



FRNet: Improving Face De-occlusion via Feature Reconstruction

Shanshan Du and Liyan Zhang^(✉)

Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
zhangliyan@nuaa.edu.cn

Abstract. Face de-occlusion is essential to improve the accuracy of face-related tasks. However, most existing methods only focus on single occlusion scenarios, rendering them sub-optimal for multiple occlusions. To alleviate this problem, we propose a novel framework for face de-occlusion called FRNet, which is based on feature reconstruction. The proposed FRNet can automatically detect and remove single or multiple occlusions through the predict-extract-inpaint approach, making it a universal solution to deal with multiple occlusions. In this paper, we propose a two-stage occlusion extractor and a two-stage face generator. The former utilizes the predicted occlusion positions to get coarse occlusion masks which are subsequently fine-tuned by the refinement module to tackle complex occlusion scenarios in the real world. The latter utilizes the predicted face structures to reconstruct global structures, and then uses information from neighboring areas and corresponding features to refine important areas, so as to address the issues of structural deficiencies and feature disharmony in the generated face images. We also introduce a gender-consistency loss and an identity loss to improve the attribute recovery accuracy of images. Furthermore, to address the limitations of existing datasets for face de-occlusion, we introduce a new synthetic face dataset including both single and multiple occlusions, which effectively facilitates the model training. Extensive experimental results demonstrate the superiority of the proposed FRNet compared to state-of-the-art methods.

Keywords: Face de-occlusion · Image inpainting · Deep learning

1 Introduction

Face de-occlusion technologies aim at automatically detecting and removing occlusions, and inpainting the occluded area simultaneously, which generally serve as a preprocessing step to assist other face-related tasks. The main idea of traditional technologies [1, 5] is to inpaint the images according to the existing information. As each part of the face image has its own characteristics,

This work was supported in part by the National Natural Science Foundation of China under Grant 62172212, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20230031.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
Q. Liu et al. (Eds.): PRCV 2023, LNCS 14435, pp. 313–326, 2024.
https://doi.org/10.1007/978-981-99-8552-4_25

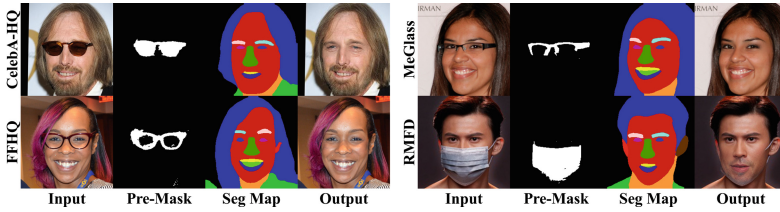


Fig. 1. The results of removing real-world face occlusions on different datasets.

their results are far from satisfactory. To overcome that, various early face de-occlusion technologies based on deep learning [2, 17] have been proposed, making it possible to leverage deep learning techniques to tackle this task. Nevertheless, most of them generate low-resolution images, which may not meet the current demands. Considering the application under real scenes, methods [4, 16, 20, 24, 26] designed for high-resolution face image de-occlusion have been proposed. Specifically, Edgeconnect [16] improves detailed information by introducing prior knowledge. Additionally, CTSDG [4] enhances results by fusing structural and texture features. Furthermore, some methods [20, 24, 26] increase flexibility in handling various-shaped occluded areas through modified convolution mechanisms. These methods have shown promising results in handling face occlusions, however, most of them are not specifically designed for face de-occlusion tasks and usually require manual marking of the occluded area, which can be time-consuming and has certain constraints in practical applications scenarios. Furthermore, there are several face de-occlusion technologies [7] designed for specific occlusions and achieved the expected results. However, they may struggle when generalizing to other types of occlusions that commonly exist in real-world scenarios.

Additionally, face attribute manipulation technologies [11, 13] and image translation technologies [21, 25] can address face de-occlusion to some extent. However, the presence of diverse types of occlusions poses labeling challenges and disrupts the feature extraction process, usually leading to unsatisfactory outcomes.

In this paper, we take inspiration from two research studies. One of these studies [2] focuses on the utilization of an occlusion-aware stage to enhance the effectiveness of face de-occlusion. The other one [16] highlights the advantages of incorporating image structure as a prior to improve image inpainting outcomes. Building upon these findings, we propose a novel framework for face de-occlusion called FRNet, which is based on feature reconstruction. The proposed de-occlusion model consists of a two-stage occlusion extractor, a two-stage face generator, and a face discriminator. To better reconstruct the global structure of the occluded face images, we introduce an Occlusion Robust Face Segmentation Module based on the PP-LiteSeg [18] network, which is utilized to obtain both occlusion location details and face structure information.

Firstly, to effectively tackle complex occlusion scenarios in the real world, the two-stage occlusion extractor utilizes the input images and the occlusion location details to obtain coarse occlusion masks and then refines the masks to acquire the final occlusion masks. **Secondly**, to address the issues of structural deficiencies and feature disharmony in the generated face images, the two-stage face generator is designed with the idea of “first reconstructing the structural features globally, and then refining important area features locally”. In the coarse stage, we adopt the U-Net structure with a large receptive field to reconstruct the global structures of the faces with the guidance of face segmentation maps. As for the refinement stage, we split it into two distinct modules: the Local Areas Refinement Module (LRM) and the Important Areas Refinement Module (IRM). The LRM extracts information from the neighboring areas by a residual network with a small receptive field, thus enhancing the local textures. While the IRM utilizes intra-feature pixel similarity to identify pixels related to the missing pixels from valid pixels of the corresponding feature, then uses them to fill in the occluded area, thereby ensuring feature harmony. We employ an adaptive merging approach to fuse the outputs from the two branches, generating the final refined face images. Meanwhile, to ensure attribute consistency before and after face de-occlusion, assisting face-related tasks, we introduce a gender-consistent loss in the coarse stage and an identity loss in the refinement stage. Both of them encourage the model to pay more attention to attribute features. **Lastly**, to enhance the model’s ability to handle multiple occlusions in a single face image during training, we propose a new synthetic face dataset based on the CelebA-HQ. This dataset includes face images with various types and quantities of occlusions in random states, providing a realistic simulation of common occlusion scenarios in real-life situations, which effectively facilitates the training and supervision of the model. Exemplar results are shown in Fig. 1, our method can effectively remove various types of face occlusions in the wild.

Our contributions can be summarized as follows: (1) We propose a novel face de-occlusion framework, which can automatically detect and remove single or multiple occlusions from the face images, achieving visually realistic results. (2) We propose the idea of “first reconstructing the structural features globally, and then refining important area features locally”, and following this idea, propose a two-stage face generator that can efficiently restore face details and preserve face

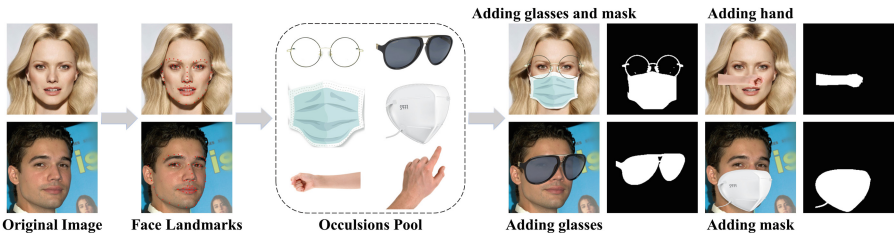


Fig. 2. The process of dataset synthesis. Based on head pose angles and face landmark points, we add occlusions to face images.

attributes. (3) We propose a new dataset dedicated to the face de-occlusion task, containing various types and quantities of occlusions. It plays a crucial role in improving the training of our model. (4) The experimental results demonstrate the good efficacy of our FRNet in eliminating various face occlusions present in the wild while preserving the essential attribute information of the face images.

2 Synthesis of Face Images with Occlusions

The key to face de-occlusion is to remove occlusions accurately while maintaining the attributes such as gender, skin color, and expression consistent with the input image. Therefore, in this research, we require a contrast dataset that includes both occluded and de-occluded face images to supervise the model’s training. However, collecting such a dataset in real life can be challenging, and a more practical solution is to create a dataset with similar characteristics of occluded face images found in the wild. By training on such a dataset, the model can more effectively perceive and remove different types of occlusions and better inpaint the de-occluded face images.

Dataset Preparation. We take the CelebA-HQ as our face image dataset. In addition, we collect 362 glasses images, 324 mask images, and 1,000 hand images as the occlusion data.

Dataset Synthesis. As shown in Fig. 2, we first use the attributes annotations of CelebA-HQ to screen out the occlusion-free face images. Then, we use dlib and OpenCV to get the face pose angles of the face images. According to the pose angles, the face images are divided into eight groups: front looking up, front

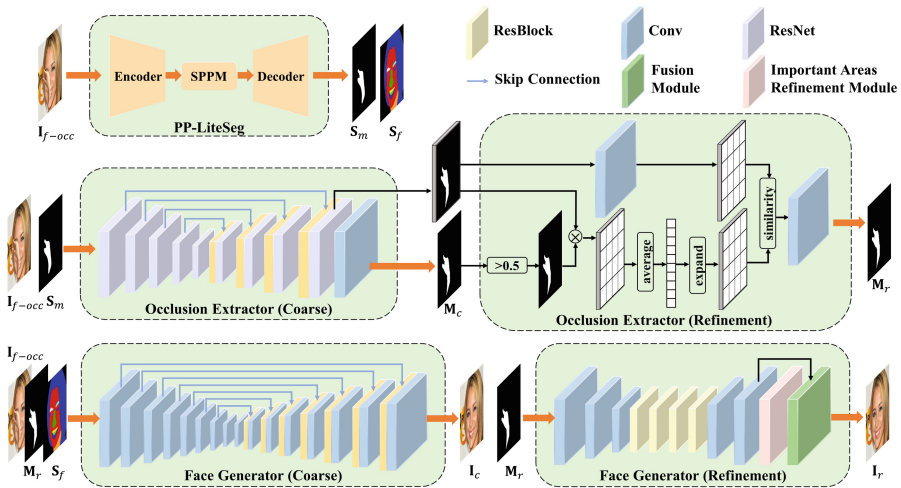


Fig. 3. The overall architecture of FRNet. It has three stages: structure prediction, occlusion extraction, and face inpainting.

looking down, the left deflection angles ranging from 10° to 40° , 40° to 60° , above 60° , and the right deflection angles ranging from 10° to 40° , 40° to 60° , and above 60° . We also classify glasses images and mask images based on the same angle ranges. Subsequently, we use face alignment to extract face features and get 68 face landmark points. Lastly, we combine the occlusion-free face images with glasses and masks based on the head poses and face landmarks. To enhance the realism of the synthetic images, we randomly vary the transparency of the glasses. To simulate real-life scenarios with various types and quantities of occlusions, we randomly add hand occlusions to some face images and both glasses and masks to others, reflecting the complexity of occlusions that may occur in real life. Following the original division of the CelebA, we collect 88,932 images as the training dataset, 11,120 images as the validation dataset, and 10,244 images as the testing dataset. Each set of data includes an original face image, an occluded face image, an occlusion image, and a binary occlusion mask image.

3 FRNet

In this section, we will introduce the overall architecture of FRNet, which is shown in Fig. 3. It has three stages: structure prediction, occlusion extraction, and face inpainting. The structure prediction stage aims to obtain prior information through the segmentation network. The occlusion extraction stage obtains the occlusion mask from the occluded face image with the guidance of occlusion position information. And face inpainting stage reconstructs the global structure with the guidance of face segment image, and refines the features by local information and feature information, to obtain an occlusion-free face image.

3.1 Occlusion Robust Face Segmentation Module

To better remove the occlusions and restore face structure, we propose the Occlusion Robust Face Segmentation Module, which aims to predict the occlusion segmentation map \mathbf{S}_m and the face segmentation map \mathbf{S}_f from the occluded face image \mathbf{I}_{f-occ} . In this paper, we leverage the PP-LiteSeg [18] to realize this module (Fig. 3). To train the module, we fuse the CelebAMask-HQ with occlusion mask images. During the face de-occlusion process, this module provides prior information about the position of occlusion and face structure, which guides occlusion extraction and face reconstruction. This approach enables our model to focus on the deep structural feature of the face.

3.2 Occlusion Extractor

The occlusion extractor consists of a coarse stage and a refinement stage, in which input occluded face image \mathbf{I}_{f-occ} and the prior information of occlusion position \mathbf{S}_m , output coarse occlusion mask \mathbf{M}_c and refinement occlusion mask \mathbf{M}_r .

In the coarse stage, we build it based on the U-Net structure and refer to the details of the architecture from [6]. The architectural detail is shown in Fig. 3. Specifically, the encoder firstly increases the number of channels of the feature map to twice by 3×3 Conv-BN-ReLU module. This is followed by using three residual blocks to increase the receptive field of the module and further extract features. Finally, the spatial size of the feature map is reduced by down-sampling through the max pooling layer. The decoder performs the reverse operation. Firstly, the feature map is up-sampled by bilinear interpolation to expand its spatial size. Then the number of channels is reduced by half through the 3×3 Conv-BN-ReLU module, and the result is connected with the output feature map of the corresponding encoder using the skip connection to obtain the concatenated feature map. The concatenated feature map is passed through another 3×3 Conv-BN-ReLU module to reduce the number of channels by half. Finally, the latest feature map passes through three residual blocks to get the new feature map. The output of the last decoder is passed through 1×1 convolution and the sigmoid activation function to obtain the coarse occlusion mask \mathbf{M}_c . We optimize it using binary cross-entropy loss:

$$\mathcal{L}_{BCE}^C = - \sum_{i,j} \left(\mathbf{M}_{gt_{i,j}} \log \mathbf{M}_{c_{i,j}} + (1 - \mathbf{M}_{gt_{i,j}}) \log (1 - \mathbf{M}_{c_{i,j}}) \right), \quad (1)$$

where $\mathbf{M}_{gt_{i,j}}$ (*resp.*, $\mathbf{M}_{c_{i,j}}$) is the (i, j) -th entry in \mathbf{M}_{gt} (*resp.*, \mathbf{M}_c).

In the refinement stage, we refer to the Self-calibrated Mask Refinement proposed in [12]. As shown in Fig. 3, we conduct similarity matching between the main feature of the coarse occlusion mask \mathbf{M}_c and features at other positions, thereby fine-tuning the occlusion mask according to its characteristics and enhancing the performance of the occlusion extractor. Through this stage, we obtain the refined occlusion mask \mathbf{M}_r . Similarly, we use binary cross-entropy loss \mathcal{L}_{BCE}^R , which is the same as Eq. (1) except for replacing $\mathbf{M}_{c_{i,j}}$ with $\mathbf{M}_{r_{i,j}}$ in Eq. (1). Additionally, we employ Intersection over Union (IoU) loss. It is defined as:

$$\mathcal{L}_{IoU}^R = \frac{\sum_{i,j} (\mathbf{M}_{gt_{i,j}} \cdot \mathbf{M}_{r_{i,j}})}{\sum_{i,j} (\mathbf{M}_{gt_{i,j}} + \mathbf{M}_{r_{i,j}} - \mathbf{M}_{gt_{i,j}} \cdot \mathbf{M}_{r_{i,j}})}, \quad (2)$$

where $\mathbf{M}_{gt_{i,j}}$ (*resp.*, $\mathbf{M}_{r_{i,j}}$) is the (i, j) -th entry in \mathbf{M}_{gt} (*resp.*, \mathbf{M}_r).

Finally, the total loss of the occlusion extractor is

$$\mathcal{L}_{mask} = \lambda_1 \cdot \mathcal{L}_{BCE}^C + \lambda_2 \cdot \mathcal{L}_{BCE}^R + \lambda_3 \cdot \mathcal{L}_{IoU}^R. \quad (3)$$

We set $\lambda_1 = 0.01$, $\lambda_2 = 1$, $\lambda_3 = 0.25$ in the experiment.

3.3 Face Generator

Coarse Face Generator

We use the occluded face image \mathbf{I}_{f-occ} and the predicted occlusion mask \mathbf{M}_r as the input. Moreover, the predicted face segmentation map \mathbf{S}_f is introduced

to facilitate the face generation process. Then, the model returns a face image where the occluded area is reconstructed.

Following the [19], the coarse face generator (Fig. 3) uses a U-Net architecture with skip connections and has a large receptive field, which can better pay attention to the global feature of the image. Specifically, the coarse face generator consists of eight encoder-decoder blocks. Each encoder down-samples the feature map through the convolution layer, increasing its number of channels to twice while decreasing its size. The decoder up-samples the feature map to the original size through the transposed convolution layer to obtain coarse face image \mathbf{I}_{out}^c . According to the mask image \mathbf{M}_r , the input image \mathbf{I}_{f-occ} and the generated face image \mathbf{I}_{out}^c are combined to obtain a merged face image \mathbf{I}_{mer}^c ,

$$\mathbf{I}_{mer}^c = \mathbf{I}_{f-occ} \odot (\mathbf{1} - \mathbf{M}_r) + \mathbf{I}_{out}^c \odot \mathbf{M}_r, \quad (4)$$

where \odot is the element-wise product operation.

At this stage, we use the weighted L_1 loss [19] as the pixel-wise reconstruction loss which is defined as:

$$\mathcal{L}_{valid}^C = \frac{1}{\text{sum}(\mathbf{1} - \mathbf{M}_r)} \|(\mathbf{I}_{out}^c - \mathbf{I}_{gt}) \odot (\mathbf{1} - \mathbf{M}_r)\|_1, \quad (5)$$

$$\mathcal{L}_{occluded}^C = \frac{1}{\text{sum}(\mathbf{M}_r)} \|(\mathbf{I}_{out}^c - \mathbf{I}_{gt}) \odot (\mathbf{M}_r)\|_1, \quad (6)$$

$$\mathcal{L}_1^C = \mathcal{L}_{valid}^C + \lambda_o \cdot \mathcal{L}_{occluded}^C, \quad (7)$$

where $\text{sum}(\mathbf{M}_r)$ (*resp.*, $\text{sum}(\mathbf{1} - \mathbf{M}_r)$) represents the number of non-zero elements in \mathbf{M}_r (*resp.*, $\mathbf{1} - \mathbf{M}_r$). And we set $\lambda_o = 6$.

Meanwhile, we add a patch-based discriminator with spectral normalization [14] and use the least square loss as the adversarial loss. It is defined as:

$$\mathcal{L}_G^C = \mathbb{E}_{\mathbf{I}_{mer} \sim p_{\mathbf{I}_{mer}}(\mathbf{I}_{mer})} \left[(D(\mathbf{I}_{mer}^c) - 1)^2 \right], \quad (8)$$

$$\mathcal{L}_D = \frac{1}{2} \mathbb{E}_{\mathbf{I} \sim p_{data}(\mathbf{I})} \left[(D(\mathbf{I}_{gt}) - 1)^2 \right] + \frac{1}{2} \mathbb{E}_{\mathbf{I}_{mer} \sim p_{\mathbf{I}_{mer}}(\mathbf{I}_{mer})} \left[(D(\mathbf{I}_{mer}^c))^2 \right]. \quad (9)$$

The main challenge of face de-occlusion is to maintain the face attributes before and after de-occlusion, which is crucial for supporting other face-related tasks. To address this, we introduce the gender-consistency loss. We obtain a gender classification model with an accuracy rate of 99.096%, by using the CelebA-HQ to perform transfer learning on the VGG-16. In the training process, the gender \mathbf{P}_{gen}^{gt} of the ground-truth \mathbf{I}_{gt} is used as the target, and the generated face image \mathbf{I}_{mer}^c is classified by using the classification model to obtain the classification result \mathbf{P}_{gen}^c . Finally, we calculate the cross entropy between the classification result and the target. So the gender-consistency loss is defined as:

$$\mathcal{L}_{gen}^C = -(\mathbf{P}_{gen}^{gt} \log \mathbf{P}_{gen}^c + (1 - \mathbf{P}_{gen}^{gt}) \log (1 - \mathbf{P}_{gen}^c)), \quad (10)$$

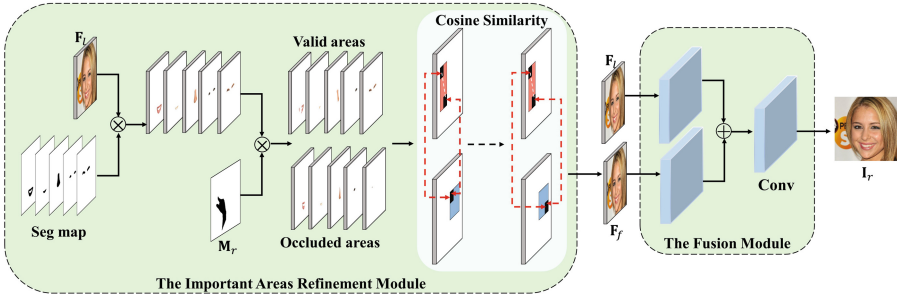


Fig. 4. The architectures of the Important Areas Refinement Module (IRM) and the Fusion Module.

where \mathbf{P}_{gen}^{gt} (*resp.*, \mathbf{P}_{gen}^c) represents the probability that the image \mathbf{I}_{gt} (*resp.*, \mathbf{I}_{mer}^c) belongs to male. Finally, the total loss for the coarse face generator is

$$\mathcal{L}_{face}^C = \mathcal{L}_1^C + \lambda_G \cdot \mathcal{L}_G^C + \lambda_{gen} \cdot \mathcal{L}_{gen}^C, \quad (11)$$

and we set $\lambda_G = 0.1$, $\lambda_{gen} = 1$.

Refinement Face Generator

Local Areas Refinement Module (LRM). As shown in Fig. 3, the main structure of the LRM is similar to the model proposed in [19]. Specifically, it consists of two down-sampled encoders, four residual blocks, and two up-sampled decoders. Besides, the feature map is padded using the reflection of its boundary before down-sampling and after up-sampling, which makes the edge information better preserved during the convolution processes. We use a shallow neural network to obtain information on surrounding pixels to locally refine the missing areas. In practice, we feed the coarse face image \mathbf{I}_{mer}^c and the predicted occlusion mask \mathbf{M}_r into this module to get the local refinement feature map \mathbf{F}_l .

Important Areas Refinement Module (IRM). By exploring the similarity of pixels in the same face feature, we propose the IRM, which uses similarity to find valid pixels related to missing areas. This module maintains harmony within the feature and achieves the goal of feature reconstruction. For example, when only one eye is occluded, we can use the features in the occlusion-free eye to inpaint the occluded one, ensuring consistency in features such as the eyeball color. As shown in Fig. 4, we feed the feature map \mathbf{F}_l into the IRM and get the segmentation map of the rough face image \mathbf{I}_{mer}^c by using the face segmentation network, then use the segmentation map to get five important areas (brow, eye, nose, lip, mouth) of the feature map. If the current important area contains both a valid area and an occluded area, we calculate cosine similarity between feature points in the valid area and those in the occluded area. Then using the most similar features to fill each feature point in the occluded area. If the current important area is all occluded area, we use the prediction information of LRM to fill them. If the current important area is all valid area, their feature points

will not be changed. Through this module, we can acquire the feature map \mathbf{F}_f of feature refinement.

Fusion Module. The LRM inpaints the occluded area using surrounding pixels through the shallow neural network, which enhances the local details. However, the features of face images have particularities, with specific correlations existing between different local areas. For example, the local features of the two eyes should exhibit similarity. However, it is difficult to obtain information from the valid area related to the occluded area in this module, leading to the output result may be discordant within the features. The IRM inpaints the occluded area using valid pixels within the features by calculating the cosine similarity, ensuring harmony within the features. But it cannot inpaint areas without valid information and has difficulty handling images with the large occluded area. Therefore, we adaptively fuse the local refined feature map \mathbf{F}_l with the feature refined feature map \mathbf{F}_f , and then up-sample them to the size of the input image to get a refined face image \mathbf{I}_{out}^r and a merged face image \mathbf{I}_{mer}^r ,

$$\mathbf{I}_{mer}^r = \mathbf{I}_{f-occ} \odot (\mathbf{1} - \mathbf{M}_r) + \mathbf{I}_{out}^r \odot \mathbf{M}_r. \quad (12)$$

Like the coarse stage, we use the weighted L_1 loss as the pixel-wise reconstruction loss \mathcal{L}_1^R , which is the same as Eq. (7) except for replacing \mathbf{I}_{out}^c with \mathbf{I}_{out}^r in Eq. (5) and Eq. (6). Meanwhile, following [19], we apply perceptual loss and style loss to the model using the VGG-16 which is pre-trained based on ImageNet. The perceptual loss is defined as:

$$\mathcal{L}_{per}^R = \sum_i \|\phi_i(\mathbf{I}_{out}^r) - \phi_i(\mathbf{I}_{gt})\|_1 + \|\phi_i(\mathbf{I}_{mer}^r) - \phi_i(\mathbf{I}_{gt})\|_1, \quad (13)$$

where ϕ_i means the feature map of i -th layer in pre-trained VGG-16 network ($i \in \{5, 10, 17\}$).

The style loss is defined as:

$$\mathcal{L}_{sty}^R = \sum_i \|\mathcal{G}_i(\mathbf{I}_{out}^r) - \mathcal{G}_i(\mathbf{I}_{gt})\|_1 + \|\mathcal{G}_i(\mathbf{I}_{mer}^r) - \mathcal{G}_i(\mathbf{I}_{gt})\|_1, \quad (14)$$

where $\mathcal{G}_i(\cdot) = \phi_i(\cdot) \phi_i(\cdot)^T$ is the Gram matrix.

Adding total variation (TV) loss as the smoothing penalty, defined as:

$$\mathcal{L}_{tv}^R = \|\mathbf{I}_{mer}^r(i, j+1) - \mathbf{I}_{mer}^r(i, j)\|_1 + \|\mathbf{I}_{mer}^r(i+1, j) - \mathbf{I}_{mer}^r(i, j)\|_1. \quad (15)$$

Like previous work [8], an identity loss is defined on the face recognition network. It is defined as:

$$\mathcal{L}_{id}^R = 1 - \frac{\mathbf{F}_{id}^{gt} \cdot \mathbf{F}_{id}^r}{\max(\|\mathbf{F}_{id}^{gt}\|_2 \cdot \|\mathbf{F}_{id}^r\|_2, \epsilon)}, \quad (16)$$

where \mathbf{F}_{id}^{gt} and \mathbf{F}_{id}^r are the output vectors of the face recognition network for \mathbf{I}_{gt} and \mathbf{I}_{mer}^r . Respectively, ϵ sets to very small values 1e-8.

Finally, the total loss for the refinement face generator is

$$\mathcal{L}_{face}^R = \mathcal{L}_1^R + \lambda_{per} \cdot \mathcal{L}_{per}^R + \lambda_{sty} \cdot \mathcal{L}_{sty}^R + \lambda_{tv} \cdot \mathcal{L}_{tv}^R + \lambda_{id} \cdot \mathcal{L}_{id}^R, \quad (17)$$

and we set $\lambda_{per} = 0.05$, $\lambda_{sty} = 120$, $\lambda_{tv} = 0.1$, $\lambda_{id} = 1$.

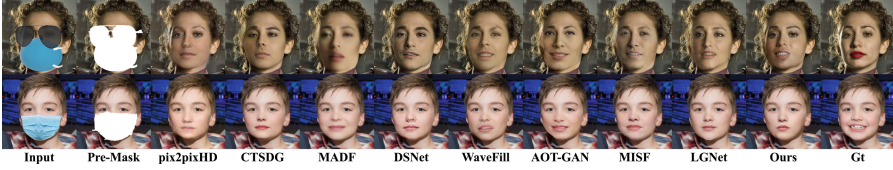


Fig. 5. Qualitative results on our synthetic dataset (top) and Gender Occlusion Data (bottom).

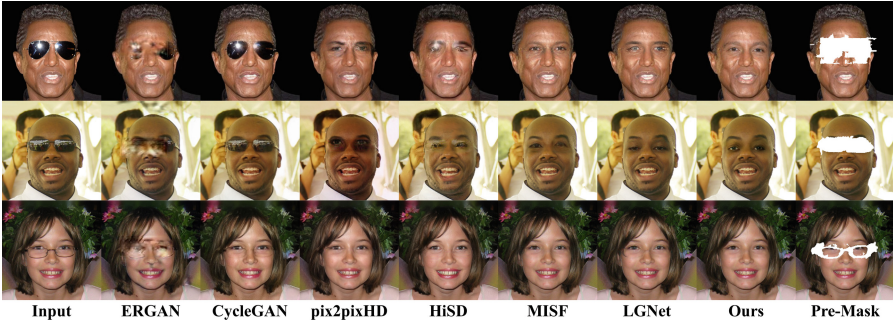


Fig. 6. Qualitative results on CelebA-HQ (top), FFHQ (middle), and MeGlass (bottom).

4 Experiments

4.1 Experimental Settings

Datasets. We train the FRNet using the synthetic dataset proposed in Sect. 2. Furthermore, we also use real-world portrait datasets including CelebA-HQ, FFHQ [9], MeGlass [3], RMFD [22], and the masked face synthesis dataset Gender Occlusion Data [15] to test the model.

Implementation Details. Our method is implemented with PyTorch 1.7.0 using a 24G NVIDIA GTX3090 GPU. And we train the model by the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. In practice, we first train the occlusion extractor for 10 epochs, then fix the parameters of the occlusion extractor and train the face generator for 40 epochs. We set the learning rate to 0.0002 when training the occlusion extractor. As for the training of the face generator, we set the learning rate to 0.0002 for the first 20 epochs and linearly decay it to zero for the next 20 epochs.

4.2 Comparisons with State-of-the-Art Methods

We compare our method with state-of-the-art image inpainting methods including CTSDG [4], MADF [26], DSNet [20], WaveFill [23], AOT-GAN [24], MISF [10], LGNet [19], image translation methods including CycleGAN [25], pix2pixHD [21], and glasses removal methods including ERGAN [7], HiSD [11].

Qualitative Comparison. As shown in Fig. 5, Fig. 6 and Fig. 8, our method can restore the global structures and details of the faces more effectively, while maintaining consistency within the features. Furthermore, our model can be directly extended to other datasets, enabling the identification and removal of multiple occlusions while producing visually realistic results without retraining. (More experimental results can be found in <https://github.com/dss9964/FRNet>.)

Table 1. Quantitative results on our synthetic dataset and Gender Occlusion Data. The best two results are shown in red and blue respectively.

Model	Our Synthetic Dataset			Gender Occlusion Data		
	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑
pix2pixHD	5.31	0.112	0.843	11.95	0.157	0.803
CTSDG	2.06	0.046	0.938	12.20	0.139	0.853
MADF	3.26	0.049	0.936	14.87	0.137	0.850
DSNet	1.94	0.043	0.939	12.08	0.137	0.855
WaveFill	2.05	0.044	0.935	14.59	0.135	0.854
AOT-GAN	2.13	0.043	0.933	11.55	0.135	0.855
MISF	1.77	0.041	0.942	11.78	0.134	0.858
LGNet	1.87	0.042	0.942	15.49	0.137	0.856
Ours	1.85	0.039	0.945	10.50	0.132	0.858

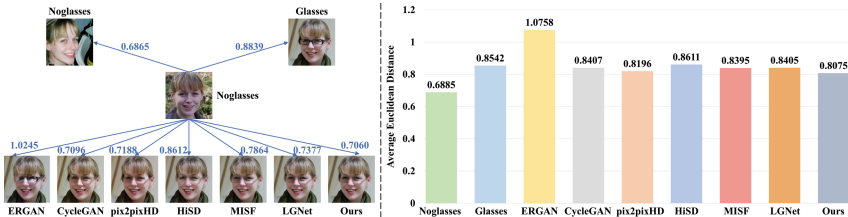


Fig. 7. The results of identity preservation.

Quantitative Comparison

Realism. As shown in Table 1, our method is comparable to MISF and LGNet on the proposed synthetic dataset and achieves the best results on the Gender Occlusion Data. These results indicate that our model possesses better generalization performance and a stronger ability to handle large-scale occlusions.

Identity Preservation. To demonstrate the positive impact of de-occlusion on face recognition, we collected 1,032 sets of face images from MeGlass. Each

set consisted of two images without glasses and one image with glasses of the same identity. Various occlusion removal methods were applied to the images with glasses to generate corresponding images without glasses. The Euclidean Distance between the first image without glasses and all other images was then calculated. As shown in Fig. 7, the presence of occlusions significantly increased the Average Euclidean Distance between the occluded face images and the target images. However, our method effectively preserves the identity information of the face images, minimizing the Euclidean Distance. It mitigated the detrimental effects of occlusion to a certain extent, ultimately enhancing the accuracy of face recognition.

4.3 Ablation Studies

In this subsection, we evaluate the performance of our key contributions in occlusion extraction and face inpainting.

Occlusion Extraction. MS and MR represent occlusion location information and the refinement stage of the occlusion extractor respectively. Table 2 shows that they both have positive effects on occlusion extraction.

Face Inpainting. \mathcal{L}_{gen} , \mathcal{L}_{id} , and FS represent the gender loss, the id loss, and the face structure information respectively. GFNet represents moving the IRM to the coarse stage. As shown in Table 2, the network with all proposed modules achieves the best performance, indicating the effectiveness of our proposed strategy and loss functions. Figure 9 demonstrates that the quality of images generated by w/o FS is the worst, with noticeable blurriness. The images generated by w/o FR exhibit internal feature inconsistency, such as the eyes in the 2nd row. w/o \mathcal{L}_{gen} & \mathcal{L}_{id} and w/o GF can effectively inpainting the images, but lack detail in some areas. These qualitative results also reflect the advantages of the FRNet.



Fig. 8. Qualitative results on RMFD.



Fig. 9. Qualitative comparison of different ablations in face inpainting on our synthetic dataset.

Table 2. Quantitative comparison of different ablations in occlusion extraction (top) and face inpainting (bottom) on our synthetic dataset.

Model	FID↓	LPIPS↓	SSIM↑
w/o MR	3.71	0.018	0.987
w/o MS	2.32	0.015	0.987
Ours	2.26	0.014	0.990
w/o IRM	1.94	0.041	0.942
w/o \mathcal{L}_{gen} & \mathcal{L}_{id}	1.89	0.041	0.943
w/o FS	2.17	0.042	0.944
GFNet	1.90	0.041	0.944
Ours	1.85	0.039	0.945

5 Conclusion

In this paper, we propose a new face de-occlusion framework based on feature reconstruction (FRNet), which consists of three stages: structure prediction, occlusion extraction, and face inpainting. In the inpainting process, the global structures are reconstructed with the guidance of the face segment images, and the important areas are refined by local information and feature information. Besides, we build a high-quality synthetic occluded face dataset, which provides supervision for the training of models. Qualitative and quantitative experiment results demonstrate that our model can effectively remove face occlusions and retain attribute information to support face-related tasks.

References

1. Barnes, C., et al.: The patchmatch randomized matching algorithm for image manipulation. *Commun. ACM* **54**(11), 103–110 (2011)
2. Dong, J., et al.: Occlusion-aware GAN for face de-occlusion in the wild. In: ICME, pp. 1–6. IEEE (2020)
3. Guo, J., Zhu, X., Lei, Z., Li, S.Z.: Face synthesis for eyeglass-robust face recognition. In: Zhou, J., et al. (eds.) CCBR 2018. LNCS, vol. 10996, pp. 275–284. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-97909-0_30
4. Guo, X., et al.: Image inpainting via conditional texture and structure dual generation. In: ICCV, pp. 14134–14143 (2021)
5. He, K., et al.: Computing nearest-neighbor fields via propagation-assisted KD-trees. In: CVPR, pp. 111–118. IEEE (2012)
6. Hertz, A., et al.: Blind visual motif removal from a single image. In: CVPR, pp. 6858–6867 (2019)
7. Hu, B., et al.: Unsupervised eyeglasses removal in the wild. *TCYB* **51**(9), 4373–4385 (2020)
8. Ju, Y.J., et al.: Complete face recovery GAN: unsupervised joint face rotation and de-occlusion from a single-view image. In: WACV, pp. 3711–3721 (2022)
9. Karras, T., et al.: A style-based generator architecture for generative adversarial networks. In: CVPR, pp. 4401–4410 (2019)
10. Li, X., et al.: MISF: multi-level interactive siamese filtering for high-fidelity image inpainting. In: CVPR, pp. 1869–1878 (2022)
11. Li, X., et al.: Image-to-image translation via hierarchical style disentanglement. In: CVPR, pp. 8639–8648 (2021)
12. Liang, J., et al.: Visible watermark removal via self-calibrated localization and background refinement. In: ACM MM, pp. 4426–4434 (2021)
13. Liu, M., et al.: STGAN: a unified selective transfer network for arbitrary image attribute editing. In: CVPR, pp. 3673–3682 (2019)
14. Miyato, T., et al.: Spectral normalization for generative adversarial networks. arXiv preprint [arXiv:1802.05957](https://arxiv.org/abs/1802.05957) (2018)
15. Modak, G., et al.: A deep learning framework to reconstruct face under mask. In: CDMA, pp. 200–205. IEEE (2022)
16. Nazeri, K., et al.: Edgeconnect: structure guided image inpainting using edge prediction. In: ICCV (2019)
17. Pathak, D., et al.: Context encoders: feature learning by inpainting. In: CVPR, pp. 2536–2544 (2016)

18. Peng, J., et al.: PP-LiteSeg: a superior real-time semantic segmentation model. arXiv preprint [arXiv:2204.02681](https://arxiv.org/abs/2204.02681) (2022)
19. Quan, W., et al.: Image inpainting with local and global refinement. *TIP* **31**, 2405–2420 (2022)
20. Wang, N., et al.: Dynamic selection network for image inpainting. *TIP* **30**, 1784–1798 (2021)
21. Wang, T.C., et al.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: *CVPR*, pp. 8798–8807 (2018)
22. Wang, Z., et al.: Masked face recognition dataset and application. arXiv preprint [arXiv:2003.09093](https://arxiv.org/abs/2003.09093) (2020)
23. Yu, Y., et al.: Wavefill: a wavelet-based generation network for image inpainting. In: *ICCV*, pp. 14114–14123 (2021)
24. Zeng, Y., et al.: Aggregated contextual transformations for high-resolution image inpainting. *TVCG* (2022)
25. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *ICCV*, pp. 2223–2232 (2017)
26. Zhu, M., et al.: Image inpainting by end-to-end cascaded refinement with mask awareness. *TIP* **30**, 4855–4866 (2021)