Multimedia Tools and Applications (2022) 81:33867–33896
https://doi.org/10.1007/s11042-022-12851-x

A new occluded face recognition framework with combination of both Deocclusion and feature filtering methods

Wang Jiang[1] · Lin Ye[1] · Zhang Yi[1] · Cheng Peng[2]

Received: 25 January 2021 / Revised: 28 January 2022 / Accepted: 9 March 2022 /
Published online: 21 April 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Face recognition plays the significant role in many human-computer interaction decvices and applications, whose access control systems are based on the verification of face biometrical features. Though great improvement in the recognition performances have been achieved, when under some specific conditions like faces with occlusions, the performance would suffer a severe drop. Occlusion is one of the most significant reasons for the performance degrade of the existing general face recognition systems. The biggest problem in occluded face recognition (OFR) lies in the lack of the occluded face data. To mitigate this problem, this paper has proposed one new OFR network DOMG-OFR (Dynamic Occlusion Mask Generator based Occluded Face Recognition), which keeps trying to generate the most informative occluded face training samples on feature level dynamically, in this way, the recognition model would always be fed with the most valuable training samples so as to save the labor in preparing the synthetic data while simultaneously improving the training efficiency. Besides, this paper also proposes one new module called Decision Module (DM) in an attempt to combine both the merits of the two mainstream methodologies in OFR which are face image reconstruction based methodologies and the face feature filtering based methodologies. Furthermore, to enable the existing face deocclusion methods that mostly target at near frontal faces to work well on faces under large poses, one

✉ Cheng Peng
chengpeng_scu@163.com

Wang Jiang
jiang.wang@stu.scu.edu.cn

Lin Ye
linlanye@sina.cn

Zhang Yi
raven.zhang@foxmail.com

[1] National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu, China

[2] College of Aeronautics and Astronautics, Sichuan University, Chengdu, China

head pose aware deocclusion pipeline based on the Condition Generative Adversarial Network (CGAN) is proposed. In the experimental parts, we have also investigated the effects of the occlusions upon face recognition performance, and the validity and the efficiency of our proposed Decision based OFR pipeline has been fully proved. Through comparing both the verification and the recognition performance upon both the real occluded face datasets and the synthetic occluded face datasets with other existing works, our proposed OFR architecture has demonstrated obvious advantages over other works.

Keywords Occluded face recognition · Face deocclusion · Generative adversarial network · Head pose estimation

1 Introduction

The task of face recognition has drawn much attention which is mainly attributed to the rapid development of deep learning techniques and the increasingly large scale public face data sets. The application of face recognition plays the significant role in human-computer interaction like the access control systems which are based on the face biometrical feature and some entertainment equipments like VR interaction devices. The existing face recognition methods which are mostly aimed at the recognition under general conditions have gained very good performance, and some works have even exceeded the human recognition ability upon some public face recognition datasets [33].

However, when the faces are under unnormalized conditions like large poses or being occluded, the recognition ability of those general recognition model would drop sharply [22, 23], this is mainly because that too large difference between the intra classes occur when the faces are occluded or under large poses. When faces are occluded by a large extent, the common face detection methods would fail, so there are also some works that are specifically aimed at improving the face detection performance under occlusions [14, 41]. With the emergence of the COVID-19 epidemic, more and more people would choose to wear masks when in public places, which has brought more challenges for those traditional face recognition systems that are not specifically designed for the occluded and masked faces, so the demand for robust occluded face recognition algorithm and system is growing more. The researches related to the fields of occluded face detection [56], dental mask detection [29] and occluded face recognition would all help promote epidemic prevention and control. Therefore, there are still large potentials and values to explore in the face recognition related tasks especially under occlusions like dental masks, sun glasses,scarfs, etc. The bottle neck in solving the problem mainly lied in two sides, one is the insufficient face datasets regarding the occlusion, and the other is the generalization ability of those existing face recognition models. There are some works that have purposely aimed to improve the robustness of their proposed methods to large poses or occlusions, and those methods that try to ease the side effect incurred by the poses mainly depend on the newly devised features that are robust and invariant to different face poses.

The methods aimed at solving the occluded face recognition problem mainly employ either deocclusion operation first before the input image being fed into the down steam recognition pipeline or try to filter the contaminated feature caused by the existence of occlusion from ID feature. These two technologies approach the occluded face recognition problem from two opposite perspectives: missing information generation and noise information elimination. Both the two different kinds of methodologies possess their own

advantages and disadvantages. Although the methods based on face deocclusion have achieved obvious improvement, the performance improvements are largely attributed to those successfully deoccluded faces that are much easier to recognize, however, the methods often fail when the input occluded faces are under large poses or occluded by the rare or strange occlusions which do not exist in the training datasets, in other words, those deocclusion based methods improve the recognition performance only by those successfully deoccluded faces, but those faces that have failed to be deoccluded would definitely keep aggravating the decline of recognition performance, it is inevitable that the methods based on the deocclusion operation have to sacrifice those failed deoccluded faces, besides, nonetheless the deoccluded faces appear to be photorealistic, but the identity feature of the deoccluded part still can not preserve well, and what is worse is, those deoccluded face part may even hinder the recognition performance, since those downstream recognition pipeline would take the whole deoccluded face as input without considering the newly deoccluded part which may contain much biased identiy information which would undoubtedly bring side effects to the final performance, as a result of this, we argue that since those failed deoccluded faces don't benefit to the recognition performance, it is reasonable to believe that the deocclusion based methodologies which are guided by the generating principle are not appropriate for those failed deoccluded faces, and those failed deoccluded faces may be more suitable for the methodologies based on the feature filtering methods which are guided by the eliminating principles to exclude the noise feature.

In order to better illustrate the rationality of the idea, we have conducted the following experiment to validate our assumptions. We sampled about 1000 occluded faces from OCC-CASIA which would be introduced in detail in the following section regarding the datasets, note that the ids of those samples are all not used in the training phase. We feed those occluded faces to the pretrained deocclusion based recognition pipeline [53] and the feature filtering pipeline [31], the deoccluded faces are fed into the general pretrained off the shelf face recognition model to extract the face identity features, the ids that are recognized correctly by different methods are distributed as shown in Fig. 2, and some of the samples recognized by only one of the methods are listed in Fig. 1.

According to Figs. 1 and 2, we could draw the conclusion that there are indeed some samples that are only recognized by one of the two methods, therefore, based on this interesting observation and inspired by the work [1] which has devised one selector to assign the best model from several candidate models to specific task, we propose to make decision first about whether the input occluded face should be deoccluded, and then the deoccluded face or original occluded face would go through the recognition pipeline which is trained on the feature filtering methods and robust to occlusions. Our proposed decision based occluded face recognition framework has combined the advantages of both mainstream occluded face recognition methods mentioned above. Besides, we also propose one new dynamic occlusion mask generator based occluded face recognition method (DOMG-OFR), which would be more robust to the occlusions especially when input faces present the occlusions that don't exist in the training data or even strange and rare. Additionally, we have also given the detailed analysis about the effects of the occlusions upon the recognition performance from the quantitative view. Furthermore, the head pose aware face deocclusion method based on the condition generative adversarial network is employed to further ensure the consistence of head poses before and after deocclusion, especially when the input occluded faces are under large poses.

Fig. 1 The face examples that are only recognized by one method, the first row shows some examples that are only recognized correctly by feature filtering method which is guided by some kind of the eliminating principles, the second row shows some examples that are only recognized correctly by deocclusion based method which is guided by some kind of the generating principles

Our proposed method mainly consists of three components, the head pose aware face deocclusion module, the dynamic occlusion mask generator based occluded face recognition module (DOMG-OFR), and the DM (Decision Module). DOMG helps offer the most informative and most hard samples for the recognition model to learn, in this way, both the labor and time of synthesizing large scale occluded face data set could be saved, and the training efficiency would also obtain some improvement compared to those methods heavily depending on the large amounts of training data among which much redundant data may

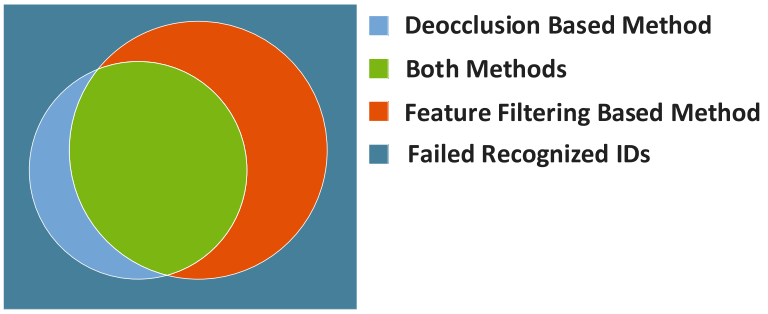

Fig. 2 The ratio of different IDs correctly recognized by feature filtering method and the deocclusion method. The bright blue color represents the Ids successfully recognized by the deocclusion based method, the green color represents the Ids successfully recognized by the both methods, the orange color represents the Ids successfully recognized by feature filtering based method, the dark blue represents those remaining Ids that are wrongly recognized by both methods

contribute less to the ability increase of final recognition model. Our method has inherited and further improved both the merits of the two main stream occluded face recogntion methods. The main contributions of this paper are summarized as follows:

- We have combined both the advantages of deocclusion based methods and feature filtering methods which approach the occlusion FR problem in the opposite way, and to the best of our knowledge, this is the first work that has made efforts in combining the two different methodologies simultaneously to mitigate the performance degrade trend upon those failed deoccluded faces, while simultaneously answering the question by giving solutions to "Do we need to deocclude all the input occluded faces before recognition?".
- Aiming at mitigating the data insufficiency of occluded face datasets, we have proposed DOMG(Dynamic Occlusion Mask Generator) to reduce the labor in synthething the datasets and save the training cost, meanwhile, the proposed DOMG also helps improve the performance of the general feature filtering based occluded face recognition pipeline.
- We quantitatively analyze the influence of occlusion location and size upon the face recognition performance, additionally, based on this analysis results, we devise the reasonable input to feed to the DM, so as to enable the DM to make better and smarter decisions.
- We have also promoted the performance of already existing works, specifically, we take the head poses into account during the deocclusion process to further improve the performance of existing deocclusion methods especially when the input faces are under large poses.
- We have conducted the comparative experiments to evaluate both the face verification performance and the face recognition performance of our method, and the results have demonstrated that our proposed method has obtained the competitive results under both the two experimental settings.

The remainder of this paper is organized as follows. We review the most related articles to the occluded face recognition in Section 2. Section 3 describes the technical details of our proposed method. Section 4 exhibits the experimental results as well as the related results analysis. Section 5 would conclude the whole work and discuss about the future work that is worth doing so as to help further improve our work.

2 Related work

Recent methods on occluded face recognition mainly vary from feature filtering based methods to deocclusion based methods. This section would first briefly review the general face recognition methods which are then followed by the review of some recent OFR works, we don't aim to cover all the works related to OFR, in other words, some works on occluded face recognition and the face deocclusion would not be mentioned here, please refer to [54] for the thorough survey about the OFR methods.

2.1 General face recognition methods

The general face recognition have been researched for a long history, the traditional features are mostly based on the hand crafted features, the hand crafted feature based methods enjoy the advantages of being easy to implement and apply [5, 10, 20],these works often

employ the PCA or traditional features like gabor features or Haar-like features to represent the faces. With the advent and rapid development of the deep learning based skills, many computer vision tasks like forgery detection [38] and object recognition [2, 35] have gained obvious performance promotion, additionally, with more and more public face training datasets available on line, the deep learning based methods have also witnessed more and more breakthroughs on performances of face related tasks such as face detection [13] and face recognition [3, 17, 18, 27, 32, 34, 40, 43, 55]. Both works [34] and [32] have been among the pioneer works that have achieved the even better performance than humans. CenterLoss [43] has been proposed to explictly constraint the distances of different classes through pushing the samples of the same class to approach towards their class centers , in an attempt to obtain the larger distance between the different classes and the smaller distance between the same classes. FaceNet [27] has proposed one new loss named tripletloss which is based on the constraints upon the distance of the input paired images that are from the same class or not, obtaining very impressive results, however, the computation efficiency during training phase in FaceNet [27] is relatively low, the preparation of the training image and the construction of training pairs is also very time consuming, based on this, L-SoftLoss [18] has been proposed to improve the computing efficiency while ensuring better performances through enlarging the margin distance based on the well devised losses. SphereFace [17] has aimed at improving the discrimination of the face features through normalizing weights and zero biases and imposing constraints upon the angular margins. CosFace [40] is the improved version of SphereFace, this work has mitigated the problem that the computation of margin in SphereFace is relatively completx, which would cause more computation in the backpropagation, CosFace has proposed to impose the margin constraint on the cosine loss to mitigate this problem. VGGFace2 [3] has proposed one large scale dataset for the training and evaluation of face recognition across different poses and ages. The deep learning based methods are better at exploiting more representative embeddings and features for recognition, and lots of impressive results have been achieved. The main contributions among these outstanding deep learning based works mainly lie on two aspects which are the network architecture settings and the devise of loss functions employed in training phase. DeepFace [34] has tried to align the input faces with the help of the 3D face models, it is the very early attempt to use CNN method in the Face recognition research. In [51], a multitask face recognition architectuer is proposed with the aim to mitigate the over fitting problems that are easy to occur in the training models, it simultaneously solved the problem of pose estimation and the face recognition problems. Recently, lots of newly devised losses have been proposed, like Range Loss [55], CosFace loss [40], Sphere Loss [17], etc. Those above methods are all not aimed specifically at the OFR, as a result of this, the performances of those works may degrade when applied in the OFR scenarios.

2.2 Deocclusion based occluded face recognition methods

Face deocclusion aims at removing the occlusion presented on faces, while one other similar research field called face inpainting mainly aims to recover the faces from the faces with some places missing. The aims of both the two research fields bear much resemblance to each other. The main difference between the two fields lies in that the face inpainting works mainly target to repair the faces with some places missing while the face deocclusion works target to get rid of the face occlusions that already occur and exist on faces.

Song et al. [30] levers the coarse face geometry information to achieve the purpose of both face deocclusion and face editing. With the advent of generative adversarial network, lots of face inpainting works have shown impressive performances [12, 16, 52, 53].Yuan

et al. [53] have exploited the 3DMM [36] to reconstruct the 3d face model, based on the reconstructed 3d face model which consists of rich geometric information, the occluded face region could be better deoccluded. However, the 3d face model which may be not very accurate when input face is occluded may bring side effects upon final deocclusion performance. Besides, there is no recognition performance report presented in [53]. Dong et al. [8] proposes the two stage method which first extracted the occlusion and then both the input occluded face and extracted occlusion would be fed together into the deocclusion pipeline. This method has aimed to ease the problem of performance drop when the new occlusion type occurs in testing phase. Zhao et al. [57] has tried to deocclude the face occlusions through LSTM, specifically, two channels are employed in this deocclusion pipeline, one for face occlusion detection and the other one is for face reconstruction, however, this work needs to split the whole face into several semantically aligned patches first and the training process of LSTM is low efficient and hard to converge. Zhao et al. [58] specifically targets the VR/AR wearing devices which could bring the occlusion into the captured face images, this work has taken the head poses into consideration during the deocclusion process and obtained good performance in the real occluded face images with VR/AR devices weared, however, this work only considers the eyes region occlusion removal which is caused by the AR head-mounted device and the detailed recognition performance comparison results on public datasets are also not reported. Li et al. [15] has firstly proposed to work out the problem of face deocclusion under large poses, however, this method only evaluates the performance upon the datasets collected in the controlled library environment, as a result of this, this method may fail when the input faces are from the in the wild environment.

2.3 Feature filtering based occluded face recognition methods

The feature filtering based methods could be simply divided into two classes, namely, the methods that mainly depend on the traditional features like Local Gabor Binary Patterns (LGBP) feature and LBP feature [9] and the methods that mainly depend on the deep learning based deep features. The methods based on the manually devised features are easy to implement, and the SVM(Support Vector Machine) is the mostly employed tool to perform classification upon those features. The sparse representation classification [46] based method is also one important class of methods in the occluded face recognition, the general pipelines of this kind of methods are to reconstruct the clean face from the occluded face by the combination of training data [46], however, the identity of reconstructed clean face can not be well preserved and the testing occluded faces are mostly at frontal view. Weng et al. [44] has proposed to improve the robustness of face feature to occlusion by aligning the different semantic face places, however, this method calls for the face landmarks detection first and the accuracy of alignment depends on the performance of the face landmarks detection. Weng et al. [45] has proposed the similar method to the work [44], but this work has also incorporated the face geometry information into the feature extraction except the original input of texture features. Yang et al. [47] mainly aimed to mitigate both the illumination and the occlusion problems in face recognition, they have proposed one new two-dimensional image-matrix-based error model , through this model, face features that are more robust to illumination and occlusion could be obtained. Yang et al. [48] has proposed a novel robust kernel representation model with statistical local features(SLF) for robust face recognition ,and the robust regression is adopted to effectively handle face occlusions. Cheng et al. [6] has tried to learn the network parameters that are more appropriate and robust for the occluded faces, to be specific, it replaces those network parameters sensitive to occlusion

with the new parameters which are obtained from the clean face data based training phase, in this way, the parameters that are more stable and robust to occlusion could be learned.

Since the rapid development of deep learning skills, there have emerged more and more deep learning based occluded face feature learning methods. Saez-Trigueros et al. [26] first tries to deploy the different sensitivity of the model to build the training datasets, and then the model sensitive data sets would be fed into the learning pipeline. The method is only tested on the AR occluded face datasets which are all captured on frontal pose and in the constrained environment. Song et al. [31] assumes that there exist some mappings between the occlusion locations and the feature filtering masks, different from other works, the masks employed in this work are not single channel but have the same number of the channels with the input features. This work has fully validated the effectiveness of employing this kind of feature masks. Yin et al. [50] proposed to enable the different feature layers to respond to different semantic feature places of input face respectively, and the output feature would be more interpretable and explainable, besides, upon this interpretable face feature, this work further proposed one new feature filtering method based on the different responsiveness of the features to different face regions which also include the occluded regions, based on this new feature filtering method, the filtered face feature would be more robust to occlusion. Wan et al. [39] proposed to insert one mask net before the feature extraction network , so as to learn the different weights of the features, however, there is not any supervision of this mask which may cause the output of futile masks. Trigueros et al. [37] proposed one new way of obtaining more occluded face training dataset, it resorted to selecting the most sensitive face places to put upon block masks, all the data synthesizing processes are at the image level , and the masks are all from the tedious block masks , therefore, the model generalization ability is severely limited when applied into the real environments. Shao et al. [28] has proposed one new way of synthesizing occluded faces through randomly putting the small colored patches onto clean faces, and also the new training strategy with biased guidance is employed, aiming at reducing the existing biases among the occluded face training data sets.

To sum up, the above methods only considered one of the both methodologies which are based either on the deocclusion principles which generate new information or on the feature filtering principles which eliminate noise information, and the advantages of combining the two approaches are not investigated at all, our proposed method has tried to fill the blanks and is able to fully exploit the potential of each method. Additionally, there are no works that have tried to mitigate the occluded face datasets' insufficiency problem on line, which means that all those works have employed the data augmentation off line, with no clear guidence to indicate the quality of those newly synthethized face datasets. Our work has been the first work that have specifically tried to simultaneoulsy mitigate this data insufficiency problem and guide the generating network to generate the most informative and training samples which are of high quality.

3 Proposed method

Figure 3 shows the overview of our proposed framework. Our proposed occlusion-FR architecture mainly comprises of three modules in all, which are head pose aware deocclusion module, DOMG-OFR, and policy decision based DM respectively. Different from general de-occlusion methods among which only near frontal faces are taken as input and the head poses are also not taken into consideration explicitly, the de-occlusion pipeline in

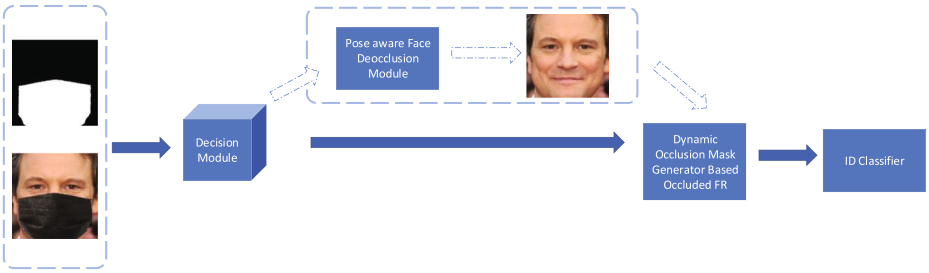

Fig. 3 The Main Architecture of the proposed framework. Our proposed framework would first decide whether to perform the deocclusion operation upon each input occluded face image through DM (Decision Module), which is followed by the DOMG-OFR (Dynamic Occlusion Mask Generator Based Occluded Face Recognition) which is trained in a adversarial way with the aim to generate more informative and valuable occluded training feature data, which would further help improve the robustness of the extracted feature to occlusion. The dotted lines denote that the deocclusion operation is not a must but selective for every input occluded face image

our method takes the face poses into consideration explicitly and the experimental results below indicate that our proposed de-occlusion module does have the advantages over the one without explicitly considering face poses. In DOMG-OFR, we lever the idea of adversarial training to dynamically obtain the most informative masks on line directly, which is more efficient than the existing method [37] which involves lots of efforts in preparing the training datasets off line. The last module is the policy decision module which is employed to select one of the two different methods which is mostly beneficial to the improvement of final face recognition performance. In testing phase, the input face image paired with the corresponding occlusion map first goes through the policy DM to decide which policy should be employed for the following downstream recognition pipeline. Then, the optimal policy about whether to do deocclusion first would be determined for the input occluded faces.

In the remaining part of this section, details of the proposed modules would be elaborated in detail. We would first show our proposed architecture and explain the reason and advantages of this architecture to employ, and then each of the modules which constitute the whole complete architecture would be explained in depth.

3.1 Policy decision based deep architecture design for occluded FR

Figure 4 shows the network architecture employed in our proposed method. Three modules are employed for our task of occluded face recognition, which include two processing modules and one DM, different from other general occluded face recognition pipelines which either consider to extract the clean feature by excluding the side impact from th occluded face region or to put the deoccluded face rather than the original occluded one directly into the general face recognition pipeline.

To combine the advantages of the both methods mentioned above, we analyze in depth how the occlusions influence final face recognition performances, there have been some works that have investigated the relationships between the occlusion and face recognition performance, but there are few works that have investigated the varying performances with occlusions at different places and ratios from a quantitative point of view. Based on our investigation, we further propose this DM which is based on our analysis results. Besides, the second module DOMG-OFR which is used to extract the filtered clean feature has also

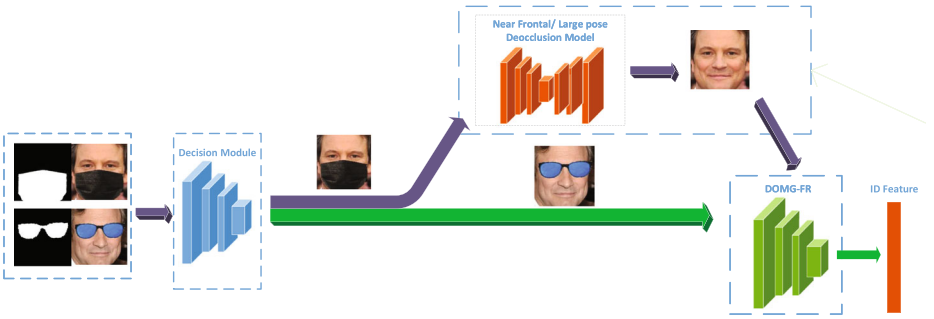

Fig. 4 The network architecture employed in the proposed network. The decision network which is used to judge upon the necessity of the deocclusion operation, this decision network contains five convolution layers followed by two fully connected layers, this network mainly employs the U-net architecture which has demonstrated its merits over other architectures in many other tasks. DOMG-OFR (Dynamic Occlusion Mask Generator Based Occluded Face Recognition) is exploited to enhance the occlusion robustness of the extracted face features

eased the labor cost and reduced the trial and error cost. This module has combined the synthesizing operation and the trail and error into one single module through adversarial training strategy, and the proposed synthething operation is on the feature level rather than the image level, in this way, the synthesizing and computing cost would be both reduced a lot. Additionally, the face deocclusion module employed in our work has also been different from the existing works, to more effectively inhibit the side effects incurred by the head poses in the face deocclusion process, we explicitly consider the head poses and impose the consistency constraint upon the head poses of the faces before and after deocclusion. To further reduce the difficulty in training, we train two deocclusion models in all among which one targets near frontal faces and the other one targets the large pose faces respectively. In this way, we could ensure that the deoccluded faces under large head poses could be more photo realistic, which would definitely bring positive impact for the following recognition pipeline. The last but not least module is the DM, the proposing of this module is based on the observation that the face recognition performance could be further improved when the input occluded faces have the right to choose which method to undergo, specifically, when we purposely assign the method for every input occluded face image based on the devised selection criteria, the recognition performance could be improved by an obvious margin, which means that there indeed exists some selection policy that could been employed directly or learned, to this end, we design this module to better improve the occluded face recognition performances.

3.2 Head pose aware face deocclusion module

Most existing face deocclusion methods have not considered the head poses explicitly, and the test data of those works mostly target the near-frontal faces, however, in real world scenarios, the faces can be in any poses which include the large and even profile head poses. In view of this, while we try to inherit the original merits of the existing works, we also try to put the head poses information into consideration in this deocclusion module, specifically, suppose the input occluded face image is denoted as I, we first use the pretraind head pose estimation method [25] to obtain the head poses of I which is denoted as C, C corresponds to poses at three different directions, which are Yaw poses Y_{pose}, Pitch poses P_{pose} and

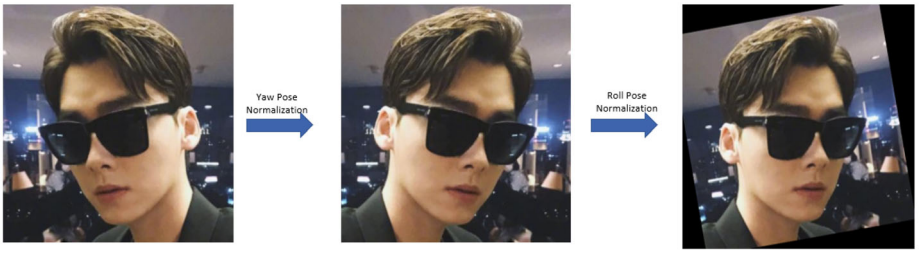

Fig. 5 The face alignment process employed in our work

Roll poses R_{pose} respectively, then, to ease the learning of the deocclusion network, we align all the input images to the left direction at yaw angle and zero degree at roll angle, specifically, we first enable all the input face images to look at left side through the mirror operation based on the positive or negative of the predicted Y_{pose}, then an in plane rotation operation of $-R_{pose}$ degrees is conducted based on predicted Roll poses to make the input face images all remain zero at roll poses. The alignment process could be seen in Fig. 5. Based on the predicted varying head poses, we would divide those input images into two classes which are the near-frontal faces classe and the large pose faces class, we define those face images with yaw angle spanned between $[0°, 45°]$ as the near-frontal faces and the remains are classified as the large pose faces, then, we would train two separate models to deal with those two different classes of face images to deocclude. The flow chart of this deocclusion module is shown in Fig. 6. Besides, we also impose the constraints that the consistency between the face images before and after deocclusion should be remained. In this way, the deocclusion model would try to fully exploit the head pose information during the deocclusion process.

The generator which is denoted as G includes one encoder and one decoder, the encoder consists of five convolutional blocks which are followed by normalization layer and LeakyRelu activation layers, and the average pooling layer is exploited to obtain the face feature embedding, the input pose constraint which is represented as the one hot vector would be concatenated with the embedding, then, the concatenated new feature that encodes both the original face feature and the pose information would be fed into the decoder architecture, and the architecture of decoder mainly includes five deconvolutional layers and outputs the image of the same size with the input image. To make the generated results more robust, the U-net architecture [24] is also employed in the generator. The discriminator enjoys the same architecture to the encoder in the generator except that one softmax layer would be appended after the last feature extraction layer. Due to that we have aligned all faces towards the left side, the head yaw would only span between $[0°, 90°]$, we split the head yaw angle range 90 into 30 classes, which means the one interval has 3 degrees, then the adversarial loss could be defined as follows:

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,c\sim p_d(x,c)}\left[\log D(x, c)\right] + \mathbb{E}_{x,c\sim p_f(x,c)}\left[\log\left(1 - D\left(G(x, c), c\right)\right)\right] \quad (1)$$

in which $p_d(x)$ is the distribution of the real clean faces and $p_f(x)$ is the distribution of the occluded faces, D is the discriminator which is employed to discriminate two aspects, one is whether the deoccluded face image is real and the other one is that whether the head pose of the deoccluded face remains same with the original input occluded face. Therefore, the adversarial loss for solving the optimal deocclusion generator and the discriminator could

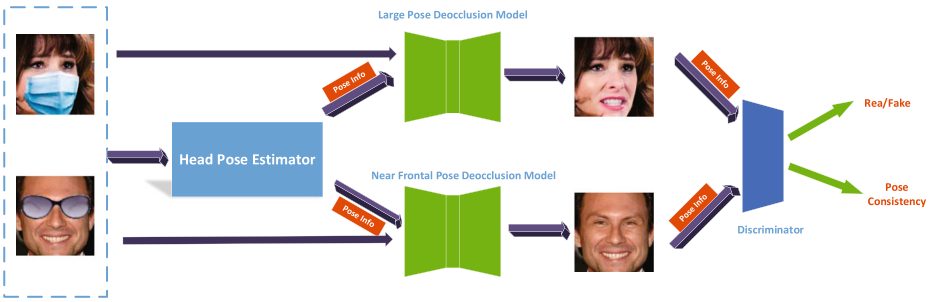

Fig. 6 Architecture of the head pose aware deocclusion pipeline. Our head pose aware deocclusion pipeline first estimates the head pose angle on yaw direction, then the corresponding occluded images at different poses would be fed into different deocclusion models based on predicted different pose classes, to further ensure the robustness of the output deoccluded faces, the head pose cues are embedded into the input of both the generator and the discriminator, the discriminator discriminates not only the reality or fakeness of the output face images, but also the reality of the head pose angles

be defined as follows :

$$G^* = \arg\min_G \mathbb{E}_{x,c\sim p_f(x,c)} \left[\log\left(1 - D\left(G(x,c),c\right)\right)\right] \tag{2}$$

$$D^* = \arg\max_D \mathbb{E}_{x,c\sim p_d(x,c)} \left[\log D(x,c)\right] + \lambda \mathbb{E}_{x,c\sim p_f(x,c)} \left[\log\left(1 - D\left(G(x,c),c\right)\right)\right] \tag{3}$$

where λ is the parameter to balance the two losses which are aimed at supervising different aspects during deocclusion. To further improve the supervision of the generating process, we also introduce the pixel wise loss into the supervision loss, which is defined as follows:

$$\mathcal{L}_{Pixel} = \mathbb{E}_{x\sim p_f(x,c)} \left[\|y - G(x,c)\|\right] \tag{4}$$

where y is the groundtruth deoccluded face image. Besides,to enable the identity feature of the deoccluded face to be better preserved, we exploit the pretrained face recognition model R to extract the id features of faces before and after deocclusion, which are denoted as $R(I)$ and $R(G(I))$, then the identity preserving loss could be defined as:

$$\mathcal{L}_{id} = |R(I) - R(G(I))| \tag{5}$$

The overall loss is as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{cGan} + \lambda_1 \mathcal{L}_{\text{Pixel}} + \lambda_2 \mathcal{L}_{id} \tag{6}$$

where λ_1 and λ_2 are the parameters to balance the three different component of the loss.

3.3 Dynamic occlusion mask generator based occluded face recognition (DOMG-OFR)

One method that is often employed in occluded face recognition is to generate many new synthetic face samples with some predefined occlusion like glasses, cups, etc. covered on, those synthetic face image samples would serve as the input of the training model. This kind of methods do have good effects when it comes to those faces with the predefined types of occlusions on, but it is unrealistic to cover all the occlusion types, so in testing phase, when it comes to the new occlusion types especially when the occlusion is so rare, the face recognition performance would undoubtedly degrade a lot. This phenomenon happens often in the existing works that only targets the synthetic occluded face images with black blocks randomly covered on. Besides, this data heavily dependent based method would consume

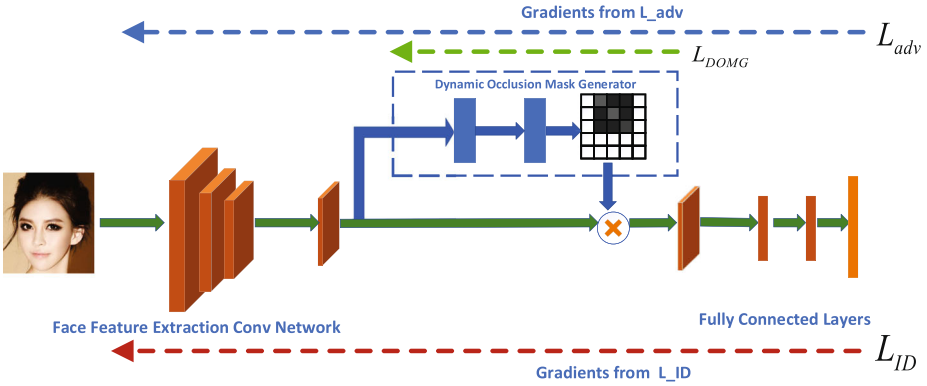

Fig. 7 Architecture of the DOMG (Dynamic Occlusion Mask Generator) based OFR(Occluded Face Recognition) and the gradients back propagation of the training losses. The training process consists of two alternative adversarial procedures, their corresponding gradient back propagation could be seen through the different colors of dot arrows. The DOMG-OFR mainly consists of two modules which are the general face feature extraction pipeline and the dynamic hard feature mask generator, both the two sub modules share the following down stream classification network, and the two sub modules are trained in a adversarial way to improve the robustness of final extracted feature to the face occlusion, while simultaneously saving the labor in synthesizing data and training time to reach converge

much labor and computing cost during the data preparation process, which is so wasteful. In a attempt to ease this problem, we propose this module which could help ahieve the goal of obtaining more robust performance while simultaneously reducing the data preparation labor and computing cost.

As illustrated in Fig. 7, this module consists of two main components which are the recognition module and the adversarial mask generator module. The recognition module could be any classical face feature extraction model like ResNet18 or ResNet50, the DOMG which aims at producing hard occluded face feature samples consists of five convolutional layers followed by the same number of deconvolution layers, the number of filters of last deconvolotion is set to one, the sigmoid function to map the output into [0, 1] is used to output the occlusion feature mask. We denote the feature extraction network as F, the remaining pipelines after the feature extraction model F which are mainly responsible for feature classifying are denoted as A, the input face image is denoted as I, and the DOMG (Dynamic Occlusion Mask Generator) is denoted as P, then the original id feature could be attained through $Y = F(I)$.

To stabilize the training process, we would pretrain the recognition model on the clean face images first and we use the loss proposed in Arcface [7] during the pretraining process, after finishing the pretrain process, the recognition model would get the general recognition ability upon those occlusion-free face images. Then, we would sample some hard occlusion samples through adding block masks at feature map level, specifically, we set one sliding block mask which is at size $\frac{w}{3} \times \frac{w}{3}$, w is the size of the feature map which is the output of the feature extraction network. While we slide this mask on the feature map, we record the recognition loss of every masked feature map, after we finish the sliding process, we select the corresponding mask position with the highest recognition loss value. We repeat this process for M clean face images, and then we would obtain M pairs of adversarial network training data for DOMG, among which one clean face image always corresponds to one specific feature mask. The training data could be denoted as $\{(Y^1, O^1), \ldots, (Y^m, O^m)\}$,

Y is the feature map which is extracted from the feature extraction model F, O is the feature mask which would cause the most dramatic performance degrade, m is the index of the training data. Then based on those hard feature samples, we could pretrain this adversarial hard sample generator network DOMG with the binary cross entropy loss which is formulated as follows:

$$\mathcal{L}_{\mathcal{DOMG}} = -\frac{1}{M}\sum_{m}^{M}\sum_{j,k}^{w}\left[O_{jk}^{m}\mathcal{P}_{jk}\left(Y^{m}\right) + \left(1 - O_{jk}^{m}\right)\left(1 - \mathcal{P}_{jk}\left(Y^{m}\right)\right)\right] \tag{7}$$

where Y is the input feature map which is extracted from F, and the $P(Y)$ is the corresponding output of the DOMG when input feature map is Y. After training the adversarial hard sample producing network P, P would have the ability to generate the masks for reducing the recognition performance of F through proposing hard feature masks.

Naturally, we could train the two networks F and P in a adversarial way, while the general recognition pipeline which involves feature extraction model F and the following classification network A always tries to improve the recognition performance, the adversarial pipeline which always tries to feed the hard occluded face feature samples to pull down the recognition performance in a adversarial way. Therefore, during the training phase, the adversarial loss for training the whole DOMG-OFR network could be defined as follows:

$$\mathcal{L}_{adv} = -\mathcal{L}\ (A(\mathcal{P}(X) \otimes X), C) \tag{8}$$

C is the true id class of the input face image and the X is the face feature of input face I extracted by F. $\otimes$ denotes the element wise multiply operation between the extracted feature by F and the occluded feature mask. Besides, the general id loss is defined as :

$$\mathcal{L}_{\mathcal{ID}} = \mathcal{L}\ (A(\mathcal{F}(I)), C) \tag{9}$$

In the adversarial training process, we keep employing the the loss in Arcface [7] as the training loss of the face recognition model, during the training process, the alternative training policy is employed, to be specific, we alternatively train one of the both networks F and P, while fixing the other network, in this adversarial way, the DOMG would learn to better generate harder feature masks while the face feature extraction network F would try to extract more robust features to the occluded faces. It is worth nothing that during this alternative training process of the whole network, only the clean face images are necessary, which is very different with other existing works which calls for large amounts of synthetic data to feed the training pipeline.

3.4 Deocclusion or not? It has the answer

After finishing training the two above modules, most existing works would just employ one of the two as their final methodology, however, there are few works that have clearly answered the question that when there comes one new occluded face image, with the intention to increase the probability of being recognized correctly, does the deocclusion operation always help improve the recognition performance? Upon this question, we have proposed this module to help us determine whether to deocclude this occluded face image or just keep putting this occluded face image into the general recognition pipeline.

The investigation results in introduction section have suggested that the deocclusion operation is not always a guarantee for recognition performance improvement, besides, we found that the performance is not only related to the ratio of the occluded face region but also related to the occluded position of face image, in the experimental part, we would show the details of these analysis. To take both the ratio and the occluded position simultaneously

into account, we concatenate both the original occluded face image and the occluded map as the input of the decision network, the occluded map plays the role of representing both the occlusion ration and occlusion position, while the original face image carries the specific information related to identity, face poses,light,etc that are not present in the occlusion map.

To obtain sufficient training samples for this decision model, we first feed the occluded face image into the deocclusion based recognition model D and the general feature filtering based face recognition model F , than two recognition results from different methods are obatained, we compare the two scores of recognition results, if both two recognition results are correct, we would assign the general recognition pipeline for this input image, if both results are wrong, we would assign the one with lower testing loss to this face image, if there is only one that is correctly recognized, we would assign this pipeline that has obtained the correct recognition result to the input face image. In this way, plenty of training samples would be collected. Through the combination of the occluded map with the original image as the input of the decision training pipeline, and the corresponding label obtained from above assignment policy would be exploited to supervise this training process. After training, this DM would have the ability to figure out whether to deocclude the input occluded face image first or just directly feed the input face image into the general feature filtering based recognition pipeline, specifically, given that the output of DM is the probability distribution vector and the two elements of the vector indicate the prediction quality of each recognition pipeline upon the same input image, in this way, the specific recognition pipeline with the greater probability that best fits to the input would be assinged more faithfully.

4 Experimental results

4.1 Datasets and evaluation protocals

The most crucial problem in the existing occluded face recognition datasets is that there are few unified public benchmarks and datasets specifically released for occluded face recognition performance evaluation. To better train our proposed network and evaluate our proposed method more faithfully, we first propose one occluded face dataset for training. The dataset is called OCC-Casia, the face images of this dataset are all from the CASIA-WEBFACE dataset [49] which owns 10575 classes with 0.49M samples, based on the original CASIA-WEBFACE dataset, we first collect two hundred kinds of different occlusion types which are splitted into two classes including 170 normal occlusion types and 30 unnormal or even strange occlusion types which would be used for the following comparison experiments, then, we randomly select one of these occlusions and patch this occlusion onto the face image in CASIA, to make the synthetic occluded face image more realistic, during the face occlusion dataset synthesizing, we would make efforts in considering the semantic position of the occlusion presented in real occluded face images, we perform the operation upon each face image in CASIA, then, the OCC-CASIA dataset is obtained. On testing phase, we employ the same method to occlude the face images in LFW [11], and then the OCC-LFW is obtained, OCC-LFW would be exploited to validate the effectiveness of our proposed method. We employ the evaluation protocal in [28] to evaluate the effectiveness of our proposed method, specifically, we employ three different protocals in all, the normal to normal protocal is just the same with the original lfw verification protocal, for the normal to occlusion protocal, we randomly occlude the left or right image of each face pair in the 6000 pairs of lfw protocal, and the Occlusion to Occlusion verification evaluation protocal is through

Fig. 8 Some samples of the synthetic datasets. We specifically select some strange occlusions like planes to be as the datsets for the following comparison experiments

occluding all the images of original face pairs. The cosine similarity distance is used in the face feature distance comparing. Some occluded samples could be seen in Fig. 8.

Besides, we further use the AR dataset [21] which is the traditional occluded face recognition performance evaluation dataset, this dataset is collected under real occlusions, and many previous works have conducted the experiments upon this dataset. The AR dataset owns about 4,000 face images of 126 people with different collecting conditions like facial expressions, occlusions and illuminations. All the occlusions presented in AR dataset are from the real world scenarios. Some examples of AR dataset are shown in Fig. 9. We employ the testing protocal used in [39] as our testing protocal on AR, to be specific, one face image without neither expression or occlusion is utilized as the gallery image, and the twelve images from the same ID with sunglasses or scarves are employed as the probe set.

Last, to validate the effectiveness of our method on masked face recognition which is also among the fied of occluded face recognition, we have conducted the experiment upon the largest occluede face dataset RMFD(Real-World Masked Face Dataset) [42], some examples of this dataset are presented in Fig. 10 this dataset consists of 5,000 pictures of 525 people wearing masks, and 90,000 images of the same 525 subjects without masks. To the best of our knowledge, this is currently the world's largest real-world masked face dataset.We have filtered those pictures with too small resolutions and too severe blurs, after this filering process, there are 4500 masked face images of the 500 subjects and 85000 images of the same 500 subjects without masks wearing. We randomly select 50 subjects as the pre training dataset and the remaining subjects are employed as the test dataset, in both the training and testing phase, the Normal to occlusion protacal is employed, that is,

Fig. 9 Some examples of AR datasets in which face images are under differfent occlusions, lights

one unmasked clean face image of each identity constitue the gallery set, and the remaining masked face images are set as probe set.

4.2 Evaluation metrics

In the following experiments, we use two different kinds of evaluation metrics which are employed to evaluate the verification performance on OCC-LFW dataset and the recognition performane on AR and RMFD datasets. Mathematically, the two metrics could be defined as follows, suppose there are N different IDs in OCC-LFW, and the M random ID paris are constructed from the N IDs as the verification protacal, in our experimental setting, M here is set as 6000, in all the selected ID pairs, the two face images of the each pair could belong to the same ID or not, we would judge whether the two face images of the pair belongs to the same ID or not, then, we would count all the correct judgements, and the sum is denoted as L, then the verification performance could be defined as :

$$Evaluation\ Metric\ of\ Verification\ Performance = L/M \tag{10}$$

. For the evaluation of recognition performance, we denote the total number of the test face images as B, and the number of test face images that are correctly recognized are denoted as T,then, the recognition performance would be defined as:

$$Evaluation\ Metric\ of\ Recognition\ Performance = T/B \tag{11}$$

In the following experiments, all the evaluation metrics emloyed are all from one of the two metrics defined above.

Fig. 10 Some examples of RMFD dataset, different rows correspond to different IDs, the first four colums correspond to the masked faces of the different IDs, and the last column shows the unmasked clean faces of each ID

Besides, ROC(Receiver Operating Characteristic) curves and CMC curves are drawn to further better demonstrate the comparison results. In face recognition related research fields, the ROC curve is often used in evaluating the verification performances, this curve shows both the relationships between the FPR(False Positive Rate) and TPR(True Positive Rate) and the performances at all possible thresholds. FPR(False Positive Rate) means the proportion of the wrongly verified pairs with different IDs in all the pairs with different IDs.TPR(True Positive Rate) means the proportion of the correctly verified pairs with same IDs in all the pairs with same IDs. CMC (Cumulative Match Characteristic) curve shows the different recognition rates under different top number settings.

4.3 Implementation details

We have implemented the proposed method based on the Pytorch framework with the computer equipped with two Nvidia 1080Ti GPUs. We employ Face Attention Network (FAN) [41] as our face detector which is retrained upon OCC-CASIA dataset and robust to occlusions. When training the pose aware deocclusion network, the off the shelf VGGFace2 face recognition model [4] has been used to preserve the identity of the deoccluded face. Resnet18 and Resnet50 are employed as the backbone network of recognition feature extraction network F, to enhance the stability of training phase, we would pretrain F and A on the clean CASIA-WEBFACE dataset first before training DOMG and the whole recognition network in the adversarial way, and both the DOMG in the Resnet18 and Resnet50 based

Fig. 11 The output occlusion maps which is one of the input of the DM

architecture would be inserted before the average pooling layer. All the input images would be resized to 224*224 to fit the network input size.The decision model network mainly consists of five convolution layers and two fully layers, which output one single two elements vector that indicates the probability distribution over the employment of two different recognition pipelines. The cross entropy loss is employed in training the decision network. We randomly sample about 10000 synthetic occluded face images from OCC-CASIA as the training datasets for pretraining DOMG and training DM. In training phase, the occlusion map would be obtained by subtracting the original clean image by the synthetic image. In testing phase, to get the occlusion map which is one indispensable part of the decision network input, we would first adopt the segmentation network in the work [19] to train the

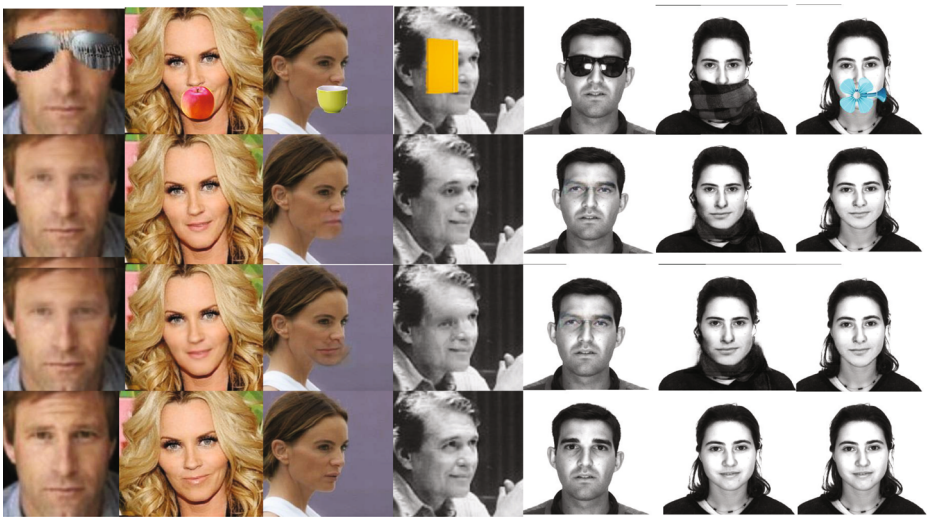

Fig. 12 The deoccluded faces under the head pose aware face deocclusion method. The first row shows the original input occluded faces. The second row shows the deoccluded faces under our proposed head pose aware face deocclusion pipeline. The third row shows the deoccluded faces under our method without considering the head poses. The last row shows the ground truth occlusion free faces

Table 1 The verification performances under normal to occlusion protocal on OCC-LFW

	With DOMG	Without DOMG†
Resnet 18	87%	83.3%
Resnet 50	89.7%	86.5%

†means the result of the baseline model which is the general recognition model trained on both the CASIA and OCC-CASIA dataset

occlusion map generator from the occluded face images, to be specific, the FCN-8s segmentation network used in the work [19] would be trained on both the data from OCC-CASIA and some AR dateaset that are not from the testing subjects, when finishing the training of this segmentation network, it would be used to infer the location of face occlusions, in this way, the occlusion map used as one input component of the decision model would be obtained. Some of the output occlusion maps are demonstrated in Fig. 11. In all the training phases, Adam optimizer has been employed. The learning rate is initially set as 1e-3 for the adversarial mask generator based occluded face recognition network and is decayed for every 10 epochs, for the deocclusion network, the learning rate is initially set as 1e-4 and is decayed for every 5 epochs, the decision network is trained with the learning rate 1e-3 for 3 epochs. The balanced parameters λ_1 and λ_2 in training loss L_{all} are set to 1 and 0.01.

4.4 Effectiveness of head pose aware deocclusion model

To validate the effectiveness of the three components of our proposed method, we would reveal the necessity of the three components from both the qualitative and quantitative views. We have compared the results of the two deocclusion methods among which one has considered the head poses explicitly while the other not. As is shown in Fig. 12, the performance of the deoccluded faces of near frontal poses under the two methods are nearly same, however, as can be drawn from the third and fourth column of the Fig. 12, the deoccluded face image under large poses with our method considering the head poses is better than the one generated from the method which doesn't take the head pose constraints into consideration. We could see that when the input faces are under large poses, the performance of the method which is not supervised by the head poses would significantly degrade. This comparison result fully demonstrate the advantage of combining the head pose constraints into the deoccluded face generation especially when input faces are under large poses. Besides, from the fifth column of Fig. 12, we could find that with the occlusion extent larger, it is harder to preserve the identity of the input occluded faces, which are also very in line with our common sense. After all, there are coutless possibilities of how the deoccluded faces present, as long as the occlusion free part of the input occluded faces remain unchanged after being deoccluded. Since that the ground truth deoccluded faces of AR dataset acturally do not exist at all, we also conduct the deocclusion process upon the AR dataset to examine the performance of our proposed method, the results are listed in the last column of Fig. 12.

Table 2 The verification performances under normal to occlusion protocal with usual occlusions

	With DOMG	Without DOMG†
Resnet 18	87.7%	86.4%
Resnet 50	90.2%	89.4%

Table 3 The verification performances under normal to occlusion protocal with unnormal or strange occlusions

	With DOMG	Without DOMG†
Resnet 18	87.4%	82%
Resnet 50	89.5%	85.1%

4.5 Effectiveness of DOMG-OFR

To show the superiority of the DOMG-OFR method over other methods which heavily depend on the synthetic datasets which are obtained through simply placing the occlusions upon faces, we compare the performance from two aspects which include face verification performance and the time consumed in training to convergence.

First, we perform two comparison experiments which include training two networks with and without adversarial occlusion mask generator. To mitigate the impact of different network architectureS and obtain more faithful result, we employ the resnet18 and resnet50 as the feature extraction network F, which could represent the shallow and deeper network respectively. Then we would compare the face verification performance under the normal to occlusion protocal on OCC-LFW as mentioned in Section 4.1, the results are shown at Table 1, to further show the advantages of our adversarial feature mask generator based method, two other experiments are conducted and the experimental results are shown at Tables 2 and 3. In the first experiment, we would obtain those synthesized testing data through those normal and usual occlusions like glasses or masks, and the rare and even unusual masks would not be exploited in the data synthesizing, we could see that when the masks are from the normal and usual ones, the performance almost remain the same, however, in the second experiment, when all the masks used in the testing dataset are all from the strange types like planes or other irregular and unnormal strange ones, DOMG-OFR would show obvious advantage over the one that depends heavily on the training data. We deduce that the main reason is that DOMG based method would try to exploit the very hard feature samples to feed to the recognition network, and the face features of those hard

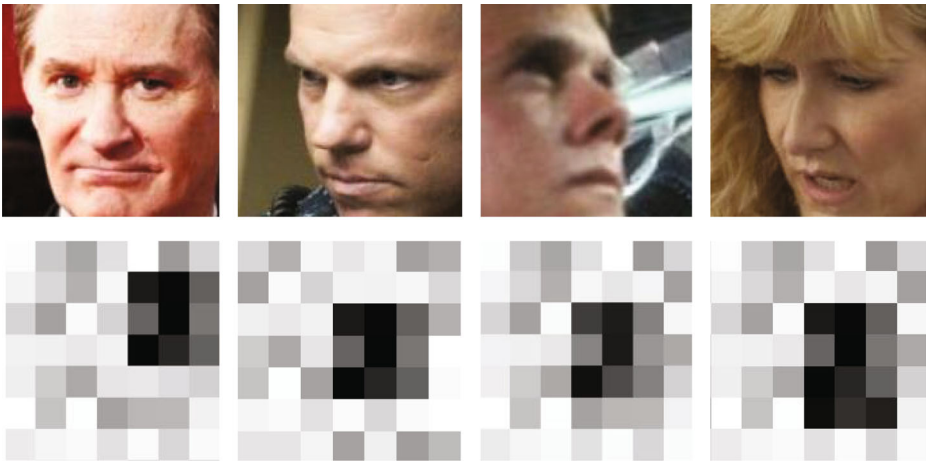

Fig. 13 The generated feature occlusion map by DOMG (Dynamic Occlusion Mask Generator)

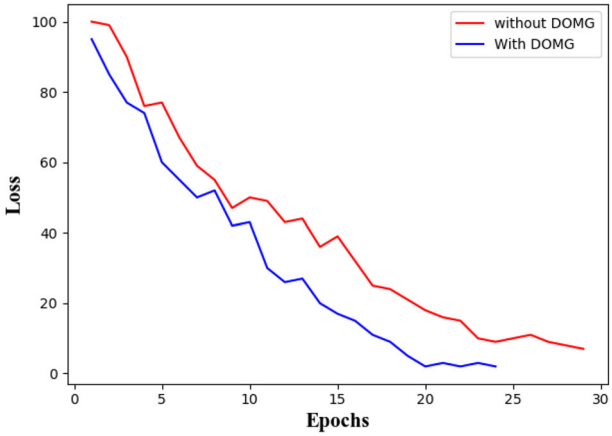

Fig. 14 Evolution of the average training losses as a function of the number of epochs

or even so strange occluded faces would also be fed into the recognition model, besides, all the masks generated are at the feature level, as a result of this, the recognition model would be more robust to the varieties of masks that may be so rare and even strange which occurred on faces. To make the results more vivid , we visualize some generated feature masks by DOMG and their corresponding original clean input face images, which are shown in Fig. 13. From Fig. 13, it can be seen that the corresponding sensitive regions of original clean face image which correspond to the near zero elements of generated feature masks all bear important and significant identity information.

Further more, we found that our method would achieve convergence more quickly than the one not equipped with DOMG.We train two occluded face recognition networks with

Fig. 15 The block occluded faces, the first row shows the faces with different semantic places occluded, the second row shows the faces with different ratio of face occluded

Table 4 The different verification performance drops with the different semantic features occluded

	Forehead	leftEye	Nose	leftCornerOfMouth	Chin	centerOfFace	Ear
Resnet 18	10.1%	7.3%	12.7%	6.6%	3.8%	12.8%	0.5%
Resnet 50	8.1%	5.1%	9.8%	5.1%	2.2%	11.2%	0.6%

and without our dynamic mask generator DOMG, when training the one without DOMG, we use both original CASIA-WEBFACE and the OCC-CASIA as the training data. The two convergence curves are shown in Fig. 14. We analyze that the reason should be that our mask generator DOMG always keep trying to feed the samples that are relatively harder but more informative and valuable, which is so different with the methods that just keep learning from the large amount of synthetic redundant occluded faces with masks randomly placed on.

4.6 Effectiveness of DM

In this part we would first investigate the relationship between the occlusion and recognition performance from the quantitative view. We would mainly aim at analyzing the two factors which are occlusion locations and the occlusion ratios of whole face. Just like the experimental settings above, we also employ the resnet18 and resnet50 as the feature extraction network F and the comparison protocal is also set as normal to occlusion upon OCC-LFW as mentioned above. For the occlusion location factor analysis, we would set the occlusion area as big as the one eight size of the original input face. Then, we would place the occlusion at different face places which mainly include the locations bearing the semantic features like eyes, noses,etc. Some occluded examples are shown in Fig. 15. The performances of the model with different semantic features being occluded are shown in Table 4, as can be seen from Table 4, the performance would be dropped at different degrees when the occlusion occurs at different places. This finding is also very intuitive, since when we people try to identify some one with face occluded, the difficulty to correctly recognize also varieties with the different face places occluded.

For the factor of occlusion ratio of the whole face, we set the occlusion size at three different sizes which include one second, one fourth, and one eighth of the size of the original input face respectively. The occlusion at three different ratios are all centered at the whole face. The performances of those different occluded faces are shown at Table 5, from Table 5, we could see that the different occlusion ratios would also bring different performance drops . In a word, it could be concluded that the recognition performance of occluded faces not only depends on the occlusion sizes but also the occlusion positions. Therefore, the occlusion map which is one of the two inputs of the DM could exactly well define the different occlusion places and the different occlusion sizes. So, it is reasonable to set the input of DM as the concatenation of both the original image and the occlusion map, the combination of both the two inputs would further help the DM make smarter decisions

Table 5 The different performance drops of occluded faces at different ratios

	1/8	1/4	1/2
Resnet 18	5.3%	25.5%	45.1%
Resnet 50	4.5%	24.1%	42.7%

Table 6 The different verification performances with and without the DM

	With DM	Without DM*
Resnet 18	87.8%	84.6%
Resnet 50	91.2%	88.1%

*indicates that when the DM is not employed , there would be two different recognition performances which are from the deocclusion based method and the dynamic occlusion masks generator based feature filtering method, and we choose the better one of the two different performances as the counterpart in our comparison experiments

on whether the input occluded face image is worth being deoccluded before being fed into the following recognition pipeline.

To examine the effectiveness of this module, we employ the method with and without DM appended respectively, and the verification results under normal to occlusion protocal on OCC-LFW are listed at Table 6. It is worth noting that in this experiment, the occlusions in the testing dataset are all randomly sampled from all the predefined occlusions which include both the usual and unusual ones. As can be seen from Table 6, the decision based model could intelligently make decisions of whether to deocclude the input face image first before forwarding it into the following recognition pipeline, which really helps improve the face recognition performance by an obvious margin. This result have fully validated the functionality of the DM.

Besides, we also conduct the comparison experiments to prove that the input occlusion map does contribute to the final performance improvement, the comparison results are shown at Table 7. We retrain the decision network with its input to be only the original face image while the other conditions like network architecture and training loss, etc. all unchanged. From Table 7, it could be seen that when the input is only the single face image, the recognition performance would drop to an obviously extent, this should be due to the more wrong decisions made by DM which only has one single input, hence, the occlusion map really plays the important role in making the smart decisions on whether to deocclude the input occluded face image.

4.7 Comparison of face verification performance upon three evaluation protocals

For that the protocals mentioned in the previous section employed in our paper are only used in BFL [28], so we would compare our results with it based on three protocals as conducted in [28]. We would continue using the resnet18 as our backbone network like previous experiments, and BFL [28] has also given the results employing the resnet18 as their feature extraction network and the Arcface [7] as the training loss, which remain the same with our experimental settings. Additinally, we also implement the method in PDSN [31] by replacing the backbone network which is originally Resnet50 with Resnet18, ensuring the fairness of the comparison results. The comparison results are shown at Table 8,

Table 7 The different verification performances with different inputs to the DM

	With Single Input	With Concatenation Input
Resnet 18	86.3%	87.8%
Resnet 50	90.1%	91.2%

Table 8 The verification performance comparison results on OCC-LFW

	Normal-Occlusion	Occlusion-Occlusion	Normal-Normal
Ours	**87.8%**	**75.23%**	**98.82%**
BFL [28]	83.91%	70.27%	98.67%
PDSN [31]	85.61%	74.29%	98.80%
MaskNet [39]	82.19%	73.25%	97.70%
DC-SSDA [6]	77.96%	69.59%	95.90%

from Table 8, it can be suggested that our method all achieve competitive results compared to the recent works BFL [28] and PDSN [31] under two occlusion protocals, however, under the normal to normal protocal, we could find that the advantage over the other two methods is not very significant, the reason should be that we put main efforts upon the occluded face recognition rather than the general face recognition problems which have been well solved by so many existing previous works. Besides, we also reimplement the previous works MaskNet and DC-SSDA and report the final performances, MaskNet [39] has shown some advantages in the occluded face recognition, however, since there do not exist any supersion on the generated masks, so the performance is still not very competitive, DC-SSDA [6] has employed the shallow network which consists only several fully connected layers, therfore, and performance under the normal to normal protacal is poor compared to other three works. To show more vividly, we have drawn the ROC(Receiver Operating Characteristic Curve) curves of our work and other related works to compare are shown in Fig. 16, from Fig. 16, we can further validate the advantages of our method more strongly.

4.8 Comparion of face recognition performance on AR dataset and RMFD(real-world masked face dataset)

AR dataset is the occluded face recognition dataset collected from the real environment which could help further examine the performance of our proposed method when applied in

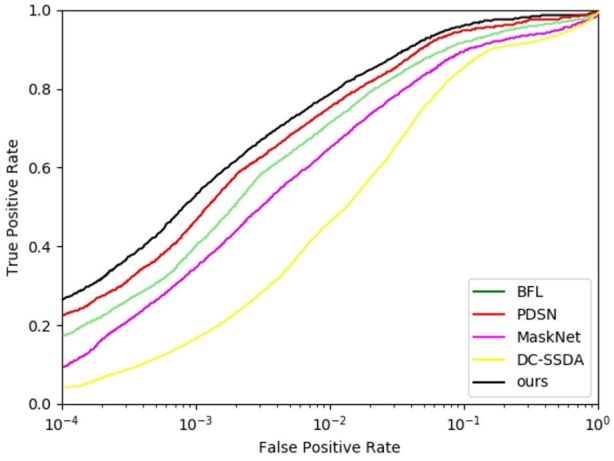

Fig. 16 ROC curves based on the verification results of our work and other works under Normal to Occlusion protacal on OCC-LFW. (The curve that is more to the top left is better)

Table 9 The recognition performance comparison results on AR dataset

	Sunglass	Scar
Ours	**99.70%**	**99.84%**
PDSN [31]	**99.70%**	99.82%
BFL [28]	99.62%	98.73%
MLERPM [44]	98.00%	97.00%
SRC,wright2008robust	87.00%	59.50%
MaskNet,wan2017occlusion	90.90%	96.70%
SCF-PKR,yang2013robust	95.65%	98.00%
NMR,yang2016nuclear	96.90%	73.50%
RPSM [45]	96.00%	97.66%

the real scenarios and applications. To compare fairly with other works, we use the Resnet50 as the backbone network of the feature filtering based method, we have retrained the network in PDSN [31] upon our training datasets and reported the final results, to make the model more robust to the extracted feature, we use the face images of subjects that are outside the training identities to fine-tune the DOMG-OFR first. The evaluation results are given in Table 9, from Table 9, we can see that our method can obviously improve the recognition performance. In addition, during the comparison experiment, we also conduct the statistics upon the decisions about whether to apply the deocclusion operation made by DM, interestingly, we found that the faces under extreme light or glass occlusions mostly call for the deocclusion operation first, and the most of the remaining faces including the scarf occlusion or the neural light are just fed straight into the feature filtering recognition pipeline like DOMG-OFR directly, we recon the reason behind this interesting phenomenon is that the glasses occlusion or extreme light are more easy to bring side effect upon the recognition performance, and the extreme light is easy to be mistaken as some kind of strange occlusion, besides, compared with scarf, the glass occlusion may cause more severe performance drop, which is also very intuitive.

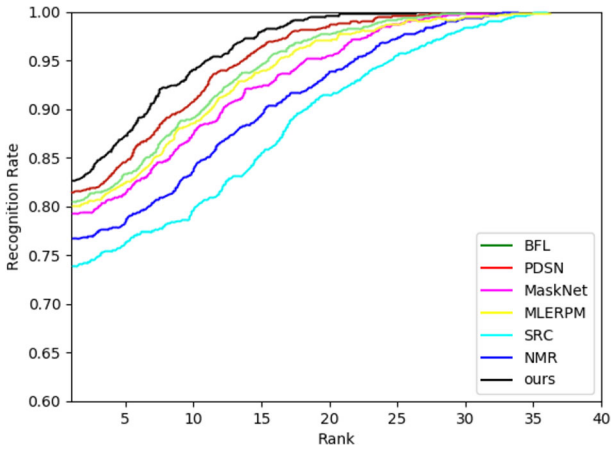

Fig. 17 CMC curves based on the recognition results of our work and other works under Normal to Occlusion protacal on RMFD dataset. (The curve that is more to the top left is better)

We have also conducted the experiments upon the RMFD dataset, this dataset has been the dataset specifically collected to evaluate the performance of masked face recognition. We have given the performance comparison of different works based on the CMC(Cumulative Match Characteristic) curve which is shown in Fig. 17, from Fig. 17, we could see that our method has achieved the best performance.

Last, we have also analyzed the computational efficiency which is of significant importance when applied into the real face recognition systems. During training phase, the average time(including the backpropagation computing time) consumed in one face image is about 0.3s, and in the testing phase, we set the batchsize to one to track the total amount of time consumed for one input face image, we have averaged the sum time of extracting the face features of 1000 input face images, and the average time to consume per image in testing is about 0.02s, which could meet requirements of most face recognition application systems.

5 Conclusions

In this paper, we have proposed one new OFR pipeline which is mainly based on head pose aware deocclusion module, occlusion robust feature extraction module DOMG-FR and the DM. We have innovatively integrated both the merits of the face deocclusion based and the face feature filtering based methods through the smart decision module DM, and both the qualitative and quantitative results have fully validated the superiority of our proposed method. However, there still exist some limitations in our work, the occlusion maps which play the important role in DM are sometimes not very accurate, how to increase the accuracy of the occlusion maps is still worth the further study.

In future, we would also consider how to leverage the reinforcement learning which is better at policy decisions to further improve the recognition performances, and one sequential deep reinforcement learning model for selecting the most informative face areas in occluded face would be proposed, additionlly, the DM would also benefit from the whole extracted sequential features and output predition results that are more robust.

Declarations

Conflict of Interests The authors have no relevant financial or non-financial interests to disclose. The authors did not receive support from any organization for the submitted work.

References

1. Ahn C, Kim E, Oh S (2019) Deep elastic networks with model selection for multi-task learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6529–6538
2. Bansal M, Kumar M, Kumar M (2021) 2d object recognition techniques: State-of-the-art work. Archives of Computational Methods in Engineering 28(3)
3. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: 2018 13Th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 67–74
4. Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: 13Th IEEE international conference on automatic face & gesture recognition, FG 2018, xi'an, china, may 15-19, 2018. IEEE Computer Society, pp 67–74. https://doi.org/10.1109/FG.2018.00020

5. Chakraborty S, Singh SK, Chakraborty P (2018) Local gradient hexa pattern: a descriptor for face recognition and retrieval. IEEE Trans Circuits Syst Video Technol 28(1):171–180. https://doi.org/10.1109/TCSVT.2016.2603535
6. Cheng L, Wang J, Gong Y, Hou Q (2015) Robust deep auto-encoder for occluded face recognition. In: Zhou X, Smeaton AF, Tian Q, Bulterman DCA, Shen HT, Mayer-Patel K, Yan S (eds) Proceedings of the 23rd annual ACM conference on multimedia conference, MM '15, Brisbane, Australia, October 26 - 30, 2015. ACM, pp 1099–1102. https://doi.org/10.1145/2733373.2806291
7. Deng J, Guo J, Xue N, Zafeiriou S (2019) Arcface: Additive angular margin loss for deep face recognition. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, long beach, CA, USA, June 16-20, 2019. Computer Vision Foundation / IEEE, pp 4690–4699. https://doi.org/10.1109/CVPR.2019.00482
8. Dong J, Zhang L, Zhang H, Liu W (2020) Occlusion-aware gan for face de-occlusion in the wild. In: 2020 IEEE International conference on multimedia and expo (ICME), pp 1–6. https://doi.org/10.1109/ICME46284.2020.9102788
9. Dosi H, Keshri R, Srivastav P, Agrawal A (2018) Comparison between LGBP and DCLBP for non-frontal emotion recognition. In: Chaudhuri BB, Nakagawa M, Khanna P, Kumar S (eds) Proceedings of 3rd International Conference on Computer Vision and Image Processing - CVIP 2018, Jabalpur, India, September 29 - October 1, 2018, vol 2, Advances in Intelligent Systems and Computing, vol 1024. Springer, pp 339–349. https://doi.org/10.1007/978-981-32-9291-8_27
10. Hu J, Lu J, Tan YP, Yuan J, Zhou J (2017) Local large-margin multi-metric learning for face and kinship verification. IEEE Trans Circuits Syst Video Technol 28(8):1875–1891
11. Huang GB, Mattar M, Berg T, Learned-miller E (2008) Labeled faces in the wild: A database forstudying face recognition in unconstrained environments
12. Iizuka S, Simo-serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph. 6(4):107:1–107:14. https://doi.org/10.1145/3072959.3073659
13. Kumar A, Kaur A, Kumar M (2019) Face detection techniques: A review. Artif Intell Rev 52(2):927–948
14. Kumar A, Kumar M, Kaur A (2021) Face detection in still images under occlusion and non-uniform illumination. Multimed Tools Appl 80(10):14565–14590
15. Li K, Zhao Q (2020) If-gan: Generative adversarial network for identity preserving facial image inpainting and frontalization. In: 2020 15Th IEEE international conference on automatic face and gesture recognition (FG 2020)(FG), pp 158–165
16. Li Y, Liu S, Yang J, Yang MH (2017) Generative face completion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3911–3919
17. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 212–220
18. Liu W, Wen Y, Yu Z, Yang M (2016) Large-margin softmax loss for convolutional neural networks. In: ICML, p 7
19. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
20. Low CY, Teoh ABJ, Ng CJ (2017) Multi-fold gabor, pca, and ica filter convolution descriptor for face recognition. IEEE Trans Circuits Syst Video Technol 29(1):115–129
21. Martinez AM (1998) The ar face database CVC Technical Report24
22. Nash S, Rhodes M, Olszewska JI (2016) Ifr: Interactively pose corrected face recognition. In: Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOSIGNALS, (BIOSTEC 2016). INSTICC, SciTePress, pp 106–112. https://doi.org/10.5220/0005857801060112
23. Olszewska JI (2016) Automated face recognition: Challenges and solutions. Pattern Recognition-Analysis and Applications, 59–79
24. Ronneberger O (2017) Invited talk: U-net convolutional networks for biomedical image segmentation. In: Maier-Hein KH, Deserno TM, Handels H, Tolxdorff T (eds) Bildverarbeitung für die Medizin 2017 - Algorithmen - Systeme - Anwendungen. Proceedings des Workshops vom 12. bis 14. März 2017 in Heidelberg, Informatik Aktuell. Springer, p 3. https://doi.org/10.1007/978-3-662-54345-0_3
25. Ruiz N, Chong E, Rehg JM (2018) Fine-grained head pose estimation without keypoints. In: 2018 IEEE Conference on computer vision and pattern recognition workshops, CVPR workshops 2018, salt lake city, UT, USA, June 18-22, 2018. IEEE Computer Society, pp 2074–2083. https://doi.org/10.1109/CVPRW.2018.00281
26. Saez-Trigueros D, Meng L, Hartnett M (2018) Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. Image Vis Comput 79:99–108. https://doi.org/10.1016/j.imavis.2018.09.011

27. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
28. Shao C, Huo J, Qi L, Feng Z, Li W, Dong C, Gao Y (2020) Biased feature learning for occlusion invariant face recognition. In: Bessiere C (ed) Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020, pp 666–672. ijcai.org. https://doi.org/10.24963/ijcai.2020/93
29. Singh S, Ahuja U, Kumar M, Kumar K, Sachdeva M (2021) Face mask detection using yolov3 and faster r-cnn models: Covid-19 environment. Multimed Tools Appl 80(13):19753–19768
30. Song L, Cao J, Song L, Hu Y, He R (2019) Geometry-aware face completion and editing. In: The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, honolulu, hawaii, USA, January 27 - February 1, 2019. AAAI Press, pp 2506–2513. https://doi.org/10.1609/aaai.v33i01.33012506
31. 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: 2019 IEEE/CVF International conference on computer vision, ICCV 2019, seoul, korea (south), october 27 - november 2, 2019. IEEE, pp 773–782. https://doi.org/10.1109/ICCV.2019.00086
32. Sun Y, Chen Y, Wang X, Tang X (2014) Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems, pp 1988–1996
33. Sun Y, Liang D, Wang X, Tang X (2015) Deepid3: Face recognition with very deep neural networks. arXiv:1502.00873
34. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708
35. Tan J, Wang C, Li B, Li Q, Ouyang W, Yin C, Yan J (2020) Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
36. Tran L, Liu X (2018) Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7346–7355
37. Trigueros DS, Meng L, Hartnett M (2018) Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. Image Vis Comput 79:99–108
38. Walia S, Kumar K, Kumar M, Gao XZ (2021) Fusion of handcrafted and deep features for forgery detection in digital images. IEEE Access 9:99742–99755
39. Wan W, Chen J (2017) Occlusion robust face recognition based on mask learning. In: 2017 IEEE International conference on image processing (ICIP). IEEE, pp 3795–3799
40. 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. In: 2018 IEEE Conference on computer vision and pattern recognition, CVPR 2018, salt lake city, UT, USA, June 18-22, 2018. IEEE Computer Society, pp 5265–5274. https://doi.org/10.1109/CVPR.2018.00552
41. Wang J, Yuan Y, Yu G (2017) Face attention network: An effective face detector for the occluded faces. arXiv:1711.07246
42. Wang Z, Wang G, Huang B, Xiong Z, Hong Q, Wu H, Yi P, Jiang K, Wang N, Pei Y et al (2020) Masked face recognition dataset and application. arXiv:2003.09093
43. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. Springer, pp 499–515
44. Weng R, Lu J, Hu J, Yang G, Tan YP (2013) Robust feature set matching for partial face recognition. In: Proceedings of the IEEE international conference on computer vision, pp 601–608
45. Weng R, Lu J, Tan YP (2016) Robust point set matching for partial face recognition. IEEE Transactions on Image Processing 25(3):1163–1176
46. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2008) Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2):210–227
47. 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 Transactions on Pattern Analysis and Machine Intelligence 39(1):156–171
48. Yang M, Zhang L, Shiu SCK, Zhang D (2013) Robust kernel representation with statistical local features for face recognition. IEEE Transactions on Neural Networks and Learning Systems 24(6):900–912
49. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv:1411.7923
50. Yin B, Tran L, Li H, Shen X, Liu X (2019) Towards interpretable face recognition. In: 2019 IEEE/CVF International conference on computer vision, ICCV 2019, seoul, korea (south), october 27 - november 2, 2019. IEEE, pp 9347–9356. https://doi.org/10.1109/ICCV.2019.00944

51. Yin X, Liu X (2018) Multi-task convolutional neural network for pose-invariant face recognition. IEEE Trans Image Process 27(2):964–975. https://doi.org/10.1109/TIP.2017.2765830
52. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. In: 2018 IEEE Conference on computer vision and pattern recognition, CVPR 2018, salt lake city, UT, USA, June 18-22, 2018. IEEE Computer Society, pp 5505–5514. https://doi.org/10.1109/CVPR.2018.00577
53. Yuan X, Park IK (2019) Face de-occlusion using 3d morphable model and generative adversarial network. In: 2019 IEEE/CVF International conference on computer vision, ICCV 2019, seoul, korea (south), october 27 - november 2, 2019. IEEE, pp 10061–10070. https://doi.org/10.1109/ICCV.2019.01016
54. Zeng D, Veldhuis RNJ, Spreeuwers LJ (2020) A survey of face recognition techniques under occlusion. arXiv:2006.11366
55. Zhang X, Fang Z, Wen Y, Li Z, Qiao Y (2017) Range loss for deep face recognition with long-tailed training data. In: IEEE International conference on computer vision, ICCV 2017, venice, italy, october 22-29, 2017. IEEE Computer Society, pp 5419–5428. https://doi.org/10.1109/ICCV.2017.578
56. Zhang Z, Shen W, Qiao S, Wang Y, Wang B, Yuille A (2020) Robust face detection via learning small faces on hard images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)
57. Zhao F, Feng J, Zhao J, Yang W, Yan S (2018) Robust lstm-autoencoders for face de-occlusion in the wild. IEEE Trans Image Process 27(2):778–790. https://doi.org/10.1109/TIP.2017.2771408
58. Zhao Y, Chen W, Xing J, Li X, Bessinger Z, Liu F, Zuo W, Yang R (2018) Identity preserving face completion for large ocular region occlusion. In: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018. BMVA Press, p 109. http://bmvc2018.org/contents/papers/0387.pdf

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