FaceInpainter: High Fidelity Face Adaptation to Heterogeneous Domains

Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, Ran He†
National Laboratory of Pattern Recognition, CASIA
Center for Excellence in Brain Science and Intelligence Technology, CAS
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
{jia.li, jie.cao}@cripac.ia.ac.cn, rhe@nlpr.ia.ac.cn,
zhaoyang0427@gmail.com, Wilson.song@hotmail.com

Figure 1: Face swapping results on wild face images by means of FaceInpainter. Our identity-guided face inpainting framework can produce high fidelity results in heterogeneous domains, *i.e*., photograph or non-photorealistic styles, *e.g*., oil painting, 3D cartoons, pencil drawing and exaggerated drawing.

Abstract

In this work, we propose a novel two-stage framework named FaceInpainter to implement controllable Identity-Guided Face Inpainting (IGFI) under heterogeneous domains. Concretely, by explicitly disentangling foreground and background of the target face, the first stage focuses on adaptive face fitting to the fixed background via a Styled Face Inpainting Network (SFI-Net), with 3D priors and texture code of the target, as well as identity factor of the source face. It is challenging to deal with the inconsistency between the new identity of the source and the original background of the target, concerning the face shape and appearance on the fused boundary. The second stage consists of a Joint Refinement Network (JR-Net) to refine the swapped face. It leverages AdaIN considering identity and multi-scale texture codes, for feature transformation of the decoded face from SFI-Net with facial occlusions. We adopt the contextual loss to implicitly preserve the attributes, encouraging face deformation and fewer texture distortions. Experimental results demonstrate that our approach handles high-quality identity adaptation to heterogeneous domains, exhibiting the competitive performance compared with state-of-the-art methods concerning both attribute and identity fidelity.

1. Introduction

Identity swapping is a technique that enables manipulating the face appearance according to the source face while preserving the attributes of the target face. DeepFakes [1], FaceSwap [2], FaceShifter [18], and SimSwap [5] are the prominent methods. These methods enable the effortless creation of fake faces, while the controllable performance and visual realism still need to be improved. Furthermore, face manipulation provides more challenging and diverse samples to facilitate face forensics [31].

The face swapping task is evaluated mainly from two aspects, *i.e*. identity and attribute fidelity for the source and target face, respectively. FaceShifter [18] adopts SPADE [28] and AdaIN [14] mechanisms to integrate spatial attributes and identity styles in the decoder, while some redundant identity features of the source, *e.g*., hair, make it difficult to preserve attributes in some challenging cases, as shown in Figure 2. SimSwap [5] uses a weak feature matching loss based on the last few layers of the discriminator, to avoid the attribute distortions during identity modifica-

†Corresponding author

tion. Despite this implicit constraint works well, it is similar to the commonly used perceptual loss for texture matching used in [38, 49, 24], which is more appropriate for paired data learning. Thus, the generated faces of SimSwap preserve lots of attributes in some challenging cases, so that they look more like the target face. In the non-photorealistic domain, we take FaceSwap [2] as an example, the main issues of this 3D-based method include low resolution and obvious modification artifacts around the facial area.

In this work, we first study the heterogeneous identity swapping problem. The heterogeneous face has been studied in some works [42, 10, 11, 41]. It is challenging to perfectly fit target faces with a new identity under heterogeneous domains, especially for an extreme pose, expression, and lighting conditions. We propose an efficient Identity-Guided Face Inpainting (IGFI) framework to improve the controllability and generalization of the face-swapping model. Facial attributes of the target face are supposed to be preserved during the IGFI process, including head pose, expression, lighting, facial conclusions, hairstyle, and other background contents. Additionally, we attempt to adaptively improve the visual quality for the target scenes with low resolution.

In our work, the identity feature is extracted from the source face via a pretrained state-of-the-art face recognition model [6]. Generally, the identity feature contains the high-level semantic representation of the source, which is used to control the identity modification. As for attribute preservation of the target, a 3D fitting model [12] is used to extract 3DMM parameters of the target, such as expression and posture, which are vital to performing IGFI based on the fixed background. For the purpose of preserving texture, a pretrained face parsing model [19] is used to extract the foreground content (*i.e*. face and neck), which is fed into a pretrained VGG network [33] to obtain the texture style. After that, the multiple factors with regard to shape and texture from the source and target faces are recombined into one style code, to achieve efficient and controllable IGFI in specific scenes, based on the Styled Face Inpainting Network (SFI-Net). Particularly, to preserve the high-fidelity non-face areas (*e.g*. background, hair, clothes, etc.), we directly add them to the last layer of SFI-Net.

To weaken the inconsistency of the target face shape and the face inpainting result for a new identity, we mask both the face and neck regions which are visually crucial for the facial shape representation in the first stage. However, for some complex scenes, the segmented background and the generated foreground can not be well integrated. Therefore, in the second stage, we propose a Joint Refinement Network (JR-Net) to jointly refine the identity, attributes, and boundary fusion, by adopting Adaptive Instance Normalization (AdaIN) [14] with respect to identity and multi-scale texture codes. With super-resolution and occlusion-aware

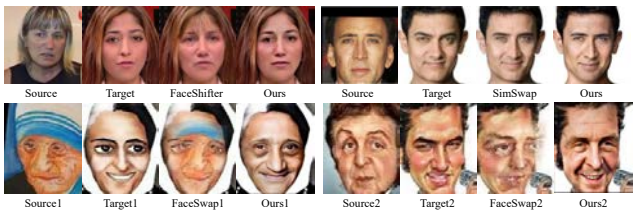

Figure 2: There are still some challenging situations, *e.g*., preserving excessive source identity (FaceShifter [18]), preserving more attributes (SimSwap [5]), and suffering from low resolution and boundary artifacts (FaceSwap [2]). Our model realizes more controllable face swapping in photorealistic (row 1) and non-photorealistic (row 2) domains.

modules, JR-Net further optimizes the visual and occlusion perception of the result from SFI-Net. For non-aligned feature matching between the target and deformed face, we adopt the contextual constraints [23] which can implicitly keep the attribute consistency. As shown in Figure 1, our method can generate high-quality identity-guided swapped faces adapted to various heterogeneous domains, even in cartoon and exaggerated styles.

Our contributions are three folds.

- We first study the heterogeneous identity swapping task, and propose a novel solution to deal with the Identity-Guided Face Inpainting (IGFI) problem. By considering high-fidelity identity and attributes, our approach FaceInpainter achieves more controllable and higher quality results than previous face inpainting methods.

- We propose an effective IGFI framework. In the first stage, we introduce a Styled Face Inpainting Network (SFI-Net) to map the identity and attribute codes to the swapped face. The second stage contains a Joint Refinement Network (JR-Net) that refines the attributes and identity details, generating occlusion-aware and high-resolution swapped faces with visually natural fused boundary.

- With achieving high fidelity of contextual attribute and identity, we achieve good generalization under heterogeneous domains both visually and quantitatively.

2. Related Work

2.1. Face Inpainting

Face inpainting is mainly applied for facial completion [22, 45, 43, 34, 44] or component editing [7]. By incorporating a VAE pipeline, Zheng et al. [46] exploited smooth priors for the latent space of the hidden partial image, from

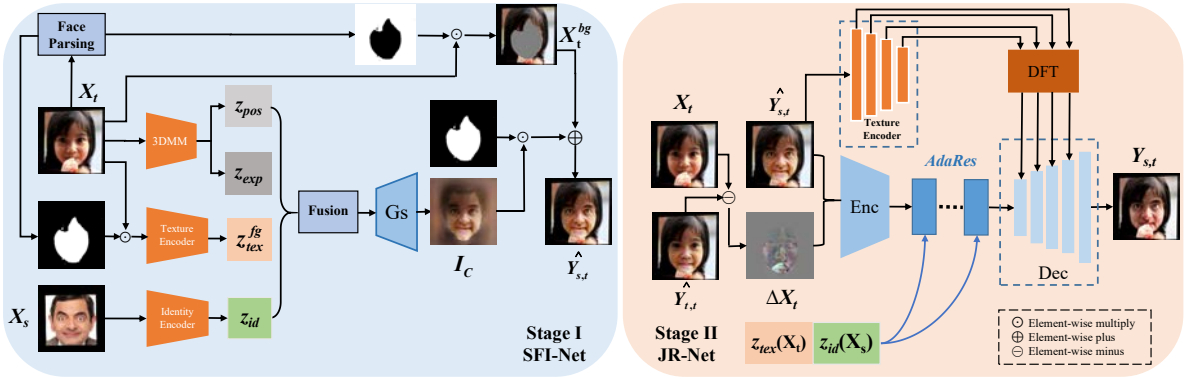

Figure 3: The framework of FaceInpainter. The first stage (SFI-Net) learns the mapping from the attribute codes of the target X_t and identity code of the source X_s to the inpainting result I_C based on StyleGAN G_s, then with the fixed background X_t^{bg} to obtain the swapped face $\hat{Y}_{s,t}$. The second stage (JR-Net) is designed for the refinement of the coarse result from SFI-Net. $z_{id}(X_s)$ and $z_{tex}(X_t)$ are fused to control the denormalization parameters in AdaIN residual blocks, which facilitates both identity and attribute refinements. $\triangle X_t$ is the residual map between X_t and $\hat{Y}_{t,t}$, used as the occlusion representation. Furthermore, the dictionary feature transfer (DFT) module is used to improve the face resolution.

which the generative pipeline inferred diverse inpainting results. To produce a controllable completed face, Song et al. [34] proposed a geometry-aware method to inpaint and edit face with the constraints of facial landmark heatmap and parsing map. Deng et al. [7] presented the r-FACE model for controlling the structure of face components based on reference images.

2.2. Identity Swapping

Identity swapping aims at exchanging the face feature without causing interference on other attributes of the target face [27, 18, 5]. FSGAN [27] incorporated cascaded modules concerning recurrent reenactment, face parsing, inpainting and blending, to achieve high-quality swapping results. FaceShifter [18] applied a two-stage framework to adaptively integrate identity and attribute embedding with the attentional map, and tackle the occlusion issue in a heuristic error acknowledging manner. SimSwap [5] utilized an ID injection module to adaptively integrate identity embedding to the target feature with the weak feature matching loss.

2.3. Computer Graphics Based Approaches

Priors from 3D Morphable Face Models (3DMM) [48, 9, 12] or nonlinear 3DMM [36, 35] have been incorporated in various face synthesis tasks [47, 8]. FaceSwap [2] applied 3D fitting and rendering to obtain the aligned face which is then blended with the target. Hang et al. [47] proposed a Rotate-and-Render framework leveraging 3D face modeling and high-resolution GAN for creating paired data and face inpainting, respectively. DiscoFaceGAN [8] disentangled the latent representations of identity, expression, pose, and illumination to precisely control the generation of faces with different attributes.

3. Approach

The IGFI task is supposed to generate the foreground content with identity modification and attribute preservation, as well as a natural fusion with the background. As shown in Figure 3, given a target face X_t, we mainly consider three important factors: expression, pose, and texture. We leverage advanced face recognition network [6], 3DMM [12] and texture encoder [33] to obtain the decomposed style codes, which are fused as one variable to control the face synthesis in Styled Face Inpainting Network. As an inpainting task, the occlusion of the foreground has been removed, and there is an inconsistency between the new face and the original target background. Therefore, the details of face shape and attributes in the generated face $\hat{Y}_{s,t}$ of SFI-Net will be refined by JR-Net, described as follows.

3.1. Styled Face Inpainting Network

We incorporate a 3DMM model [3] to learn the mapping from multiple codes of source and target faces to the IGFI result. As a parametric model, 3DMM is used to fit the shape $\mathbf{S}$ and texture $\mathbf{T}$ from a single image. Specifically, $\mathbf{S}$ is denoted as:

$$\mathbf{S} = \overline{\mathbf{S}} + \mathbf{B_{id}}\alpha_{id} + \mathbf{B}_{\exp}\alpha_{\exp}, \tag{1}$$

where $\overline{\mathbf{S}}$ is the mean 3D shape, $\mathbf{B_{id}}$ is the 3D shape bases and α_{id} is the corresponding coefficient vector. The same goes for expression. 3D fitting for a single image is defined as:

$$\mathbf{F}(\mathbf{S}) = f * \mathbf{P_r} * \mathbf{R}(p, y, r) * \mathbf{S} + \mathbf{t}_{2d}, \tag{2}$$

where f is the scale factor, $\mathbf{P_r}$ is the orthographic projection matrix, $\mathbf{R}(p, y, r)$ is the rotation matrix constructed by pitch, yaw, and roll representations, and $\mathbf{t}_{2d}$ is the translation vector. Face posture is defined as $z_{pos} = [f, \mathbf{R}(p, y, r), \mathbf{t}_{2d}]$. We denote $z_{exp} = \alpha_{exp}$ as the expression code from target face. Generally, the identity and expression bases of $\mathbf{S}$ are entangled with each other, especially for nonlinear 3DMM [36, 35] with one shape vector representing identity and expression factors. Instead of α_{id}, we use a pretrained state-of-the-art face recognition model [6] to obtain identity latent code z_{id}.

Given a target image $X_t \in \mathbb{R}^{3\times H\times W}$, with the corresponding K-channel heat map $M \in \mathbb{R}^{K\times H\times W}$ extracted by an off-the-shelf face parser, we separate it into background X_t^{bg} (*e.g.* hair, clothes, and other non-human areas) and foreground X_t^{fg} (*e.g.* face, neck). In our SFI-Net training, X_t^{fg} is fed into a texture encoder E_{vgg}, *i.e.* a pretrained VGG19 network [33], to acquire texture code z_{tex} of the target face, which contains the skin color and illumination information. With $C = [z_{id}(X_s), z_{exp}(X_t), z_{pos}(X_t), z_{tex}(X_t^{fg})]$, we get the style code z_C of IGFI via a fusion module consisting of one fully connected layer, and then inject z_C to StyleGAN [17] by implementing multi-scale AdaIN.

The encoded features of the identity recognition network [6] generally contain the hairstyle, which may have an impact on the target background preservation while modifying the target face guided by $z_{id}(X_s)$. To maintain the high fidelity scene, background area in X_t is directly fused with the output of StyleGAN I_C via

$$\hat{Y}_{s,t} = X_t \odot M_{bg} + I_C \odot (1 - M_{bg}), \tag{3}$$

where $\odot$ means the element-wise product, and M_{bg} is the mask indicating the target background.

We use multiple code consistency losses on $\hat{Y}_{s,t}$ for styled face synthesis as follows. First, we calculate the distance between the identity features of $\hat{Y}_{s,t}$ and X_s via

$$\mathcal{L}_{id} = 1 - \langle z_{id}(\hat{Y}_{s,t}), z_{id}(X_s)\rangle, \tag{4}$$

where $\langle\cdot,\cdot\rangle$ means cosine similarity.

The expression and pose codes are supposed to maintain consistency for $\hat{Y}_{s,t}$ and X_t, denoted by

$$\mathcal{L}_{\text{exp}} = \left| z_{\text{exp}}(\hat{Y}_{s,t}) - z_{\text{exp}}(X_t) \right|_1, \tag{5}$$

$$\mathcal{L}_{pos} = \left| z_{pos}(\hat{Y}_{s,t}) - z_{pos}(X_t) \right|_1. \tag{6}$$

Let $\mathcal{L}_{GAN}$ be the adversarial loss to discriminate the generated face $\hat{Y}_{s,t}$ and the real face X_s via

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_X[\log D(X_s)] + \mathbb{E}_Y[\log(1 - D(\hat{Y}_{s,t}))]. \tag{7}$$

We utilize a semi-supervised learning strategy, *i.e.* the reconstruction loss is used to penalize the pixel-level distances between $\hat{Y}_{s,t}$ and X_t in the case of the same identity. It is formulated as

$$\mathcal{L}_{rec} = \begin{cases} \frac{1}{2}\left\|\hat{Y}_{s,t} - X_t\right\|_2^2 & if\ X_s = X_t \\ 0 & otherwise \end{cases}. \tag{8}$$

FaceShifter [18] designs the Adaptive Attentional Denormalization (AAD) layer to integrate attribute and identity embeddings, where the attentional mask is learned under the constraints of attribute and identity preservation losses. Note that $z_{id}(X_s)$ encoded by the face recognition model [6] usually contains additional style belonging to X_s^{bg}, *e.g.* hairstyle, which may result in uncontrollable appearance (Figure 2 row 1). It is essential to adopt a strong texture preservation function, to mitigate the influence of redundant identity information on the swapped face. To this end, we utilize contextual loss [23] to measure the feature similarity between $\hat{Y}_{s,t}$ and X_t. This loss allows face shape deformation and reduces the texture distortions after face swapping. It is formulated as

$$\mathcal{L}_{CX} = -\log(CX(F_{vgg}^l(\hat{Y}_{s,t}), F_{vgg}^l(X_t))), \tag{9}$$

where l means *relu*{3_2,4_2} layers of the pretrained VGG19 network.

Moreover, we combine path length regularization [17] to optimize the training of SFI-Net. It is formulated as

$$\mathcal{L}_{ppl} = \mathbb{E}_{w,y\sim N(0,I)}(\left\|J_w^T y\right\|_2 - a)^2, \tag{10}$$

where $J_w = \frac{\partial \hat{Y}_{s,t}}{\partial w}$, $w \sim f(C)$ means the intermediate latent code of the fused style code z_C through the mapping network, y is a random image with normal distribution and a is set as the exponential moving average of the perceptual path lengths. The total loss of SFI-Net is as follows

$$\begin{aligned}\mathcal{L}_{\textit{SFI-Net}} = \mathcal{L}_{GAN} + \lambda_{id}\mathcal{L}_{id} + \lambda_{exp}\mathcal{L}_{exp} + \lambda_{pos}\mathcal{L}_{pos} + \\ \lambda_{rec}\mathcal{L}_{rec} + \lambda_{CX}\mathcal{L}_{CX} + \lambda_{ppl}\mathcal{L}_{ppl}.\end{aligned} \tag{11}$$

3.2. Joint Refinement Network

SFI-Net can generate face with the identity style of X_s and other identity-irrelevant styles (*e.g.* expression, pose and texture) of X_t. Moreover, the result of SFI-Net preserves most of the background contents except for the facial occlusions, due to the mistakes of the parsing model. Other concerns are mainly some low-resolution situations, and the boundary artifacts when blending X_t^{bg} and I_C^{fg}.

We propose the Joint Refinement Network (JR-Net) to refine the face inpainting result of SFI-Net, as shown in

Stage II of Figure 3. For the purpose of making generated image occlusion-aware, we take the residual map of X_t and $\hat{Y}_{t,t}$ as the occlusion representation, and feed it into JR-Net:

$$\triangle X_t = X_t - \hat{Y}_{t,t}. \tag{12}$$

Attribute embedding of JR-Net comes from the high-level texture feature of X_t. Specifically, we use a pretrained VGG to extract the texture feature of $conv4\_1$ layer, and obtain the style code with 512 channels using adaptive average pooling. It is fused with the identity embedding of X_s, then injected to the AdaIN [14] residual blocks $AdaRes$ to adaptively transfer feature styles of $\{\hat{Y}_{s,t}, \triangle X_t\}$. It is formulated as:

$$AdaIN(h^i, \gamma^i, \beta^i) = \gamma^i_{\{id,att\}} \odot \frac{h^i - \mu^i}{\sigma^i} + \beta^i_{\{id,att\}}, \tag{13}$$

where $h^i \in \mathbb{R}^{C^i_h \times H^i \times W^i}$ is the embedding of $\{\hat{Y}_{s,t}, \triangle X_t\}$ encoded by Enc. μ^i and σ^i are the means and standard deviations of h^i, and they are used to perform instance normalization. Let $z^i_{tex} \in \mathbb{R}^{C^i_{tex} \times 1}$ and $z^i_{id} \in \mathbb{R}^{C^i_{id} \times 1}$ be the attribute and identity embedding. In the AdaIN form, $\gamma^i_{\{id,att\}}$ and $\beta^i_{\{id,att\}} \in \mathbb{R}^{C^i_h \times H^i \times W^i}$ are obtained from $\{z^i_{tex}(X_t), z^i_{id}(X_s)\}$ using several fully connected layers.

Moreover, we adopt a super-resolution refinement scheme based on multi-scale component dictionaries from high-quality reference images [20]. Specifically, in the dictionary feature transfer (DFT) module, the offline generated dictionaries are modulated via AdaIN, based on texture features of the corresponding components (*i.e.* left eye, right eye, nose, mouth) from $\{\hat{Y}_{s,t}, \triangle X_t\}$. Then the matched restored features are used for feature modulation in the decoder Dec, via the corresponding spatial feature transform (SFT) layer [40].

In the second stage, we obtain the refined face via

$$Y_{s,t} = \textit{JR-Net}(\triangle X_t, \hat{Y}_{s,t}, z_{tex}(X_t), z_{id}(X_s)). \tag{14}$$

To better preserve attributes, we utilize contextual loss to measure the feature similarity between $Y_{s,t}$ and X_t

$$\mathcal{L}^{'}_{CX} = -\log(CX(F^l_{vgg}(Y_{s,t}), F^l_{vgg}(X_t))). \tag{15}$$

In order to improve the performance of identity refinement, we impose an identity consistency loss via

$$\mathcal{L}^{'}_{id} = 1 - \langle z_{id}(Y_{s,t}), z_{id}(X_s) \rangle. \tag{16}$$

In a similar way, we combine a reconstruction loss between $Y_{s,t}$ and X_t, when X_s and X_t are the same identity

$$\mathcal{L}^{'}_{rec} = \begin{cases} \frac{1}{2}\|Y_{s,t} - X_t\|^2_2 & if\ X_s = X_t \\ 0 & otherwise \end{cases}. \tag{17}$$

We further add the perceptual loss to improve the high-resolution swapped face restoration

$$\mathcal{L}_{vgg} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} \|F^{(i)}_{vgg}(Y_{s,t}) - F^{(i)}_{vgg}(X_t)\|_2 & if\ X_s = X_t \\ 0 & otherwise, \end{cases} \tag{18}$$

where $F^{(i)}_{vgg}$ denotes the i-th convolution layer of VGG19 model. We set N equal to 4.

The adversarial loss in JR-Net ensures the image quality via

$$\mathcal{L}^{'}_{GAN}(G^{'}, D^{'}) = \mathbb{E}_X[\log D^{'}(X_s)] + \mathbb{E}_Y[\log(1 - D^{'}(Y_{s,t}))]. \tag{19}$$

Finally, the total loss of JR-Net is formulated as:

$$\mathcal{L}_{\textit{JR-Net}} = L^{'}_{GAN} + \lambda^{'}_{id}\mathcal{L}^{'}_{id} + \lambda^{'}_{rec}\mathcal{L}^{'}_{rec} + \lambda_{vgg}\mathcal{L}_{vgg} + \lambda^{'}_{CX}\mathcal{L}^{'}_{CX}. \tag{20}$$

4. Experiment

Implementation Details. Both SFI-Net and JR-Net are trained with the CelebA-HQ [15], FFHQ [16] and VGGFace [29] datasets. We align and crop the source and target faces using five point landmarks extracted by [4]. We train SFI-Net with 256×256 resolution and the final resolution of JR-Net is 512×512. More results can be found in the supplementary material. We acquire 3DMM parameters using 3DDFA2 [12]. The dimension of $z_{exp}(X_t)$ and $z_{pos}(X_t)$ is 10 and 12, respectively. The dimensions of $z_{id}(X_s)$, $z_{tex}(X^{fg}_t)$ and $z_{tex}(X_t)$ are 512. The output dimension of the fusion module in SFI-Net is 1024. In Equation 11 and 20, $\lambda_{id} = \lambda^{'}_{id} = 20$, and other weights are set to 10.

In JR-Net, we warp $\hat{Y}_{t,t}$ according to the facial landmarks of X_t, making $\triangle X_t$ focus more on the occlusion. We use a total of 8 AdaIN residual blocks. The number of downsampling and upsampling blocks in JR-Net is set to 3 and 4, respectively. As for data augmentation, we adopt multiple random operations, *e.g.* cropping, flipping, rotation, blur, variations on brightness, saturation and contrast, as well as color jittering.

4.1. Comparison with Other Methods

Comparison with Previous Inpainting Tasks: General inpainting tasks focus on the completion of hidden regions. For face inpainting tasks, previous methods [43, 44, 46] are not competent to generate the identity-guided completed images, and cause some artifacts in the facial area, as shown in Figure 4. Our inpainting results have better performance with respect to high fidelity attributes and identity factors.

Qualitative Comparison on FaceForensics++: FaceForensics++ [31] is a large-scale face forensics dataset that consists of 5,000 video clips. We compare FaceInpainter

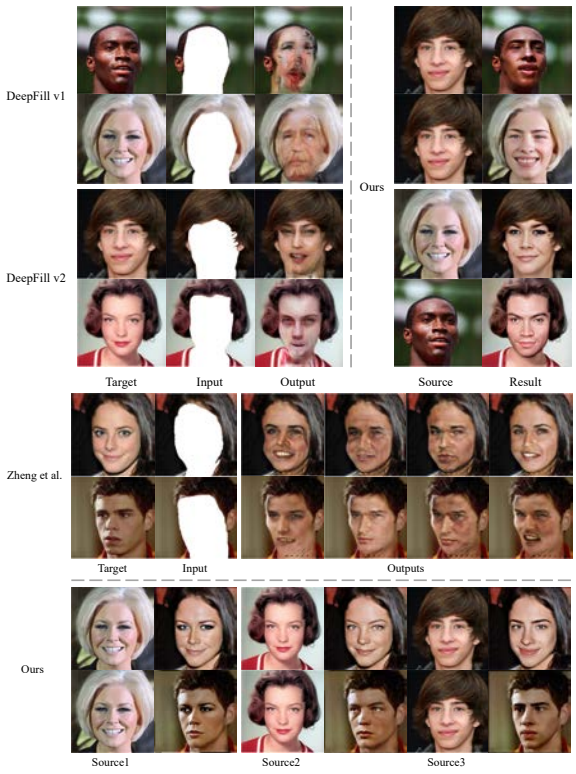

Figure 4: Comparison results with DeepFill v1 [43], DeepFill v2 [44] and Zheng et al. [46]. Our approach can generate high-quality identity-guided face inpainting results.

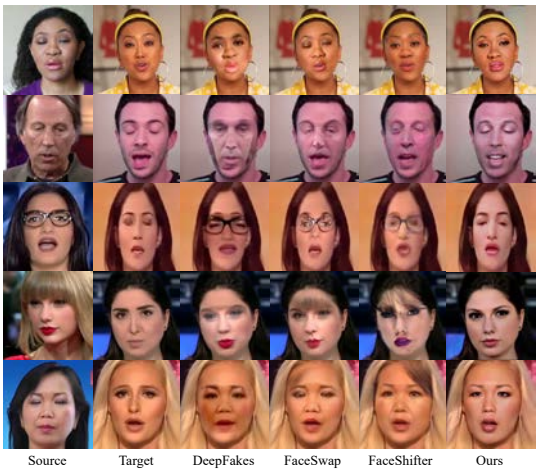

Figure 5: Comparison with FaceSwap [2], DeepFakes [1], and FaceShifter [18] on FaceForensics++ [31]. Our results better realize controllable attribute preservation and effective identity modification.

with DeepFakes [1], FaceSwap [2], and FaceShifter [18] on this benchmark. As shown in Figure 5, FaceShifter concentrates on more identity modification so that the source style is uncontrollable on the swapped face (*e.g.* eyebrow in row 1, eyes in row 2, eyeglass in row 3, and hair in rows 4&5). Similar artifacts are shown in the results of DeepFakes and FaceSwap. Our results strike a better balance between new identity and original attribute styles for the target face, and possess high-quality details on facial areas.

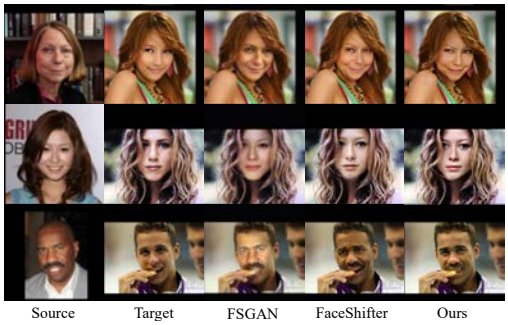

Figure 6: More comparison results with FSGAN [27] and FaceShifter [18]. Our results have competitive visual performance with respect to both the identity and attributes.

Figure 7: More comparison results with SimSwap [5]. Our method can generate swapped faces with fewer artifacts on the cheek and forehead areas, correctly tracking the gaze direction and expression of the target simultaneously.

Comparison with FaceShifter: As shown in Figure 6, FaceShifter is capable of generating highly identifiable face-swapping results, but not good enough at maintaining the attribute details, *e.g.*, gaze direction and skin color in row 3. In rows 1&2, the aligned attribute consistency loss proposed in [18] considers the low-level features and brings face shape style of the target to the generated face to some extent (col 4). Our swapping approach conducts high-level feature matching employing $\mathcal{L}_{CX}$, and the results preserve higher-fidelity target attributes as well as source face shape, with high-resolution appearance.

Comparison with SimSwap: SimSwap utilizes a weak feature matching loss [5] to efficiently improve the ability to preserve the facial attributes, *i.e.*, the attribute feature extractor abandons the first few layers of the discriminator. SimSwap is competent to modify identity and preserve attributes, but there are obvious artifacts (*e.g.* in the forehead and cheek) around the face boundary. Our results correctly

Methods	ID retrieval ↑	Pose ↓	Expression ↓	SMD2 ↑	Overall ↑
DeepFakes [1]	81.96	4.14	0.187	8.835	86.47
FaceSwap [2]	54.19	2.51	0.148	**9.573**	61.11
FaceShifter [18]	**97.38**	2.96	**0.136**	8.483	**102.77**
SimSwap [5]	92.83	**1.53**	-	-	-
w/o $z_{tex}(X_t^{fg})$	97.23	2.10	0.121	8.499	103.51
w/o $z_{exp}(X_t)$	**98.43**	1.93	0.161	8.660	105.00
w/o $z_{pos}(X_t)$	97.93	4.97	0.140	8.428	101.25
w/o $\mathcal{L}_{CX}$	98.13	2.31	0.135	8.602	104.29
w/ $C^{'}$	0.10	29.57	0.273	9.195	-20.55
SFI-Net	97.93	**1.68**	**0.117**	8.407	104.54
w/o DFT	**98.43**	2.18	0.139	8.773	104.88
w/o $\triangle X_t$	98.30	2.13	0.134	8.654	104.69
w/o AdaRes	97.68	1.85	0.123	8.791	104.50
FaceInpainter	97.63	2.21	0.141	**11.708**	**106.99**

Table 1: Quantitative comparison with DeepFakes [1], FaceSwap [2], FaceShifter [18], and SimSwap [5] on FaceForensics++ [31]. As a comprehensive comparison, we obtain the overall score by subtracting others from ID retrieval and Sum of Modulus of gray Difference (SMD2) [21] score.

	Official code	Should need image sequence for training	Public swapped results		
			FaceForensics++	IIIT-CFW	CUHK Sketch
DeepFakes	✓	✓	✓	×	×
FaceSwap	✓	×	✓	×	×
FaceShifter	×	×	✓	×	×
SimSwap	×	×	×	×	×

Table 2: The reason why we only compare with FaceSwap in Table 3.

Methods	ID retrieval ↑	Pose ↓	Exp ×10 ↓	FID ×0.1 ↓	SMD2 ×0.1 ↑	Overall ↑
FaceSwap_IIIT-CFW	59.00	3.66	**2.93**	7.236	1.705	46.88
FaceInpainter_IIIT-CFW	**89.13**	**1.90**	3.01	**3.646**	**2.802**	**83.38**
FaceSwap_CUHK	86.10	1.15	**2.13**	2.407	1.581	81.99
FaceInpainter_CUHK	**88.36**	**0.99**	2.23	**2.148**	**2.843**	**85.84**

Table 3: Quantitative evaluation on IIIT-CFW [25] and CUHK [39] dataset.

Figure 8: More qualitative results for identity frontalization and profile on CelebA-HQ dataset [15]. Our method can generate the swapped face for different identities both in front and extreme pose, given the target face.

track the gaze direction and expression, simultaneously better transfer new identity without visual artifacts, as shown in Figure 7.

Additional Qualitative Results: As shown in Figure 8, all the results of IGFI are generated by means of well-trained FaceInpainter, from various identities (row 1) to the corresponding front view (row 2), as well as to the profile view with the extreme posture (row 3). Based on the target attributes, FaceInpainter can generate photo-realistic faces with corresponding identities.

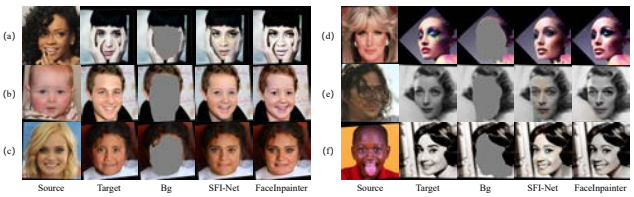

Figure 9: JR-Net handles some face-swapping synthesis issues, *e.g.*, fused boundaries, occlusions, face shape, skin color and low resolution.

Quantitative Comparison: We follow the experimental protocol adopted in [18, 5] for quantitative evaluation of identity transfer and pose preservation. Concretely, we apply CosFace [37, 26] to compute the ID retrieval score. Our method achieves a better score compared to other algorithms, as shown in Table 1. We extract pose features using [32]. SimSwap has a lower posture score based on $\mathcal{L}$-2 distance between the swapped and target face, but a relatively poor performance on ID retrieval, which demonstrates that it preserves excessive target attributes. As for expression estimation, we extract the coefficients using another 3DMM model [9]. Our performance is comparable with FaceShifter, in spite of only using expression code in the IGFI framework. In Table 1, w/o DFT means w/ $\triangle X_t$ and w/ AdaRes, w/o $\triangle X_t$ means w/ AdaRes and w/o DFT, w/o AdaRes means w/ $\triangle X_t$ and w/o DFT, FaceInpainter means w/ $\triangle X_t$, w/ AdaRes, w/ DFT.

Considering the swapped face has no paired high-resolution face as the reference image to operate objective metrics, *i.e.*, PSNR and SSIM, we implement an evaluation based on gray scale difference [21]. With DFT module, our results achieve a better score of the image sharpness.

IIIT-CFW [25] is used in the cartoon face classification and Photo2Cartoon tasks. CUHK [39] is for sketch synthesis and face sketch recognition. The two datasets have no image sequences or videos. Different from the above tasks, we first study the heterogeneous identity swapping problem. As for IIIT-CFW and CUHK Face Sketch, we collect N= 1000 and 606 images for quantitative experiments, respectively (Table 3). The ith target face is swapped with the identity of the $N-i$ th face. The Frechet Inception Distance (FID) [13] is used to measure both the quality and diversity of the swapped faces.

Human Evaluation: After a brief introduction to the IGFI task, 10 users are instructed to observe the source, target and swapped faces. The participants need to give scores according to their own experiences. Scores are recorded as

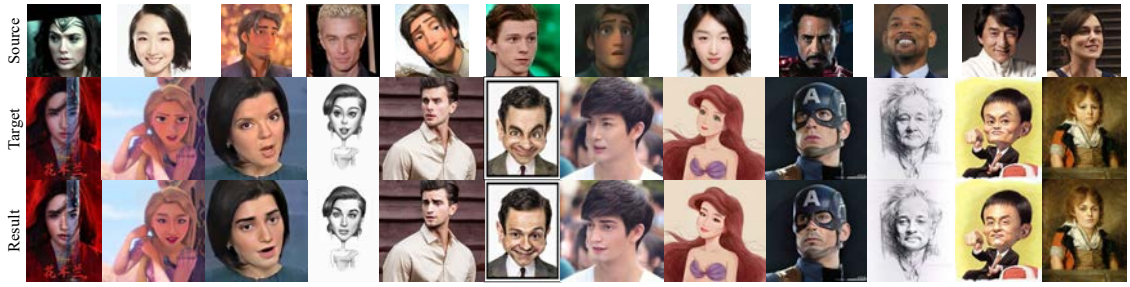

Figure 10: IGFI results on various wild face images based on FaceInpainter. Face shape is an important attribute for the target face, especially when it comes to exaggerated drawing and 3D cartoons (cols 4&6). While we implement face adaptation to heterogeneous domains, the unique styles are maintained.

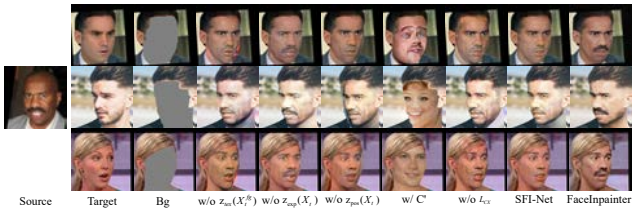

Figure 11: Ablation study of SFI-Net.

Methods	Identity	Attribute	Boundary	Occlusion	Realism
DeepFakes [1]	43.96	34.14	22.57	44.11	20.56
FaceSwap [2]	31.64	41.22	50.55	42.25	47.89
FaceShifter [18]	93.64	81.22	84.34	**94.11**	85.66
SFI-Net	90.83	71.53	55.78	40.22	50.58
FaceInpainter	**95.83**	**89.53**	**88.34**	93.67	**92.50**

Table 4: User study results.

1-100 from five aspects: (a) identity perception, (b) attribute preservation, (c) boundary naturalness, (d) occlusion awareness, and (e) overall realism, where a higher score means higher quality. There are 1000 observed images from FaceForensics++ [31]. We collect 10000 human decisions, and the average scores are shown in Table 4.

4.2. Analysis of FaceInpainter

The necessity of Style Code: As shown in Figure 11, we show the corresponding results of different style code settings. There is texture distortion in the completed face area (col 4) in the case of IGFI without the texture code of the foreground. The gaze direction or mouth motion are not fitted well, if without expression code (col 5). As for posture, in the case of profile face inpainting, an obvious gap between the inner face and background can be found in the synthesized face when discarding the pose code (col 6). Moreover, inspired by [30], we use a learnable encoder to exploit the multi-level attribute code C_{att}, instead of using 3D priors in SFI-Net. C_{att} is supposed to contain all the attribute style codes in C of SFI-Net and even more details of X_t. Then the fused style code $C^{'} = \{C_{att}(X_t), z_{id}(X_s)\}$ is sent to StyleGAN [17] for styled face inpainting. However, we find the results can not correctly fit the identity, expression, texture and posture, as shown in col 7. The results of SFI-Net indicate that all style codes are necessary for the IGFI framework (col 9).

Efficient Contextual Loss: As shown in Figure 11, contextual loss helps to efficiently preserve the texture of the target (cols 9&10). If discarding $\mathcal{L}_{CX}$, the result will exhibit a poor performance with lots of texture distortions (col 8).

Refinement of JR-Net: In the first stage, the facial occlusions have been removed (Figure 9 (a), (d) and (e)), and there are shape and appearance inconsistency between the transferred face and original target background, therefore the details concerning identity and attributes in the generated face of SFI-Net need to be refined by means of JR-Net.

5. Conclusion

We have proposed FaceInpainter to implement effective identity swapping, achieving high fidelity identity and attributes. Different from previous identity swapping methods, FaceInpainter adopts an inpainting way for swapping so that it can better keep various heterogeneous scene information. SFI-Net uses 3D priors to make the representation of target attributes more accurate during rendering. Moreover, contextual structure and texture details are well refined in JR-Net. Extensive experiments demonstrate that our approach can handle high-quality and occlusion-aware identity adaptation to heterogeneous domains.

6. Acknowledgment

This work is partially funded by Beijing Natural Science Foundation (Grant No. JQ18017), Youth Innovation Promotion Association CAS (Grant No. Y201929), and National Natural Science Foundation of China (Grant No. U20A20223).

References

[1] Deepfakes github. https://github.com/deepfakes/faceswap.

[2] Faceswap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski, Accessed: September 30, 2019.

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[4] Dong Chen, Shaoqing Ren, Yichen Wei, Xudong Cao, and Jian Sun. Joint cascade face detection and alignment. In *European conference on computer vision*, pages 109–122. Springer, 2014.

[5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020.

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

[7] Qiyao Deng, Jie Cao, Yunfan Liu, Zhenhua Chai, Qi Li, and Zhenan Sun. Reference guided face component editing. *arXiv preprint arXiv:2006.02051*, 2020.

[8] Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[10] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. In *NeurIPS*, 2019.

[11] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *TPAMI*, 2021.

[12] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. *arXiv preprint arXiv:2009.09960*, 2020.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[18] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Peipei Li, Yinglu Liu, Hailin Shi, Xiang Wu, Yibo Hu, Ran He, and Zhenan Sun. Dual-structure disentangling variational generation for data-limited face parsing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 556–564, 2020.

[20] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020.

[21] YF Li, NN Chen, and JC Zhang. Fast and high sensitivity focusing evaluation function. *Application Research of Computers*, 27(4):1534–1536, 2010.

[22] Y. Li, S. Liu, J. Yang, and M. Yang. Generative face completion. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5892–5900, 2017.

[23] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 768–783, 2018.

[24] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020.

[25] Ashutosh Mishra, Shyam Nandan Rai, Anand Mishra, and CV Jawahar. Iiit-cfw: A benchmark database of cartoon faces in the wild. In *European Conference on Computer Vision*, pages 35–47. Springer, 2016.

[26] MuggleWang. Cosfacepytorch. https://github.com/MuggleWang/CosFace_pytorch, Accessed: 2018.

[27] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE international conference on computer vision*, pages 7184–7193, 2019.

[28] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[29] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[30] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[31] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019.

[32] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[34] Linsen Song, Jie Cao, Lingxiao Song, Yibo Hu, and Ran He. Geometry-aware face completion and editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2506–2513, 2019.

[35] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019.

[36] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, June 2018.

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

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[39] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1955–1967, 2008.

[40] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[41] Xiang Wu, Huaibo Huang, Vishal M Patel, Ran He, and Zhenan Sun. Disentangled variational representation for heterogeneous face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9005–9012, 2019.

[42] Aijing Yu, Haoxue Wu, Huaibo Huang, Zhen Lei, and Ran He. Lamp-hq: A large-scale multi-pose high-quality database and benchmark for nir-vis face recognition. *International Journal of Computer Vision*, pages 1–17, 2021.

[43] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019.

[45] S. Zhang, R. He, Z. Sun, and T. Tan. Demeshnet: Blind face inpainting for deep meshface verification. *IEEE Transactions on Information Forensics and Security*, 13(3):637–647, 2018.

[46] C. Zheng, T. Cham, and J. Cai. Pluralistic image completion. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1438–1447, 2019.

[47] Hang Zhou, Jihao Liu, Ziwei Liu, Yu Liu, and Xiaogang Wang. Rotate-and-render: Unsupervised photorealistic face rotation from single-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[48] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):78–92, 2017.

[49] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019.