Masked Face Recognition with Generative Data Augmentation and Domain Constrained Ranking

Mengyue Geng
Department of Computer Science and Technology, Peking University

Peixi Peng*
Department of Computer Science and Technology, Peking University
Peng Cheng Laboratory

Yangru Huang
Department of Computer Science and Technology, Peking University

Yonghong Tian*
Department of Computer Science and Technology, Peking University
Peng Cheng Laboratory

ABSTRACT

Masked faces recognition (MFR) aims to match a masked face with its corresponding full face, which is an important task especially during the global outbreak of COVID-19. However, most existing face recognition models generalize poorly in this case, and it is hard to train a robust MFR model due to two main reasons: 1) the absence of large scale training data as well as ground truth testing data, and 2) the presence of large intra-class variation between masked faces and full faces. To address the first challenge, this paper firstly contributes a new dataset denoted as *MFSR*, which consists of two parts. The first part contains 9,742 masked face images with mask region segmentation annotation. The second part contains 11,615 images of 1,004 identities, and each identity has masked and full face images with various orientations, lighting conditions and mask types. However, it is still not enough for training MFR models with deep learning. To obtain sufficient training data, based on the *MFSR*, we introduce a novel Identity Aware Mask GAN (IAMGAN) with segmentation guided multi-level identity preserve module to generate the synthetic masked face images from the full face images. In addition, to tackle the second challenge, a Domain Constrained Ranking (DCR) loss is proposed by adopting a center-based cross-domain ranking strategy. For each identity, two centers are designed which correspond to the full face images and the masked face images respectively. The DCR forces the feature of masked faces getting closer to its corresponding full face center and vice-versa. Experimental results on the *MFSR* dataset demonstrate the effectiveness of the proposed approaches.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Image representations**; **Object identification**.

*Corresponding author: Peixi Peng, Yonghong Tian
Email: pxpeng@pku.edu.cn, yhtian@pku.edu.cn

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413723

Figure 1: Masked faces (first row) lost most of the facial cues compare with corresponding full faces (second row).

KEYWORDS

Masked Face Recognition; Identity Aware Mask GAN; Domain Constrained Ranking

ACM Reference Format:
Mengyue Geng, Peixi Peng, Yangru Huang, and Yonghong Tian. 2020. Masked Face Recognition with Generative Data Augmentation and Domain Constrained Ranking. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3394171.3413723

1 INTRODUCTION

Face recognition has received much attention in recent years, and has been well addressed due to the great success of convolutional neural networks (CNNs) [3, 20, 25, 29, 34, 39]. However, most existing works are designed based on full faces, which can be easily violated in several real-world applications, especially when more and more people are wearing facial masks in their daily life with the global outbreak of diseases such as COVID-19. Since most facial cues are occluded by the mask, most existing full face recognition models generalize poorly in this case. In order to make the face recognition method more scalable, this paper focuses on the masked face recognition (MFR) task, which aims to match masked faces among a set of full faces.

Despite the importance of MFR, it is still a very challenging task because of two main reasons. The **first issue** is the absence of large scale training data. Different with full face recognition datasets [1, 6, 40] which can be semi-automatically collected from Internet, it is extremely hard to collect large scale mask face images with identity information. For each target identity, the masked face either does not exist, or needs to be carefully selected, which can be notoriously time consuming. To the best of our knowledge, MAFA [4] is one of the biggest datasets with masked faces, yet it is collected for masked face detection and does not contain any identity information. The

second issue is the learning method. As shown in Figure 1, the masks greatly affect the global visual appearance of a face and increases intra-class variance, which makes it hard to train an effective MFR model.

To handle the first challenge, this paper presents a novel Masked Face Segmentation and Recognition (*MFSR*) dataset which consists of two parts. The first part contains 9,742 web-collected masked face images with manually labeled mask region segmentation annotation, and the second part contains 11,615 images collected from 1,004 identities, in which 704 are real-world collected and the rest 300 identities are obtained from Internet. We ensure each identity has at least one full face image and one masked face image. As is known, a robust face recognition model [1, 29] needs millions identities to train and 1,004 identities is not enough. Although the MFR training data is hard to collect, there exists several large scale full face recognition datasets [1, 6, 40]. Motivated by the success of synthetic training data [22], we propose a novel Identity Aware Mask GAN (IAMGAN) based on *MFSR* to generate synthetic masked face images from standard full face images. The IAMGAN consists of two modules, the first module is a cyclic generator [44] that converts full face images from standard face recognition datasets into corresponding masked faces. Because of the huge domain difference and lacking of paired training data between masked images and full face images, the cyclic generator alone may generate images that completely lose the identity related facial details. To address this problem, we further add a multi-level identity preserve module. Specifically, the generated masked face image is fed into a segmentation network to get its mask region, then a semantic region guided multi-level idenity preserve loss is applied on the full face image and generated masked image to ensure the generated image keeps the identity information on both identity level and pixel level. Compared with using only full face training data, we show that models trained with the generated masked face image can achieve significantly better recognition performance.

In order to tackle the large intra-class variation between masked faces and full faces, we take advantage of the idea of learning class centers [39]. Instead of treating all samples from the same label equally and forming a single center for each class, we assume that features of masked faces often contain mask region related information that should be separately modeled. To this end, we learn two centers for each class, one for full face images and the other for masked face images. Then, a Domain Constrained Ranking Loss (DCR) is used to force the feature of masked faces getting closer to its corresponding full face center and vice-versa. By doing so the model is able to simultaneously separate different identities and learn to extract identity specific feature from both masked face and full face images.

The main contributions of this paper are three folds: 1) We build a Masked Face Segmentation and Recognition (*MFSR*) dataset which is used for developing and benchmarking masked face recognition models; 2) A novel Identity Aware Mask GAN is proposed to generate synthetic masked face images from standard face images as a remedy of training data shortage; and 3) A novel Domain Constrained Ranking Loss is designed to learn discriminative deep features for masked face recognition.

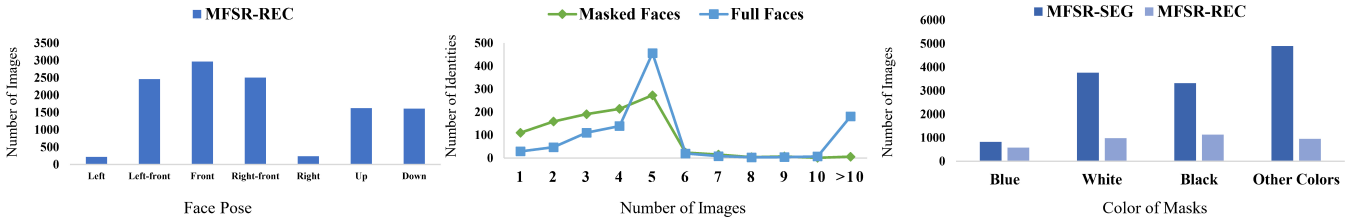
2 RELATED WORKS

Deep Face Recognition. Deep face recognition has achieved encouraging performance due to two main reasons: 1) the publication of large-scale face recognition datasets *e.g.* CASIA-Webface [40], VGGFace2 [1] and MS-Celeb-1M [6] and 2) the improvements of deep learning techniques *e.g.* DeepFace [35], Center Loss [39], SphereFace [20], CosFace [37], ArcFace [3] and CurricularFace [14]. However, these datasets and methods are used to learn discriminative features from full face images. Differently, the proposed dataset and methods focus on a novel masked face recognition (MFR) task, where most of facial cues are occluded by the masks.

Occlusion Robust and Partial Face Recognition. Occlusion robust face recognition (OFR) aims to match face images with random occlusions. Earlier works [15, 28, 32] are designed based on the hand-craft low-level features and the discriminative ability are limited. Based on the deep CNN, a pairwise differential siamese network (PDSN) is proposed in [33] to explicitly find correspondence between occluded facial blocks and corrupted feature elements for deep CNN models. Yin [41] *et al.* propose a spatial activation diversity loss to encourage learning interpretable face representations and a feature activation diversity loss to enhance discrimination and robustness to occlusions. Intuitively, OFR and MFR have similar task, yet the mechanisms of two types of methods are essentially different: In OFR, the occlusions are unknown and uncertain, hence the most of OFR methods aims to learn a robust feature to defense any potential occlusions. On the contrary, the occlusions (*i.e.*, masks) are known in MFR task and MFR methods can benefit from the prior knowledge of masks. Our methods for MFR differ significantly from existing OFR methods in the generative data augmentation and domain constrained ranking approach. We find that these approaches are sufficient to produce state-of-the-art performance without resorting to complicated feature matching and network architecture designs.

Partial face recognition (PFR) aims to match partial face images with full face images. Earlier works [16, 24, 26, 31] rely on the hand-craft local descriptors. Recently, a multi-scale region-based CNN (MR-CNN) model is proposed in [9] to achieve highly compact and discriminatory features. He *et al.* [8] propose a dynamic feature matching (DFM) method with a sliding loss to address partial face images regardless of size. There are two disadvantages when employing these methods on the MFR task directly: Firstly, the PFR methods require the pre-defined partial face images as inputs, which are hard to be detected or semantically defined in the masked faces. Secondly, the PFR methods ignore the global appearance of masked face images such as face contours. Different with these methods, the proposed methods directly take masked face images as input and can benefit from global visual cues.

Image to Image Translation by GAN. Generative Adversarial Net (GAN) has become a powerful technique to perform unsupervised generation of new images and is extremely effective in many tasks, such as facial attribute editing [11, 19] and make up transfer [18]. Several GANs are also proposed to facilitate full face recognition tasks by generating identity-preserved face images. Luan *et al.* [36] propose DR-GAN to frontalize or rotate a face with an arbitrary pose by predicting the identity and pose of generated faces in discriminator. FaceID-GAN [30] is proposed to



(a) The number of face images with different poses in MFSR-REC

(b) The number of identities that have certain amount of masked/full face images in MFSR-REC

(c) The number of faces with different colors of masks in MFSR-SEG and MFSR-REC

Figure 2: Statistics of *MFSR*.

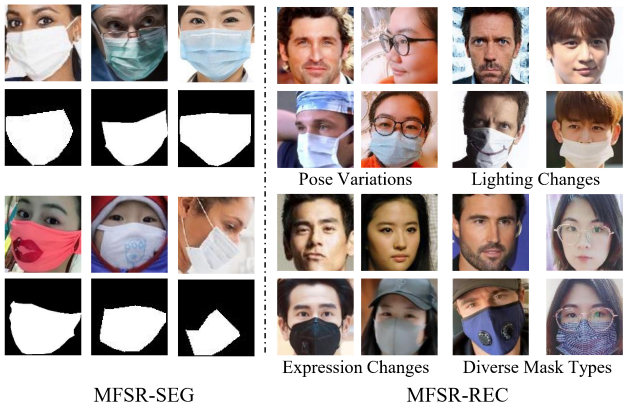


Figure 3: Examples of face images in *MFSR*. Left part shows images and corresponding segmentation annotation in *MFSR-SEG*. Right part shows images in *MFSR-REC* with a variaty of challenging features, each column gives full face and masked face images of the same identity.

generate faces of arbitrary viewpoint while preserve identity using a three-player framework. However, these methods take full faces as inputs, thus are not suitable for generating masked faces. For example, FaceID-GAN uses a facial shape estimator to estimate shape information from generated faces, which cannot be applied to masked faces that loses much facial attributes. In comparison, our proposed IAMGAN adds extra constraints on face identity by using a segmentation guided identity preserve module to ensure the transferred masked face images can be used for MFR model training. Wei *et al.* [38] also proposed PTGAN to generate new samples for person re-identification using segmentation approach, but it does not consider the partial appearance change on foreground as in masked face generation. As far as we know, this is the first work on masked face recognition with GAN based generative data augmentation.

3 *MFSR* DATASET

3.1 Dataset Construction

MFSR consists of two subsets: A facial mask segmentation subset (*MFSR-SEG*) and a masked face recognition benchmark set (*MFSR-REC*).

Facial Mask Segmentation Subset. To construct *MFSR-SEG*, we firstly collect nearly 200K images from image search engines such as Google and Bing using keyword phrases such as "face with mask", "doctors with mask", *etc.*. After that, about 10K face images wearing facial masks such as surgery masks or N95 masks are selected and cropped manually, and other images with no faces, faces with no mask or unusual masks such as respirator mask that cover the whole face are removed. Based on these selected images, the mask regions are manually labeled. The labeled annotation is cross-validated and images with bad label quality are discarded, leaving 9,742 well annotated images in the end. As shown in Figure 3, *MFSR-SEG* includes face images with various poses, age, gender and lighting condition, and these faces wear various types of masks with different textures. These two features make the *MFSR-SEG* potential useful in practical applications.

Masked Face Recognition Benchmark Set. The *MFSR-REC* contains 11,615 images collected from 1,004 identities, in which 704 identities are from real-world and the rest 300 identities are obtained from Internet. Some sample images of *MFSR-REC* are shown Figure 3. For the real-world collected data, we have got the permission of nearly 1,000 subjects and invited each one to take their full face images and masked face images. To increase the data diversity, each subject is asked to take images under 7 face orientations, including left, left-front, front, right-front, right, up and down. For full face images, we encourage each subject to provide photos taken in different time and/or places with the masked ones. After getting the raw images, we manually crop the faces and discard the corrupt images to ensure each identity has at least one full face image and one masked face image. For the Internet collected data, we firstly obtain a name list of over 20K celebrities, and employ the keyword "name with medical mask" for each celebrity to download images from image search engines. After getting 1.6 million raw images, we did a quick selection to filter celebrities with no mask image, which leaves us with nearly 8K celebrities and 0.5 million images. Then the second selection is performed for each celebrity to check whether the masked face is indeed belonging to the celebrity. Finally, 300 celebrities with 5,940 confirmed masked face and full face images are remained. Note that the annotation process for Internet data is extremely time consuming, yet only a small portion (300/20K) of identities have masked face images on the web. Meanwhile, it is also hard to use semi-automatic methods to collect masked face recognition data due to the loss of facial cues. Our annotation experience again proves that gathering large scale labeled data for MFR is very difficult at present.

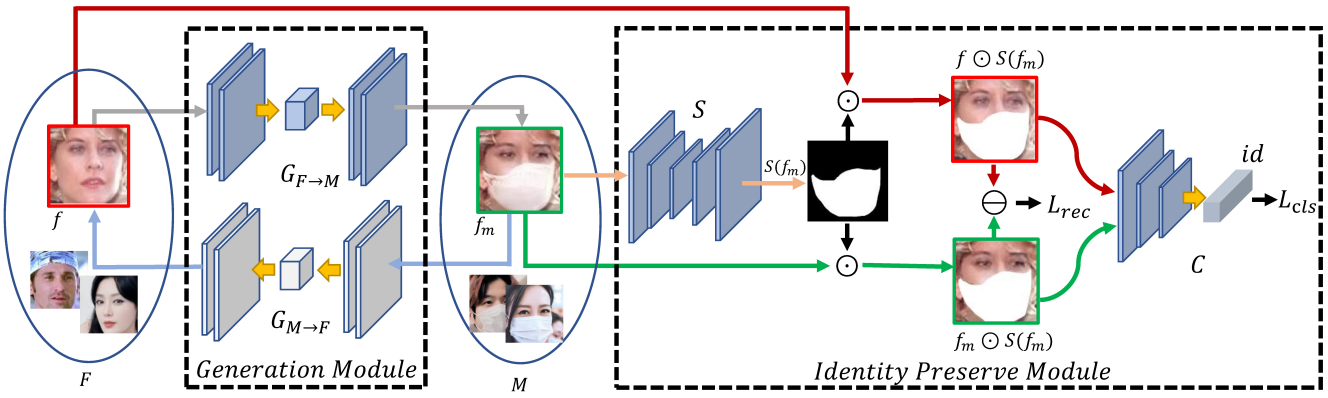


Figure 4: The overall framework of IAMGAN. Given a full face image $f \in F$, $G_{F\to M}$ generates masked face image f_m from f. S predicts mask region $S(f_m)$. L_{cls} classifies the identity of $f \odot S(f_m)$ with classifier C and forces $f_m \odot S(f_m)$ to retain the same identity. L_{rec} is applied to keep the image-specific details at pixel level.

Some statistics of *MFSR* are shown in Figure 2, and the features are be summarized as follows:

1) Sufficient number of identities. *MFSR-REC* contains 1,004 identities with confirmed masked face images, which are collected through complex annotation process. Although there is still not enough ground truth training data, the number of identites is sufficient for benchmarking MFR models.

2) Challenging faces with diverse pose, lighting, expression and mask type changes. As shown in Figure 3, face images in *MASR-SEG* and *MASR-REC* are either collected from Internet or taken by subjects under different backgrounds with different types of masks, bringing a severe change in poses, lighting, expressions and mask types.

3.2 Evaluation Protocol

For model training purpose, we encourage using *MFSR-SEG* as well as existing full face recognition datasets to generate synthesis masked face images, thus we use all 11,615 images in *MFSR-REC* as test data. Note that there is no overlap data between *MFSR-REC* and *MFSR-SEG*. Similar to LFW [13], verification accuracy is applied to evaluate MFR performance. To be specific, 10 sets of image pairs are randomly selected, each contains 5,000 matched and 5,000 unmatched image pairs. Each pair contains one full face image and one masked face image. The accuracy is reported by performing 10 separate experiments in a leave-one-out cross validation scheme.

Despite verification performance, we also test MFR models under masked face retrieval scenario. Since the Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) are broadly used in retrieval tasks such as person re-identification [43], we utilize CMC and mAP as evaluation metrics. When calculating CMC curve and mAP, we use all 3,648 masked face images in *MFSR-REC* as probe images and the rest 7,967 full face images as gallery images.

4 METHODS

We resolve the masked face recognition problem from both data and model aspects. In this section we first describe the details of masked face image generation method using Identity Aware Mask GAN (IAMGAN), and then introduce our proposed Domain Constrained Ranking Loss.

4.1 Identity Aware Mask GAN for Masked Face Generation

Given a face recognition dataset F with full face images and a masked face dataset M (such as *MFSR-SEG*) without any identity information, the goal of IAMGAN is to generate synthetic masked face images from corresponding full face images in F using the mask information of M. The generated face images should preserve identity-specific details except for the masked regions. As illustrated in Figure 4, the proposed IAMGAN consists of two parts: a cyclic masked face generation module and a segmentation guided multi-level identity preserve module.

Masked Face Generation Module. Given a full face image $f \in F$, the masked face generation module aims to generate masked face images from f. Since there are no paired images (*i.e.*, full face and masked face images of the same identity) between F and M, we regard this as an unpaired image-to-image translation task and employ CycleGAN [44] following [21, 23]. Suppose $G_{F\to M}$ and $G_{M\to F}$ are the domain mapping functions from F to M and M to F, respectively. D_F and D_M are two domain discriminators for F and M. The object function of the masked face generation module can be formulated as follows:

$$\begin{aligned}\mathcal{L}_{MFG} = \mathcal{L}_{adv}(G_{F\to M}, D_M) + \mathcal{L}_{adv}(G_{M\to F}, D_F) \\ + \lambda_1 \mathcal{L}_{cyc}(G_{F\to M}, G_{M\to F}),\end{aligned} \tag{1}$$

where $\mathcal{L}_{adv}$ is the standard adversarial loss [5] and $\mathcal{L}_{cyc}$ is the cycle consistency loss [44]. λ_1 is a hyper parameter controlling the importance of $\mathcal{L}_{cyc}$.

Identity Preserve Module. Facial identity information in generated masked face image $f_m = G_{F\to M}(f)$ may become ambiguous with only $\mathcal{L}_{MFG}$ translation. To keep as much identity-related information as possible, three components are introduced in the identity preserve module: 1) a **mask region segmentation network** S to obtain the mask region of f_m, 2) an **identity classifier** C with loss $\mathcal{L}_{cls}$ to reduce the appearance variation of f and f_m *outside the*

masked region at identity level, and 3) a **reconstruction loss** $\mathcal{L}_{rec}$ to further alleviate the detail variation at pixel level. To this end, the objective of the identity preserve module is formulated as:

$$\mathcal{L}_{IP} = \mathcal{L}_{cls} + \mathcal{L}_{rec}. \tag{2}$$

For the mask region segmentation network S, we pretrain a Unet [27] on *MFSR-SEG*. The parameters of S are then fixed and do not propagate gradients during the optimization of $\mathcal{L}_{IP}$. Given a masked face image f_m, S predicts a binary segmentation map $S(f_m)$, where pixel value 0 and 1 represent the mask and non-mask region, respectively.

The identity classifier C aims to classify the out-of-mask region of f to its identity label l_f and force the out-of-mask region of f_m to retain the same identity. The loss function $\mathcal{L}_{cls}$ is defined as

$$\mathcal{L}_{cls} = \mathcal{L}^C_{cls} + \mathcal{L}^G_{cls}, \tag{3}$$

where

$$\mathcal{L}^C_{cls} = -\sum_i [\![l_f = i]\!] \log(\{C(f \odot S(f_m))\}_i), \tag{4}$$

and

$$\mathcal{L}^G_{cls} = -\sum_i [\![l_f = i]\!] \log(\{C(f_m \odot S(f_m))\}_i). \tag{5}$$

are cross-entropy losses, $\{C(\cdot)\}_i$ is the estimated probability of input belonging to identity i and $\odot$ represents the element-wise multiplication. During training, the parameters of C are learned by minimizing $\mathcal{L}^C_{cls}$, while $\mathcal{L}^G_{cls}$ is optimized together with $\mathcal{L}_{MFG}$ to learn the parameters of $G_{F\to M}$.

$\mathcal{L}_{cls}$ forces the generator $G_{F\to M}$ to preserve coarse identity related information such as face contour and shape of eyebrows, *etc.*, while the finer and image specific details such as wrinkles are not preserved. To address this issue, the pixel level reconstruction loss $\mathcal{L}_{rec}$ is formulated as

$$\mathcal{L}_{rec} = \|f \odot S(f_m) - f_m \odot S(f_m)\|_1 . \tag{6}$$

Overall Objective Function. The overall objective function of IAMGAN is defined as

$$\mathcal{L}_{IAMGAN} = \mathcal{L}_{MFG} + \lambda_2 \mathcal{L}_{IP}, \tag{7}$$

where λ_2 is used to balance the loss weight.

Figure 5 gives some samples generated by CycleGAN and the proposed IAMGAN using CAISA-Webface dataset as F and *MFSR-SEG* as M. Note that we crop the original image in F using a fixed bounding box to avoid background interference. Compared with CycleGAN, faces generated by IAMGAN preserve more identity related details from input images due to the usage of identity preserve module. It can be seen from figure 5 that IAMGAN also works well with large head poses. More detailed evaluation of IAMGAN will be given in the experiment section.

4.2 Domain Constrained Ranking for Masked Face Recognition

Once we have the generated masked face data and corresponding full face data with associated identity labels, the MFR models can be learned in a supervised manner. To address the large intra-class variance issue, we define two sets of domain specific identity centers C_f and C_m, in which $C^{\ell_i}_f$ and $C^{\ell_i}_m$ stands for the centers of full face and masked face images for identity ℓ_i, respectively. Let $\phi(I)$ be the deep feature representation of input image I, the proposed DCR contains a single domain ranking loss and a cross domain ranking loss.

Figure 5: Comparision of masked face images generated by CycleGAN and IAMGAN. Results are shown in dashed bounding boxes. The third row gives the segmentation map calculated by UNet S.

Single Domain Ranking Loss. Single domain ranking loss aims to distinguish identities within the same domain. For image $I^{\ell_i}_x$ with domain label $x \in \{f, m\}$ and identity label ℓ_i, single domain intra-class distance is defined as $D_s(C^{\ell_i}_x, I^{\ell_i}_x) = \left\| C^{\ell_i}_x - \phi(I^{\ell_i}_x) \right\|_2$. Similarly, for each center $C^{\ell_i}_x$, the inter-class distance is defined as $D_s(C^{\ell_i}_x, I^{\ell_j}_x) = \left\| C^{\ell_i}_x - \phi(I^{\ell_j}_x) \right\|_2$, where $I^{\ell_j}_x$ is an image from the same domain x but with different identity label ℓ_j. For a minibatch B, hard-negative mining strategy [12] is adopted to obtain the smallest inter-class distance $D_s(C^{\ell_i}_x, I^*_x) = \min_{I^{\ell_j}_x \in B, i\neq j} \left\| C^{\ell_i}_x - \phi(I^{\ell_j}_x) \right\|_2$.

With the intra and inter class distance, the single domain ranking loss is defined as

$$\mathcal{L}_s(I^{\ell_i}_x) = \max\{D_s(C^{\ell_i}_x, I^{\ell_i}_x) + \alpha_s - D_s(C^{\ell_i}_x, I^*_x), 0\}, \tag{8}$$

in which $\alpha_s > 0$ is a distance margin.

Cross Domain Ranking Loss. Cross domain ranking loss aims at matching images from one domain to the other. Given an image $I^{\ell_i}_x$, the cross-domain intra-class distance D_c is calculated by $D_c(I^{\ell_i}_x, I^{\ell_i}_y) = \left\| \phi(I^{\ell_i}_x) - \phi(I^{\ell_i}_y) \right\|_2$, with $y \in \{f, m\}$ and $x \neq y$. Similar to $D_s(C^{\ell_i}_x, I^*_x)$, the largest distance $D_c(I^{\ell_i}_x, I^{\ell_i *}_y)$ across a minibatch B is used, *i.e.* $D_c(I^{\ell_i}_x, I^{\ell_i *}_y) = \max_{I^{\ell_i}_y \in B} \left\| \phi(I^{\ell_i}_x) - \phi(I^{\ell_i}_y) \right\|_2$.

To better distinguish different identities across two domains, our idea is to let the largest cross domain intra-class distance less than the smallest single domain inter-class distance within a minibatch. Hence the inter-class distance is defined as:

$$D_c(I^{\ell_i}_x, C^*_x) = \min_{I^{\ell_j}_x \in B, i\neq j} \left\| \phi(I^{\ell_i}_x) - C^{\ell_j}_x \right\|_2 . \tag{9}$$

Note that $I^{\ell_i}_x$ and C^*_x have the same domain label. The cross domain ranking loss is then defined as

$$\mathcal{L}_c(I^{\ell_i}_x) = \max\{D_c(I^{\ell_i}_x, I^{\ell_i *}_y) + \alpha_c - D_c(I^{\ell_i}_x, C^*_x), 0\}, \tag{10}$$

Table 1: Performance of state-of-the-art methods on *MFSR-REC*.

Methods	CAISA-Webface					VGGFace2				
	Acc	Rank-1	Rank-5	Rank-10	mAP	Acc	Rank-1	Rank-5	Rank-10	mAP
Softmax	78.3	37.2	54.3	61.2	21.9	82.2	40.3	57.9	64.3	24.9
Softmax+Triplet[12]	79.0	40.8	57.7	64.0	22.8	82.9	43.4	61.9	67.7	26.4
Center Loss[39]	79.2	40.5	57.3	64.0	22.9	83.2	43.8	62.3	68.5	26.9
SphereFace[20]	78.4	42.9	58.5	64.2	22.8	83.3	44.2	62.2	68.9	27.1
CosFace[37]	79.1	44.0	59.8	65.7	23.8	83.6	45.4	62.3	68.2	28.4
ArcFace[3]	79.4	43.2	59.3	64.8	23.3	83.8	43.6	61.8	68.1	27.8
Interpret-FR [41]	78.9	42.3	59.0	65.0	22.5	82.6	44.5	61.6	68.0	26.8
DFM [8]	79.0	43.3	58.1	64.3	23.1	83.3	45.7	62.4	68.4	27.4
Softmax+IAMGAN	80.3	60.4	73.0	77.2	33.8	84.0	61.7	75.2	78.7	40.7
Softmax+Triplet+IAMGAN	81.1	64.5	75.2	78.6	36.2	86.1	65.6	77.1	80.3	41.1
Center Loss+IAMGAN	81.7	64.0	75.1	79.0	36.6	85.6	65.3	76.9	80.2	41.6
SphereFace+IAMGAN	80.8	60.2	70.9	74.5	31.7	84.9	59.9	73.1	77.8	37.8
CosFace+IAMGAN	81.6	59.2	70.9	74.9	32.0	85.9	61.0	73.6	77.5	39.8
ArcFace+IAMGAN	81.7	59.3	70.7	74.9	31.9	86.1	60.0	73.1	77.0	39.6
DCR + IAMGAN	**82.3**	**67.4**	**76.5**	**79.6**	**37.5**	**86.5**	**68.1**	**77.4**	**80.6**	**42.7**

where $0 < \alpha_c < \alpha_s$ considering the domain variance between I_x and I_y.

With the above losses, given training image $I_x^{\ell_i}$, the domain constrained ranking loss is defined as

$$\mathcal{L}_{DCR}(I_x^{\ell_i}) = \mathcal{L}_s(I_x^{\ell_i}) + \gamma \mathcal{L}_c(I_x^{\ell_i}), \tag{11}$$

with $0 < \gamma < 1$ balancing the importance of single domain and cross domain losses. Similar to Center Loss, we also use joint supervision of $\mathcal{L}_{DCR}$ and softmax loss to avoid the degrad of centers.

Updating of Identity Centers. During training, the identity centers are updated by a minibatch momentum manner:

$$\{C_x^{\ell_i}\}^{t+1} = \beta\{C_x^{\ell_i}\}^t + (1-\beta)\frac{1}{\left\|\mathcal{B}_x^{\ell_i}\right\|}\sum_{I\in\mathcal{B}_x^{\ell_i}}\phi(I), \tag{12}$$

where β is the momentum factor and $\mathcal{B}_x^{\ell_i}$ is the set of samples with domain label x and identity label ℓ_i in minibatch $\mathcal{B}$.

5 EXPERIMENTS

5.1 Datasets and Settings

IAMGAN is trained using *MFSR-SEG* as M and large-scale face recognition datasets as F. For the latter, two publicly available datasets are used: 1) CAISA-WebFace [40] which consists of 494,414 images from 10,575 subjects and 2) VGGFace2 [1] with over 3.31 million images from 9,131 subjects. Full faces in F are aligned with MTCNN following [3] and masked faces in M are not aligned. We conduct evaluation of different methods on *MFSR-REC*.

5.2 Implementation Details

For IAMGAN, the generation module uses similar network architecture with the one in Cycle-GAN [44]. S employs UNet [27] architecture and C employs ResNet-50 [7]. Input images are resized to 256×256 and the batch size is set to 1. Adam optimizer [17] is applied with decay rates 0.5 and 0.999. The model is trained for 500K steps, the initial learning rate is 0.0002 and is decreased linearly to 0 after 350K steps. We set $\lambda_1 = 10$. λ_2 is set to 0 at the beginning and linearly increased to 1.0 after 100K steps.

For training MFR models, ResNet50 [7] is used as feature extractor. Each minibatch contains 32 identities, with 6 randomly selected images for each identity. When using CAISA-Webface as training data, all models are trained with Adam optimizer for 60K steps. The initial learning rate is 0.0001 and is multiplied by 0.1 at 30K and 45K steps. For training with VGGFace2, all models are trained with Adam optimizer for 200K steps. The initial learning rate is 0.0001 and is multiplied by 0.1 at 100K and 150K steps. For DCR Loss, we set $\alpha_c = 0.5$ and $\alpha_s = 0.7$. γ is set to 0.75 and β is fixed to 0.5.

5.3 Comparisons with State-of-the-Art Methods

We conduct a comprehensive benchmark of our approaches and state-of-the-art face recognition methods on *MFSR-REC*, including 6 standard face recognition methods, one occlusion robust face recognition method and one partial face recognition method. For standard face recognition methods, we evaluate the performance of Softmax loss, the combination of Softmax and advanced Triplet loss [12] , Center loss [39], Sphereface [20], CosFace [37] and Arcface [3]. All the methods are trained with and without IAMGAN generated masked face data for comparision. For occlusion robust and partial face recognition methods, we test the state-of-the-art Interpret-FR [41] and DFM [8], respectively. Finally, we evaluate our proposed DCR loss trained with IAMGAN generated data. From the results in Table 1 we have the following observations:

1) Standard face recognition models trained with full face data cannot be directly applied to MFR. These models focus on learning

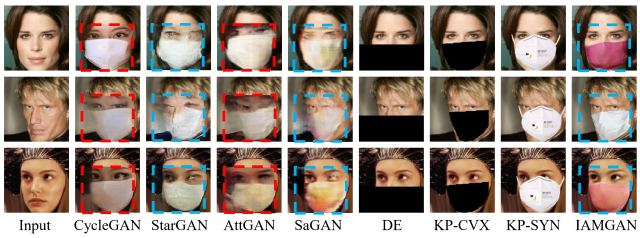


Figure 6: Comparision of masked face images generated by IAMGAN and other methods. IAMGAN is able to generate masks as well as preserve identity-related facial details.

discriminative features on full faces, thus are inevitably affected by the missing of facial cues and the disturbance of mask regions. For example, the model trained with softmax loss on CAISA-WebFace only achieves verification accuracy 78.3%, mAP 21.9% and Rank-1 accuracy 37.2%. The performance is better when advanced losses are used, however, the best result is still substantially lower than the reported performance on full face datasets, *e.g.* accuracy 99.83% on LFW achieved by ArcFace [3].

2) Occlusion robust and partial face recognition methods show little improvement over standard full face recognition methods on the MFR task. Although occlusion robust methods can beat standard face recognition methods when facing arbitary potential occlusions, the advantage is not reflected when the occlusion type is known and fixed. Partial face recognition methods use partial face patches as input and ignore global visual appearance of faces such as face contours, as a result, the performance of PFR methods are still limited on the MFR task.

3) Models trained with IAMGAN generated masked face data achieve significant performance boost. As shown in Table 1, all 6 standard face models trained with generated masked face data perform better than corresponding full face trained ones, *e.g.*, Rank-1 accuracy of Softmax is improved from 37.2% to 60.4% on CAISA-Webface. The performance improvements clearly indicate that the generated masked face images by IAMGAN can effectively bridge the domain gap between training and testing data, thus is beneficial to the learning of MFR models.

4) The proposed DCR loss outperforms all other losses. As shown in Table 1, the model trained on CAISA-WebFace with DCR loss is able to achieve accuracy 82.3%, mAP 37.5% and Rank-1 accuracy 67.4%, when trained on VGGFace2, it achieves accuracy 86.5%, mAP 42.7% and Rank-1 accuracy 68.1%, which is superior to other state-of-the-art losses. It is because DCR loss not only keeps the features of different classes separable by minimizing single-domain intra-class distance, but also explicitly minimizes the cross-domain intra-class difference between full faces and masked faces.

5.4 Ablation Study

The Effectiveness of IAMGAN. Using CAISA-Webface and *MFSR-SEG* as training data, we compare IAMGAN with two domain translation methods (**CycleGAN** [44] and **StarGAN** [2]), two facial attribute editing methods (**AttGAN** [10] and **SaGAN** [42]) and the following three ad-hoc methods:

Table 2: Performance comparision of different masked face image generation methods on *MFSR-REC*.

Methods	Acc	Rank-1	mAP
CycleGAN [44]	79.6	50.9	26.9
StarGAN [2]	79.3	48.8	25.5
AttGAN [10]	79.5	48.1	24.7
SaGAN [42]	79.2	51.4	27.3
DE	79.0	57.6	31.1
KP-CVX	79.8	58.1	31.6
KP-SYN	79.5	57.7	31.5
IAMGAN	**80.3**	**60.4**	**33.8**

1) **Directly Erasing (DE).** In this method, the lower 40% of the full face image is directly erased by setting pixel value to 0, leaving only the upper part of the image.

2) **Facial Key-point Convex Hull Erasing (KP-CVX).** In this method, the facial key-points of full faces are first detected, after locating the key-points in the lower face region, the convex hull formed by these key-points are erased.

3) **Facial Key-point based synthetic mask (KP-SYN).** In this method, several mask images (*e.g.* N95 mask) are prepared. Given a input full face image, the facial key-points are first detected to estimate the size and orientation of lower part face region. Then, the prepared mask images are resized, rotated and patched directly on the input image to synthesize masked face.

The results of the above mask generation methods are shown in Figure 6. It can be observed that, IAMGAN clearly generates the most natural-looking masks and is able to preserve the face identity compared with AttGAN, CycleGAN and StarGAN. SaGAN treats regions with non-zeros attention values as attribute-specific region. However, due to the domain divergence and lacking of paired training data, attention values are not accurately calculated. Images generated by DE and KP-CVX are able to preserve identity information, yet they lack proper mask textures. KP-SYN can also preserve identity and generate real masks, however, the texture of mask is simple and unnatural. Moreover, facial key-point based methods like KP-CVX and KP-SYN are limited by the robustness of facial key-point detectors, *e.g.* sometimes eye region can be occluded due to key-point detection errors.

To quantitatively evaluate the above methods, we generate masked face images with them on CAISA-Webface, then several MFR models (with Softmax loss) are trained on the generated data and evaluated on *MFSR-REC*. As shown in Table 2, the model trained with IAMGAN generated data outperforms all other models by a clear margin. The performance of other GAN based methods are lower than ad-hoc methods because of the missing of facial identity information, while IAMGAN can preserve facial identity and generate input-specific mask textures, thus outperforms ad-hoc methods. As shown in Figure 6 and Table 2, two aspects are important for the generation of masked face data: 1) the preserve of facial identity information and 2) the diverse appearance of generated masks.

Impact of Each Modules in IAMGAN. To analyze the function of different modules in IAMGAN, we train three variants of IAMGAN by removing $\mathcal{L}_{IP}$, $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$, respectively. Figure 7

Table 3: Performance comparision of different variants of IAMGAN on *MFSR-REC*.

Methods	Acc	Rank-1	mAP
w/o $\mathcal{L}_{IP}$	79.6	50.9	26.9
w/o $\mathcal{L}_{rec}$	80.0	58.5	32.1
w/o $\mathcal{L}_{cls}$	80.1	58.3	32.3
with all	**80.3**	**60.4**	**33.8**

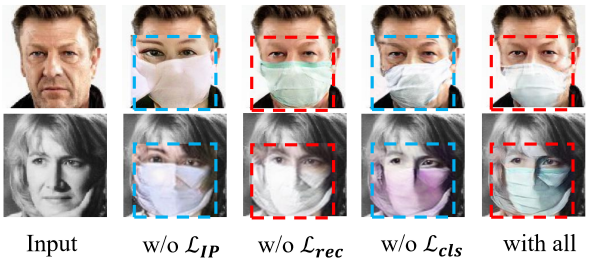


Figure 7: Different variants of the IAMGAN without certain loss functions.

Table 4: Performance of DCR without certain loss terms on *MFSR-REC*.

Methods	Acc	Rank-1	mAP
w/o $\mathcal{L}_c$	81.5	64.7	36.3
w/o $\mathcal{L}_s$	81.9	66.9	37.0
with all	**82.3**	**67.4**	**37.5**

shows a comparison of these variants and IAMGAN trained with all the loss functions. Without the proposed facial identity preserve losses, the model is not able to preserve the identity. Without the pixel level reconstruction loss $\mathcal{L}_{rec}$, the model tends to generate images that miss fine details such as wrinkles and skin colors. The model without $\mathcal{L}_{cls}$ tends to overemphasize the facial details by generating longer eyebrows and sharper textures. By jointly learning $\mathcal{L}_{rec}$ and $\mathcal{L}_{cls}$, IAMGAN is able to preserve facial identity-related information well. To quantitively prove this, the three variants are then used to generate masked face images on CAISA-Webface to train MFR models with Softmax loss. As shown in Table 3, IAMGAN trained with all losses outperforms other variants.

The Impact of Different Loss Terms in DCR. We verify the influence of different terms in DCR loss by training variant MFR models on CAISA-Webface+IAMGAN without $\mathcal{L}_s$ or $\mathcal{L}_c$ and testing on *MFSR-REC*. As shown in Table. 4, $\mathcal{L}_s$ and $\mathcal{L}_c$ are complementary to each other. The former models inter-class distance and intra-class distance from homogeneous faces (full-full and masked-masked faces), and the latter models them from heterogeneous faces (full-masked faces).

Visualisation. To gain a better understanding of MFR models learned with IAMGAN generated masked face data and DCR loss, we show some sample retrieval results of different MFR models and

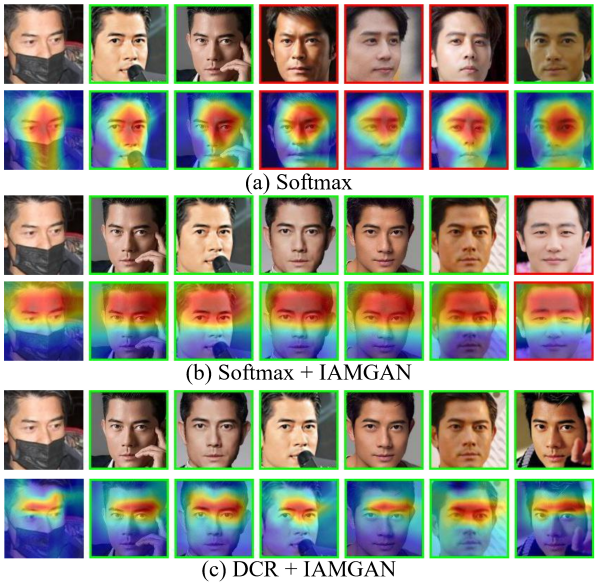


Figure 8: Visualisation of different MFR models: (a) The model trained with only Softmax loss on full faces. (b) The model trained with Softmax loss on both full faces and generated masked faces. (c) The model trained with the proposed DCR loss on full faces and generated masked faces. The left most image in each row is the probe image and top-6 retrival results are shown on the right, with true positives and false positives in green and red boxes, respectively.

visualize the activate part of feature maps. As shown in Figure 8, the model trained with only full face data inevitably pays attention on mask region and lower part of faces. Models trained with full and masked faces together focus on the top half of the whole image. Compared with Softmax loss, the model trained with DCR loss focuses more on identity-related region *e.g.* eyes and brows. The visualization results clearly show that IAMGAN and DCR loss can increase the discriminative ability of MFR models.

6 CONCLUSION

This paper contributes a Masked Face Segmentation and Recognition (*MFSR*) dataset for developing and benchmarking masked face recognition models. *MFSR* is a challenging dataset as it presents huge variants on poses, lightings, expressions and mask types. In addition, IAMGAN and DCR loss are proposed in this paper to tackle the data shortage problem and enhance the discriminative power of MFR models. Extensive experiments on *MFSR* have demonstrated the effectiveness of the proposed approaches.

ACKNOWLEDGMENTS

This work is partially supported by grants from the National Key R&D Program of China under grant 2017YFB1002400, the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010153002, and the National Natural Science Foundation of China under contract No.61825101, No.61702515 and No.U1611461.

REFERENCES

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *IEEE International Conference on Automatic Face & Gesture Recognition*. 67–74.

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4690–4699.

[4] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. 2017. Detecting masked faces in the wild with lle-cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2682–2690.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. 2672–2680.

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision*. 87–102.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[8] Lingxiao He, Haiqing Li, Qi Zhang, and Zhenan Sun. 2018. Dynamic feature learning for partial face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 7054–7063.

[9] Lingxiao He, Haiqing Li, Qi Zhang, Zhenan Sun, and Zhaofeng He. 2016. Multi-scale representation for partial face recognition under near infrared illumination. In *IEEE International Conference on Biometrics Theory, Applications and Systems*. 1–7.

[10] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2017. Arbitrary facial attribute editing: Only change what you want. *arXiv preprint arXiv:1711.10678* (2017).

[11] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. 2019. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing* (2019), 5464–5478.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[13] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 2007. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Technical Report*.

[14] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. 2020. CurricularFace: Adaptive Curriculum Learning Loss for Deep Face Recognition. *arXiv preprint arXiv:2004.00288* (2020).

[15] Hyun Jun Oh, Kyoung Mu Lee, and Sang Uk Lee. 2008. Occlusion invariant face recognition using selective local non-negative matrix factorization basis images. *Image Vision Computing* (2008), 1515–1523.

[16] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2013. Robust partial face recognition using instance-to-class distance. In *Visual Communications and Image Processing*.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. 2018. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *Proceedings of the ACM International Conference on Multimedia*. 645–653.

[19] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. 2019. STGAN: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3673–3682.

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 212–220.

[21] Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. 2018. Conditional cyclegan for attribute guided face image generation. (2018).

[22] Iacopo Masi, Anh Tuan Tran, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. 2016. Do we really need to collect millions of faces for effective face recognition?. In *Proceedings of the European Conference on Computer Vision*. 579–596.

[23] Sveinn Palsson, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. 2018. Generative adversarial style transfer networks for face aging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2084–2092.

[24] Pan Ke, ShengCai Liao, Zhijian Zhang, Stan Z. Li, and Peiren Zhang. 2007. Part-based face recognition using near infrared images. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[25] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).

[26] Renliang Weng, Jiwen Lu, and Yap-Peng Tan. 2016. Robust point set matching for partial face recognition. *IEEE Transactions on Image Processing* (2016), 1163–1176.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*. 234–241.

[28] Rui Min, Abdenour Hadid, and Jean-Luc Dugelay. 2011. Improving the recognition of faces occluded by facial accessories. In *Face and Gesture*. 442–447.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 815–823.

[30] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. 2018. Faceidgan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 821–830.

[31] Shengcai Liao, Anil K Jain, and Stan Z Li. 2013. Partial face recognition: alignment-free approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013), 1193–1205.

[32] Sohee Park, Hansung Lee, Jang Hee Yoo, Geonwoo Kim, and Soonja Kim. 2015. Partially occluded facial image retrieval based on a similarity measurement. *Mathematical Problems in Engineering* (2015), 1–11.

[33] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. 2019. Occlusion Robust Face Recognition Based on Mask Learning With Pairwise Differential Siamese Network. In *Proceedings of the IEEE International Conference on Computer Vision*. 773–782.

[34] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*. 1988–1996.

[35] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1701–1708.

[36] Luan Tran, Xi Yin, and Xiaoming Liu. 2017. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1415–1424.

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5265–5274.

[38] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 79–88.

[39] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*. 499–515.

[40] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).

[41] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. 2019. Towards interpretable face recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. 9348–9357.

[42] Gang Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. 2018. Generative adversarial network with spatial attention for face attribute editing. In *Proceedings of the European Conference on Computer Vision*. 417–432.

[43] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*. 1116–1124.

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 2223–2232.