FReeNet: Multi-Identity Face Reenactment

Jiangning Zhang[1*†] Xianfang Zeng[1†] Mengmeng Wang[1†] Yusu Pan[1]
Liang Liu[1] Yong Liu[1‡] Yu Ding[2] Changjie Fan[2]
[1] Zhejiang University [2]Fuxi AI Lab, NetEase
{186368, zzlongjuanfeng, mengmengwang, corenel, leonliuz}@zju.edu.cn
yongliu@iipc.zju.edu.cn, {dingyu01, fanchangjie}@corp.netease.com

Abstract

This paper presents a novel multi-identity face reenactment framework, named FReeNet, to transfer facial expressions from an arbitrary source face to a target face with a shared model. The proposed FReeNet consists of two parts: Unified Landmark Converter (ULC) and Geometry-aware Generator (GAG). The ULC adopts an encode-decoder architecture to efficiently convert expression in a latent landmark space, which significantly narrows the gap of the face contour between source and target identities. The GAG leverages the converted landmark to reenact the photorealistic image with a reference image of the target person. Moreover, a new triplet perceptual loss is proposed to force the GAG module to learn appearance and geometry information simultaneously, which also enriches facial details of the reenacted images. Further experiments demonstrate the superiority of our approach for generating photorealistic and expression-alike faces, as well as the flexibility for transferring facial expressions between identities.

1. Introduction

Face reenactment is a task to transfer the facial expression from one source face to a target face, which has vast promising applications such as film-making, facial animations, and augmented reality. Moreover, a live video of a particular person can be generated under the control of another person, which can attack or strengthen the bioassay system. In this work, we focus on solving a more challenging task: multi-identity face reenactment, where the source face is from an arbitrary person and the target person is not specific. This task is distinguished from one(many)-to-one face reenactment tasks in whether the target is specific, which is more general and flexible in practical applications.

Benefiting from the release of the large-scale face datasets [11, 20, 8, 41], many high-accuracy and reliable face detection algorithms are proposed [38, 21, 4, 9]

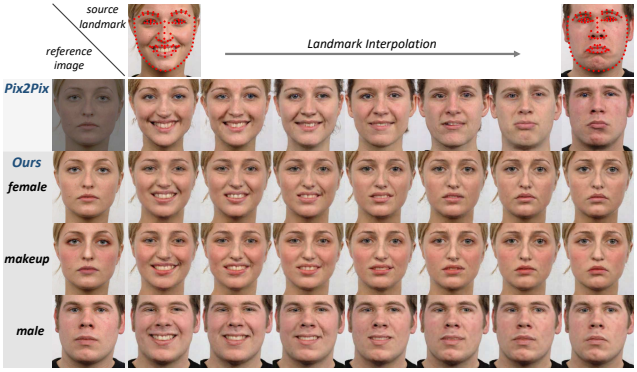

Figure 1. A toy experiment of the face reenactment task. The results of the first row are from *Pix2Pix* [12] method with the only landmark as input, while the second to fourth rows are from our method with different reference faces but the same geometry information. Note that the reference image of the third row is the makeup of the girl in the second row.

that contribute to the development of the face reenactment task. Many excellent facial expression migration and face reenactment methods have been proposed in the last decade, which summarily fall into two main categories: three dimensional (3D) model based synthesis and GAN based generation. For the 3D model based synthesis methods [34, 17, 5, 31, 33], the person is represented by a predefined parametric model. Generally, the methods firstly capture the facial movement of a source video that will be fitted into a parametric space over the predefined model and then render the target video through morphing. These techniques are famous for the animation of computer graphics (CG) avatars in both games and movies [26] because they have high-quality and high-resolution face reenactment ability. However, these methods generally suffer from big-budget model making and are computationally expensive.

Recently, many GAN based methods have achieved a significant improvement [6, 13, 40, 39, 26, 37] due to the natural advantages in learning distribution patterns from a large-scale dataset. Generally, the encoder-decoder structure is adopted to reenact target face in an adversarial idea [12], and some further works [43, 13, 30, 40] introduce a powerful cycle consistent loss to achieve unpaired face

*Work mainly done during an internship at NetEase Fuxi AI Lab.
†Equal contribution.
‡Corresponding author.

reenactment. However, they can only reenact faces between two specific identities once the network is trained. Subsequently, the many-to-one face reenactment task sets out to solve the above problem, which can reenact one target face from multiple persons using only one network. Recent Nirkin *et al*. [25] derive a recurrent neural network based approach for the subject agnostic face reenactment, but it requires extra segmentation information and has much more parameters than common methods. ReenactGAN [39] introduces a *Transformer* module to adapt source facial movements from multiple persons to the target person in a latent facial boundary space, and then decodes the target face. Nonetheless, it is still network-inefficient in practical applications for requiring a transformer and a decoder for each target person. So it is of considerable significance to implement a multi-identity face reenactment task, also called many-to-many face reenactment, in a unified network, where both the source and reference faces can be from multiple persons. X2Face [40] achieves the task by first adopting an embedding network to encode an embedded face and then using a driving network to reenact the target face, but the generative images are not satisfying in quality and facial details. In summary, there are still two main challenges in this task: (1) How to convert multi-identity facial expression by a unified network, because there exists a gap in facial contours between source and target persons. (2) How to reenact photorealistic and identity-consistent target face as the reference face while keeping consistent with the pose, hue, and illumination.

To address the issues above, we propose a multi-identity face reenactment framework named FReeNet, which can efficiently transfer expressions from arbitrary source identities to target identities. Firstly, a landmark detector [9] is leveraged to encode faces into a latent landmark space. This latent space serves as a high-quality bridge for the next facial converting step, where the facial geometry information is efficiently preserved while the appearance information is omitted. Subsequently, a unified landmark converter module is applied to effectively convert the expression of an arbitrary source person to the target person in latent landmark space. Then the geometry-aware generator simultaneously extracts geometry information from the converted landmark and appearance information from the reference person to reenact the target face. Such a decoupling design can generate distinguishing faces with similar or same landmark inputs, because different persons with different appearances are used as the reference face. Furthermore, we combine the triplet loss [28] and the perceptual loss [14] to form a novel triplet perceptual loss, which helps generate more detailed images as well as decouple the appearance and geometry information. As shown in Figure 1, our approach can effectively preserve the reference identity while converting the geometry information.

To our best knowledge, the proposed FReeNet is the first to successfully perform multi-identity face reenactment task using a unified model, while keeping the pose, hue, and illumination information consistent with the reference face. Specifically, we make the following four contributions:

- A unified landmark converter is proposed to convert the expression from the source identity to the target identity, and both the source and target identities are from multiple persons.
- A geometry-aware generator is proposed to reenact the photorealistic target face, which is designed in a decoupling idea and extracts the appearance and geometry information from separate paths.
- A new triplet perceptual loss is proposed to enrich the facial details of the reenacted face.
- Experimental results indicate that the proposed approach implements the many-to-many face reenactment task and can generate high-quality and identity-consistent face images.

2. Related Work

Image Synthesis. Driven by remarkable generation effects of GAN [7], researchers have achieved excellent results in various domains, such as image translation [27, 12, 43, 35], person image synthesis [3, 23], and face generation [2, 26, 15, 16]. Mehdi *et al*. [24] designed a cGAN structure to condition on to both the generator and discriminator for a more controllable generation of attributes. Subsequently, the Pix2Pix [12] achieved incredible results in paired image translation tasks by using L1 and adversarial losses between the generated image and the ground-truth. Zhu *et al*. [43] subsequently proposed a new cycle consistency loss for unpaired image translation between two domains, which dramatically reduces the requirement of the data annotation. DualGAN [42] analogously learns two translators from one domain to the other and hence can solve general-purpose image-to-image translation tasks. Furthermore, StarGAN [2] proposed a unified model for multi-domain facial attribute transferring and expression synthesis. Recently, some methods can generate vivid faces directly from the latent input code. Tero *et al*. [15] described a new progressive growing training methodology for the face generation from an underlying code. StyleGAN [16] proposed a style-based generator that embeds the latent input code into an intermediate latent space, which can control the weights of image features at different scales and synthesize extremely naturalistic face images. However, using latent code as the input is an uncontrollable generation process that is not suitable for the face synthesis task, and they have limited scalability in handling the many-to-many face reenactment task. Our method introduces a landmark space for expression transferring among multiple persons and uses converted landmark images as the guidance to reenact target faces, which is not the same as the existing methods.

Face Reenactment. Benefiting from large-scale face database collections [11, 20, 41] and reliable landmark detectors [9, 38, 4, 21], numerous impressive face reenactment methods have been proposed in recent years. These methods can be roughly categorized into either 3D model based method or GAN based method. In the 3D model based branch, Volker *et al*. [1] proposed the morphable 3D model to estimate the parameters of the shape-vector and texture-vector, which are then used to recover full 3D shape and texture of the face. Subsequent Face2face [32] applied an efficient deformation transfer to tracked facial expressions of both source and target videos, and then re-rendered the synthesized target faces for better fitting the retrieved and warped mouth interior. Ma *et al*. [22] reconstructed high-resolution facial geometry and appearance by capturing an individual-specific face model with fine-scale details. Those 3D model based methods generally require delicately designed models, which are time and money consuming, and they are also computationally expensive. Therefore, Albert *et al*. [26] introduced GANimation to control the magnitude of activation of each AU and then combine several of them to synthesize target faces, which only requires images without other procedures. Jin *et al*. [13] directly applied CycleGAN to transfer facial expressions between two identities. Recently, Wu *et al*. proposed ReenactGAN [39] that is capable of transferring facial movements from one monocular image of multiple persons to a specific target person. However, our framework aims at solving the harder many-to-many face reenactment problem using a unified concise framework, which has more promising applications.

3. FReeNet

In this paper, a novel framework named FReeNet is proposed to complete the multi-identity face reenactment task efficiently. As depicted in Figure 2, we first adopt the face landmark detector [9] to encode two input images $\boldsymbol{I}_{T,r}$ and $\boldsymbol{I}_{S,n}$ ($\in \mathbb{R}^{3\times256\times256}$) to a latent landmark space $\boldsymbol{l}_{T,r}$ and $\boldsymbol{l}_{S,n}$ ($\in \mathbb{R}^{106\times2}$), where the first subscript means identity (T means target person and S means source person) and the second is expression (r means the reference expression and n means an arbitrary expression). For example, the $\boldsymbol{l}_{T,r}$ represents the landmark from the target person with the reference expression (the neutral expression is used as the reference expression in the paper) that can be in a different pose. The unified landmark converter subsequently adapts the source expression to the target, denoted as $\boldsymbol{\psi} : (\boldsymbol{l}_{T,r}, \boldsymbol{l}_{S,n}) \rightarrow \hat{\boldsymbol{l}}_{T,n}$. Finally, the geometry-aware generator simultaneously leverages the converted geometry information $\hat{\boldsymbol{L}}_{T,n} \in \mathbb{R}^{1\times64\times64}$ and the appearance information $\boldsymbol{I}_{T,r} \in \mathbb{R}^{3\times256\times256}$ to reenact the target face $\hat{\boldsymbol{I}}_{T,n} \in \mathbb{R}^{3\times256\times256}$, denoted as $\boldsymbol{\phi} : (\hat{\boldsymbol{L}}_{T,n}, \boldsymbol{I}_{T,r}) \rightarrow \hat{\boldsymbol{I}}_{T,n}$. $\hat{\boldsymbol{L}}_{T,n}$ represents plotted landmark image from generated landmark vector $\hat{\boldsymbol{l}}_{T,n}$. Moreover, a new triplet perceptual loss is introduced to boost the performance of the GAG.

3.1. Unified Landmark Converter

As mentioned in [39], it may lead to artifacts if we directly apply ill-suited facial contour to synthesis the target image. In contrast to existing methods, we design a unified landmark converter module (ULC) to adapt the source expression from an arbitrary person to the target person. It can significantly alleviate the geometrical gap between the source and target faces. As shown in Figure 2 (top), the proposed ULC module contains two landmark encoders ($\boldsymbol{\psi}_{\alpha_1}$ and $\boldsymbol{\psi}_{\alpha_2}$) and a landmark shift decoder ($\boldsymbol{\psi}_{\alpha_3}$). Encoders $\boldsymbol{\psi}_{\alpha_1}$ and $\boldsymbol{\psi}_{\alpha_2}$ extract landmark features of the target and source faces respectively, and then the decoder $\boldsymbol{\psi}_{\alpha_3}$ fuses them to estimate the landmark shift $\boldsymbol{l}_{shift}$. After that, we add $\boldsymbol{l}_{T,r}$ and $\boldsymbol{l}_{shift}$ in a point-wise manner to get the converted landmark $\hat{\boldsymbol{l}}_{T,n}$, which has the facial contour with $\boldsymbol{I}_{T,r}$ while keeping the expression information of $\boldsymbol{I}_{S,n}$. This process is denoted as:

$$\begin{aligned}\hat{l}_{T,n} =& \boldsymbol{l}_{T,r} + \boldsymbol{l}_{shift} \\ =& \boldsymbol{l}_{T,r} + \boldsymbol{\psi}_{\alpha_3}(\boldsymbol{\psi}_{\alpha_1}(\boldsymbol{l}_{T,r}), \boldsymbol{\psi}_{\alpha_2}(\boldsymbol{l}_{S,n})).\end{aligned} \tag{1}$$

During the training phase, the overall loss function $\mathcal{L}_{ULC}$ is defined as:

$$\mathcal{L}_{ULC} = \lambda_1\mathcal{L}_{L1} + \lambda_2\mathcal{L}_{cyc} + \lambda_3\mathcal{L}_D, \tag{2}$$

where $\lambda_i (i = 1, 2, 3)$ are weights of the three loss functions.

Point-wise L1 Loss. The first term $\mathcal{L}_{L1}$ is defined by the point level $l1$ loss function to calculate errors of the landmark coordinates:

$$\mathcal{L}_{L1} = ||\hat{\boldsymbol{l}}_{T,n} - \boldsymbol{l}_{T,n}||_1. \tag{3}$$

Cycle Consistent Loss. The second term $\mathcal{L}_{cyc}$ constrains that the converted $\hat{L}_{T,n}$ is capable of converting back again:

$$\mathcal{L}_{cyc} = ||\boldsymbol{\psi}(\boldsymbol{l}_{S,r}, \boldsymbol{\psi}(\boldsymbol{l}_{T,r}, \boldsymbol{l}_{S,n})) - \boldsymbol{l}_{S,n}||_1, \tag{4}$$

where $L_{S,r}$ represents the reference expression of person S.

Adversarial Loss. Here we regard the ULC $\boldsymbol{\psi}$ as a generator, and the third term $\mathcal{L}_D$ contains two discriminators (D_{TF} and D_S) to make $\boldsymbol{\psi}$ more accurate and robust. The discriminator D_{TF} is used to judge $real$ or $fake$ of the landmark, and D_S is used to estimate the identity similarity score of the landmark pair. Two discriminator losses are defined as:

$$\begin{aligned}\mathcal{L}_{D_{TF}} =& \mathbb{E}_{x\sim p_{data}(x)}[\log(D_{TF}(x))] + \\ & \mathbb{E}_{z\sim p_{data}(z)}[\log(1 - D_{TF}(\boldsymbol{\psi}(z)))],\end{aligned} \tag{5}$$

$$\begin{aligned}\mathcal{L}_{D_S} =& \mathbb{E}_{x_1,x_2\sim p_{data}(x)}[\log(D_S(x_1, x_2))] + \\ & \mathbb{E}_{z\sim p_{data}(z),x_1\sim p_{data}(x)}[\log(1 - D_S(x_1, \boldsymbol{\psi}(z)))],\end{aligned} \tag{6}$$

where x indicates real landmark data space, and z indicates an input space of $\boldsymbol{\psi}$.

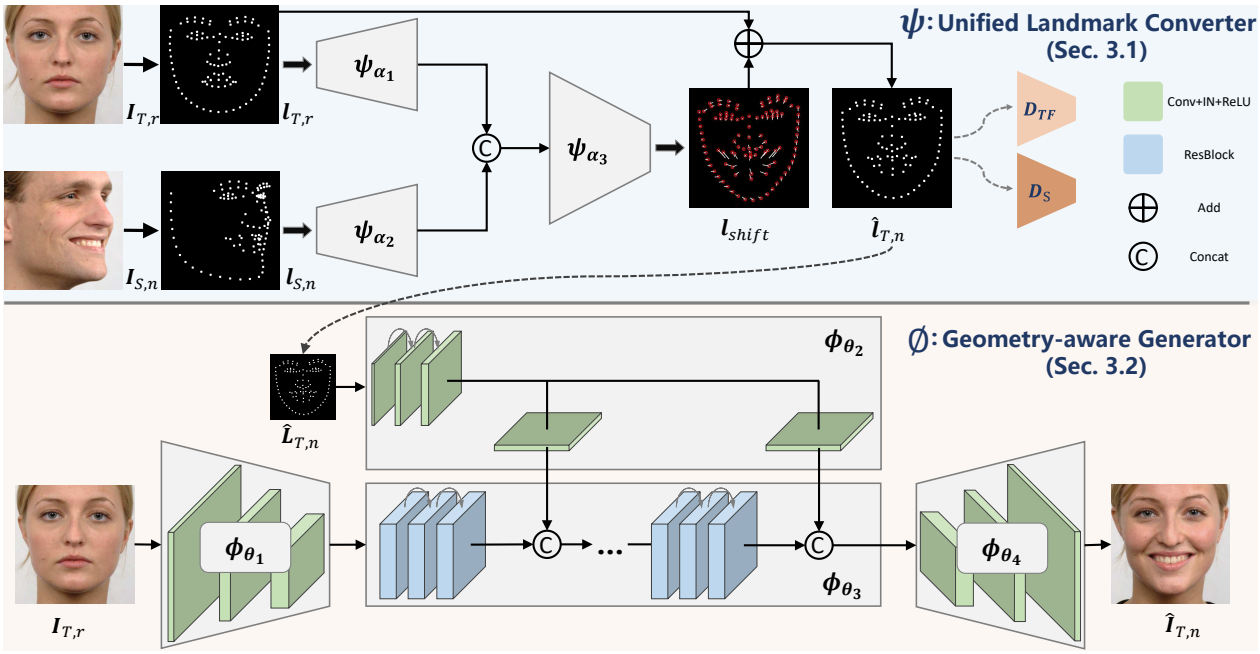

Figure 2. Overview of the proposed FReeNet. The net consists of a unified landmark converter ψ and a geometry-aware generator ϕ. Given a source person with an arbitrary expression $\boldsymbol{I}_{S,n}$ and a target person with the reference expression $\boldsymbol{I}_{T,r}$, the converter ψ uses extracted landmarks $\boldsymbol{l}_{S,n}$ and $\boldsymbol{l}_{T,r}$ to regress a landmark shift $\boldsymbol{l}_{shift}$, and then constructs the converted landmark $\hat{\boldsymbol{l}}_{T,n}$. Two discriminators $\boldsymbol{D}_{TF}$ and $\boldsymbol{D}_S$ are adopted for adversarial training. After that, geometry-aware generator ϕ reenacts target face $\hat{\boldsymbol{I}}_{T,n}$ under the guidance of $\hat{\boldsymbol{L}}_{T,n}$, where $\boldsymbol{I}_{T,r}$ is used as the reference image. $\alpha_i (i = 1, 2, 3)$ and $\theta_j (j = 1, 2, 3, 4)$ represent partial parameters of the network.

3.2. Geometry-aware Generator

Given a target reference face image $\boldsymbol{I}_{T,r}$ and a converted landmark $\hat{\boldsymbol{L}}_{T,n}$, the GAG reenacts the target face $\hat{\boldsymbol{I}}_{T,n}$ which is in the same expression with the source face $\boldsymbol{I}_{S,n}$. Specifically, the GAG is designed based on the commonly used *Pix2Pix* framework [12] in a decoupling thought. It simultaneously learns the appearance information from $\boldsymbol{I}_{T,r}$ and the geometry information from $\hat{\boldsymbol{L}}_{T,n}$ in different paths. As shown in Figure 2 (bottom), the GAG consists of an image encoder ϕ_{θ_1}, a landmark encoder ϕ_{θ_2}, a transformer ϕ_{θ_3}, and an image decoder ϕ_{θ_4}. The transformer consists of three ResBlock parts, with each part connecting to the output of the landmark encoder. This kind of design ensures that the geometric information ($\hat{\boldsymbol{L}}_{T,n}$) will be enhanced for the output image, and the process can be described as:

$$\begin{aligned}\hat{\boldsymbol{I}}_{T,n} &= \phi(\boldsymbol{I}_{T,r}, \hat{\boldsymbol{L}}_{T,n}) \\ &= \phi_{\theta_4}(\phi_{\theta_3}(\phi_{\theta_1}(\boldsymbol{I}_{T,r}), \phi_{\theta_2}(\hat{\boldsymbol{L}}_{T,n}))).\end{aligned} \tag{7}$$

GAG is designed in a decoupling idea and can efficiently reenact identity-reserved and expression-converted face images among multiple persons. During the training phase of the GAG, the full loss function $\mathcal{L}_{GAG}$ is defined as:

$$\mathcal{L}_{GAG} = \lambda_{pix}\mathcal{L}_{pix} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{TP}\mathcal{L}_{TP}, \tag{8}$$

where λ_L, λ_{adv}, and λ_{TP} represent weight parameters.

Pixel-wise L1 Loss. The first term $\mathcal{L}_{pix}$ calculates $l1$ errors between generated and supervised images:

$$\mathcal{L}_{pix} = ||\hat{\boldsymbol{I}}_{T,n} - \boldsymbol{I}_{T,n}||_1. \tag{9}$$

Adversarial Loss. The second term $\mathcal{L}_{adv}$ introduces the discriminator to improve the realism of the generated images in an adversarial idea:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{x\sim p_{data}(x)}[\log(D(x))] + \\ & \mathbb{E}_{k\sim p_{data}(k), l\sim p_{data}(l)}[\log(1 - D(\phi(k, l)))],\end{aligned} \tag{10}$$

where x indicates real image data space, and k and l represent input image and landmark space respectively of ψ. Discriminator D is similar to the work[43].

Triplet Perceptual Loss. The third term $\mathcal{L}_{TP}$ does duty for intra-class and inter-class evaluations, which helps generate images with more details as well as decouple the appearance and geometry information. It will be specified in the following section 3.3.

3.3. Triplet Perceptual Loss

During the training phase, we find that the GAG module is ill-conditioned to learn a mapping between the input landmark and the generated image if only under the supervision of the adversarial and L1 losses. This problem is caused by the different distributions between the RGB and landmark images. The generator tends to only learn from the landmark since its distribution is simple. In order to conquer this problem, we combine the triplet loss [28] and the perceptual loss [14] to form a novel triplet perceptual (TP) loss, which can maximize inter-class and minimize intra-class perceptual variations.

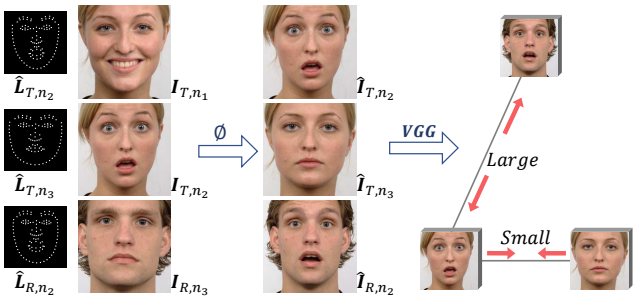

Figure 3. Schematic diagram of the triplet perceptual loss. Simultaneously maximizing inter-class perceptual variation of reenacted images ($\hat{\boldsymbol{I}}_{T,n_2}$ and $\hat{\boldsymbol{I}}_{R,n_2}$) and minimizing intra-class perceptual variation ($\hat{\boldsymbol{I}}_{T,n_2}$ and $\hat{\boldsymbol{I}}_{T,n_3}$).

As shown in Figure 3, two images $\boldsymbol{I}_{T,n_1}$ and $\boldsymbol{I}_{T,n_2}$ with arbitrary expressions (n_1 and n_2) are randomly selected within the target person T, while the third image $\boldsymbol{I}_{R,n_3}$ is randomly selected within another person R in an arbitrary expression n_3. Images $\hat{\boldsymbol{I}}_{T,n_2}, \hat{\boldsymbol{I}}_{T,n_3}, \hat{\boldsymbol{I}}_{R,n_2}$ are generated by the GAG with inputs $(\hat{\boldsymbol{L}}_{T,n_2}, \boldsymbol{I}_{T,n_1})$, $(\hat{\boldsymbol{L}}_{T,n_3}, \boldsymbol{I}_{T,n_2})$, and $(\hat{\boldsymbol{L}}_{R,n_2}, \boldsymbol{I}_{R,n_3})$ respectively. $\hat{\boldsymbol{L}}_{R,n_2}$ indicates converted landmark image is from identity T to R with expression n_2, and so do $\hat{\boldsymbol{I}}_{T,n_2}$ and $\hat{\boldsymbol{I}}_{T,n_3}$. Then the TP loss is applied to distinguish images generated by similar landmarks but different reference persons denoted as:

$$\mathcal{L}_{TP}(\hat{\boldsymbol{I}}_{T,n_2}, \hat{\boldsymbol{I}}_{T,n_3}, \hat{\boldsymbol{I}}_{R,n_2}) = \Big[m + D\left(\kappa(\hat{\boldsymbol{I}}_{T,n_2}), \kappa(\hat{\boldsymbol{I}}_{T,n_3})\right) - D\left(\kappa(\hat{\boldsymbol{I}}_{T,n_2}), \kappa(\hat{\boldsymbol{I}}_{R,n_2})\right)\Big]_+, \tag{11}$$

where m is the margin for controlling intra and inter gaps; $\kappa(\cdot)$ means features extraction operation by VGG [29]; $D(\cdot,\cdot)$ means L2 distance; $+$ means the value is positive.

Without TP loss, $\hat{\boldsymbol{I}}_{T,n_2}$ is only a converted intra-class image under the supervision of its ground-truth, which means the GAG tends to couple the landmark and the generated face naturally. By comparison, the GAG has additional inter-class or intra-class constrains for the generated $\hat{\boldsymbol{I}}_{T,n_2}$ from $\hat{\boldsymbol{I}}_{T,n_3}$ and $\hat{\boldsymbol{I}}_{R,n_2}$ when using the TP loss. In this case, the TP loss forces one landmark to participate in the face reenactment of all target persons in the dataset. Thus the GAG has to extract features from the reference face and the landmark image simultaneously to reenact the target person.

3.4. Training Scheme

The training process of the FReeNet consists of two phases. In the first phase, we train the *ULC* module from scratch with the loss function defined in Eq. 2, where the corresponding loss weights are set as $\lambda_1 = 100$, $\lambda_2 = 10$, and $\lambda_3 = 0.1$. Then we fix the parameters of the trained *ULC* module and learn the parameters of *GAG* module in the second phase, where the loss weights λ_{pix}, λ_{adv}, and λ_{TP} are 100, 1, and 0.1, respectively. The margin value m of the TP loss is set 0.3 in all experiments.

4. Experiments

In this section, we evaluate our approach on the aforementioned RaFD and Multi-PIE datasets and make a contrastive analysis with the state-of-the-art methods. Moreover, some ablation studies on the RaFD dataset are conducted to illustrate the effect of each proposed component in FReeNet, and additional images in the wild are further tested. Finally, a series of landmark interpolation and manipulation experiments on the RaFD dataset are performed to highlight the decoupling superiority of our approach. We also supply a demo video in the supplementary file for more generative results.

4.1. Datasets and Implementation Details

RaFD. The Radboud Faces Database (RaFD) [19] consists of 8,040 images collected from 67 participants. Each participant makes eight facial expressions in three different gaze directions and five different angles, and all $45°$, $90°$, and $135°$ face images are used in the paper. Images are cropped to 416×416 with face-centered and then resized to 256×256. The landmark with 106 key points for each face image is provided by HyperLandmark [9].

Multi-PIE. A total of 337 subjects (more than 750,000 images) are recorded inside the CMU 3D room using a hardware-synchronized network of 15 high-quality video cameras and 18 flashes. The detailed processing of faces is similar to the *RaFD* dataset.

Evaluation Metrics. We use *Amazon Mechanical Turk* (AMT) to evaluate the visual quality of reenacted images, *Structural Similarity* (SSIM) [36] to measure the structural similarity between generated and real images, and *Fréchet Inception Distance* (FID) [10] to measure the realism and variation of generated images.

Implementation Details. We follow our training scheme described in Section 3.4. For the ULC, we use Adam [18] optimizer for all modules and set $\beta_1 = 0.99$, $\beta_2 = 0.999$. The initial learning rate is set to $3e^{-4}$, and it decays by ten every 300 epochs. We train the converter for 1,000 epochs, and the batch size is 16. For the GAG, we use Adam optimizer and set $\beta_1 = 0.5$, $\beta_2 = 0.999$. The initial learning rate is set to $2e^{-4}$, and it decays by ten every 120 epochs. We train the converter for 400 epochs, and the batch size is 4. PatchGAN proposed in [12] is used as the discriminator, and the training setting is the same as the generator. We further test the inference speed of the FReeNet that the ULC can efficiently run with CPU at a speed of *878 FPS* and the inference time of the proposed GAG model is around 13.5 ms with a 2080 Ti GPU. Detailed structure and parameters of FReeNet can be found in the supplementary material. For the *baseline* in the paper, we choose a modified *Pix2Pix* [12] that the landmark is seen as the fourth channel concatenated to the input RGB image.

Figure 4. Generative results compared with the baseline on the RaFD dataset. The first column and row are reference and source images, respectively. The right four source images are in the wild. Images on the second and third rows are generated by the baseline, while the other images are generated by our method. Please zoom in for more details.

4.2. Qualitative Results

We conduct and discuss a series of qualitative experiments on the RaFD and Multi-PIE datasets to demonstrate the high quality of generated images and the flexibility of the proposed framework. As shown in Figure 4, we randomly choose eight identities with different expressions from the training dataset and four identities with random expressions outside the dataset as source images. Then their facial expressions and movements are transferred to three target persons in three poses. The results show that our proposed FReeNet can preserve geometry information of the reference images and reenact high-quality target face images. For example, the generated images of the baseline at the column (k) are unable to keep the facial contour as the reference images, while our method can achieve it. Moreover, our model performs better at details such as the upper lip (the column (l)), nose (the column (m)), and teeth (the column (h)). The generated faces of our method are photo-realistic and expression-alike, where the facial appearances and contours are consistent with the reference images.

Some experiments on the Multi-PIE dataset are further conducted to demonstrate the effectiveness of our proposed method. As shown in Figure 5 and Figure 6, the experimental results show that our approach can well transfer expression from the source persons to the target persons while simultaneously maintain the same pose and illumination information of the reference images.

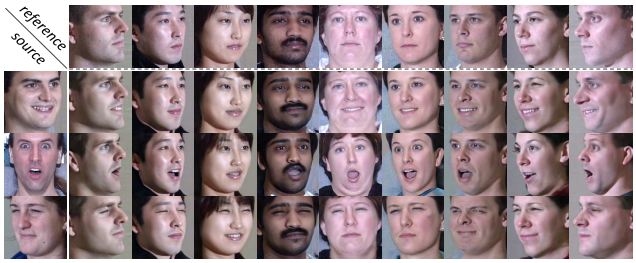

Figure 5. Experimental results in the same illumination on the Multi-PIE dataset. The first column contains three randomly selected source identities with different poses and expressions, and the first row presents nine reference identities in different poses. The rest faces are reenacted by our approach.

Figure 6. Experimental results in the varying illumination on the Multi-PIE dataset. The first row shows three reference persons in different illumination, and the first column contains three randomly selected source identities with different poses and expressions. The rest faces are reenacted by our approach.

Table 1. Metric evaluation results of the reproduced baseline and our method with different components on the RaFD dataset. Missing entry (-) means that the model is not evaluated by the metric.

Model	SSIM ↑	FID ↓	AMT
Pix2Pix [12]	0.629	12.84	41.3%
GAG	0.659	**11.67**	-
GAG+ULC	0.711	13.26	-
GAG+ULC+TP (full)	**0.717**	12.17	**74.9%**

Table 2. Model parameter comparison when learning all transformations among n persons. Missing entry (-) means that the model has no corresponding component.

Model	Parameters (M)		Speed
	Transfer	Generator	(FPS)
Pix2Pix [12]	-	16.7×n(n-1)	75
Xu *et al*. [40]	-	16.7×n(n-1)	73
ReenactGAN [39]	7.8×n	61.1×n	48
X2Face [37]	-	108.8	16
Ours	4.5	17.3	57

4.3. Quantitative Results

We choose SSIM and FID metrics to evaluate the effectiveness of our proposed method on the RaFD dataset quantitatively. During the experiment, we generate 100 reenacted images for each reference identity (67 identities totally) where corresponding 100 source images are randomly selected from other identities (6,700 images totally). In this way, diversified images can be generated because different identities (used as the source image) have different face attributes, *e.g*. face contour and interorbital distance. From the comparison results shown in Table 1, the proposed GAG outperforms the baseline in two metrics. However, both the two models can not keep the identity consistent for no landmark adaptation operation, which we call an identity shift problem. So we design the ULC module to alleviate the problem. As a result, the metric scores have a significant improvement in SSIM for the identity preserving capacity of the ULC, but a little descend in FID. We analyze it reasonable for that the FID metric judges both variety and reality of the image, and GAG model can generate more various images because various contour-inconsistent landmark images of other persons are used for one identity. The last row indicates the proposed TP loss brings a little increase in SSIM (0.006 ↑) and an obvious benefit on FID (0.96 ↓) at the same time. The reason can be intuitively found from 4.5 that the TP loss boosts the quality of the reenacted faces for having more facial details.

We further perform a user study on Amazon Mechanical Turk (AMT). For each of 30 testers, 67 real and 67 fake images of different identities are shown in random order with unlimited decision time. The result shows that our generated images confuse testers on 74.9% trials, *i.e.*, 74.9% generated images are recognized as real, while this value in the baseline is only 41.3%. As a reference, the percentage of true positive samples is 71.1%, which is unexpectedly slightly lower than our reenacted images.

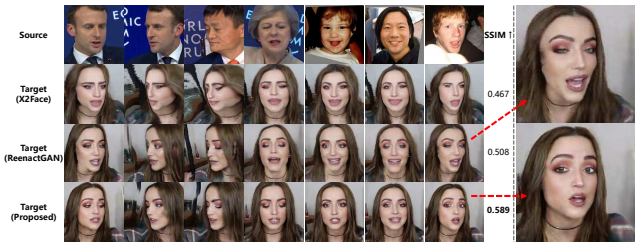

Figure 7. Result comparisons with state-of-the-art methods on the CelebV dataset. The first row is source images. The results of the second and third rows are from X2Face and ReenactGAN. The results of the last row are ours. Please zoom in for more details.

Moreover, we compare several most relevant works on model parameters (M: million) and inference speed (FPS: frames per second) to further prove the efficiency of our method, as shown in Table 2. The model parameter of our method is much lower than other methods especially when the identity number n is large, because our method only requires one unified model whatever n is that reduces the occupancy of space. Also, our approach has the approximate time consumption (57 FPS) with other methods (*e.g*. Pix2Pix, Xu *et al*., and ReenactGAN) while nearly a third of the time against X2Face (16 FPS) when reenacting the special person, which means that our method is efficient for practical application.

For one identity generation, our approach has the approximate time consumption with other methods(*e.g*. Pix2Pix and ReenactGAN), while takes less than half time against X2Face. For multiple identities generation, our approach uses only one unified model while other methods have to reload the model of the corresponding identity, which consumes extra time and space.

4.4. Comparison with State-of-the-art

As shown in Figure 7, we also conduct a contrast experiment with most related methods on the CelebV dataset [39] that contains five celebrities. We first choose 15 images of each person in this dataset to build paired images, which are used to train the ULC module. The generative results indicate our method can reenact more photorealistic and detail-abundant faces, such as teeth and hair. In detail, our approach gains 26.1% and 15.9% improvements compared to X2Face and ReenactGAN, respectively. Note that the *ULC* module is slightly modified to regress $\hat{\boldsymbol{l}}_{T,n}$ that has the same pose with $\boldsymbol{l}_{S,n}$, so as to compare with those methods.

Table 3. ACE results of different loss terms on the RaFD dataset.

Losses	$\mathcal{L}_{L1}$	$+\mathcal{L}_{cyc}$	$+\mathcal{L}_{cyc}, \mathcal{L}_{D}$
ACE	7.236 ± 0.015	4.526 ± 0.015	0.895 ± 0.010

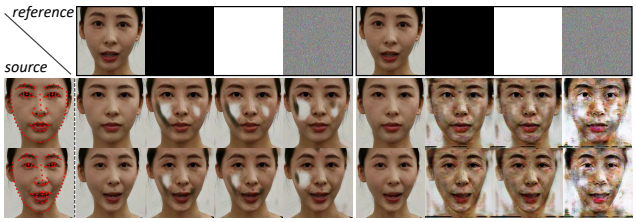

Figure 8. A toy experiment for testing the effect of *TP loss*. The second and third rows are in different source images. The second to fifth columns are results without TP loss while the rest columns use TP loss. Please zoom in for more details.

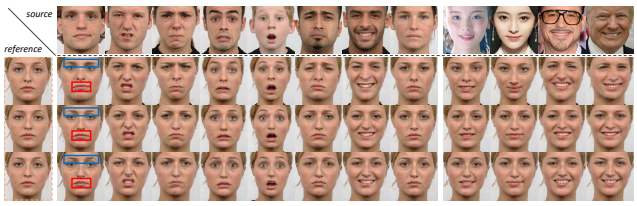

Figure 9. Ablation study on the RaFD dataset. The results of the second row are only generated by GAG. The results after adding ULC (GAG+ULC) are shown in the third row, and the last row shows the results by our complete method (GAG+ULC+TP). Please zoom in the blue and red rectangles for more details.

4.5. Ablation Study

Loss Functions of the ULC. We test the average coordinate-wise error (ACE) of the converted landmarks in different loss functions when training the ULC module. As shown in Table 3, ACE value gradually decreases when different loss terms are added, which proves the effectiveness of discriminators.

Effect of the TP Loss. To further illustrate the effectiveness of the *TP* loss, we conduct experiments that only contain two persons, because the triplet loss requires at least two persons. As shown in Figure 8, experimental results show that the *TP* loss can well separate appearance and geometry information to a certain extent. For example, when feeding the reference image with a black or gaussian noise image, the reenacted face with TP loss contains more abstract features rather than nearly the full face. It means the GAG itself contains less appearance information and can capture more appearance features from the reference image.

Components of the FReeNet. We conduct an ablation study to evaluate the impact of each component on our proposed approach. As shown in Figure 9, evaluation results of models with different components are reported. The first row shows that the proposed GAG can generate images in a good-quality whether the source images are in the dataset or not, but it is unable to preserve the geometry information of the target person. Comparing the results of the second and the third rows, we can observe that adding ULC module can significantly enhance the performance. The SSIM score meanwhile increases by a large margin, as shown in Table 1. Moreover, the effectiveness of the proposed TP loss is evaluated in the last row. It shows that the generated images can maintain more opulent facial details, e.g., brow, wrinkle, and mouth, and also have a better score on FID.

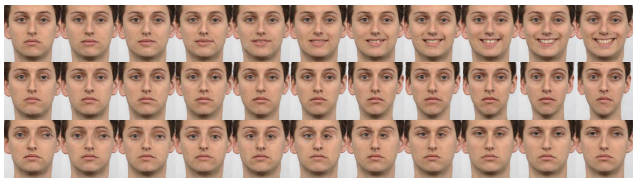

Figure 10. Landmark manipulation experiment on the RaFD dataset. From top to bottom, we manipulate mouth shape (close to open), facial contour (wide to thin), and eye position (rotation) when reenacting target faces. Please zoom in for more details.

4.6. Landmark Manipulation

We further present a landmark manipulation experiment to highlight the advantage of the decoupling design of our model in the RaFD dataset. Specifically, we control the geometry of the generated image by directly modifying the coordinates of the input landmark, which provides a flexible way to adjust the geometric position for the reenacted face. As shown in Figure 10, three groups of manipulation experiments are conducted that only change the partial attribute of the generated face, *e.g.* mouth shape, facial contour, and eye position. The results show that our approach can well keep the identity and unchanged attributes of the reference person when manipulating a specific attribute, which intuitively confirms the advantage of the decoupling idea.

5. Conclusion

In this paper, we propose a novel FReeNet to address the multi-identity face reenactment task, which aims at transferring facial expressions from source persons to target persons while keeping the identity and pose consistency to the reference images. Specifically, a ULC module is proposed to effectively convert the expression of an arbitrary source person to the target person in the latent landmark space. Then the GAG module input the reference image and the converted landmark image to reenact photorealistic target image. Moreover, a TP loss is proposed to help the GAG to decouple geometry and appearance information as well as generate detail-abundant faces. Extensive experiments demonstrate the efficiency and flexibility of our approach.

We hope our work will help users to achieve more effective and efficient works in the face reenactment task. And our approach can be easily transferred to other domains, such as gesture migration or posture migration of the body.

Acknowledgment We thank anonymous reviewers for their constructive comments. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61836015 and Key R&D Program Project of Zhejiang Province (2019C01004).

References

[1] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Siggraph*, volume 99, pages 187–194, 1999. 3

[2] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, June 2018. 2

[3] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, pages 474–484, 2018. 2

[4] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *CVPR*, pages 379–388, 2018. 1, 3

[5] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM TOG*, 35(3):28, 2016. 1

[6] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. In *SIGGRAPH Asia 2018 Technical Papers*, page 231. ACM, 2018. 1

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 2

[8] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 1

[9] Xiaojie Guo, Siyuan Li, Jiawan Zhang, Jiayi Ma, Lin Ma, Wei Liu, and Haibin Ling. Pfld: A practical facial landmark detector. *arXiv preprint arXiv:1902.10859*, 2019. 1, 2, 3, 5

[10] 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 *NeurIPS*, 2017. 5

[11] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1, 3

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 4, 5, 7

[13] Xiaohan Jin, Ye Qi, and Shangxuan Wu. Cyclegan face-off. *arXiv preprint arXiv:1712.03451*, 2017. 1, 3

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 2, 4

[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. 2

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 2

[17] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM TOG*, 37(4):163, 2018. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[19] Oliver Langner, Ron Dotsch, Gijsbert Bijlstra, Daniel HJ Wigboldus, Skyler T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 5

[20] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 1, 3

[21] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, pages 3317–3326, 2017. 1, 3

[22] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *ACM SIGGRAPH*, page 11. ACM, 2019. 3

[23] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, pages 406–416, 2017. 2

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[25] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, pages 7184–7193, 2019. 2

[26] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *ECCV*, pages 818–833, 2018. 1, 2, 3

[27] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 2, 4

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[30] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *ACM*, pages 627–635. ACM, 2018. 1

[31] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM TOG*, 36(4):95, 2017. 1

[32] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*, pages 2387–2395, 2016. 3

[33] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: Real-time reenactment of human portrait videos. *ACM TOG*, 37(4):164, 2018. 1

[34] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM TOG*, volume 24, pages 426–433. ACM, 2005. 1

[35] 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 *CVPR*, 2018. 2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[37] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using images, audio, and pose codes. In *ECCV*, pages 670–686, 2018. 1, 7

[38] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *CVPR*, pages 2129–2138, 2018. 1, 3

[39] Wayne Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. In *ECCV*, pages 603–619, 2018. 1, 2, 3, 7

[40] Runze Xu, Zhiming Zhou, Weinan Zhang, and Yong Yu. Face transfer with generative adversarial network. *arXiv preprint arXiv:1710.06090*, 2017. 1, 2, 7

[41] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *CVPR*, pages 5525–5533, 2016. 1, 3

[42] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876. IEEE, 2017. 2

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 1, 2, 4