Toward High-quality Face-Mask Occluded Restoration

LU FEIHONG, Xidian University, PRC
CHEN HANG, Xidian University, PRC
LI KANG, Xidian University, PRC
DENG QILIANG, Xidian University, PRC
ZHAO JIAN, Institute of North Electronic Equipment, PRC
ZHANG KAIPENG, The University of Tokyo, PRC
HAN HONG*, Xidian University, PRC

Face-mask occluded restoration aims to restore the masked region of a human face, which has attracted increasing attention in the context of the COVID-19 pandemic. One major challenge of this task is the large visual variance of masks in the real world. To solve it we first construct a large-scale Face-mask Occluded Restoration (FMOR) dataset, which contains 5,500 unmasked images and 5,500 face-mask occluded images with various illuminations, and involves 1,100 subjects of different races, face orientations and mask types. Moreover we propose a Face-Mask Occluded Detection and Restoration (FMODR) framework, which can detect face-mask regions with large visual variations and restore them to realistic human faces. In particular, our FMODR contains a self-adaptive contextual attention module specifically designed for this task, which is able to exploit the contextual information and correlations of adjacent pixels for achieving high realism of the restored faces, which are however often neglected in existing contextual attention models. Our framework achieves state-of-the-art results of face restoration on three datasets, including CelebA, AR and our FMOR datasets. Moreover, experimental results on AR and FMOR datasets demonstrate that our framework can significantly improve masked face recognition and verification performance.

CCS Concepts: • **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**; • **Reconstruction**;

Additional Key Words and Phrases: Face-mask Occluded Dataset, Face Restoration, Self-adaptive Contextual Attention, Masked Face Recognition and Verification

1 INTRODUCTION

Face restoration, aiming at filling the missing regions of a human face, benefits many face analysis applications, such as occluded face detection [6] and occluded face recognition [3]. Since the catastrophic outbreak of COVID-19 pandemic, more and more people have to wear face-masks every day when outing. Masks of various types would result in loss of most features of a face, which greatly reduces the accuracy of face analysis tasks. See Fig. 1 (b) for an illustration. So, faces appearing under occlusion is a major hindrance for accurate face analysis tasks,

*Corresponding author.

Authors' addresses: Lu Feihong, Xidian University, No.2,South Taibai Road, Xi'an, Shaan Xi, PRC, 710071; Chen Hang, Xidian University, No.2,South Taibai Road, Xi'an, Shaan Xi, PRC, 710071; Li Kang, Xidian University, No.2,South Taibai Road, Xi'an, Shaan Xi, PRC; Deng Qiliang, Xidian University, No.2,South Taibai Road, Xi'an, Shaan Xi, PRC; Zhao jian, Institute of North Electronic Equipment, 226 North Fourth Ring Road middle, Beijing, PRC; Zhang Kaipeng, The University of Tokyo, No.1, 3-fan, 7-d-mu, moto, Tokyo, PRC; Han Hong*, Xidian University, No.2,South Taibai Road, Shaan Xi, Texas, PRC, 710071, hanh@mail.xidian.edu.cn.

1551-6857/2022/3-ART $15.00
https://doi.org/10.1145/3524137

which has been far from being solved. Therefore, it is a practical and feasible method to use the face restoration method for masked face recognition as the pre-processing of occluded face analysis tasks without losing useful face information. For example, Baidu, SenseTime, face++ and many other APP for are difficult to process masked face tasks.

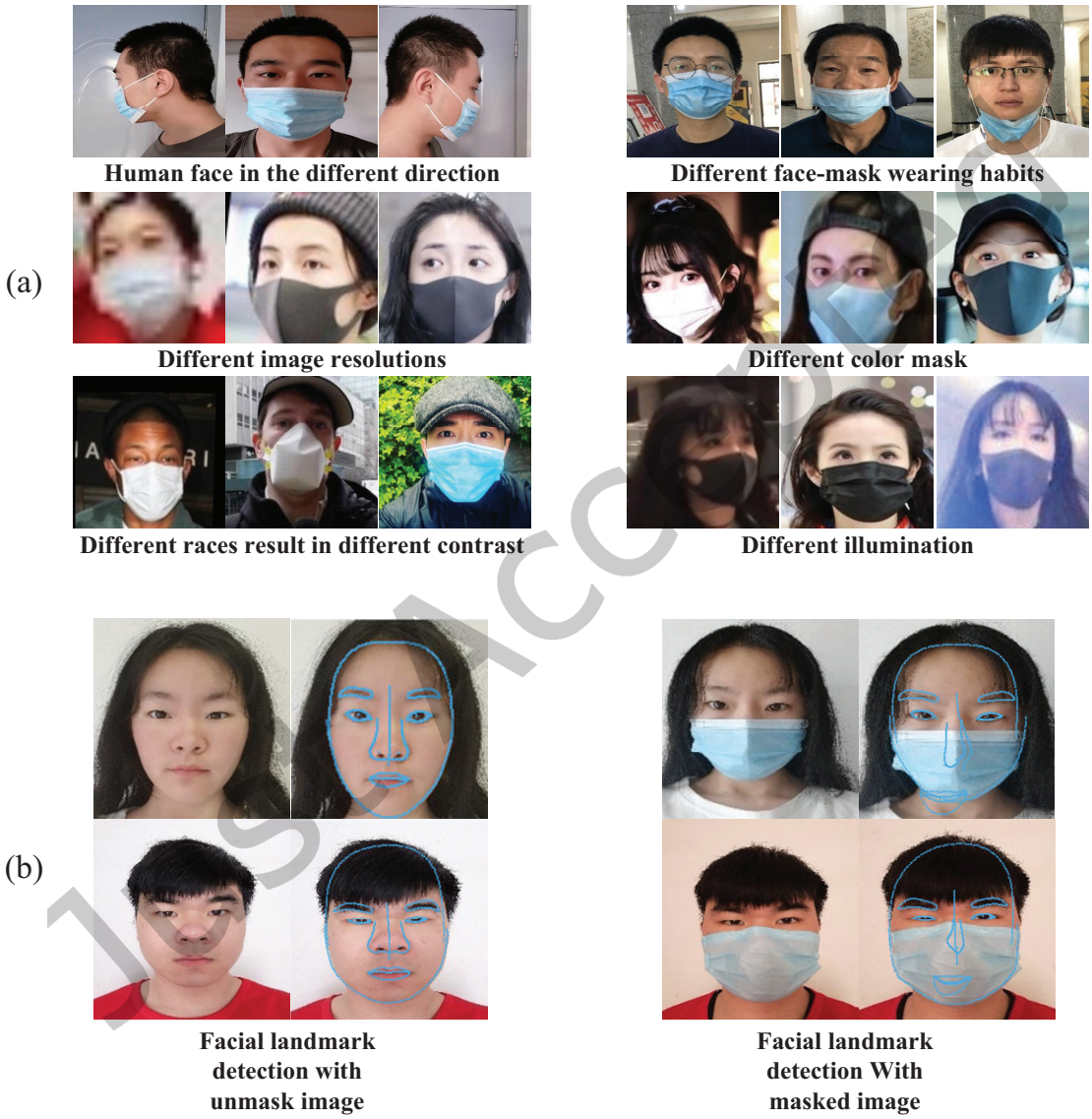


Fig. 1. (a) Face-mask restoration challenges. (b) Results of facial landmark detection with or without face-mask occlusion. Face-mask occlusion will seriously affect accuracy of facial landmark detection.

Previous face restoration methods have achieved remarkable success with advanced deep learning-based architectures [7, 28, 41, 45, 53–55] or loss functions [30] in recent years. However, they tend to perform unsatisfactorily on the faces with challenging poses, bad illumination and occlusion, and many of them can only deal with regular occlusion problem. Some methods [3, 8] apply Generative Adversarial Networks (GAN) for tackling occluded face restoration. However, these GAN-based methods bring artifacts in boundaries, distorted structures, or inconsistently blurred textures with surrounding areas. Some other methods [23, 34, 40, 47, 48, 56] try to generate more realistic images by adopting non-local schemes to exploit contextual relations of no missing areas to fill in missing pixels for a face. However, they can only deal with occluded images in human-controlled environments. In this paper, we tackle in-the-wild face-mask occluded restoration. See Fig. 1 (a) for some examples, which are challenging due to the large visual variances, such as various face-masks types, bad illumination, and face poses.

Beyond methodology, existing mask occlusion face datasets are so rare. Each dataset contains a small number of mask occlusion face images, which is a tremendous challenge for masked face analysis. To our best knowledge, RMFRD [42] is the publicly available face-mask occluded dataset, containing 2,203 masked images of 525 people and 90,000 images of the same subjects without masks. The images in RMFRD are clipped and aligned. Besides, the background (indoor scene) is very monotonous, and all the people are Asian. Hence, its data diversity is limited, while current deep learning-based models require more data and greater variety. The models trained on RMFRD are less likely to show satisfactory results in real-world scenarios.

We propose a new face-mask dataset to solve the problem of insufficient data quantity and diversity of existing mask occlusion datasets. We address in-the-wild face-mask occluded restoration by presenting a new dataset, called Face-Mask Occluded Restoration (FMOR) dataset. FMOR consists of 1,100 subjects, including 5,500 free images and 5,500 face-mask occluded images. It offers great data diversity in age, race, pose, mask type, and illumination. Each subject in FMOR contains at least one mask-wearing face image and a corresponding face image without the mask. We manually annotate the masked region for each face-mask occluded image by LabelMe [35]. Compared with the RMFRD dataset, our FMOR contains samples in more diverse scenes, closer to the real scenarios. FMOR can be used for all tasks related to the occlusion of the human face, so we believe this large-scale and challenging FMOR dataset will significantly contribute to the face analysis community.

To solve the problem that the existing methods are difficult to restore the mask occluded face image accurately. We further develop a Face-Mask Occluded Detection and Restoration (FMODR) framework for detecting and restoring the face-mask regions. It mainly consists of three modules. First, a face-mask detector is used to detect the region of a face mask, which uses very few convolution layers. After that, a U-net architecture is employed to generate a coarse output from the face-mask occlusion image input. This coarse restoration module can restore the face image occluded by the face mask, accelerating the refine network's convergence speed. Finally, a refine network is applied to refine the coarse result, which takes the coarse output and the region of a face mask as inputs. Particularly, we incorporate a novel self-adaptive contextual attention module and global-masked region discriminators in the refine network architecture. The self-adaptive contextual attention model exploits the features of the unmasked patches to refine those of the masked region in a convolution fashion. Better restoration results can be attained with guidance from the correlations of adjacent pixels in the occluded region and the perceptual loss of the generated image. The global-masked region discriminators use adversarial learning to make the restored images more realistic.

Our contributions are summarized as follows:

1) We propose a large-scale face-mask occluded restoration dataset called FMOR and will be released soon. FMOR contains 11,000 images with various visual variances. To our best knowledge, it is the largest dataset for addressing face-mask occluded restoration.
2) We propose a Face-mask Occluded Detection and Restoration Approach (FMODR). It is a two-stage GAN-like image restoration framework, which contains a specially designed self-adaptive contextual attention model

to exploit the contextual information from non-occluded regions' features to enhance those of the occluded region. Besides, we also devise a global-masked region discriminator to generate more realistic results.

3) Quantitative and qualitative comparisons demonstrate that our proposed approach outperforms state-of-the-art methods on CelebA [32] and AR-dataset [16]. Besides, FMODR is proved to generate higher-quality restoration results than existing methods on FMOR, even for boosting the performance of masked face recognition and verification on AR and FMOR datasets.

2 RELATED WORK

2.1 Datasets for mask occluded face restoration

Existing face restoration datasets include AR [16] (4,000 face images of 126 subjects), CelebA [32] (202,599 face images of 10,177 subjects), CAS-PEAL [13] (9,594 face images of 1,040 subjects. 1,299 face images with real occlusion), CMU-PIE [17] (41,368 face images of 68 subjects with rectangles occlusions), NIR-Distance [20] (4,300 face images of 276 subjects. Some of them are occluded), and MAFA [14] (35806 face images of 30811 subjects). Early datasets like AR [16], FaceDB [39] are collected from the controlled environments; Recent datasets like CelebA [32] collect images from the Internet and serve as challenging benchmarks to evaluate face restoration performance in the wild. To the best of our knowledge, MAFA is the largest face occlusion dataset. MAFA consists of 30,811 images, each containing at least one occluded face. There are 35,806 occluded faces in MAFA. However, it is built for face detection without identity annotations and thus not applicable to occluded face recognition and restoration tasks.

RMFRD [42] is the publicly available face-mask occluded dataset collected from the real world, with 2,203 masked images and 90,000 unmask images of 525 people. As all images are clipped, it is hardly applicable to face-mask occluded restoration. Besides, all images in RMFRD are frontal images of Asian people. Thus its data diversity is insufficient. Webface260M [58] is the largest face recognition dataset in the world recently, which consists of real-world masked faces in addition to providing features to generate synthetic masks on real-world non-masked faces. This dataset focuses on face detection and recognition, even though there are some occluded face images, and the types of occlusion are complex (not only does it include face-mask occlusion, but also various occlusion such as a scarf, hand, and so on), which is challenging to be used in real-world face-mask restoration. We propose a large-scale face-mask occluded dataset for the face restoration research community.

2.2 Image restoration

Existing image restoration methods can be divided into hand-crafted and deep-learning-based methods. The first category copies patches similar to the masked area from the unmasked area to fill in missing pixels [11, 12, 19, 26]. They are generally effective for restoring small or narrow holes but tend to fail when the holes are large, or the texture changes fast. Besides, they often produce noticeable visual artifacts. The second category applies a deep-learning-based encoder-decoder structure to predict each pixel of the missing area. For example, Pathak *et al.* [34] present an unsupervised visual feature learning algorithm driven by context-based pixel prediction, where the context encoder learns representations capturing not only appearance information but also the semantics of visual structures, but it heavily relies on post-processing to reduce artifacts. [8] uses the U-net network to restore missing images; however, it only uses a generative model, which generated very fuzzy images. Yu [50] et al. proposes a Region Normalization method, which divided the spatial pixels into different regions according to the input mask, and calculated the mean and variance of each region for normalization, thereby overcoming the limitations of Feature Normalization on the training of the image restoration network. Although this method has achieved good results for the natural occlusion restoration problem, for highly structured face images, the restoration results are still somewhat blurred. In addition, some methods of text image generation [18, 44, 49] can also be used for image restoration.

2.3 Face restoration

Face restoration methods holes by propagating neighborhood appearance [1, 2], and often producmainly include traditional patch-based [5, 9, 31] and deep learning methods [47, 48]. Traditional methods fille noticeable visual artifacts which are not desired in face restoration. These methods use pixels around the undesired or non-missing region to restore those in the missing area, costing high computation and memory usage. PatchMatch [4] uses a fast nearest-neighbor field algorithm to significantly accelerate calculation and produce high-quality restoration results. However, it suffers severe artifacts in case of complex and diverse facial structures.

Recently, face restoration based on Variational AutoEncoder (VAE) [52] and Generative Adversarial Network (GAN) [8, 15, 51] has achieved remarkable results. Paper [29] applied a generative adversarial network to perform face completion, which restores the content under the mask and eliminates appearance ambiguity. However, in the face of the problem of mask occlusion in complex scenes, it is difficult to restore the occluded face accurately with GAN. These methods can synthesize or generate new samples from the same distributed training dataset, but they do not well leverage contextual information of the images. Yu *et al.*. [47] propose an end-to-end image and face restoration model by using contextual attention layer and dilated convolution. Although performing better than many existing methods, it can only deal with the occlusion of regular shapes. Yu *et al.*. [48] later use gated dilated convolution to enlarge the receptive field and can restore arbitrary face occluded images, but it needs artificial inputs of occluded regions and cannot detect face occluded regions automatically. Therefore, the occluded region detector can be used to get arbitrary face-mask regions. The improved self-adaptive attention module can then be used to restore the image accurately.

2.4 Attention mechanism

Attention is also widely used for image restoration, but we only review a few representative ones related to our self-adaptive contextual attention model. Yu *et al.* [47, 48] propose a contextual attention model to learn correlations between missing and unmasked regions. This method is better for simple image restoration (*e.g.* natural scenes); for complex images (like the human face), it is challenging to exploit correlations between pixels in the occluded area to make generated results smoother and reduce artifacts because the learned attention lacks direct supervision. Zhou *et al.* [56] use the association of self-attention module and cross-attention module to restrict the generation of the attention map, which works well for regular occlusion restoration but not for face-mask and other occlusions with different shapes.

We propose a self-adaptive contextual attention model, which directly supervised and updates the attention map through the perceptual loss of the generated image to make the generated image more natural and clear.

2.5 Challenging face recognition

The large discrepancy between two face images is one of the key challenges in face recognition. The reasons for this difference include occlusion, posture changes, and so on. With the development of the representation learning [46, 57], some researchers try to use features that are not affected by occlusion or pose to complete the task of face recognition. Tran [37] et al. use the disentangled representation learning-Generative Adversarial Network to jointly optimize the task of face frontalization and learning pose invariant representation, so that it can robustly recognize face images with different poses. Yin [46] et al. design a activation diversity loss to learn interpretable face representations. In addition, the feature activation diversity loss was introduced to enhance the discrimination and robustness of occlusion features. However, for large-area occluded face images with masks, it is difficult to accurately recognize the face due to the loss of facial information.

3 FMOR DATASET

We collect a new dataset, Face-mask Occluded Restoration (FMOR), with the expectation to promote masked face restoration and recognition research. It contains 11,000 face images from 1,100 subjects, including 5,500 masked faces and 5,500 mask-free images, larger and more diverse than previous similar attempts [55].

Most of the images in FMOR are collected from the natural world and the Internet, and some images of Caucasian wearing face masks were synthesized using collected images. Compared with previous datasets, FMOR contains more various and complex scenarios. Please see Fig. 2 for some example images.

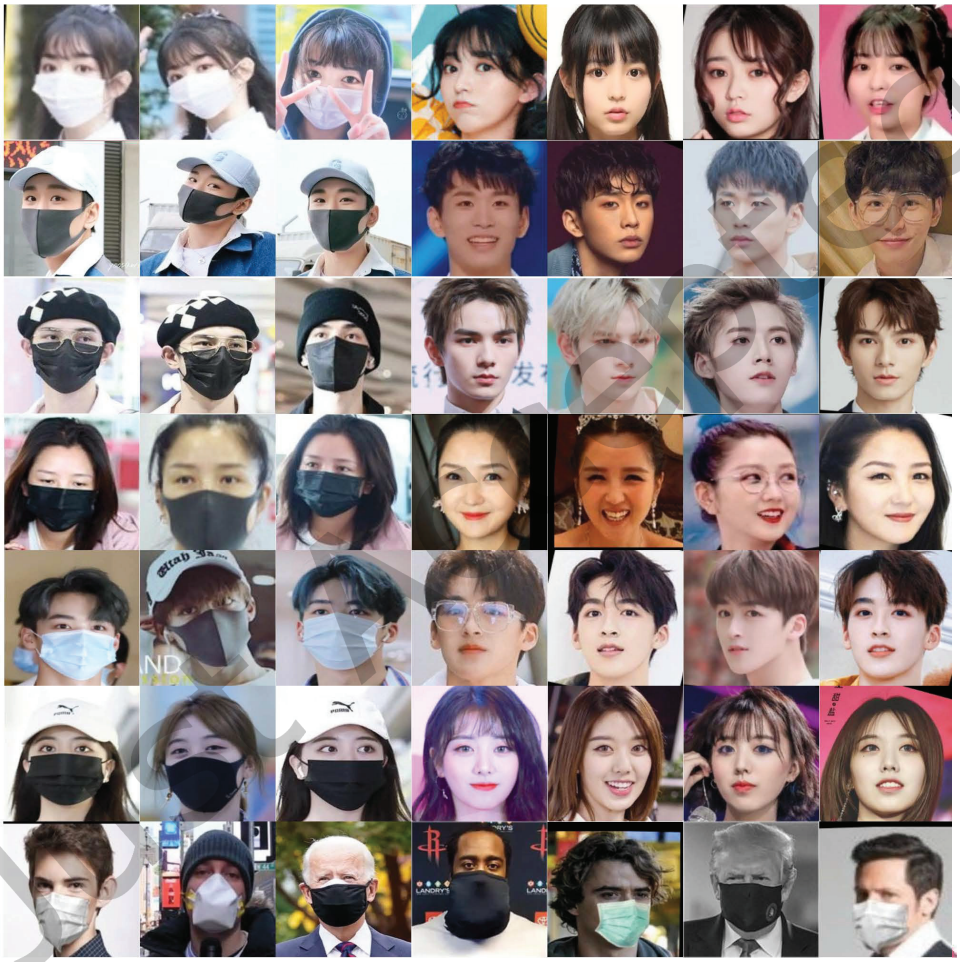


Fig. 2. Samples from FMOR with various face orientations, different races and scenes in face-mask occluded images.

Image collection and cleaning. We collect face images from different resources. On the one hand, we take photos of 230 students in our campus. On the other hand, we select nearly 4,000 celebrities from manmankan.com and then use crawler to crawl relevant images with keywords such as star names and wearing-masks with the following rules: 1) each person should have at least one masked image; 2) we consider both the gender balance and ethnic diversity. In this way, we collect 100,000 images of 2,500 persons in total. The average accuracy of

Table 1. Distributions on age, race, posture, mask type, illumination, gender in the FMOR dataset. Besides, the sixth column indicates the ratio of the synthesis images to the natural images in the Caucasians masked face images.

Race		Age		Illumination		Pose		Gender		Caucasian masked face image		Mask types
Asian	62.56%	**0-20**	13.48%	**Bright**	25.66%	**Front**	89.63%	**Male**	56%	**real-world**	18.13%	27
Caucasians	32.40%	**20-50**	80.10%	**Normal**	66.59%	**Side**	10.37%	**Female**	44%	**synthesis**	81.87%	
The black	5.04%	**50+**	6.42%	**Dark**	7.75%							

Table 2. Comparing statistics of the most recently occluded face dataset. "Mask Images." denotes Face-mask occluded images; "Real Occ." denotes whether the real occlusion is included while "Syn. Occ." denotes synthesized occlusion.

Dataset	#Subjects	#Images	#Mask Images	#Mask/Subject	Real Occ.	Syn. Occ.
AR [16]	126	4,000	×	×	Yes	×
MAFA [14]	30,811	38,506	×	×	Yes	×
RMFRD [42]	525	95,000	2,203	avg. 4	Yes	×
FMOR	**1,100**	11,000	**5,500**	**avg. 5**	**Yes**	**Mask**

image search results from the Internet is around 10% [36]. Some images do not meet our requirements, and some are damaged, which are discarded after manual screen to ensure accuracy and diversity of the attained images.

To further enlarge and diversify the subject pool, especially to expand the amount of data for masked Caucasians, we collect images of Caucasian without masks and manually add face masks to their faces. We first collected some face-mask images of different styles from the Internet and cut out some face-mask occlusion images by using Meitu Xiu Xiu [25], which is shown in Fig. 3. After that, we manually adjust the face-masks size and position by the software Meitu Xiu Xiu and splice them to the Caucasians images without face-masks originally. The synthetic image is shown in Fig. 4. For instance, we collect some images of face-masks and then manually adjust their size and position by the software Meitu Xiu Xiu, and then splice them into Caucasians images without face-masks.

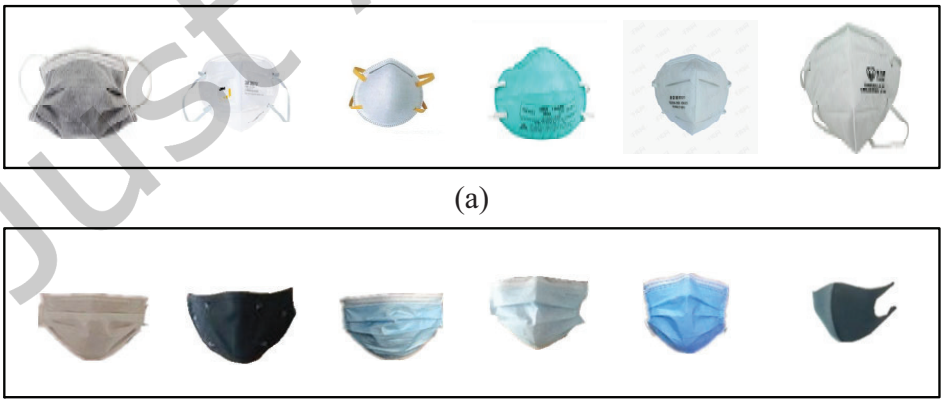


Fig. 3. Some examples of face-mask images that we collected. (a) shows the face mask collected from the internet, (b) presents the face-mask cut from the face image covered by the face-mask. Best viewed with zoom-in.

Fig. 4. Some examples of synthetic images on FMOR. Best viewed with zoom-in.

Clipping images. In this step, we crop the face area in each image and resize it into 256 × 256. Then, we put the occluded images and their corresponding unmasked images in two folders, A and B. Folder A stores images covered by face-masks while folder B stores corresponding unmasked images.

Face-mask region annotation. We use LabelMe [35] annotation software to annotate pixel-wise face-make regions for each image in Folder A. We save the mask annotation images in the face-mask folder.

Manual inspection. After annotation, we perform a manual inspection under all images. We recheck each image to ensure the subject is correct. The entire work has taken around five months.

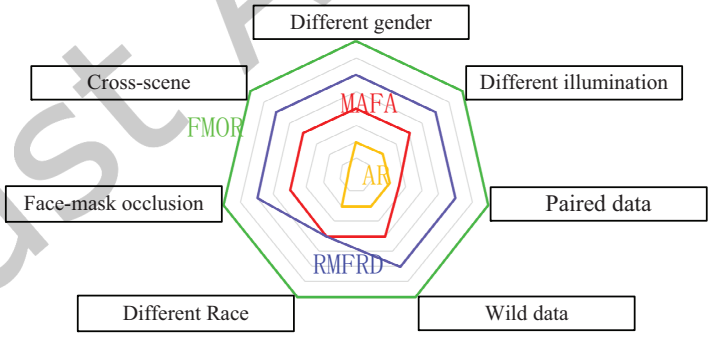


Fig. 5. Comparison of included attributes between FMOR and other occluded face datasets. Different radial axis represents different attribute. Best viewed in color.

Dataset split and statistics. In total, there are 11,000 face images with 1,100 subjects in FMOR dataset. Each subject has at least one masked face image with one corresponding unmasked image. We show the distribution of the FMOR in Tab. 1, and we give statistical comparisons between FMOR dataset and existing face-mask occluded datasets AR [16], RMFRD [42] and MAFA [14] in Tab. 2 and Fig. 5, which can see FMOR is a face-mask restoration

dataset that contains the most subjects and masked images compared with AR [16], RMFRD [42] and MAFA [14] datasets.

4 FMODR FRAMEWORK

We elaborate on the proposed face-mask occluded restoration method, named Face-Mask Occluded Detection and Restoration (FMODR). It contains a face-mask region detector, an coarse network, and a refine network. We start by introducing the FMODR and then proposed techniques to improve our generator's generalization, making restored images more realistic.

4.1 Overall architecture

The network architecture of our proposed model is shown in Fig. 6, which follows a coarse-to-refine face-mask occluded restoration network architecture. We first input the face-mask occluded image into the face-mask region detector and get the masked region image. The structure of the face-mask region detector is shown in Fig. 6 (a). It is a full convolution neural network with convolutions and up-sampling structure. For a given image, we calculate its pixel-wise L2 loss function as

$$\mathcal{L}_{pixel} = \sum_{i=1}^{n}((\mathcal{I}mg_i^{mask}) - 1)^2 + (\mathcal{I}mg_i^{unmask} - 0)^2), \tag{1}$$

where i indicates the index of the pixel, n represents the total number of pixels in the segmentation image, $\mathcal{I}mg_i^{mask}$ is the pixel value of the face-mask region of the generated segmentation image and $\mathcal{I}mg_i^{unmask}$ is the pixel value of the unmask region of the generated segmentation image. The value in the face-mask regions are 1, and the other regions are 0.

After that, we multiply the face-mask occluded image and the unmask region(1-mask region) image, and the result is input into the coarse network and obtain a coarse restored image without occlusion, in which we apply a U-Net architecture. For a given image, we calculate its L1 loss function as follows:

$$\mathcal{L}_{rec} = \sum_{i=1}^{n}|\mathcal{I}_i^{pred} - \mathcal{I}_i^{gt}|, \tag{2}$$

where i indicates the index of the pixel, n represents the total number of pixels in the restored image, $\mathcal{I}_i^{pred}$ is the pixel value of the generated image and $\mathcal{I}_i^{gt}$ is that of the ground truth.

Then, the refine network takes the coarse image and the masked region image as inputs and outputs a refined image. As shown in Fig. 6 (c), the refine network is a two-stream network. The coarse image and mask region are multiplied, which is added to the original image without mask region and then the result input into the two branches of the refine network. After that, the combined features are convoluted and up-sampled to get the restored image. Finally, the restored image and the original image are input into the discriminators for adversarial training to refine the restored image. In the two-stream network, for two given input images, we calculate its loss function as follows:

$$\mathcal{L}_{per} = \sum_{i=1}^{n}(\mathcal{F}_{refine_i}^{pred} - \mathcal{F}_{refine_i}^{gt})^2, \tag{3}$$

where $\mathcal{F}_{refine_i}^{pred}$ is the feature map value obtained by refine restoration image through a relu3-3 layer of VGG16, $\mathcal{F}_{refine_i}^{gt}$ is that of the ground truth, i is the value of the i-th feature pixel, and n represents the total number of pixels in the feature map. Besides, the loss function of discriminators is shown in Eq. (5). In addition, self-adaptive contextual attention and global-masked region discriminators are also essential modules of the FMODR, and we will describe them in detail below.

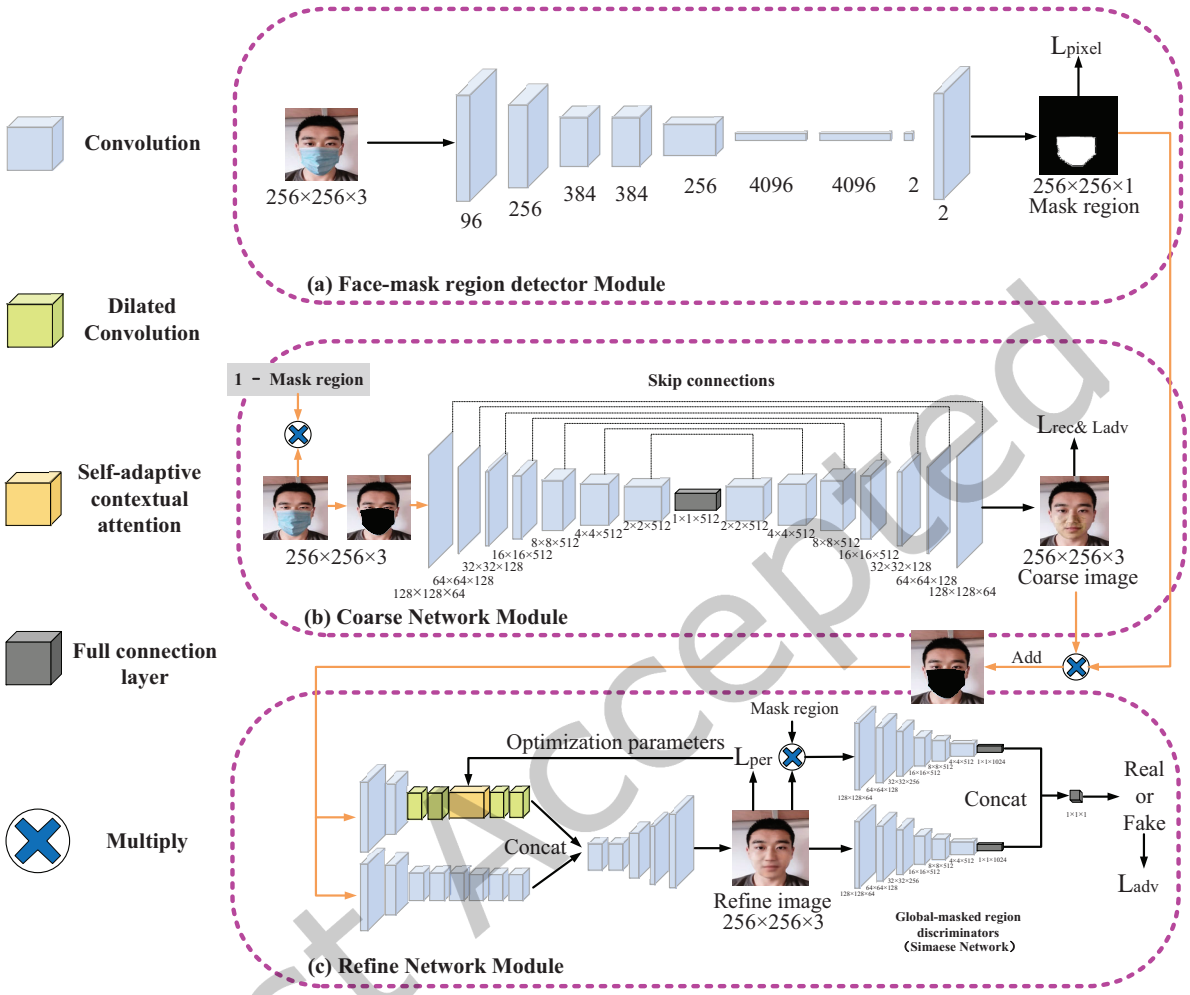


Fig. 6. Overview of FMODR. It consists of a face-mask region detector, a coarse network, and a refine network. (a) A face-mask region detector is applied to get the masked region. (b) The face-mask occluded image and the unmask region image are multiplied, and the result is input into the coarse network and obtains a coarse restored image without occlusion. (c) The coarse image and mask region are multiplied, which is added to the original image without mask region and then input into the two branches of the refine network.

4.2 Self-adaptive contextual attention model

In the refine network, we use a self-adaptive contextual attention model to borrow and copy the feature information from background patches to generate masked patches. The structure of this self-adaptive contextual attention model is shown in Fig. 7.

The architecture of our model is improved based on [48]. In [48], background patches are used as convolution filters to compute matching scores with the foreground. The kernel of attention propagation is an identity matrix and invariant, which tends to yield good restoration in smooth scenes with slight changes but often fails

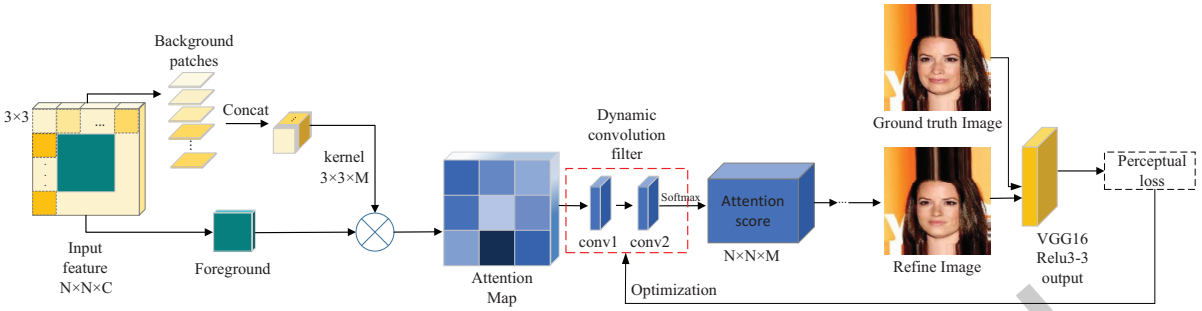


Fig. 7. Structure of self-adaptive contextual attention model. The Foreground represents the occluded part of the image, and the Background patches are the 3×3 patches extracted from the background image, "..." represents the process from the output of self-adaptive contextual attention module to the final output of the refine network.

for human masked face restoration due to the complex structure of human faces and fast changes of the face texture. In addition, due to lack of adequate supervision on the attention model, the generated attention map is not accurate on complex scenes. We use two-layer convolution to replace the identity matrix to supervise the attention map and make it more accurate to address these issues.

Different from Gconv [48], to get the attention map, we first calculate the cosine similarity between the foreground blocks and the background blocks, the background block which has the largest similarity with the foreground block is selected, and then multiplied by the cosine similarity. Next, the foreground blocks are processed one by one to obtain the attention map, instead of which is obtained by convoluting the foreground and background regions in Gconv [48].

After getting the attention map, we adjust the attention map and let it go through two layers of convolution. The convolution layer parameters are optimized by calculating the perceptual loss of the natural non-occluded image and the generated refine restored image, which can dynamically adjust the values of the attention map constrained by the neighborhood pixels. In this way, the restoration accuracy can be improved while keeping the face highly structured. Finally, we reuse the extracted patches as deconvolution kernels and attention scores as input features to restore the face-mask region. FMODR further encourages optimizing convolution kernel parameters and updating the attention map by dynamic constraints for adjacent pixels. The update is done by

$$\mathcal{S}_{x,y}^{update} = \sum_{i\in[-k,k]} \sum_{j\in[-k,k]} (w_{i,j} \times \mathcal{S}_{x+i,y+j}), \tag{4}$$

where $\mathcal{S}_{x,y}^{update}$ represents the updated attention map, k is the distance from the center of the kernel, i and j indicate the positions of different elements on the kernel, $w_{i,j}$ is learnable parameters of the convolution kernel, which is only changeable during training and fixed during testing, x and y indicate the center position of the kernel. Firstly, the convolution kernel of the filter is initialized with the size of 3×3. After that, we calculate the perceptual loss $\mathcal{L}_{per}$ as shown in Eq. 3 between the ground truth unmasked image and restored image, then update the convolution kernel filter by adopting backpropagation and gradient descent.

4.3 Global-masked region discriminators

Global-masked region discriminator is improved on the basis of the discriminator [23], which is used to judging whether the image is real or fake. The network is based on a convolution network, and the feature is compressed into a small vector by a full connection layer. The network's output is fused by a concatenation layer, which

predicts the continuous values corresponding to the real probability of the image. An overview is shown in Fig. 6 (c).

The structure of the global-masked region discriminators is a siamese network (the global part and the mask part have the same structure) with six convolution layers. Finally, the results of global-masked region discriminators are concatenated into a 2,048-dimensional vector and then pass through the full connection layer and a sigmoid transfer function to output 1-dimensional values between 0 and 1. The adversarial loss is shown in

$$\begin{aligned}\mathcal{L}_{adv} &= (\mathbb{E}_{x-\mathbb{P}_{gt}}[\log \mathcal{D}_{global}(x)] - \mathbb{E}_{x-\mathbb{P}_{gen}}[\log(1-\mathcal{D}_{global}(G(x)))]) \times \frac{1}{2} \\ &+ (\mathbb{E}_{x-\mathbb{P}_{gt}}[\log \mathcal{D}_{masked-region}(x)] - \mathbb{E}_{x-\mathbb{P}_{gen}}[\log(1-\mathcal{D}_{masked-region}(G(x)))]) \times \frac{1}{2},\end{aligned} \tag{5}$$

where $\mathcal{D}_{global}$ and $\mathcal{D}_{masked-region}$ are global and masked region discriminators respectively. $\mathbb{E}_{x-\mathbb{P}_{gt}}$ indicates that x belongs to the real images, and $\mathbb{E}_{x-\mathbb{P}_{gen}}$ indicates that x belongs to the restored images. G denotes the images restore model.

4.4 Improve image quality via special training strategy

To ensure that the restored image is more realistic and has fewer artifacts, after lots of para-selection experiments, we found that, the below strategy can be helpful to get better results: for coarse network, the overall loss function is shown in Eq. 6. We first train the coarse network (λ_1=1) and global-masked region discriminators (λ_2=1). After that, the coarse network and global-masked region discriminators are optimized jointly , where λ_1=1 and λ_2=0.05. If the image has been restored well in the coarse network, i.e., if the loss of the result obtained in the coarse network is less than 0.1, the optimization in refine network will not be performed.

$$\mathcal{L}_{coarse} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{adv}, \tag{6}$$

For a refine network, the overall loss function as Eq. 7. We first train the refine network (λ_5=1) and global-masked region discriminators (λ_4=1). After that, the refine network and global-masked region discriminators are optimized jointly, the three trade-off parameters λ_3, λ_4 and λ_5 are set to 0, 0.1 and 0.5, respectively. After 20,000 iterations, the three trade-off parameters λ_3, λ_4 and λ_5 are set to 1, 0.1 and 0, respectively.

$$\mathcal{L}_{fine} = \lambda_3 \mathcal{L}_{per} + \lambda_4 \mathcal{L}_{adv} + \lambda_5 \mathcal{L}_{rec}. \tag{7}$$

Reconstruction loss in the previous iterations can ensure that the global features of the generated image (such as edge contour) are more accurate and has fewer artifacts. And the later iterations make the generated image more precise and more realistic with the constraint of perceptual loss.

5 EXPERIMENTS

5.1 Settings

Dataset processing. In the FMOR dataset experiment, we select 2,990 face images without face-mask occlusion from the FMOR dataset and artificially add different mask occlusion on them, which called "FMOR-VM". To do the quantitative experiment, we select 2,276 images as training set and the rest act as testing samples. Besides, to verify the qualitative performance of image restoration, all images of "FMOR-VM"(2,990 images) are used as training set to train the model, and 5,500 real scene masked images from FMOR are used as testing samples. Considering there are few existing open-source face-mask occlusion datasets, and the scarf's masking effect is similar to that of the face-mask, we also use the scarf occlusion images of AR dataset to test our framework. A total of 600 pairs of scarf occlusion images and no occluded images are used from the AR dataset, with 450 pairs for training and 150 pairs for testing. Besides, we also extract a synthetic face-mask restoration dataset based on CelebA, denoted as "S-CelebA", by randomly selecting 2,000 images from CelebA as testing samples and using rest as training samples. For synthesis, we reshape the size of these images to 256×256 and manually put

face-masks on each face image in CelebA. "S-CelebA" in qualitative and quantitative experiments is utilized for image restoration. For the face-mask restoration task, we use the above three datasets for experiments and report the Mean Absolute Error (MAE or L1 err.), Mean Square Error (MSE or L2 err.), Structural SIMilarity (SSIM), and Peak Signal to Noise Ratio (PSNR). To compare fairly, the training and testing dataset composition and partition are the same as FMODR in all comparative experiments.

We also evaluate face recognition and verification based on restored images on the AR and FMOR datasets. For the FMOR dataset, 2,000 face-mask occluded images are utilized as the testing set and the rest act as the training samples for face recognition. In addition, for the above FMOR face recognition testing set, we randomly produce 3,000 pairs for face verification, including 1,500 pairs with the same subjects and 1,500 pairs with different subjects for face verification experiments. For the AR dataset, 1200 images of scarf occlusion are used, with 900 images for training and 300 images for testing, form face recognition set. After that, we randomly select and produce 300 pairs from the face recognition testing set, including 150 pairs with the same subjects and 150 pairs with different subjects to test the face verification task. To verify the effectiveness of FMODR, CTSDG, and Gconv methods, we first restore masked face images and test them with the trained face recognition model. Rank-1 Recognition rate is used to evaluate the face recognition models, and face verfication models are evaluated by Detection Error Trade-off (DET) curve.

Implementation details. Parameters λ_1-λ_5, are adjusted according to the results of lots restoration experiments. All input images are resized to 256×256. For the coarse network, we first train the coarse network and perform 10,000 iterations, and the parameters λ_1=1 and λ_2=0. After that, the global-masked region discriminators are performed 1,000 iterations. The parameters λ_1=0 and λ_2=1. Next, the coarse network and discriminators are optimized jointly for 24,000 iterations, where λ_1=1, and λ_2 =0.05. At the same time, for refine network, we first train the refine network and perform 10,000 iterations, and the parameters λ_5=1, where λ_3=0 and λ_4=0. Then, the global-masked region discriminators are performed 1,000 iterations. The parameters λ_4=1 (λ_3, λ_5 = 0). After that, the refine network and discriminators are optimized jointly for 20,000 iterations (λ_4 =0.1, λ_5=0.5 and λ_3=0). Finally, the refine network and discriminators are optimized jointly for 80,000 iterations, the three trade-off parameters λ_3,λ_4 and λ_5 are set to 1, 0.1, and 0, respectively. During training, Adam is used as the optimizer with a learning rate of 0.000015. We try different levels of the learning rate, from 0.1 to 0.000001. Finally, find that the convergence speed is the fastest when the learning rate is 0.000015. On a single NVIDIA RTX 2080Ti (11GB), we train our model on AR dataset, which cost about 8 hours, and also on S-CelebA and FMOR datasets, which cost 96 hours with a batch size of 4. We use the triplet learning [22] method to evaluate the impact of our restored image on the face-mask occlusion recognition and verification task.

Table 3. Quantitative results for FMODR, CTSDG and Gconv on FMOR-VM. Higher SSIM and PSNR indicate better performance; lower L1 error and L2 error indicate better performance.

	FMOR-VM			
Method	L1 err.	L2 err.	SSIM	PSNR
Gconv[48]	3.49%	1.61%	0.89	24.79
CTSDG [18]	3.56%	1.20%	0.91	26.09
FMODR	**3.24%**	1.45%	**0.92**	**26.14**

5.2 Quantitative results

To verify the effectiveness of the proposed FMODR, we compare it with state-of-the-art methods including CTSDG [18] and Gconv [48] on FMOR-VM, S-CelebA, and AR datasets. Besides, we also compare it with GCF [30]

Table 4. Quantitative results for Gconv, CTSDG, GCF, Pix2Pix and our FMODR on AR and S-CelebA datasets. Higher SSIM and PSNR values are better; lower L1 error and L2 error are better.

AR dataset				
Method	L1 err.	L2 err.	SSIM	PSNR
Gconv [48]	24.97%	16.05%	0.59	14.44
CTSDG [18]	22.99%	13.48%	0.64	15.40
FMODR	7.69%	4.78%	**0.84**	**20.19**
S-CelebA				
GCF [30] (Rec)	–	–	0.84	20.20
Pix2Pix[24](Fm)	7.97%	2.62%	0.83	22.90
Pix2Pix[24](Rec)	5.87%	1.68%	0.89	24.60
Gconv[48](Fm)	3.20%	1.41%	0.91	25.22
Gconv[48](Rec)	2.33%	1.14%	0.94	25.70
CTSDG [18] (Fm)	3.43%	0.89%	0.91	27.30
CTSDG [18] (Rec)	2.21%	0.33%	0.95	31.52
FMODR (Fm)	2.27%	0.81%	0.94	27.87
FMODR (Rec)	1.01%	**0.29%**	**0.96**	32.17

and Pix2Pix [24] on S-CelebA dataset. "Rec" indicates rectangle occlusion added randomly(64×64) on the CelebA dataset at lower face, and "Fm" denotes different kinds of face-mask occlusion added randomly on the CelebA dataset. The experimental results are averaged on all testing sets and summarized in Tab. 3 and Tab. 4. From the tables, our framework outperforms its competitors on these metrics.

From the experiments on the FMOR-VM dataset in Tab. 3, it can be seen that the FMODR outperforms Gconv in all metrics and superior to the CTSDG in L1 error, SSIM, and PSNR. The proposed FMODR outperforms the CTSDG and Gconv for the AR dataset on four metrics as shown in Tab. 4; because of the extensive range of scarf occlusion, the Gconv and CTSDG methods cannot restore the occluded area effectively. In contrast, our FMODR can restore the occluded area of any size. On the S-CelebA(Fm) and S-CelebA(Rec) datasets, FMODR outperforms all the above methods on these four metrics. All the results well prove the effectiveness of our framework.

5.3 Qualitative results

To check visual effects of the restored faces from mask occlusion with our proposed framework, we compare our FMODR with CTSDG [18], Gconv [48], Pix2Pix [24], PatchMatch (PM) [4] and Baidu's existing image restore tool(called Baidu in this paper) on FMOR and S-CelebA datasets, respectively. Some examples are shown in Fig. 8. It can be seen FMODR is able to restore face-mask occluded images effectively, demonstrating notable effectiveness on real world face-mask occluded restoration (8th column in Fig. 8 from left). Comparatively, Baidu, PM, Pix2Pix, Gconv and CTSDG have difficulty restoring the face's occlusion part. For example, using Pix2Pix method(5th column in Fig. 8 from left) will produce color inconsistency. Even though Gconv uses contextual attention to complete image restoration, it still produces semantically inconsistent structure or texture (6th column in Fig. 8 from left). In addition, these five methods (3rd, 4th, 5th, 6th and 7th columns in Fig. 8 from left) produce severe artifacts for the face area, possibly because they do not well leverage the guidance from the components and the structural symmetry of the human face. These results well prove FMODR produces better visual effects than the state-of-the-art methods for face-mask occluded restoration.

Fig. 8. Qualitative comparison of face-mask occluded restoration on FMOR and S-CelebA. Best viewed with zoom-in.

We visualize the results of facial landmark detection for different restoration methods, as shown in Fig. 9. For the occluded face image, the facial landmark detection algorithm is difficult to apply effectively. This is because face-mask occlusion loses a lot of helpful face information. In comparison, the restored image is similar to the natural face to detect the facial landmark accurately. In addition, we also show the restoration results of masked

Fig. 9. Results of facial landmark detection with different types of images. Columns 1st and 2nd represent face-mask occluded images. The 3rd and 4th columns represent Pix2Pix restored images. The 5th and 6th columns represent FMODR restored images, and the 7th and 8th columns represent non-occluded images.

Fig. 10. Restoration results of real face-mask occluded images.

face images collected in real scenes, as shown in Fig. 10. Not only does it can see that FMODR can accurately restore images taken in real scenes, but also proves the effectiveness of the face restoration method from the side.

5.4 Face recognition

Face recognition systems underperform when the faces are occluded. Our framework restores the non-occluded face image from the face image occluded by the face-mask, facilitating efficient performance for the face recognition tasks. Performance of recent face restoration techniques, viz. CTSDG [18] and Gconv [48] have been compared with our proposed framework FMODR for generation of the faces, evaluated using state-of-the-art benchmark face recognition system, like PCA [38], LPP [21], Sparse Representation [43], GPCA [27], VGG-Face [33], Triplet-loss learning [22] and SCA [10]. The results in Tab. 5 show that the rank-1 accuracy for the AR and FMOR datasets. From Tab.5, we can see that with FMODR + SCA [10] method, the accuracy is better than other face recognition methods, which indicates the power of the face image restoration-based techniques has ability to overcome occluded face problem, and our proposed method (FMODR) is able to generate a better face recognition part than other restoration methods.

We randomly select a face image without occlusion for each subject, calculating the cosine similarity between itself and the testing set. It can be seen from Tab.6 that the similarity ratio between the FMODR, CTSDG restored image and the non-occluded image is higher than that between the face-mask occluded and non-occluded image, which proves that it is feasible and effective to recognize the face by restored image. However, the similarity of images restored by Gconv on FMOR dataset is lower than that of images with occlusion, which indicates that the restoration effect of this method on FMOR dataset is not good. So we can see that some image restoration methods are not suitable for mask occlusion face restoration and recognition tasks.

Table 5. Rank-1 Recognition rates (in %) exhibiting a higher performance for Face Recognition by FMODR+Triplet learning, compared with several state-of-the-art face recognition techniques on AR and FMOR datasets. The results in bold demarcate the best performance (row-wise).

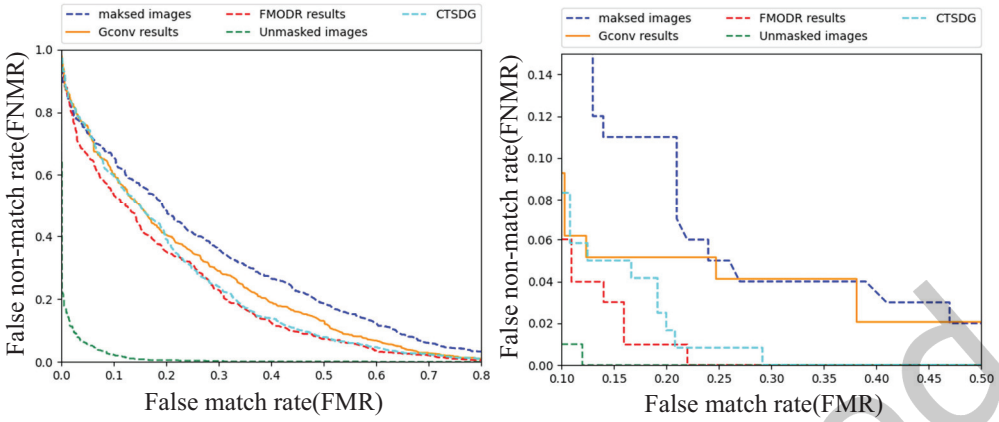
AR	
Model	Rank-1(%)
PCA [38]	30.2
LPP [21]	43.9
SR [43]	51.8
GPCA [27]	61.3
VGG-Face [33]	73.8
Triplet learning [22]	77.3
Gconv [48]+Triplet learning [22]	85.3
SCA [10]	87.3
CTSDG [18]+Triplet learning [22]	91.9
Gconv [48]+SCA [10]	94.0
CTSDG [18]+SCA [10]	95.9
FMODR+Triplet learning [22]	96.7
FMODR+SCA [10]	**97.3**
FMOR	
VGG-Face [33]	8.1
Triplet learning [22]	33.5
Gconv [48]+Triplet learning [22]	34.1
CTSDG [18]+Triplet learning [22]	34.6
SCA [10]	35.3
FMODR+Triplet learning [22]	35.5
Gconv [48]+SCA [10]	36.8
CTSDG [18]+SCA [10]	39.0
FMODR [18]+SCA [10]	**47.6**

Table 6. Comparison of cosine similarity between face-mask occluded image, restored face image and non-occluded face image.

	Masked image	Gconv	CTSDG	FMODR
AR	0.234	0.227	0.241	**0.245**
FMOR	0.523	0.535	0.592	**0.631**

5.5 Face verification

Face verification experiments have been done on AR and FMOR datasets. It can be seen from Fig. 11 that, the verification results of the image restored by FMODR on FMOR and AR datasets are better than those of Gconv, CTSDG, and the face-mask images. Also the accuracy of Gconv+Triplet learning method is lower than that of



(a) DET curve of FMOR testing set (b) DET curve of AR testing set

Fig. 11. DET curve of FMOR and AR testing set.

masked images+Triplet learning method on the FMOR dataset. From above experimental results, we can see that some face restoration methods, such as FMODR and CTSDG, can effectively improve the accuracy of masked face verification, while some other methods, like Gconv, are challenging to improve accuracy of face verification tasks.

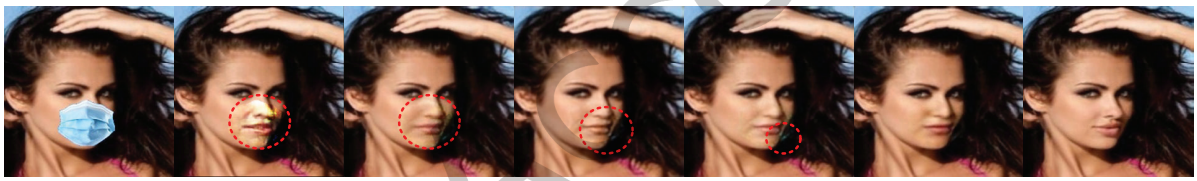


Fig. 12. Visualization results of ablation experiments on different modules. Best viewed with zoom-in.

5.6 Ablation study

Effect of self-adaptive contextual attention. In order to verify the effect of the self-adaptive contextual attention model, we replace it with the contextual attention model [48]. Baseline means the base model without additional self-adaptive contextual attention module and Global-masked region discriminators, while "SCA" means the model using the self-adaptive contextual attention layer, "CA" is the model using contextual attention layer [48]. We randomly selected 2,000 images from S-CelebA, including 1000 for training and 1000 for testing. As shown in tab. 7 and the 2nd, 3rd, and 4th columns of Fig. 12, on the S-celebA dataset, the model with self-adaptive contextual attention model obtains better and smoother results than the model with contextual attention. This may be because the restored pixels can be dynamically constrained by adjacent pixels.

Effect of global-masked region discriminators. To verify effectiveness of global-masked region discriminators, ablation experiments are performed on global-masked region discriminators. "GM" is the model using global-masked region discriminators. We randomly selected 2,000 images from S-CelebA, including 1,000 for training and 1,000 for testing. We use the model without global-masked region discriminators as a control. Tab. 7 and the 3nd, 4rd, 5th, and 6th columns of Fig. 12 show that the model with the global-masked region discriminators

Table 7. Ablation results on self-adaptive contextual attention module and global-masked region discriminators.

Ablation on global-mask region discriminators				
Method	L1 err.	L2 err.	SSIM	PSNR
Baseline	3.11%	0.44%	0.90	30.19
Baseline+CA [48]	3.08%	0.43%	0.91	30.27
Baseline+SCA	2.57%	0.36%	0.92	31.11
Baseline+CA [48]+GM	2.46%	0.35%	0.93	31.24
Baseline+SCA+GM	**2.28%**	**0.35%**	**0.94**	**31.36**

Table 8. Ablation experiments on CelebA and FMOR datasets.

Dataset	L1 err.	SSIM
Data-cop2	3.68%	0.90
Data-cop1	**3.40%**	**0.91**

obtain more realistic and fewer artifacts results than that without the discriminators on S-CelebA dataset. We attribute better realism to the discriminant function of global-masked region discriminators.

Effect of our FMOR dataset. From Tab. 3 and Tab. 4, we can see that all methods work better on CelebA, but on the FMOR dataset, the performances are not good enough. This reveals that FMOR is more challenging than CelebA dataset. To evaluate our FMOR dataset further, we expand the CelebA dataset by adding samples from FMOR. We first selected 2,000 occlusion images from CelebA as the testing samples and then randomly selected 2,000 images from CelebA and 2,000 images from FMOR, which is called Data-cop1. Next, 4,000 images were randomly selected from CelebA, called Data-cop2. And then FMODR is trained on Data-cop1 and Data-cop2, respectively, and tested on above test set. In Tab. 8, it can be seen that the model trained with Data-cop1 works better than that of Data-cop2. That means, the diversity of FMOR is better than other available datasets, such as CelebA. That is, there are more complex scenes and conditions in FMOR.

Effect of loss functions. We visualized the effects of L1 loss and perceptual loss on image generation. From Fig. 13 (a), we can see that the image generated by L1 loss is very fuzzy. Especially for the teeth, which can't see the details of it, as shown in the 2nd and 3rd column from left in Fig. 13 (a)). However, the image generated by perceptual loss is clear and realistic, which is closer to the natural scene. In addition, different loss functions are utilized to update the attention map, as shown in Fig. 13 (b). Updating of the attention map with L1 and L2 Loss will blur the generated image, and lose texture detail information, which can be solved with perceptual loss. The image generated by the perceptual loss function is more legible, even the tooth part can be generated well.

5.7 Failure cases

We show some failure cases of our model in Fig. 14. It can be seen that the restoration results of some side face images are poor, and the generated images are blurred. Because of glasses, the illumination of this part is different with unmasked regions, so in the restored results, there is a gap between the part of glasses and face mask. Therefore, it is difficult to restore these challenging images accurately.

5.8 Running time of model reasoning

In the reasoning stage, we calculate the running time of FMODR on S-CelebA, AR, and FMOR datasets, respectively. As shown in Tab. 9, FMODR framework can meet the real-time requirements of the face restoration task. In

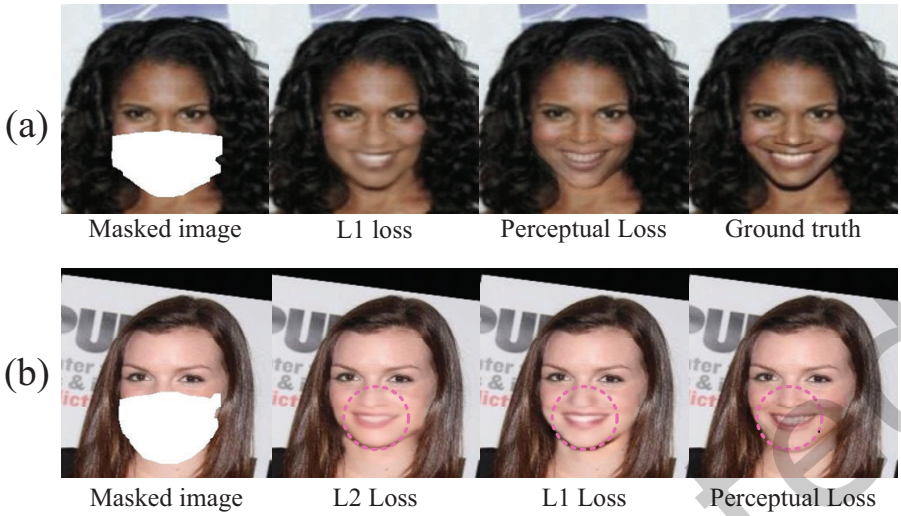


Fig. 13. (a) The visual effect of the image generated by different loss functions. (b) We are using different loss functions to update the attention map to obtain the visualized results of the restored image. Please zoom in to see the details.

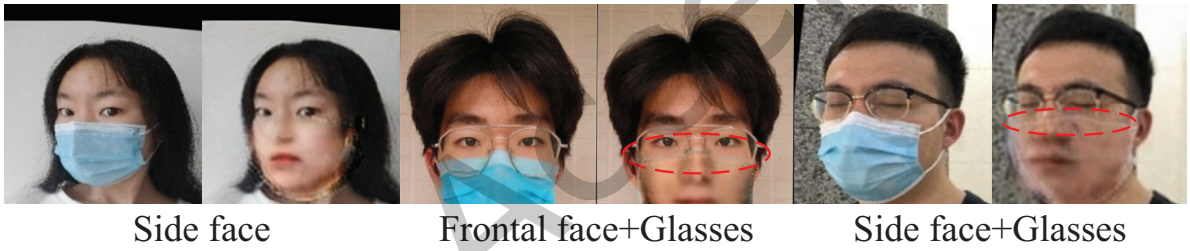


Fig. 14. Failure cases. Please zoom in to see the details.

Table 9. The running time of FMODR on different data and the memory occupied by the graphics card, MS / IMG represents the time required for each image processing, Ms represents the millisecond.

	Running time of reasoning phase			
	AR	S-CelebA	FMOR	Memory space
FMODR	69MS / IMG	80MS / IMG	63MS / IMG	8773 MB

addition, we also show the memory space occupied by the model in the process of reasoning. It can be seen that the memory space occupied by the model is small and has practical application value.

6 CONCLUSIONS

In this work, we build a face-mask occluded restoration dataset named FMOR and propose a new framework called FMODR for face-mask occluded restoration. The new FMOR dataset is the largest face-mask restoration dataset with 1,100 subjects corresponding to 5,500 free images and 5,500 face-mask occluded images, featuring

various illuminations, different races, face orientations, and mask types. It is more consistent with the real-world scenarios and is expected to facilitate and advance research progress in face-mask restoration, masked face recognition and verification, which is more challenging than existing masked face datasets. The proposed FMODR framework consists of a coarse network, a face-mask detector, and a refine network containing a specifically designed self-adaptive contextual attention module and global-masked region discriminators.

The experiments demonstrate that the proposed model FMODR outperforms the state-of-the-art methods on our new FMOR dataset and other datasets. Besides, FMODR method also helps to improve the accuracy of masked face recognition and verification tasks.

REFERENCES

[1] Sunil Arya and DM Mount. 1998. ANN: library for approximate nearest neighbor searching. In *Rep Cucurbit Genet Coop*.

[2] Michael Ashikhmin. 2001. Synthesizing natural textures. In *Proceedings of the 2001 symposium on Interactive 3D graphics*. 217–226.

[3] Samik Banerjee and Sukhendu Das. 2020. SD-GAN: Structural and Denoising GAN reveals facial parts under occlusion. *arXiv preprint arXiv:2002.08448* (2020).

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans Graph* 28, 3 (2009), 24.

[5] Raphaël Bornard, Emmanuelle Lecan, Louis Laborelli, and Jean-Hugues Chenot. 2002. Missing data correction in still images and image sequences. In *Proceedings of the tenth ACM international conference on Multimedia*. 355–361.

[6] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. 2011. Robust principal component analysis? *JACM* 58, 3 (2011), 1–37.

[7] Chaofeng Chen, Xiaoming Li, Lingbo Yang, Xianhui Lin, Lei Zhang, and Kwan-Yee K Wong. 2021. Progressive semantic-aware style transformation for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11896–11905.

[8] Xiang Chen, Linbo Qing, Xiaohai He, Jie Su, and Yonghong Peng. 2018. From eyes to face synthesis: a new approach for human-centered smart surveillance. *IEEE Access* 6 (2018), 14567–14575.

[9] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans Image Process* 13, 9 (2004), 1200–1212.

[10] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. 2020. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *European Conference on Computer Vision*. Springer, 741–757.

[11] Alexei A Efros and William T Freeman. 2001. Image quilting for texture synthesis and transfer. In *CVPR*. ACM, 341–346.

[12] Alexei A Efros and Thomas K Leung. 1999. Texture synthesis by non-parametric sampling. In *ICCV*, Vol. 2. IEEE, 1033–1038.

[13] Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2007. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans Syst Man Cybern Syst* 38, 1 (2007), 149–161.

[14] Shiming Ge, Jia Li, Qiting Ye, and Zhao Luo. 2017. Detecting masked faces in the wild with lle-cnns. In *CVPR*. 2682–2690.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*. 2672–2680.

[16] Ralph Gross. 2005. Face databases. In *Handbook of face recognition*. Springer, 301–327.

[17] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image Vis Comput* 28, 5 (2010), 807–813.

[18] Xiefan Guo, Hongyu Yang, and Di Huang. 2021. Image Inpainting via Conditional Texture and Structure Dual Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14134–14143.

[19] Kaiming He and Jian Sun. 2012. Statistics of patch offsets for image completion. In *ECCV*. Springer, 16–29.

[20] Lingxiao He, Haiqing Li, Qi Zhang, Zhenan Sun, and Zhaofeng He. 2016. Multiscale representation for partial face recognition under near infrared illumination. In *BTAS*. IEEE, 1–7.

[21] Xiaofei He and Partha Niyogi. 2004. Locality preserving projections. In *Adv Neural Inf Process Syst*. 153–160.

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

[23] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–14.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *CVPR*. 1125–1134.

[25] Wen Jian-Ke. 2013. Ways to retouch photos. *Laboratory ence* (2013).

[26] Jino Lee, Dong-Kyu Lee, and Rae-Hong Park. 2012. Robust exemplar-based inpainting algorithm using region segmentation. *IEEE Transactions on Consumer Electronics* 58, 2 (2012), 553–561.

[27] Zhen Lei, Shengcai Liao, Ran He, Matti Pietikainen, and Stan Z Li. 2008. Gabor volume based local binary pattern for face representation and recognition. In *Proc Int Conf Autom Face Gesture Recognit*. IEEE, 1–6.
[28] Ang Li, Jianzhong Qi, Rui Zhang, and Ramamohanarao Kotagiri. 2019. Boosted gan with semantically interpretable information for image inpainting. In *IJCNN*. IEEE, 1–8.
[29] Chenyu Li, Shiming Ge, Daichi Zhang, and Jia Li. 2020. Look through masks: towards masked face recognition with de-occlusion distillation. In *Proceedings of the 28th ACM International Conference on Multimedia*. 3016–3024.
[30] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. 2017. Generative face completion. In *CVPR*. 3911–3919.
[31] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. 2001. A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *CVPR*, Vol. 1. IEEE, I–I.
[32] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *ICCV*. 3730–3738.
[33] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep face recognition. (2015).
[34] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *CVPR*. 2536–2544.
[35] Bryan Christopher Russell, Antonio J Torralba, Kevin Patrick Murphy, and William T Freeman. 2008. LabelMe. *IJCV* (2008).
[36] Antonio Torralba, Rob Fergus, and William T Freeman. 2008. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans Pattern Anal Mach Intell* 30, 11 (2008), 1958–1970.
[37] 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.
[38] Matthew A Turk and Alex P Pentland. 1991. Face recognition using eigenfaces. In *CVPR*. IEEE Computer Society, 586–587.
[39] Qiong Wang and Jingyu Yang. 2006. Eye detection in facial images with unconstrained background. *JPRR* 1, 1 (2006), 55–62.
[40] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9168–9178.
[41] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. 2018. Image inpainting via generative multi-column convolutional neural networks. In *NeurIPS*. 331–340.
[42] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. 2020. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093* (2020).
[43] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. 2008. Robust face recognition via sparse representation. *IEEE Trans Pattern Anal Mach Intell* 31, 2 (2008), 210–227.
[44] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Towards Open-World Text-Guided Face Image Generation and Manipulation. *arXiv preprint arXiv:2104.08910* (2021).
[45] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. 2017. High-resolution image inpainting using multi-scale neural patch synthesis. In *CVPR*. 6721–6729.
[46] Bangjie Yin, Luan Tran, Haoxiang Li, Xiaohui Shen, and Xiaoming Liu. 2019. Towards interpretable face recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9348–9357.
[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *CVPR*. 5505–5514.
[48] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *ICCV*. 4471–4480.
[49] Lingyun Yu, Hongtao Xie, and Yongdong Zhang. 2021. Multimodal Learning for Temporally Coherent Talking Face Generation with Articulator Synergy. *IEEE Transactions on Multimedia* (2021).
[50] Tao Yu, Zongyu Guo, Xin Jin, Shilin Wu, Zhibo Chen, Weiping Li, Zhizheng Zhang, and Sen Liu. 2020. Region normalization for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 12733–12740.
[51] Xian Zhang, Canghong Shi, Xin Wang, Xi Wu, Xiaojie Li, Jiancheng Lv, and Imran Mumtaz. 2021. Face inpainting based on GAN by facial prediction and fusion as guidance information. *Applied Soft Computing* 111 (2021), 107626.
[52] Fang Zhao, Jiashi Feng, Jian Zhao, Wenhan Yang, and Shuicheng Yan. 2017. Robust lstm-autoencoders for face de-occlusion in the wild. *IEEE Trans Image Process* 27, 2 (2017), 778–790.
[53] Jian Zhao, Yu Cheng, Yan Xu, Lin Xiong, Jianshu Li, Fang Zhao, Karlekar Jayashree, Sugiri Pranata, Shengmei Shen, Junliang Xing, et al. 2018. Towards pose invariant face recognition in the wild. In *CVPR*. 2207–2216.
[54] Jian Zhao, Jianshu Li, Xiaoguang Tu, Fang Zhao, Yuan Xin, Junliang Xing, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. 2019. Multi-prototype networks for unconstrained set-based face recognition. *arXiv preprint arXiv:1902.04755* (2019).
[55] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. 2017. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *NeurIPS*. 66–76.
[56] Tong Zhou, Changxing Ding, Shaowen Lin, Xinchao Wang, and Dacheng Tao. 2020. Learning Oracle Attention for High-fidelity Face Completion. In *CVPR*. 7680–7689.

[57] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. 2018. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing* 11, 3 (2018), 542–552.

[58] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. 2021. WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10492–10502.