authors: sundandan2021@126.com; hitliyunhong@163.com; 273775143@qq.com; Matthias.Raetsch@Reutlingen-University.DE;
†These authors contributed equally to this work.

Abstract

Wearing masks in accordance with scientific guidelines is the most economically-effective protective measure for preventing respiratory diseases, such as COVID-19 and influenza, by interrupting viral transmission and safeguarding one's own health. In fact, researchers found that wearing masks can effectively reduce the infection rate of COVID-19 by 65%. However, this brings great inconvenience to face recognition, and human face recognition accuracy under occlusion conditions is low.Therefore, this paper proposes masked face recognition with BF-FaceNet and multi-view features. Firstly, in order to extract fine-grained features of the face region, ResNet50 is replaced by BoT-Net as the backbone network of FaceNet. Secondly, a non-masked map is generated to accurately locate the non-masked area. Meanwhile, the face attention augmentation model (FAAM) is designed to extract local face features of the non-masked map. Thirdly, by combining loss function $\boldsymbol{L_{triplet}}$ with $\boldsymbol{L_{attention}}$, joint loss function $\boldsymbol{L_{face}}$ is proposed to improve the accuracy of masked recognition. Finally, experimental results on the publicly-available masked face

dataset, SMFRD, demonstrate a significant improvement in recognition accuracy using our proposed algorithm compared to other methods.

Keywords: masked face recognition; BoTNet; face attention augmentation model; multi-view features

1 Introduction

The rapid spread of COVID-19 in 2020 led the World Health Organization (WHO) to declare it a global pandemic. The wearing of a mask is essential in effectively reducing the spread of this new epidemic. Indeed, research has shown that if all household members use masks prior to the onset of symptoms, the effectiveness of prevention is 79%[1]. Additionally, according to findings published in Proceedings of the National Academy of Sciences (PNAS), the use of masks is considerably more crucial than social distancing and home-isolation policies in preventing the spread and transmission of COVID-19[2]. The use of masks while traveling is now commonplace, leading to a loss of facial feature information, which poses a challenge to facial recognition. Consequently, scholars have dedicated themselves to researching occlusion facial recognition, which involves both traditional methods and deep-learning techniques[3].

Despite the facile implementation of traditional masked face recognition methods, manual feature annotation remains both time-consuming and laborious, highlighting the need for alternative solutions. Deep-learning approaches have the ability to generate robust facial representations, but the resulting low recognition accuracy, significant parameter size, and high computational complexity have restricted their wider application. Therefore, this paper proposes masked face recognition with BF-FaceNet and multi-view features, as follows:

1. In order to extract fine-grain features of the face region, ResNet50 is replaced by BoTNet as the backbone network of FaceNet;
2. A non-masked map is generated to accurately locate the non-masked area. Meanwhile, the face attention augmentation model (FAAM) is designed to extract local face features of the non-masked map;
3. By combining loss function Ltriplet with Lattention, joint loss function Lface is proposed to improve model convergence speed and performance.

The remaining sections of this paper proceed as follows. Section 2 elaborates on relevant research on masked face recognition and attention mechanisms. Section 3 provides a detailed description of the principle and framework of the proposed mask-obstructed face recognition algorithm using joint multi-view features. Section 4 introduces the dataset, performance metrics, experimental settings, experimental results, and analysis. Section 5 concludes the paper and provides future research directions.

2 Related Work

Face recognition refers to the identification and verification of identity based on optical facial images, while masked face recognition is the process of identifying faces wearing masks based on the eyes and forehead regions. The incomplete facial features due to obstruction make masked face recognition a challenging task in the field of face recognition. Masked face recognition can be mainly divided into traditional methods and deep-learning-based methods.

The mainstream traditional algorithms include the following: (1) algorithms based on sparse representation[4]: LASRC[5], CRC-RLS[6], RSC[7], RASR[8], etc. These models are robust to noise, but require all face images to be aligned. Otherwise, it is difficult to satisfy the sparsity requirement, making it unsuitable for generalization. (2) Algorithms based on collaborative representation[6]: CRC[9], ProCRC[10], PK-PCRC[11], CCRC[12], etc. These models are based on global representation for classification, which results in fast calculation speed, but changes in local information can lead to performance degradation. (3) Feature analysis-based algorithms[13]: LFA[14], AMM[15], LS-ICA[16], PSVM[17], etc. These algorithms are suitable for cases with a small number of samples, with small computational complexity, and the locally-emphasized algorithms are less sensitive to partial occlusions.

The mainstream deep-learning algorithms include: (1) algorithms based on convolutional neural networks[18]: CNN[19], PCANet[20], and combining CNN with LBP[21], etc. These models can effectively recognize occluded faces, but possess a large number of parameters and high computational complexity, which makes them unsuitable for mobile devices. Moreover, when CNN models extract features of occluded faces, the obstructed parts are embedded in the latent space representation[22]. (2) Algorithms based on generative adversarial networks[23]: GAN[24], WGAN[25], WGAN-GP[26], LSGAN[27], DCGAN[28], etc. These models generate clearer and more realistic samples, but suffer from poor training stability, gradient vanishing, and mode collapse issues. (3) Algorithms based on attention mechanisms[29]: AM[30], SANs[31], GATs[32], CBAM[33], etc. These types of models enhance the discriminative visual information extracted from images by establishing global dependencies and expanding the receptive field[34]. However, they require a large amount of training data and have high computational complexity. Among them, the BoTNet[35] model proposed in 2021 integrates various self-attention mechanisms and replaces spatial convolution with global self-attention, which achieves significant improvements in instance segmentation and object detection while reducing the number of parameters.

Overall, traditional masked face recognition algorithms often use shallow structures to extract facial image features, which may overlook image details and lead to performance degradation when faces are obstructed by masks. Deep-learning-based masked face recognition algorithms, on the other hand, extract features through deep structures and have broader applications, but

possess a large number of parameters, high computational complexity, and cannot simultaneously consider global and local information. Therefore, this paper proposes masked face recognition with BF-FaceNet and multi-view features.

3 Methods

The overall technical framework of this algorithm is shown in Fig.1,the main steps of which are as follows: (1) face detection and non-masked map generation: if the input is a face wearing a mask, then determine the mask area to regenerate a non-masked map; if the input is a face that is not wearing a mask, the simulated mask occlusion is regenerated into a non-masked map. (2) Feature extraction: fine-grained features of the face region are extracted by BoTNet, and local features of the non-masked map region are extracted by FAAM. (3) Masked face recognition: the joint loss function is used to accelerate the convergence speed of the model, and recognize faces that are obstructed by masks.

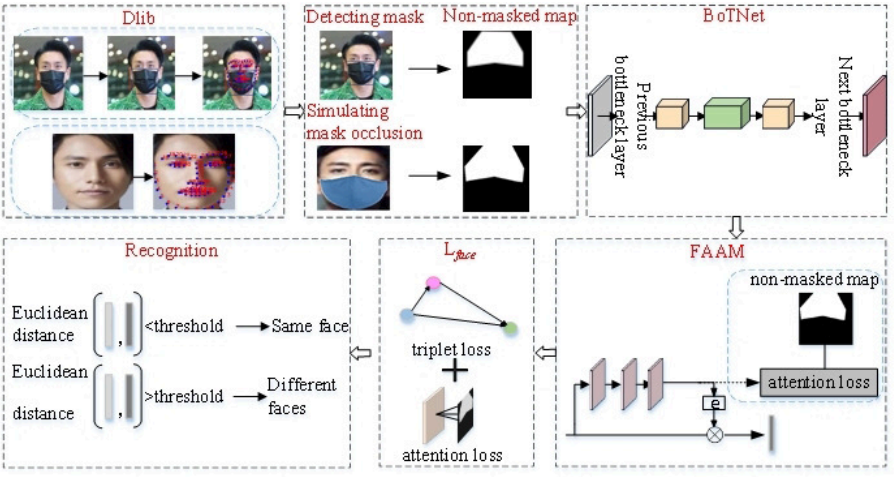

Fig. 1 Overall technical framework diagram.

3.1 Face detection and non-masked map generation

3.1.1 Face detection

Lightweight Dlib uses the method of HOG+SVM[36] to detect faces, which offers strong robustness to posture changes. It can detect positive faces, faces with small rotation angles, faces of different scales, and even faces under occlusion. Therefore, this paper uses Dlib to detect faces. If the input is a face wearing a mask, as shown in Fig.2, the mask is detected by Efficient-YOLOV3[37] with high detection accuracy, rapid speed and strong model generalization ability, and then 68 key coordinate points of the face are

detected by Dlib and normalized. The face input is not wearing a mask as shown in Fig.3, and 68 key coordinate points of the face are detected by Dlib and normalized.

Fig. 2 The detection of faces obstructed by masks.

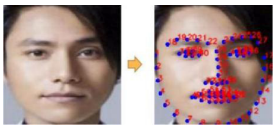

Fig. 3 The detection of faces without obstruction from masks.

3.1.2 Generate non-masked maps

In order to accurately locate the face features of the non-mask area, this paper generates the non-masked map.

The input is a face wearing a mask, and the generation steps are as follows: (1) sequentially connect the key coordinate points 1, 2, 30, 16, and 17 of the face to generate the lower boundary of the non-masked area. (2) Calculate the upper boundary of the non-mask area by using the eyebrow key points 20 and 25. (3) Connect coordinate points 1, 20, 25 and 17 in sequence, and finally a non-masked map of the non-mask area (mainly including the area around the eyes, eyebrows and forehead, as shown in Fig.4) is generated.

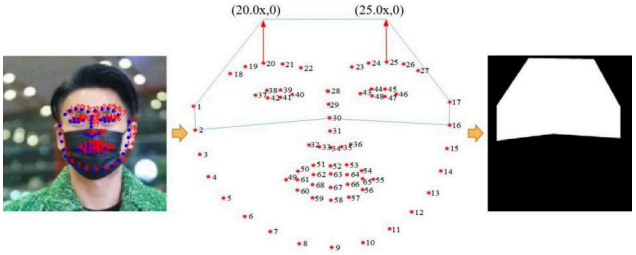

Fig. 4 Non-masked map generation of masked faces.

The face input is not wearing a mask as shown in Fig.5, and Mask The Face[38] is used to simulate mask occlusion. Mask The Face is based on Dlib to identify the face inclination and the six key features of the face needed to apply the mask, and then the template mask is transformed according to the six key features. The mask generated by this method can fit the face perfectly.

The step of generating the non-masked map of the non-mask area is consistent with the step of generating the face with the mask input.

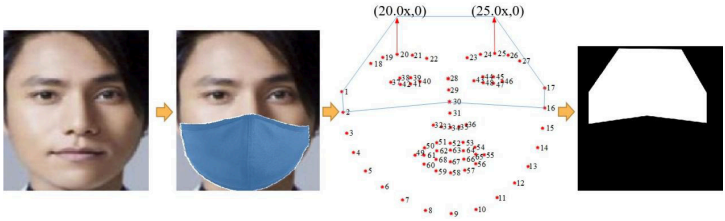

Fig. 5 Non-masked map generation of non-masked faces.

3.2 Feature extraction

3.2.1 BoTNet

FaceNet[39] uses end-to-end training, which both simplifies the setup and shows that directly optimizing a loss relevant to the task at hand improves performance. FaceNet adopts the Residual Neural Network (ResNet) as the backbone network. ResNet solves the degradation problem caused by the increase of network depth and achieves high classification accuracy, but it requires large capacity memory and long training time. In reference[35], BoTNet was designed, and multi-head self-attention (MHSA) was used to replace the 3×3 convolution of ResNet50, which solved the problems of large redundancy, high computational energy consumption and over-fitting, and effectively obtained local information, with few parameters, high speed and good effect. Therefore, this paper uses BoTNet as the backbone network, and its network architecture is shown in Fig.6.

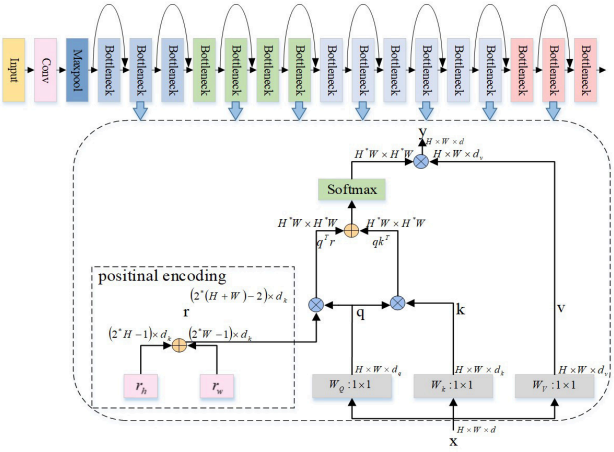

Fig. 6 BoTNet architecture diagram[35].

3.2.2 FAAM

In real-world scenarios, there is a high degree of individual variation, commonly referred to as "thousands of faces, thousands of looks." However, when people wear masks, the highly discriminative facial features such as the nose and the mouth are obscured, leading to the loss of critical information in that region. In addition, the feature discrimination of the same type of mask area is small, which decreases the accuracy of face recognition. BoTNet only extracts the fine-grained features of the face area, and does not effectively utilize the local features of the face in the non-masked area. Therefore, this paper designs FAAM to effectively extract the face's local features of the non-masked map area.

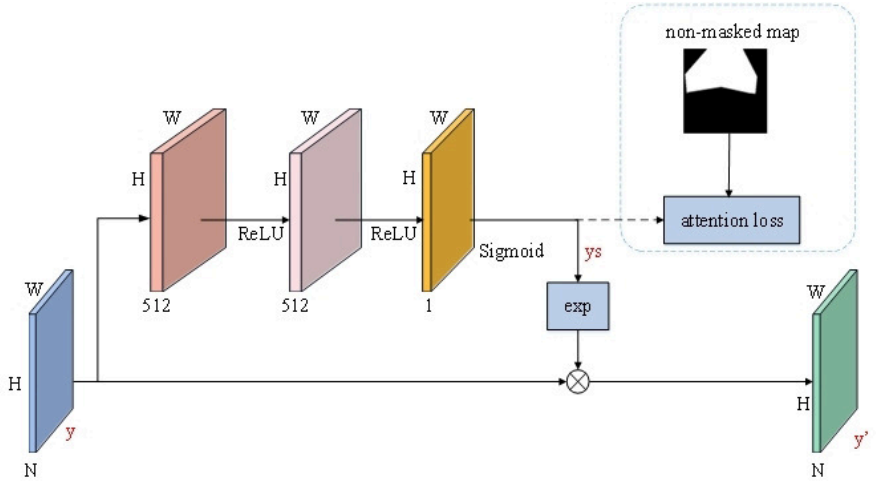

Fig. 7 FAAM structure diagram.

The FAAM structure is presented in Fig.7. The network is composed of three convolution layers, and the input feature vector W×H×N is denoted as y. The first two convolution layers are convolved with 512 convolution kernels with the size of 3×3 and activated by ReLU, and the feature map of W×H×512 is obtained. In the third convolution layer, a 3×3 convolution kernel with a step size of 1 is used for dimensionality reduction, and Sigmoid function is used for activation, so that the enhanced area of the learned attention feature map is set to 1 and the useless area is set to 0. The dimension reduction feature is denoted as ys for E-exponent operation, and the result is point-multiplied with y to obtain the enhanced feature map y'. The calculation formula is shown in (1).

$$y' = \mathbf{y}_{\otimes} exp(ys) \tag{1}$$

3.3 Masked face recognition

3.3.1 Joint Loss Function

The triplet loss function is superior at learning subtle features, and input vectors with smaller differences can learn better representations. Its model training is unstable, however, and converges slowly, which requires constant adjustment of parameters, and over-fitting occurs easily. The binary cross-entropy (BCE) loss function can measure the prediction accuracy of the model and make the model converge faster. Consequently, this paper uses the combined loss function (combining $L_{triplet}$ with $L_{attention}$) for face recognition training to assist the model to rapidly learn face features.

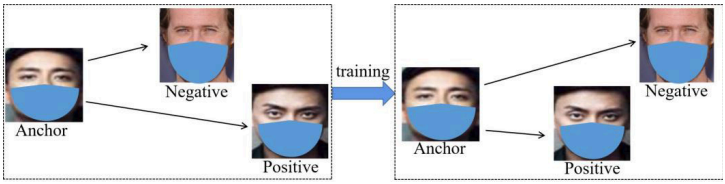

Fig. 8 An example of triplet selection.

Triplet loss selects three face sample images from the face dataset each time, as shown in Fig.8: (1) anchor, which is the reference face; (2) positive, which is a face that belongs to the same class as the reference face; and (3) negative, which is the face of a different class from the reference face. Euclidean distance is selected as the similarity of different face images, in which if the distance between anchor and positive and anchor and negative is greater than the set threshold m, it is the same type of face. However, if the distance is less than the set threshold m, it is not the same type of face. Then the $L_{triplet}$ calculation formula is:

$$L_{triplet} = \sum_{i \in N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \omega \right] \tag{2}$$

where N is the total number of triplets; and f(x_i^a), f(x_i^p) and f(x_i^n) represent the feature vector of face anchors, face positive examples and face negative samples extracted by the feature network respectively. When creating face triplet samples, three samples of the same kind of people may be put together to calculate the loss function, resulting in the calculated loss function being 0. In order to avoid this situation, this paper introduces the weight coefficient ω to determine whether it is the same person, and the value of ω is set to 0.5. As shown in Table1, if the value of ω is either too large or too small, it will affect the final recognition rate.

Similarly, BCE loss selects three face images each time, the cross-entropy loss of key pixels is calculated by using the heat map extracted by BoTNet and the non-masked map of the input FAAM, and the mean of the three losses is taken as $L_{attention}$. Its calculation formula is written as:

Table 1 Recognition accuracy under different ω

ω	0	0.25	0.5	0.75	1
Accuracy	51.6%	62.5%	**74.7**%	63.2%	62.6%
Precision	59.6%	66.5%	**72.5**%	62.5%	60.7%
Recall	52.6%	55.6%	**81.1**%	63.8%	74.3%

$$L_{attention} = \sum_{i \in N} \{ \frac{1}{3} [BCE(x_i^a) + BCE(x_i^p) + BCE(x_i^n)] \} \tag{3}$$

$$BCE(x) = -\sum_{i}^{M} \sum_{j}^{M} [m_{ij} \log(h_{ij}) + (1 - m_{ij}) \log(1 - h_{ij})] \tag{4}$$

where BCE(x) represents the cross entropy loss function; and m_{ij} is the pixel value of the non-masked map. The non-masked map is normalized to the attention feature heat map size, then the pixel point of the mask area is 0, and the face area after removing the mask area the pixel point is 1. h_{ij} is the extracted attention eigenvalue, and M is the feature map size. After calculation by formula (3) and (4), irrelevant features such as the mask area and the face background can be removed, and the face non-mask area features can be located and enhanced. In summary, the final loss function in this paper can be expressed as follows:

$$L_{face} = L_{triplet} + L_{attention} \tag{5}$$

In addition, in the training stage, if the reference face of the sample pair is extremely similar to the positive example or the reference face is markedly different from the negative example, it will make it difficult for the model to learn the distinguishing features of the face. Therefore, after screening the difficult samples, this paper trains the model again to learn more complex and discriminating face features. The selected sample pairs mainly comprise two types: sample pairs with large differences between the reference face and the positive example; and sample pairs with a similar reference face and negative sample. Its calculation is expressed as:

$$\|f(x_i^a) - f(x_i^n)\|_2^2 - \|f(x_i^a) - f(x_i^p)\|_2^2 < \omega \tag{6}$$

Where ω is a hyperparameter, with the value of 0.5 in this paper. The meanings of other variables are shown in Equation (2). The sample pairs after conditional screening can avoid the situation that the gradient of the model does not decrease, which makes the model more stable and converge quickly.

3.3.2 Recognition

FaceNet only needs to crop the face area as model input and extract features to directly calculate the distance, which is simple and effective, and is invariant

to both lighting and posture. As a consequence, this paper uses the improved FaceNet for masked face recognition, the steps of which are as follows: (1) input a face image; (2) extract face region features through the above BoTNet and extract local features of non-masked map regions through the FAAM network; (3) apply the L2 standardization feature vector; and (4) compare with the face images in the dataset and obtain the prediction result.

4 Experiment

4.1 Dataset

In this paper, the CASIA-WebFace-mask dataset[40] and the VGGFace2 dataset[41] are used as training sets. Among them, the CASIA-WebFace-mask dataset contains 445,466 face images wearing masks. The mask types involved are: surgical (white medical surgical mask); surgical blue (blue medical surgical mask); N95; KN95; and cloth (black cloth mask). As shown in Fig.9, the distribution types of masks are all generated by distribution uniform stochastic. The VGGFace2 dataset contains 3 million face images of approximately 8,000 people in different poses and lighting backgrounds, as shown in Fig.10. In this paper, Mask The Face is used to generate a dataset simulating mask occlusion based on the VGGFace2 dataset, as shown in Fig.11. When the required dataset is generated, the images with low pixels and large face offset angle in the dataset are cleaned, and the high-quality face dataset is saved and generated, thus realizing data enhancement, increasing training data, and improving model generalization ability.

Fig. 9 CASIA-WebFace-mask dataset.

Fig. 10 VGGFace2 dataset.

Fig. 11 Simulated mask occlusion dataset.

The test set adopts the SMFRD dataset[42], which contains 500,000 face images of 10,000 people. As shown in Fig.12, it uses the images in the LFW dataset of GAN to generate the face images with masks.

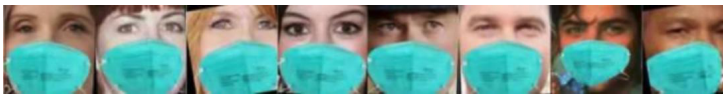

Fig. 12 SMFRD dataset.

4.2 Evaluation index

The evaluation indexes of face recognition are accuracy, precision, and recall. The calculation is written as follows:

$$Accuracy = \frac{TP + TN}{(TP + FP) + (TN + FN)} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

where TP indicates that both prediction and reality are true; TN indicates that both prediction and reality are false; FP indicates that prediction is true, but it is actually false; and FN indicates that prediction is false, but it is actually true.

4.3 Experimental settings

This paper trains the model on an NVIDIA GeForce RTX 3090 Ti. The dataset is divided into a training set, a validation set, and a test set with ratios of 0.7, 0.1, and 0.2, respectively. The algorithm is developed using the Pytorch deep-learning framework and implemented based on Python. It is found through experiments in which, when training for 145 cycles, the ROC reaches the maximum. To prevent over-fitting, this article sets the epoch to 145(as shown in Table2). The size of each input image is 256×256×3, and a total of 100,000 pairs of triple face samples are generated in each training. Each time, 30 pairs of training set faces and test set faces are selected from the samples for training. In order to make the network loss designed in this paper reach a faster convergence speed, different learning rates are selected for different periods, and the selection of learning rate lr is shown in Equation (10):

$$lr = \begin{cases} 0.1250 & 0 < epoch < 30 \\ 0.0625 & 30 \leq epoch < 60 \\ 0.0155 & 60 \leq epoch < 90 \\ 0.0030 & 90 \leq epoch < 120 \\ 0.0001 & epoch \geq 120 \end{cases} \tag{10}$$

Table 2 ROC data under different epochs

Epoch	130	135	140	145	150
ROC	0.784	0.788	0.891	**0.893**	0.892

4.4 Results and analysis

In order to verify the effectiveness of the masked face recognition with BF-FaceNet and multi-view features proposed in this paper, we chose to compare with FaceNet, CosFace[43], ArcFace[44], and PDSN[45] algorithms. Among them, FaceNet is a classic face recognition algorithm, which realizes end-to-end training, and the algorithm in this paper is an improvement on FaceNet. ArcFace is a superior performance algorithm, building on the previous SoftmaxLoss, Center Loss, A-Softmax Loss, and Cosine Margin Loss. Similarly to ArcFace, CosFace introduces a cosine margin to increase decision margin in the angular space, thus minimizing intra-class variations and maximizing inter-class variations. PDSN, unlike other occlusion face recognition algorithms, utilizes the differences between top convolutional features of occluded and unoccluded faces to build a dictionary for recognition. The experimental results of masked face recognition with different algorithms on the SMFRD dataset are shown in Fig.13 and Fig.14.

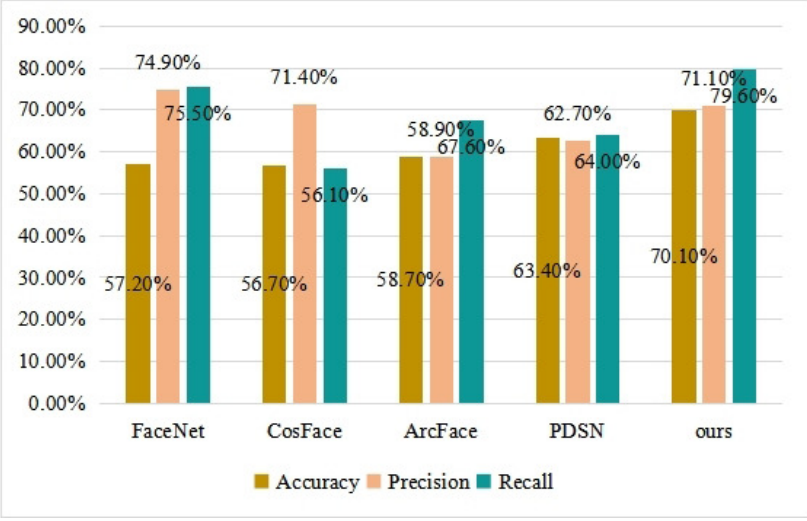

Fig. 13 Experimental results after input non-masked images training.

As shown in Fig.13 and Fig.14, the recognition rates of the algorithms proposed in this paper for mask-occluded faces are superior to other models, regardless of whether or not the input includes simulated mask occlusion. The recognition rate can reach 74.7%, and the recognition rate is further improved when images with simulated mask occlusion are included as input. In the

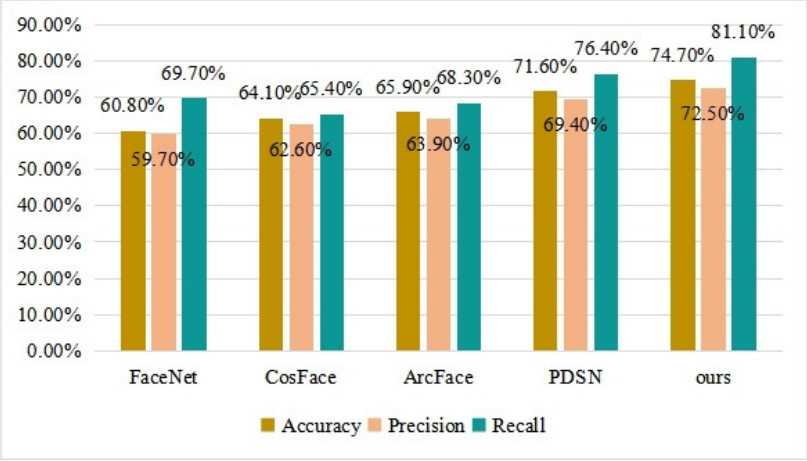

Fig. 14 Experimental results after input simulated mask occlusion images training.

conventional face recognition algorithms, the recognition rate of FaceNet is only 60.8%, the approximation of CosFace and ArcFace recognition rates are 64.1% and 65.9%, respectively, while the more advanced algorithm PDSN in masked face recognition has a recognition rate of 71.6%. Compared with the other four algorithms, the recognition rate of this paper is improved by 13.9%, 10.6%, 8.8% and 3.1% respectively. The reason for this is that the algorithm proposed in this paper highlights the features of the area around the eyes with more face information, and reduces the parameters while accurately extracting features. In summary, the algorithm in this paper can achieve the best accuracy of masked face recognition.

In order to deeply analyze the effect of the improved strategy proposed in this paper on the performance of masked face recognition, a method similar to the control variable method was used to conduct ablation experiments by controlling different network modules. The ablation experiments on the SMFRD dataset are shown in Table 3 and Table 4. The feature map visualization results of the same face are shown in Fig.15, and the FPN layer feature visualization results are presented in Fig.16.

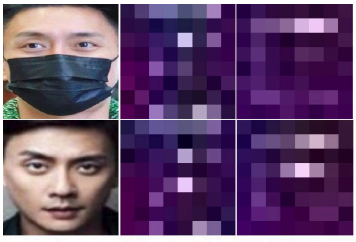

Fig. 15 Feature map visualization.

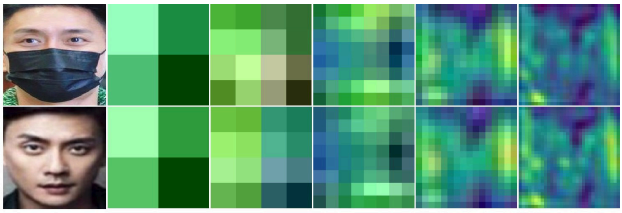

Fig. 16 FPN layer feature visualization results.

In Fig.15, (a) is the input face image, (b) is the feature visualization result extracted by the BoTNet, and (c) is the feature visualization result extracted by the FAAM. Among them, the face image wearing a mask is input by the first behavior and its feature map, and the face image without a mask is input by the second behavior and its feature map. It can be seen that the BoTNet extracts the fine-grained features of the face area, while the FAAM accurately locates the eye area and focuses the attention of the network on the features of the area around the eyes, eyebrows, and forehead as high as possible. It can be seen in Fig. 16 that the extracted features pay more attention to detailed features, among which the first behavior inputs the face image of wearing a mask and its feature map, and the second behavior inputs the face image of not wearing a mask and its feature map.

Table 3 Input the experimental results of image ablation without mask occlusion

ResNet50	BoTNet	FAAM	L_{face}	Accuracy	Precision	Recall
√				57.2%	74.9%	75.5%
	√			58.6%	63.6%	61.5%
√		√		65.4%	71.2%	69.3%
√			√	59.3%	66.6%	71.7%
	√	√	√	**70.1**%	**71.1**%	**79.6**%

Table 4 Input the experimental results of image ablation simulating mask occlusion

ResNet50	BoTNet	FAAM	L_{face}	Accuracy	Precision	Recall
√				70.3%	70.8%	75.5%
	√			71.5%	69.4%	72.4%
√		√		72.8%	71.8%	73.2%
√			√	70.9%	71.9%	69.9%
	√	√	√	**74.7**%	**72.5**%	**81.1**%

It can be obtained from Table3 and Table4 that when ResNet50 is also selected as the feature extraction network, if the simulated mask occlusion

images are not input during training, the face image identification rate is only 57.2%. However, if the simulated mask occlusion images are input during training, the face image identification rate can reach 70.3% constituting a difference of 13.1%. In the future, irrespective of which improvement strategy is added, the recognition rate after the input of the face image training network that does not simulate wearing a mask is lower than the result obtained by the input of the face image training network wearing a mask. It can be seen that the mask image that simulates wearing can assist the network to locate the key areas of the face and improve accuracy. It can also be seen from Table4 that the accuracy is only 70.3% when ResNet50 is selected as the main feature extraction network, and the accuracy is increased by 1.2% after changing the ResNet50 model to BoTNet, which verifies that the approach presented in this paper enhances the face in the non-masked area through the attention mechanism features effectively. Moreover, using FAAM alone can improve the accuracy by 2.5%, which is because the FAAM input non-masked map discards the features of the masked area, and accurately localizes and removes the face contour outside of the mask by calculating the loss value of the original input photo of the face, focusing the features of the network on the face in the non-masked area. After using the joint loss function, the accuracy is improved by 0.6%, which shows that the difficult sampling model can learn more complex and differentiated face features.

In aggregate, BoTNet assists the network to enhance the face features around the eyes and extract the fine-grained features of the face area, FAAM accurately locates the non-masked area of the face to help the network to extract the local features of the non-masked map, and L_{face} improves the model convergence speed and performance. The masked face recognition with BF-FaceNet and multi-view features proposed in this paper can achieve excellent recognition performance for faces occluded by masks.

5 Conclusion

In this paper, masked face recognition with BF-Face Net and multi-view features is proposed, which takes the BoTNet network as the main network, adds mixed multi-head attention to each residual convolution block to enhance the features of non-mask face regions, and extracts fine-grained features of face regions. Secondly, FAAM is added after the last convolution block of BoTNet to locate the non-masked face region and extract the local features of the non-masked map region. Finally, the joint loss function, L_{face}, is used to improve the convergence speed and performance of the model. Experiments confirm that the algorithm can effectively identify faces shielded by masks, which provides an effective solution for masked face recognition shielding under epidemic prevention and control. However, the categories of masks in real scenarios are markedly different, and dissimilarities also exist in the scope of facial occlusion. Since this paper only simulates the generation of one type of mask occlusion, the next step will be to simulate the generation of different types of mask areas,

and use real masks to cover faces. Testing should then be performed to find the optimal solution for masked face recognition. Furthermore, compared with unobstructed environments, face recognition accuracy in occluded environments is still relatively low, and thus new technologies need to be investigated in order to improve face recognition accuracy in occluded environments.

6 Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant no. 61902301, the Shaanxi Natural Science Basic Research Project under Grant no. 2022JM-394, the Natural Science Basic Research Key Program funded by Shaanxi Provincial Science and Technology Department n.2022JZ-35 and the Xi'an Science and Technology Bureau Science and Technology Innovation Leading Project under Grant no.21XJZZ0020, .

References

[1] Howard, J., Huang, A., Li, et al: An evidence review of face masks against covid-19. Proceedings of the National Academy of Sciences **118**(4), 2014564118 (2021)

[2] Boutros, Damer, N., Kirchbuchner, F., Kuijper, A.: Self-restrained triplet loss for accurate masked face recognition. Pattern Recognition **124**, 108473 (2022)

[3] Lal, M., Kumar, K., Arain, R.H., Maitlo, A., Ruk, S.A., Shaikh, H.: Study of face recognition techniques: a survey. International Journal of Advanced Computer Science and Applications **9**(6) (2018)

[4] Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE transactions on pattern analysis and machine intelligence **31**(2), 210–227 (2008)

[5] Ortiz, E.G., Becker, B.C.: Face recognition for web-scale datasets. Computer Vision and Image Understanding **118**, 153–170 (2014)

[6] Zhang, L., Yang, M., Feng, X.: Sparse representation or collaborative representation: Which helps face recognition? In: 2011 International Conference on Computer Vision, pp. 471–478 (2011). IEEE

[7] Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: CVPR 2011, pp. 625–632 (2011). IEEE

[8] Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: Robust alignment and illumination by sparse representation. IEEE transactions on pattern analysis and machine intelligence **34**(2), 372–386 (2011)

[9] Peng, Y., Li, L., Liu, S., Lei, T., Wu, J.: A new virtual samples-based crc method for face recognition. Neural Processing Letters **48**, 313–327 (2018)

[10] Cai, S., Zhang, L., Zuo, W., Feng, X.: A probabilistic collaborative representation based approach for pattern classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2950–2959 (2016)

[11] Lan, R., Zhou, Y., Liu, Z., Luo, X.: Prior knowledge-based probabilistic collaborative representation for visual recognition. IEEE transactions on cybernetics **50**(4), 1498–1508 (2018)

[12] Yuan, H., Li, X., Xu, F., Wang, Y., Lai, L.L., Tang, Y.Y.: A collaborative-competitive representation based classifier model. Neurocomputing **275**, 627–635 (2018)

[13] Min, R., Hadid, A., Dugelay, J.-L.: Efficient detection of occlusion prior to robust face recognition. The Scientific World Journal **2014** (2014)

[14] Penev, P.S., Atick, J.J.: Local feature analysis: A general statistical theory for object representation. Network: computation in neural systems **7**(3), 477–500 (1996)

[15] Martinez, A.M.: Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. IEEE Transactions on Pattern analysis and machine intelligence **24**(6), 748–763 (2002)

[16] Kim, J., Choi, J., Yi, J., Turk, M.: Effective representation using ica for face recognition robust to local distortion and partial occlusion. IEEE transactions on pattern analysis and machine intelligence **27**(12), 1977–1981 (2005)

[17] Jia, H., Martinez, A.M.: Support vector machines in face recognition with occlusions. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 136–141 (2009). IEEE

[18] Lawrence, S., Giles, C.L., Tsoi, A.C., Back, A.D.: Face recognition: A convolutional neural-network approach. IEEE transactions on neural networks **8**(1), 98–113 (1997)

[19] Siradjuddin, I.A., Muntasa, A., *et al.*: Faster region-based convolutional neural network for mask face detection. In: 2021 5th International Conference on Informatics and Computational Sciences (ICICoS), pp. 282–286 (2021). IEEE

[20] Yu, H., Zhao, J., Zhu, Y.: Research on face recognition method based

on deep learning. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (2019)

[21] Vu, H.N., Nguyen, M.H., Pham, C.: Masked face recognition with convolutional neural networks and local binary patterns. Applied Intelligence **52**(5), 5497–5512 (2022)

[22] Zeng, D., Veldhuis, R., Spreeuwers, L., Arendsen, R.: Occlusion-invariant face recognition using simultaneous segmentation. IET biometrics **10**(6), 679–691 (2021)

[23] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)

[24] Zhao, J., Xiong, L., Karlekar Jayashree, P., Li, J., Zhao, F., Wang, Z., Sugiri Pranata, P., Shengmei Shen, P., Yan, S., Feng, J.: Dual-agent gans for photorealistic and identity preserving profile face synthesis. Advances in neural information processing systems **30** (2017)

[25] Adler, J., Lunz, S.: Banach wasserstein gan. Advances in neural information processing systems **31** (2018)

[26] Hu, M., He, M., Su, W., Chehri, A.: A textcnn and wgan-gp based deep learning frame for unpaired text style transfer in multimedia services. Multimedia Systems **27**, 723–732 (2021)

[27] Wang, C., Cao, Y., Zhang, S., Ling, T.: A reconstruction method for missing data in power system measurement based on lsgan. Frontiers in Energy Research **9**, 651807 (2021)

[28] Wu, Q., Chen, Y., Meng, J.: Dcgan-based data augmentation for tomato leaf disease identification. IEEE Access **8**, 98716–98728 (2020)

[29] Qian, H., Zhang, P., Ji, S., Cao, S., Xu, Y.: Improving representation consistency with pairwise loss for masked face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1462–1467 (2021)

[30] Chaudhari, S., Mithal, V., Polatkan, G., Ramanath, R.: An attentive survey of attention models. ACM Transactions on Intelligent Systems and Technology (TIST) **12**(5), 1–32 (2021)

[31] Biesialska, M., Biesialska, K., Rybinski, H.: Leveraging contextual embeddings and self-attention neural networks with bi-attention for sentiment analysis. Journal of Intelligent Information Systems **57**(3), 601–626

(2021)

[32] Xie, Y., Zhang, Y., Gong, M., Tang, Z., Han, C.: Mgat: Multi-view graph attention networks. Neural Networks **132**, 180–189 (2020)

[33] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)

[34] Wang, W., Hu, H.: Image captioning using region-based attention joint with time-varying attention. Neural Processing Letters **50**, 1005–1017 (2019)

[35] Srinivas, A., Lin, T.-Y., Parmar, N., Shlens, J., Abbeel, P., Vaswani, A.: Bottleneck transformers for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16519–16529 (2021)

[36] Jacob, M.P.: Comparison of popular face detection and recognition techniques. International Research Journal of Modernization in Engineering Technology and Science, e-ISSN, 2582–5208 (2021)

[37] Su, X., Gao, M., Ren, J., Li, Y., Dong, M., Liu, X.: Face mask detection and classification via deep transfer learning. Multimedia Tools and Applications, 1–20 (2022)

[38] Anwar, A., Raychowdhury, A.: Masked face recognition for secure authentication. arXiv preprint arXiv:2008.11104 (2020)

[39] Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)

[40] Wang, W., Zhao, Z., Zhang, H., Wang, Z., Su, F.: Maskout: a data augmentation method for masked face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1450–1455 (2021)

[41] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74 (2018). IEEE

[42] Wang, Z., Huang, B., Wang, G., Yi, P., Jiang, K.: Masked face recognition dataset and application. IEEE Transactions on Biometrics, Behavior, and Identity Science (2023)

[43] Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274 (2018)

[44] Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)

[45] Song, L., Gong, D., Li, Z., Liu, C., Liu, W.: Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 773–782 (2019)