Neural Processing Letters
https://doi.org/10.1007/s11063-022-10861-2

Learning Unoccluded Face Texture Completion from Single Image in the Wild

Yongtang Bao[1] · Pengfei Zhou[1] · Peng Zhang[1] · Yue Qi[2,3]

Accepted: 20 April 2022

Abstract
In recent years, single tasks such as face frontalization, image inpainting, and glasses removal have improved face de-occlusion. However, there is little work on joint learning of multiple de-occlusion tasks. To achieve multi-task learning, we propose an unoccluded face synthesis (UFS) framework for multi-tasks such as face frontalization, image inpainting, and glasses removal, which can remove glasses, face self-occlusion, and external occlude. Our UFS framework consists of an encoder, an image reconstruction module, a decoder, and an image discriminator. First, Gaussian random noise extracts high-dimensional features from images in the encoder module. Next, the image reconstruction module includes multi-scale feature fusion, residual hole block, and self-attention network. As a result, it can strengthen the learning of multi-level fine-grained features and achieve better results in face restoration and face frontalization tasks. Then, we synthesize unoccluded face textures from multi-level fine-grained elements in the decoder. Finally, the image discriminator learns the global information structure of the synthesized image, preventing problems such as distortion and blurring of the picture. Experiments show that our UFS framework can achieve better results on single tasks such as face frontalization, image inpainting, and glasses removal. It also can obtain acceptable results on multiple tasks such as face frontalization and glasses removal simultaneously.

✉ Yongtang Bao
baozi0221@sdust.edu.cn

✉ Yue Qi
qy@buaa.edu.cn

Pengfei Zhou
pengfeiz96@163.com

Peng Zhang
pengzhang_skd@sdust.edu.cn

1 College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

2 State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

3 Virtual Reality Research Institute, Beihang University Qingdao Research Institute, Qingdao 266100, China

Published online: 15 May 2022

Keywords Multi-task learning · Face frontalization · Image inpainting · Face synthesis

1 Introduction

Due to the occlusion of faces and other objects, it isn't easy to find full-view unoccluded face textures in photos. In the wild, non-frontal face images are prevalent. People often pose various gestures poses and wear different glasses when taking pictures. These images often have various occlusions, challenging to synthesize unoccluded face textures. Although the PCA-based methods [1–3] can remove other face occlusions, it relies heavily on the prior knowledge of face scanning, resulting in the loss of many local details in the synthesized face image. As a result, it cannot guarantee the high fidelity of the synthesized face. Tasks such as face frontalization and face removal have made breakthroughs in occlusion removal, but it is still impossible to remove multiple occlusions. Although the various occlusions of the face can be removed by connecting numerous networks, the high computational cost and excessive resource consumption lead to training difficulties.

In response to the above problems, this paper proposes an unoccluded face synthesis (UFS) frameworka that combines face frontalization, image inpainting, and face glasses removal to remove face occlusion. Our framework can generate unoccluded face textures using only one network, reducing network overhead. The UFS framework includes four parts such as an encoder module, an image reconstruction module, a decoder module, and an image discriminator module. The encoder and decoder are used to synthesize unoccluded face textures. The image reconstruction module adopts a multi-scale feature fusion module, a hole residual block module, and a self-attention network module. Multi-scale feature fusion extracts features of different scales and global features from the encoder. The residual hole block expands the receptive field and performs feature fusion with multi-scale feature fusion to obtain fine-grained features. The self-attention network generates spatial attention maps for fine-grained features and enhances face region learning. Finally, multi-level fine-grained features are generated through the momentum formula. The image discriminator module captures the global information structure, making the synthesized images more realistic.

In summary, this study makes the following contributions to existing literature:

- We propose a UFS framework that combines multiple tasks of face frontalization, image inpainting, and glasses removal to synthesize unoccluded face textures.
- Our proposed framework jointly employs an image reconstruction module with multi-scale feature fusion, an atrous residual block, and an attention network to obtain multi-level fine-grained feature representations. The image discriminator learns the overall structure of the face, and the synthesized face is more realistic. As a result, we can achieve better visual results on multiple tasks such as face frontalization, image inpainting, and glasses removal.
- We adopt the high-frequency focal loss to make the synthesized image closer to the target image in frequency, recover the high-frequency details that are difficult to synthesize and make the picture is more realistic. We also construct a face dataset and train the network to synthesize transparent face textures without occlusion.

2 Related Work

2.1 Face Frontalization

Face frontalization aims to synthesize the frontal face image from the profile image. Hassner et al. [4] employed a mean 3D face model to recover the frontalized face image in the traditional method. Still, this method ignores the texture details, resulting in a severe loss of texture details of faces. With the development of deep learning, many GAN-based approaches [5, 6] have been present. Tran et al. [7] proposed a disentangled representation learning generative adversarial network (DR-GAN) to decouple various pose information to synthetic frontal faces. However, this method has poor generalization ability and is prone to image degradation in unconstrained environments. Two-pathway generative adversarial network (TP-GAN) was proposed in [8] to apply two-pathway GAN to learn global face structure and local area structure, preventing distortion and artifacts during image frontalization. This method provides a new idea for face frontalization, and most of the subsequent face frontalization work employs the global and local overall structure. However, this method still cannot eliminate the limitation of the environment, and the effect is poor in the unconstrained environment. Zhao et al. [9] presented a pose invariant model (PIM) using a domain adaptation strategy for face recognition in extreme poses. The Dual-attention generative adversarial network (DA-GAN) [10] was proposed to adopt a self-attention mechanism to learn rich feature representations and preserve identity consistency. Still, the frontal face images generated by this method cannot guarantee the consistency of illumination information. In response to the above problems, we adopt an image reconstruction module in the UFS framework to recover the face's local details and maintain the illumination information's consistency. The image reconstruction module employs multi-scale features to perceive illumination information and uses a self-attention network to enhance face regions' learning.

2.2 Image Completion

Unocclusion face synthesis is also an image completion problem. Recently, image completion has become a research hotspot in computer vision. Some deep learning-based methods have been applied to image completion. Context-encoder in [11] was the first to propose using GAN for image inpainting. It jointed reconstruction and adversarial loss to repair the structure of missing regions. Iizuka et al. [12] presented atrous convolution to enlarge the receptive field. To address the irregular mask problem, Liu et al. [13] proposed partial convolution to perform convolution operations on valid pixels to reduce the difference between repaired and entire regions. Joint global and local discriminator methods [3] were proposed to improve the image quality. The global discriminator learns the global information structure. The local discriminator learns the information of the local image to ensure the consistency of the context information. Huang et al. [14] presented a range scaling global U-Net, which used global features in the U-Net framework to reduce visual artifacts and realize the transformation from low-quality images to high-quality images. We apply the image discriminator in the UFS framework to learn the global information of the picture. We use atrous residual blocks to enlarge the receptive field and employ a spatial self-attention network in the image reconstruction module to enhance the learning of local face regions. As a result, we achieve good results on image inpainting tasks.

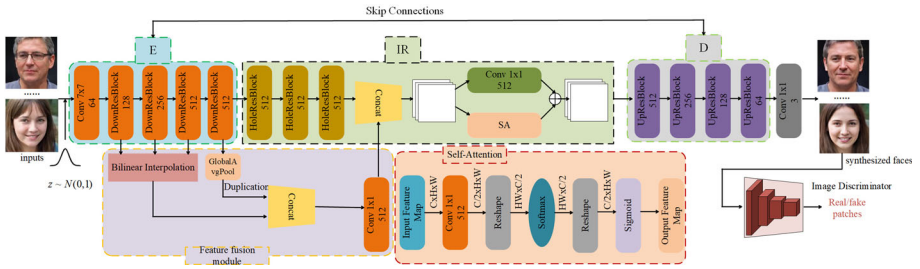


Fig. 1 Our proposed UFS framework consists of an encoder module, a decoder module, an image reconstruction module, and an image discriminator. An encoder-decoder module is used to generate unoccluded face textures. The image reconstruction module can obtain feature representations with different semantics to recover texture details. The feature fusion module is used to get fine-grained features, and the self-attention network is used to enhance the learning of face details in the image reconstruction module. Finally, the image discriminator contains global structural information to improve the quality of face generation

2.3 Face Attributes Editing

Face attribute editing is a research direction that has received much attention. People have proposed deep learning-based face attribute manipulation methods. Xiao et al. [15] presented an elegant to decouple multiple attributes of a face and implement multi-attribute manipulation based on reference images. However, the generalization ability of this method is poor. Moreover, the effect of removing glasses and bangs on the face is not so good. Ergan et al. [16] pulled the occlusion of glasses by transforming the face encoder and the glasses region encoder. Cyclegan et al. [17] proposed an unsupervised way to achieve image translation, in which face glasses and bangs are removed based on reference images. Pix2pix in [18] offered a unified framework to solve the image translation problem. However, these methods still cannot altogether remove the occlusion of the bangs of the glasses. Li et al. [19] proposed hierarchical style disentanglement (HiSD) to achieve independent separation of single-attribute or multiple-attribute removal. This method removes glasses and bangs altogether but ignores the image quality, resulting in low image synthesis quality and blurring and artifacts. We constructed a paired face dataset to eliminate the occlusion of glasses and cracks. We learned the unoccluded face texture in a supervised way, removing the occlusion of glasses and bangs and realizing the editing of face attributes.

In conclusion, we propose a UFS framework to achieve multi-task removal of face occlusion and perform cooperative learning for multi-tasks such as face frontalization, face image inpainting, and face attribute editing. The image reconstruction module we propose realizes the generation of facial texture details and solves the inconsistency between artifacts and illumination information generated by the face frontalization process. We implement image inpainting using atrous residual blocks and self-attention networks.

3 Method

In previous work, tasks such as face frontalization and face de-glassing worked as separate tasks, and few jobs jointly learned these tasks. Therefore, we propose combining multiple tasks such as face frontalization and face de-glasses to generate unoccluded face textures. To achieve multi-task learning, we present the UFS framework. Our framework is outlined in Fig. 1, which contains an encoder module (E), an image reconstruction module (IR), a

decoder module (D), and an image discriminator module. The encoder module extracts rich feature representations for different types of face images (non-frontal, wearing glasses, with bangs, etc.). The image reconstruction module first employs a hole residual block to enlarge the receptive field and interpolates to fuse the multiscale feature from the downsampling residual block of the encoder in the feature fusion module. Then the self-attention network is used to obtain the spatial attention map. Finally, the momentum formula is used to improve the quality of face reconstruction. The decoder module recovers the high-frequency information of the face using the up-sampling residual block and the skip connection with the encoder, which further improves the face generation ability. Finally, the image discriminator is used to learn the overall structure of the face texture and mitigate problems such as distortion and artifacts in the generated face. Detailed processes are outlined in various sections.

In this section, we first introduce the network architecture in Sect. 3.1, the encoder and decoder are introduced in Sect. 3.1.1, and the image reconstruction module is presented in Sect. 3.1.2. We then introduce image discriminators in Sect. 3.2. Finally, the loss formula is introduced in detail in Sect. 3.3.

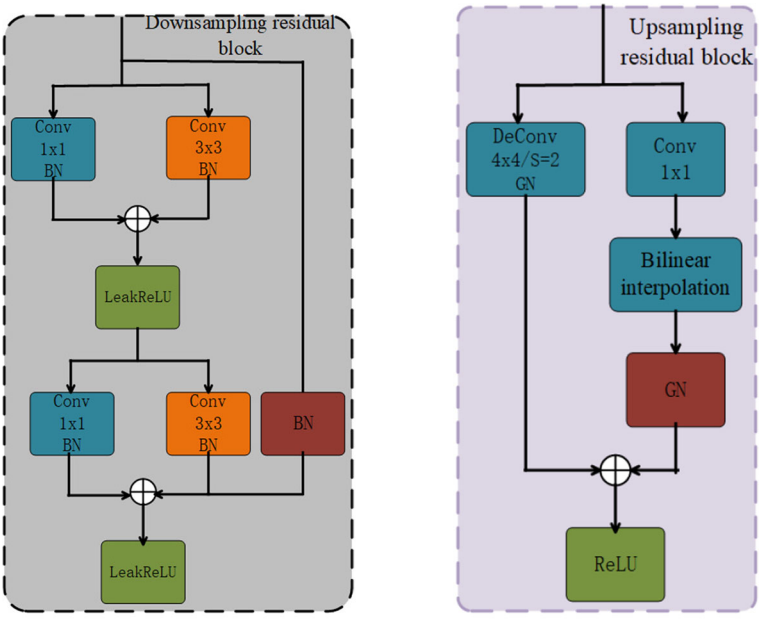
3.1 Framework and Network Architecture

3.1.1 Encoder–Decoder Module

In this paper, the encoder-decoder module generates unoccluded face textures. Various types of train images I fused with Gaussian random noise are the input images $\tilde{I}$ of the UFS framework. Gaussian random noise is significant to recover the high-frequency details of face images. We first use a 7×7 convolutional block in the encoder to obtain a large receptive field and then apply four downsampling residual blocks to learn rich feature representations. As shown in Fig. 2a, each block employs batch normalization layers and LeakReLU activation function. The downsampling residual block uses 3×3 and 1×1 convolution at each step to speed up inference, and batch normalization is used in the residual branch to improve the network performance. Although the four downsampling residual blocks have only 16 layers of structure, the encoder can also learn rich feature representations. The feature $e = E\left(\tilde{I}\right)$ can be obtained by the encoder. In the decoder, we use a combination of four residual upsampling blocks and skip-connected residual blocks, each of which uses group normalization and ReLU activation function. As shown in Fig. 2b, the upsampling residual block uses transposed convolution and bilinear interpolation to obtain enlarged feature maps, respectively. Among them, the 1×1 convolution layer reduces the dimension during bilinear interpolation, reducing the calculation amount of the bilinear interpolation, and finally, the two perform the residual fusion. The loss of feature information is further compensated by upsampling the residual layer to learn different enlarged feature maps. In addition, skip residual blocks learn high-frequency details of latent textures to improve the quality of generated face textures. After that, we synthesize the unoccluded face image $\tilde{y}$ through the decoder.

3.1.2 Image Reconstruction Module

The image reconstruction module can learn the texture details of the invisible face area and occlusion area (glasses and bangs occlusion, etc.) of the training image, making the generated image more realistic. The image reconstruction module can be divided into hole residual modules, feature fusion modules, and self-attention modules. Hole convolution can expand the receptive field without increasing the amount of computation, but hole convolution



(a) Downsampling residual block

(b) Upsampling residual block

Fig. 2 Upsampling and downsampling residual structures

will cause the loss of local details. Therefore, we employ hole residual blocks to reduce the loss of local details while enlarging the receptive field. As we can see from Fig. 3, the hole residual blocks all use 3×3 convolutions, and the hole steps are set to 2, 3, 1, and 5, respectively. We perform feature fusion on the features of different atrous convolutions to obtain fine-grained features. The residual branch applies 1×1 convolutional layers and batch normalization to reduce the loss of local details. To improve the quality of the generated images, we employ multi-scale feature fusion to obtain fine-grained features. Multi-scale feature fusion consists of four steps. First, we perform a linear interpolation operation on the first three downsampling residual blocks of the encoder to obtain a feature map of the same size as the hole residual block. We then perform a global average pooling operation on the last downsampling residual block to obtain global features. The generation of artifacts can be eliminated and reshaped to the feature map of the same size as the hole residual block. After that, we perform feature fusion on these four feature maps and reduce the amount of computation through dimensionality reduction. Finally, feature fusion with the hole residual block is performed to obtain fine-grained features.

We use a self-attention module to enhance the learning of face regions and mitigate the impact of harsh environments (low resolution, bright and dark light) on the generated images. First, the convolution used in this paper reduces the dimensionality of the fine-grained features. Then, the softmax operation is performed on the fine-grained features to obtain the spatial attention map. Finally, the sigmoid function is performed to output the feature map. The momentum method is used to fuse the features of the attention module and the fine-grained features of dimensionality reduction. The final fine-grained features are given as

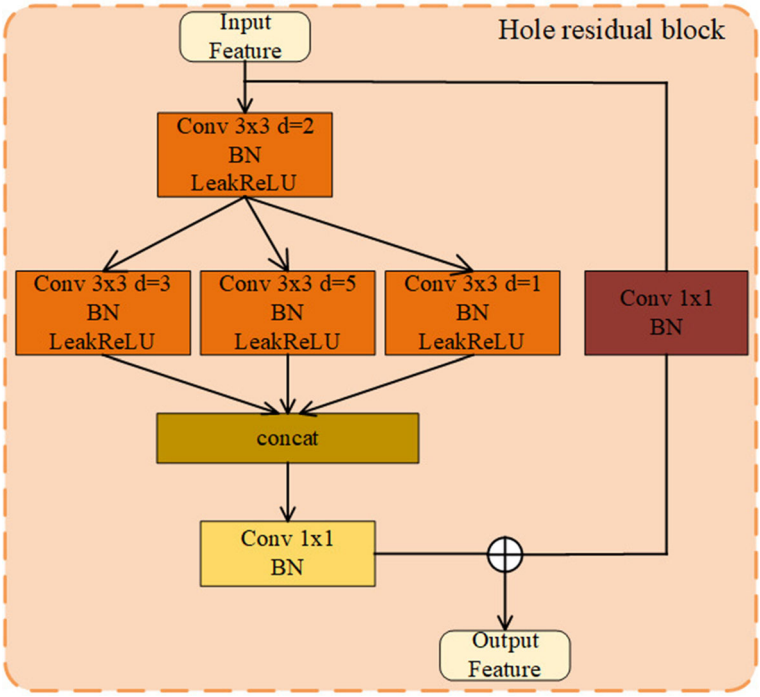


Fig. 3 We use different atrous convolutions for both layers in the atrous residual block to expand the receptive field

$$f_n = (1 - m) \cdot f_{sa} + m \cdot f_{n-1}, \tag{1}$$

where m is the momentum parameter we set to 0.5, f_{sa} is the feature map after attention network, f_{n-1} is the feature map after 1×1 convolution layer dimensionality reduction. Through the momentum method, the features can obtain multi-level feature information, improve the network's performance, and prevent the degradation of images during the training process.

3.2 Discriminator

We use the image discriminator to learn the overall structure of the face, strengthen the learning of the texture of the occluded area, and reduce the distortion and blurring of the occluded region. The image discriminator consists of four residual blocks and a convolution. In the residual blocks, spectral normalization is used to prevent training instability. We also use an attention network to enhance the discrimination of inpainted face regions. First, we input the training and target images into the discriminator, which outputs two feature maps as the predicted and actual labels, respectively. We then use the least square GAN as the adversarial loss, which is given by

$$L_{adv} = \min_{P} \max_{D} \mathrm{E}_I \left[\log D\left(I\right)\right] - \mathrm{E}_{P(X)} \left[1 - D(P(X))\right], \tag{2}$$

where $D(\cdot)$ is the discriminator, $P(\cdot)$ is the UFS framework, X is the training image, and I is the label image. In this paper, an image discriminator is used to learn global structure

information, repair texture information of occluded areas, and prevent image distortion and artifacts, which plays a vital role in image inpainting and face frontalization tasks.

3.3 Loss Formula

This section introduces further losses to constrain the UFS framework to generate unoccluded face textures. Our objective function consists of pixel-wise loss, identity loss, lpips loss, and focal frequency loss. They are described in detail below.

3.3.1 Pixel-Wise Loss

To make the generated image more approximate to the target image, we use pixel loss to reduce the difference between the generated image and the target image. It can be defined as

$$L_{pixel} = \|I^f - I^{gt}\|_1, \tag{3}$$

where I^f is the generated unoccluded image, I^{gt} is the target image, and $\|\cdot\|_1$ is the L1 paradigm.

3.3.2 Identity Loss

To make the generated unoccluded images retain more training image identity features, we adopt the ArcFace network [20] to extract the high-dimensional identity features of the developed and training images. We also use the cosine similarity of the two high-dimensional identity features as the identity loss. It can be given by

$$L_{id} = 2 - 2 \times \frac{\langle F(I^f), F(I^{gt})\rangle}{\|F(I^f)\| \cdot \|F(I^{gt})\|}, \tag{4}$$

where $F(\cdot)$ is pretrained ArcFace model taken from TreB1eN, I^f is the generated unoccluded image, I^{gt} is the target image, $\langle\cdot,\cdot\rangle$ is the vector inner product.

3.3.3 Lpips Loss

We use Learned perceptual image patch similarity (Lpips) [21] loss to learn perceptual similarity. The Lpips loss obtains the perceptual similarity of the image by calculating the channel cosine distance of the output of the network model layer by layer and averaging all the cosine distances. We find that Lpips loss generates higher quality images than standard perceptual losses. It is more in line with human visual cognition than other losses. It prevents image blurring. The Lpips loss is described as

$$L_{lpips} = \|L(I^f) - L(I^{gt})\|_2, \tag{5}$$

where $L(\cdot)$ is a deep feature extractor that uses the Alexnet backbone network to extract features. I^f is the generated unoccluded image and I^{gt} is the target image.

3.3.4 Focal Frequency Loss

To narrow the frequency gap between the generated image and the target image and improve the quality of the generated image, we strengthen the learning of frequency components that are difficult to synthesize through the focal frequency loss [22]. Then the generated image is closer to the target image in frequency. It can be defined as

$$\begin{aligned} L_{ffl} &= \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} w(u,v) \| F_r(u,v) - F_f(u,v) \|^2, \\ w(u,v) &= \| F_r(u,v) - F_f(u,v) \|^{\alpha}, \\ F(u,v) &= \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-i2\pi(\frac{ux}{M} + \frac{vy}{N})}, \end{aligned} \tag{6}$$

where $f(x, y)$ is the pixel value, M and N are the height and width of the trained image, respectively, u and v are the coordinates of the corresponding frequency space. $F_r(u, v)$ is the spatial frequency value of the real image at the spectral coordinate (u, v), and the corresponding $F_f(u, v)$ is the spatial frequency value of the fake image at the spectral coordinate (u, v).

Finally, the objective function for all losses is given by

$$L_{loss} = \lambda_{pixel} L_{pixel} + \lambda_{id} L_{id} + \lambda_{lpips} L_{lpips} + \lambda_{ffl} L_{ffl}, \tag{7}$$

where λ_{pixel}, λ_{id}, λ_{lpips}, and λ_{ffl} are weights corresponding to each loss formula, respectively, and their values will be illustrated in Sect. 4.

4 Results

The UFS framework proposed in this paper generates clear face images through multi-tasks such as face frontalization, glasses, and bangs. We first made with recent methods for single tasks such as removing glasses. We then show the multi-task experimental results in quantitative and qualitative experiments, such as eliminating glasses or bangs while face frontalization. And we also use the generated unoccluded face images for 3D face reconstruction. Our method has a significant improvement in texture compared to previous work. Finally, to verify the effectiveness of our proposed modules, we conduct ablation experiments on each module. The robustness of our framework is demonstrated through multiple experiments.

This section presents the implementation details and datasets in Sect. 4.1, followed by quantitative and qualitative experiments compared with state-of-the-art methods in Sect. 4.2. Precise details are introduced in individual sections.

4.1 Experimental Settings

4.1.1 Implementation Details

We propose a UFS framework to generate unoccluded face textures, implemented using PyTorch. We use Adam optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the initial learning rate is set to 4×10^{-4}. Furthermore, we use a fixed step size learning-rate-decay strategy, gamma is set to 0.7, step size is set to 30. In the training process, we employ RTX3060 GPU, the

batch size is set to 8, and 200 echoes are trained. For the weight of the loss function, we set $\lambda_{pixel} = 10$, $\lambda_{id} = 1$, $\lambda_{lpips} = 1$, and $\lambda_{ffl} = 10$.

4.1.2 Datasets

To achieve multi-task work, we construct a face dataset, which includes various face images with glasses, different bangs, and profile faces. The corresponding label images are unoccluded frontal face images. There are 30,000 face images in our dataset, consisting of the FFHQ dataset [23] and the Celeba-HQ dataset [24]. We then split this dataset into 28,000 images in the training set and 2,000 in the test set. We first align and crop the face images according to the face detection algorithm for the constructed high-resolution face dataset. The size of the cropped image is 256×256. We then perform data augmentation by randomly rotating and adding noise to the training data. Finally, we apply the LFW dataset [25] and Celeba dataset [26] for single-task comparative experiments. The Celeba is a public largescale face dataset containing 202,599 face images of 10,177 Celebrity identities. The LFW is a face recognition dataset containing 13,233 face images collected from the websites. The LFW is often used to evaluate frontalization performance in uncontrolled settings.

4.2 Comparison to State-of-the-Art Methods

4.2.1 Qualitative Evaluation

Image Inpainting We train the image inpainting task separately due to the difference between image inpainting and other tasks. We do two-fold work on the image inpainting task. On the one hand, for the external occlusion of the image (hand, microphone, sunglasses, etc.), we add a small area mask to the external occlusion area to obtain the mask image. We then input it into the UFS framework, which outputs an unoccluded face image. Although the images synthesized by our method are more realistic, our approach still has some problems. When the mask area we add is too large, there will be a difference between the synthesized face area and the source image. On the other hand, like other image inpainting methods [27–32], we add different large region masks on face images. We hope that our approach can repair the face texture of the large-area mask even when covering more areas of the face under the large mask. Therefore, the synthesized face texture has a high sense of realism. The images we test are all from the test set of our constructed face dataset. Due to the different image inpainting tasks, we increase the weight of the identity loss and the high-frequency focal loss and reduce the importance of the Lpips loss during training. we set $\lambda_{id} = 2$, $\lambda_{ffl} = 50$, $\lambda_{lpips} = 0.1$.

As shown in Fig. 4, we add a mask to the face area covered by the microphone, hands, and sunglasses to generate a mask image. As we can see from the synthesizing face images, our method fixes the microphone, hands, sunglasses, etc. The occluded texture gives the synthetic face a certain sense of realism. However, we found in our experiments that if the mask area we add is too large, it will cause the synthesized texture information to be different from the source image. For example, we added a mask area covering the entire mouth in the last picture. As a result, the expression of the synthesized face has changed, and we will solve this problem in future work.

As shown in Fig. 5, we select five kinds of large-area masks to add to the test set to generate mask images, and our composite face images are consistent with the actual pictures. Although the faces synthesized by our method are somewhat different from the authentic images in some details, they are acceptable in terms of visual effects.

Fig. 4 Occlusion-based image inpainting results

Fig. 5 Image inpainting results

Face Frontalization Face frontalization is challenging, and all face frontalization work is trained on the multi-pie dataset. However, the datasets are collected in closed indoor scenes, lacking face data in wild environments. Testing unconstrained face images in the wild often leads to degradation (blur, artifacts) in synthetic frontal face images. To solve this problem, we construct a face dataset consisting of all wild data, which can adapt to various unconstrained environments to slow down image degradation. On the single task of face frontalization, we conduct a two-part experiment. In the first section, we compare with state-of-the-art face frontalization methods. In the second part, we conduct experiments on an unconstrained test set to verify the effectiveness of our approach on the face frontalization task.

We perform qualitative comparisons with face frontalization methods such as PIM [9], TP-GAN [8], DR-GAN [7], and Hassner et al. [4]. As shown in Fig. 6, our method is robust to unconstrained face images on the single task of face frontalization. The frontal face synthesized by the TP-GAN method seriously loses texture details. The synthesized result is very blurred. The frontalized face images synthesized by the PIM method are not high quality and lose some texture details. The frontalized faces synthesized by DR-GAN and Hassner et al. deviate significantly from the accurate frontal pose.

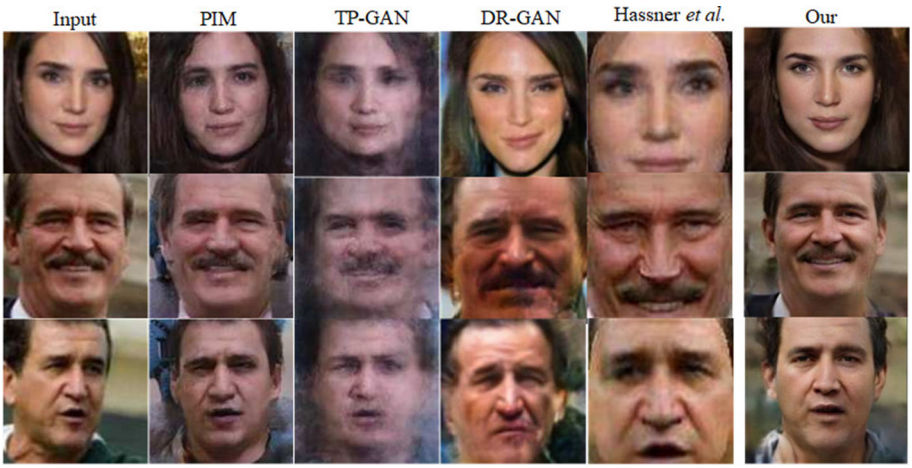


Fig. 6 Comparison of different face frontalization methods

Fig. 7 Face frontalization results

In contrast, our method to synthesize frontalized face images preserves the high-frequency details of the face and maintains the identity consistency with the input image. Furthermore, we find that our composite face images have higher image quality, and the synthesized results are sharper than the input images. In conclusion, our method can realistically recover high-fidelity frontal face textures.

We selected eight non-frontal face images on the test set for face frontalization experiments. As shown in Fig. 7, our method can recover the missing face texture. Although our method is not exactly the same in texture as the input image, our synthesized frontal images preserve the identity consistency of the input image. Therefore, the synthetic result is more realistic.

Glasses Removal Wearing glasses will block parts of the face, which will lead to the inability to express some high-frequency details and reduce the accuracy of face recognition. On the single task of face glasses removal, we conduct two-part experiments. In the first part, we compare the effect of glasses removal with state-of-the-art methods. In the second part, we conduct experiments on the test set to demonstrate the robustness of our approach on the glasses removal task. To be consistent with the results of other methods on a single job, we select appropriate data for training to prevent the synthetic face texture from performing face frontalization while removing the glasses.

Our method visually compares with CycleGAN [17], ELEGANT [15], ERGAN [16], ByeGlassGAN [33], pix2pix [18], and StarGAN [34] on a single task. ERGAN and ByeGlassGAN are glasses removal methods. CycleGAN, ELEGANT, pix2pix, and StarGAN are

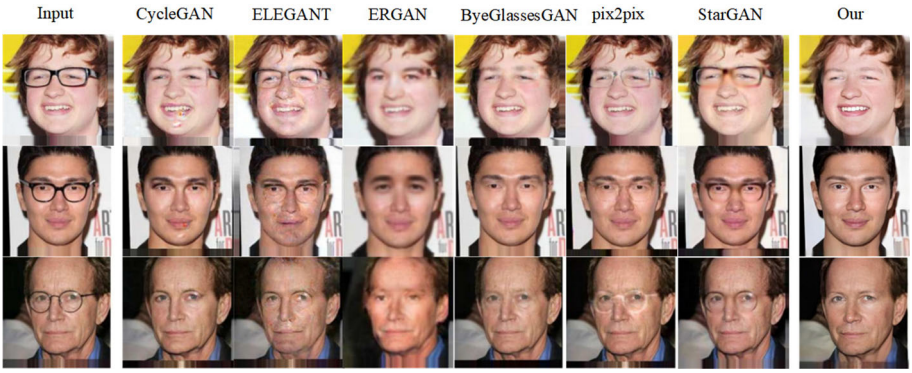


Fig. 8 Comparison of different eyeglass removal methods

Fig. 9 Glasses removal results

face editing methods. As shown in Fig. 8, the results of the pix2pix and StarGAN methods show a lot of artifacts. The results of CycleGAN and ELEGANT methods show a lot of noise speckles. The ERGAN method removes the occlusion of the glasses, but the synthesized images become blurred and lose some high-frequency details. The results of the ByeGlassesGAN method achieve better results, but the results of our method are visually better than their method, and the synthetic images are sharper. Our method removes the occlusion of glasses and preserves more texture details to make the synthesized images more realistic.

To demonstrate the capability of our method on the task of glasses removal, we select eight images of different glasses from the Celeba dataset, including small glasses frames, round glasses frames, square glasses frames, large glasses frames, and sunglasses frames. As shown in Fig. 9, our method can altogether remove the glasses frame without losing high-frequency details. In addition, as shown in the input image in the first row and seventh column, the glasses can create shadows on the face when illuminated by one-sided light. However, our method can also remove such shadows when removing glasses. Our process does not produce some sunglasses artifacts like other methods in removing sunglasses. Our approach can altogether remove the sunglasses and maintain the identity consistency of the synthesized image and the input image.

Multi-task To synthesize unoccluded face textures, we jointly learn multiple face de-occlusion methods. We perform multi-task learning on the constructed face dataset and conduct experiments on the test set. It isn't easy to include all de-occlusion tasks in typical face datasets. Our multi-task experiment is divided into the following parts. First, we combine the face frontalization task with the glasses or bangs removal task to remove the occlusion of the glasses or bangs. Then, we synthesize unoccluded high-resolution faces for blurred occlusion face images. Finally, we can use the generated unoccluded face texture for 3D

Fig. 10 Face frontalization and glasses removal results

Fig. 11 Results from blurred image to sharp image

face reconstruction. The experimental results show that the face texture synthesized by our method is better than the previous methods.

We first perform the test set's dual face frontalization and glasses removal tasks. We select five face images with different glasses and small poses for dual-task experiments. As shown in Fig. 10, our method recovers the frontalized face texture while completely removing the glasses, synthesizing an unoccluded face image. Then, we perform the dual tasks of face frontalization and bang removal on the test set. We select six face images with bangs and side faces for dual-task experiments. As shown in Fig. 11, our method removes most of the occlusion of bangs while restoring the frontalized face texture. The removal of the contralateral bangs by our approach works best. In conclusion, our method can achieve face frontalization and remove glasses or bangs to synthesize unobstructed face textures. However, the results of our strategy are not exactly the same texture as the input images, which we will address in future work.

In the face frontalization single-task experiment, we are surprised to find that when synthesizing a frontal face image from a blurred non-frontal face image, the synthesized face image becomes very clear. As shown in Fig. 12, we choose five non-frontal, blurred face images with bangs and glasses for experimentation. The occlusion-free face texture synthesized by our method makes the image sharper without losing face details. Therefore, our approach is more robust.

The synthesized unoccluded face is used for 3D face reconstruction. First, we adopt the method of Deng et al. [1] to generate 3D face geometry. We then generate unoccluded frontalized face textures and use texture mapping to create textures for 3D face geometry. As shown in Fig. 13, we present the results generated by our method at two angles with realistic

Fig. 12 Face frontalization and bangs removal results

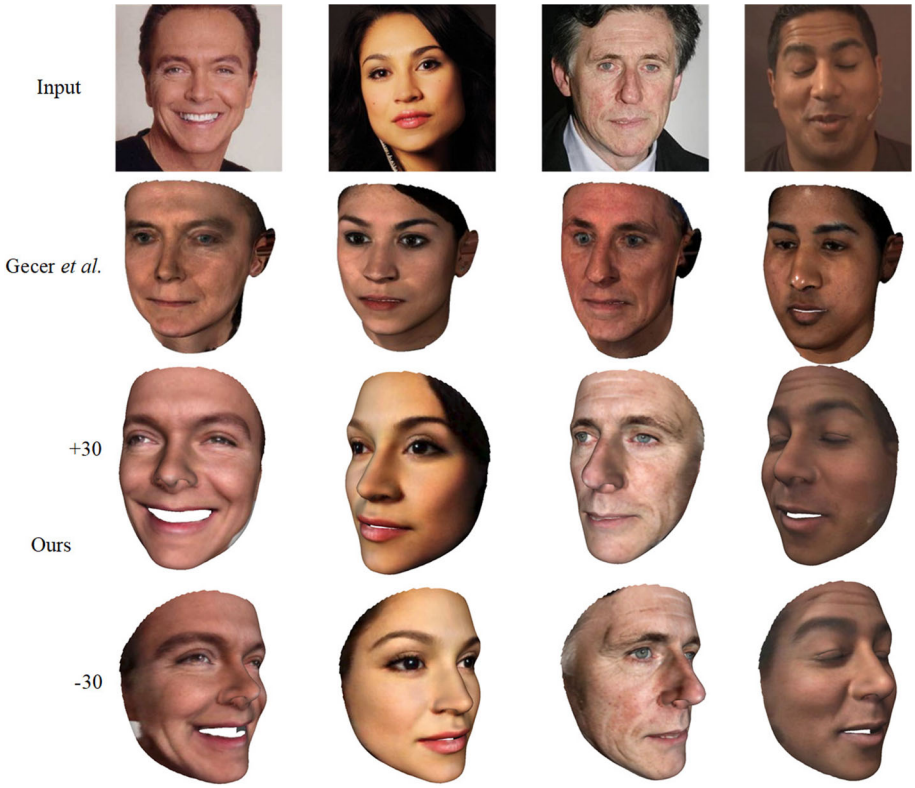


Fig. 13 Results from blurred image to sharp image

textures. The third row is the result with a right bias of 30, and the fourth row is the result with a left bias of 30. Comparing our approach with Gecer et al. [2], their approach cannot preserve the high-frequency details of the face texture, and the generated texture is also far from the input image. On the contrary, our approach keeps the high-frequency components of faces and recovers the texture information of self-occlusion regions with high fidelity.

Table 1 Comparison of FID (↓) under different tasks

Methods	Glasses test set	Glasses and bangs test set
Hisd	12.0989	14.0735
w/o IR UFS	12.8567	15.4167
UFS	11.0810	13.9924

Table 2 Comparison of SSIM (↑) and PSNR (↓) under different datasets

Datasets	SSIM	PSNR
Glasses test set	0.8253	22.1256
Glasses and bangs test set	0.7993	22.7240
Masks test set	0.9395	20.6068

4.2.2 Quantitative Evaluation

To quantitatively evaluate the advantages of our method, we divide the test set into three sub-test sets, namely the test set with glasses, the test set with glasses and bangs, and the test set with masks. We utilize the Frechet inception distance (FID) to measure the performance of glasses and bangs removal. FID is a measure of calculating the distance between the actual image and the feature vector of the generated image. It uses the inception model to obtain the feature map. The lower the scores of the two sets of FID feature maps, the more similar the two sets of images. We compare FID with the Hisd method [19] on the single task of removing glasses and the dual task of removing glasses and removing bangs. To demonstrate the effectiveness of the image reconstruction module, we add the UFS framework without the image reconstruction module to the comparative experiments. As shown in Table 1, we use the test set with glasses and the test set with glasses and bangs for experiments. The UFS frame achieves the lowest FID score in the single task of removing glasses, indicating that the glasses-free images generated by the UFS frame are more accurate. Closer to the target image, it is more capable in the single task of glasses removal. The result without the image reconstruction module achieves the worst score, proving that the image reconstruction module can improve the performance of UFS. Our UFS framework and the Hisd method achieve similar scores in the dual-task of removing glasses and bangs. Because our method employs multi-scale feature fusion, atrous residual blocks, and self-attention networks to obtain multi-level feature representations, it can handle multi-task occlusion problems and can generate better results. The Hisd method uses the idea of image translation to inject style parameters into the feature map to realize the attribute editing of the face, but this style injection method is difficult to achieve the effect of our method in the single task of de-occlusion. However, in the multi-task of de-occlusion, the Hisd method improves the ability of de-occlusion by injecting style parameters multiple times. The Hisd method finally achieves a similar effect to our method. However, UFS without the image reconstruction module still gets the worst score, and the single task score is significantly lower than the dual-task score. The dual-task of removing glasses and bangs is comparable to the Hisd method. Therefore, our UFS frame has a more vital ability to remove glasses and bangs. Removing glasses for a single task is stronger than eliminating glasses and bangs for dual tasks.

We use the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) to measure the generative ability of our UFS framework. We conduct experiments with the

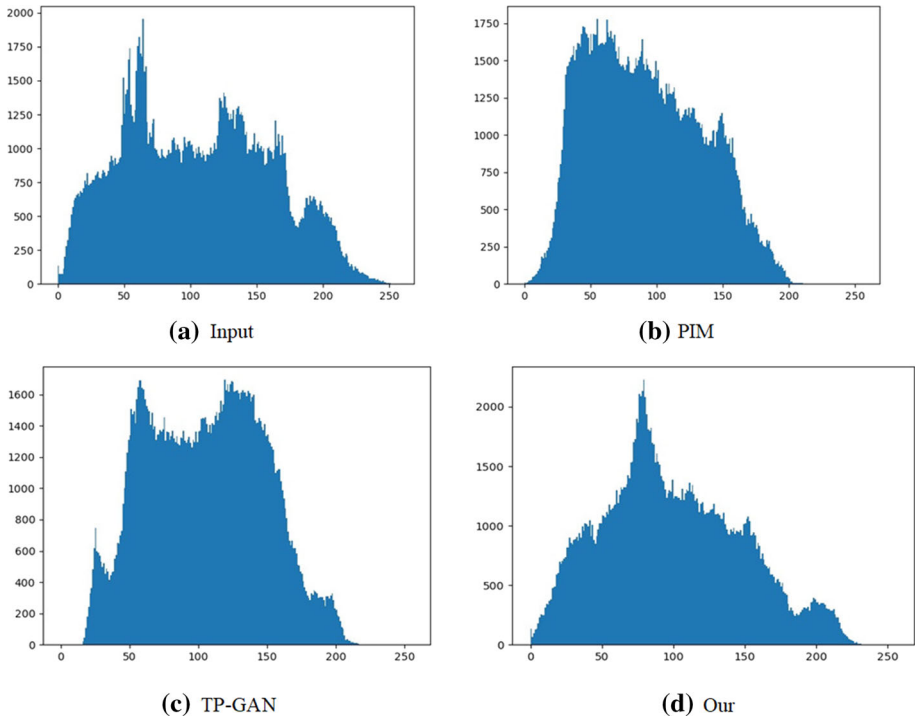


Fig. 14 Comparison of histograms with different methods. **a** Histogram of the input image, **b** histogram of the image generated by the PIM method, **c** histogram of the image generated by the TP-GAN method, **d** histogram of the image generated by our method

glasses dataset, the glasses and bangs dataset, and the masks dataset. As shown in Table 2, the image inpainting task achieves the best results on the SSIM and PSNR metrics in multiple datasets, followed by the glasses removal task, and finally, the glasses and bangs dual removal task. Therefore, our UFS framework can achieve the best results on a single job. However, in the face of multitasking, our UFS frame generation ability becomes weak, but our method can still achieve better results visually. In the future, we will improve the generative capability of multitasking.

To verify that our method can generate sharper unocclusion results, we select blurred images in the LFW dataset for face frontalization and display the generated results in a histogram. As shown in Fig. 14, the pixel value distribution in the histogram of the results generated by the b and c methods is around 30–150, which is somewhat different from the pixel distribution of the input image. The histogram pixel distribution of b and c is lost at 200 Pixel points after the pixel point, which causes the generated result to lose some details. The histogram generated by our method is more similar to the input image's histogram. The pixels expressing face information are denser, and the pixels after 200-pixel values are retained, which allows some details to be preserved. In conclusion, our method can generate sharper results.

5 Conclusion

This paper proposed a novel UFS framework that combines face frontalization, glasses removal, and image inpainting to synthesize unoccluded face textures. We adopted an image reconstruction module in the UFS framework to obtain multi-level fine-grained features to improve the quality of image synthesis. The image reconstruction module used a multi-scale feature fusion and self-attention network to enlarge the receptive field and strengthen the learning of face regions. We also used focal frequency loss to enhance the synthesis of complex frequencies. Therefore, our synthesized unoccluded faces retain more local high-frequency details. In addition, we used an image discriminator to learn the overall structure of the face, preventing image distortions and artifacts. Our UFS framework achieves good results in the single task of face frontalization, glasses removal, and image inpainting in all experiments. However, in the case of multitasking face fronting and removing glasses or bangs, the results are not quite perfect. We will address the instability of multi-task synthetic images in future work and address inconsistencies between synthetic textures and input images.

Acknowledgements We would like to thank the anonymous reviewers for their valuable suggestions. This research was funded by the Shandong Provincial Natural Science Foundation (ZR2020MF132); the National Natural Science Foundation of China (62072020); and the Qingdao Leading Scholars Project on Innovation and Entrepreneurship 2019 (No.19-3-2-21-zhc).

References

1. Deng Y, Yang JL, Xu SC et al (2019) Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set. In: CVPR2019
2. Gecer B, Ploumpis S, Kotsia I et al (2019) Ganfifit: generative ad-versarial network fitting for high fidelity 3D face reconstruc-tion. In: CVPR2019
3. Deng J, Cheng S, Xue N et al (2018) UV-GAN: adversarial facial UV map completion for pose-invariant face recognition. In: CVPR2018
4. Hassner T, Harel S, Paz E et al (2015) Effffective face frontalization in unconstrained images. In: CVPR2015
5. Wei Y, Liu M, Wang H et al (2020) Learning flow-based feature warping for face frontalization with illumination inconsistent supervision. In: CVPR2020
6. Yin X, Yu X, Sohn K et al (2017) Towards large-pose face frontalization in the wild. In: ICCV2017
7. Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: CVPR2017
8. Huang R, Zhang S, Li T et al (2017) Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: CVPR2017
9. Zhao J, Cheng Y, Xu Y et al (2018) Towards pose invariant face recognition in the wild. In: CVPR2018
10. Yin Y, Jiang S, Robinson JP et al (2020) Dual-attention gan for large-pose face frontalization. arXiv:2002.07227
11. Pathak D, Krahenbuhl P, Donahue J et al (2016) Context encoders: feature learning by inpainting. In: CVPR2016
12. Iizuka S, Simo-Serra E, Ishikawa H (2017) Globally and locally consistent image completion. ACM Trans Graph 36(4):107
13. Liu G, Reda FA, Shih KJ et al (2018) Image inpainting for irregular holes using partial convolutions. In: ECCV2018
14. Huang J, Zhu P, Geng M et al (2018) Range scaling global u-net for perceptual image enhancement on mobile devices. In: ECCV2018
15. Xiao T, Hong J, Ma J (2018) Elegant: exchanging latent encodings with gan for transferring multiple face attributes. In: ECCV2018
16. Hu B, Yang W, Ren M (2019) Unsupervised eyeglasses removal in the wild. arXiv:1909.06989

17. Zhu JY, Park T, Isola P et al (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV2017
18. Isola P, Zhu JY, Zhou T et al (2017) Image-to-image translation with conditional adversarial networks. In: CVPR2017
19. Li X, Zhang S, Hu J et al (2021) Image-to-image translation via hierarchical style disentanglement. In: CVPR2021
20. Deng J, Guo J, Xue N, Zafeiriou S (2019) ArcFace: additive angular margin loss for deep face recognition. In: CVPR2019
21. Zhang R, Isola P, Efros AA et al (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR2018
22. Jiang L, Dai B, Wu W et al (2021) Focal frequency loss for image reconstruction and synthesis. In: ICCV2021
23. Karras T, Aila T, Laine S et al (2018) Progressive growing of GANs for improved quality, stability, and variation. In: ICLR2018
24. Karras T, Laine S, Aila T (2019) A stylebased generator architecture for generative adversarial networks. In: CVPR2019
25. Huang G, Ramesh M, Berg T et al (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst 2007
26. Liu Z, Luo P, Wang X et al (2015) Deep learning face attributes in the wild. In: CVPR2015
27. Ren Y, Yu X, Zhang R et al (2019) Structureflow: image inpainting via structure-aware appearance flow. In: ICCV2019
28. Yi Z, Song W, Li S et al (2022) Automatic image matting and fusing for portrait synthesis. Sci China Inf Sci 65:124101. https://doi.org/10.1007/s11432-021-3279-y
29. Chen Y, Hu H (2019) An improved method for semantic image inpainting with GANs: progressive inpainting. Neural Process Lett 49:1355–1367. https://doi.org/10.1007/s11063-018-9877-6
30. Chao G Q, Mao S H, Wang F et al. Supervised nonnegative matrix factorization to predict ICU mortality risk. In: BIBM2018
31. Chao GQ (2019) Discriminative k-means Laplacian clustering. Neural Process Lett 49:393–405
32. Zhao L, Mo Q, Lin S et al (2020) UCTGAN: diverse image inpainting based on unsupervised cross-space translation. In: CVPR2020
33. Lee Y H, Lai S H (2020) ByeGlassesGAN: identity preserving eyeglasses removal for face images. In: ECCV2020
34. Choi Y, Choi M, Kim M et al (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR2018