# Designing One Unified Framework for High-Fidelity Face Reenactment and Swapping

Chao Xu[1⋆], Jiangning Zhang[1,3⋆], Yue Han[1], Guanzhong Tian[2], Xianfang Zeng[1], Ying Tai[3], Yabiao Wang[3], Chengjie Wang[3], and Yong Liu[1†]

[1] APRIL Lab, Zhejiang University
[2] Ningbo Research Institute, Zhejiang University
[3] YouTu Lab, Tencent
{21832066, 186368, 22132041, gztian, zzlongjuanfeng}@zju.edu.cn
{yingtai, caseywang, jasoncjwang}@tencent.com yongliu@iipc.zju.edu.cn

**Abstract.** Face reenactment and swapping share a similar identity and attribute manipulating pattern, but most methods treat them separately, which is redundant and practical-unfriendly. In this paper, we propose an effective end-to-end unified framework to achieve both tasks. Unlike existing methods that directly utilize pre-estimated structures and do not fully exploit their potential similarity, our model sufficiently transfers identity and attribute based on learned disentangled representations to generate high-fidelity faces. Specifically, *Feature Disentanglement* first disentangles identity and attribute unsupervisedly. Then the proposed *Attribute Transfer* (AttrT) employs learned *Feature Displacement Fields* to transfer the attribute granularly, and *Identity Transfer* (IdT) explicitly models identity-related feature interaction to adaptively control the identity fusion. We joint AttrT and IdT according to their intrinsic relationship to further facilitate each task, *i.e.*, help improve identity consistency in reenactment and attribute preservation in swapping. Extensive experiments demonstrate the superiority of our method. Code is available at https://github.com/xc-csc101/UniFace.

**Keywords:** Face Swapping, Face Reenactment, Unified Framework

## 1 Introduction

Recent research has witnessed the development of face reenactment and swapping due to their extensive applications in the metaverse. Face reenactment aims to transfer the attributes (*e.g.*, pose and expression) from target to another source face while keeping the identity of the source face unchanged. Similarly, face swapping aims to transfer the identity of the source face into the target face while keeping the attributes of the target face unchanged. Although these two tasks have the same pattern, *i.e.*, recombining the corresponding identity and attribute from source and target faces, current methods treat them independently

⋆ indicates equal contributions.
† indicates corresponding author.

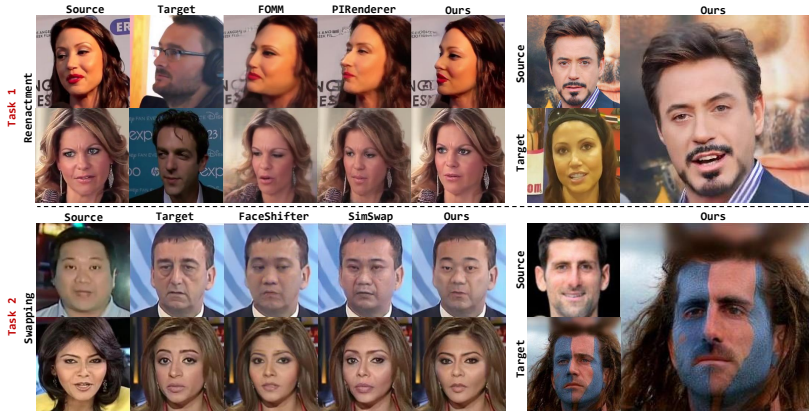


**Fig. 1. Comparison with SOTA methods. Top part** shows the results of challenging situations in face reenactment, *e.g.*, large pose, occlusion, and extreme lighting. Our method is significantly better than FOMM [34] and PIRenderer [31] with higher realism and better source identity preservation. **Bottom part** shows that our method achieves better performance both on the source identity integration and target attributes preservation (*i.e.*, especially on mouth and eyes regions) than SOTA FaceShifter [23] and SimSwap [5] in face swapping. Images are from official attached results or released codes for fair comparisons. We further present the results of *in-the-wild* situations in the right part. Our method could generalize to out-of-domain pairs and generate high-fidelity faces. Please zoom in for more details.

and design specific networks for each task, which lack universal applicability. In this paper, we focus on exploiting a unified end-to-end framework for both tasks.

Some previous works have the same spirits as ours. Ngo *et al*. [26] learn isolated disentangled representations from input face and its corresponding 2D landmarks to guide the generation of the transformed face. Other works borrow help from the 3D information. Peng *et al*. [29] and Cao *et al*. [4] incorporate 3DMM [2, 9] to decompose a face into pose, expression, and identity coefficients, and then recombined factors of the specific task are used to synthesize transformed face. However, **there are two continuously critical issues:** *1) How to get rid of pre-trained structure information.* Obtaining the landmarks and 3DMM coefficients requires excessive annotation effort and complicated preprocessing. Besides, as shown in the top part of Fig. 1, when under some challenging situations, *e.g.*, occlusion, extreme lighting, and large pose, the structure representations are ambiguous and would cause degradation problems. *2) How to sufficiently transfer the identity and attribute representations and facilitate corresponding tasks in a joint framework.* The above methods just recombine different facial parameters for face swapping or reenactment accordingly, ignoring the intrinsic relationship between two tasks and resulting in poor performance.

In this paper, we are dedicated to solving the above problems. First, instead of relying on fixed identity and attribute models to extract corresponding embeddings, we design a *Feature disentanglement* module that consists of two embed-

ders to decouple identity and attributes by imposing a set of reconstruction losses in face reenactment. Our attribute embedder learns only attributes-related descriptors, and identity embedder encodes face to low-resolution semantic feature maps with sufficient identity information. Second, to produce more high-fidelity transformed faces based on disentangled representations, we devise *Attribute Transfer* (AttrT) and *Identity Transfer* (IdT) to process attributes and identity, respectively. For face reenactment, recent works [26, 44, 3, 28] implicitly extract attribute information into the latent vector space, which significantly leads to spatial information loss. Consequently, other works [34, 31, 11] explicitly predict flow fields to transform the source face spatially, but they are still dependent on structure information. To align the attributes of the source face with the target more efficiently, our AttrT learns *Feature Displacement Fields* (FDF) from attribute representation and applies them to source identity features. A powerful learned decoder with rich facial prior is followed to eliminate unnatural texture introduced by the warping operation and generate contents that do not exist in the source faces end-to-end. For face swapping, to align the identity of the target face with the source face, mainstream methods [23, 5, 39] employ AdaIN [16] to transfer the identity information of the source face into the target. However, lack of semantic interaction makes these methods prone to aggregate inappropriate identity cues to the target, leading to low identity consistency. Inspired by self-attention [45], we design an efficient IdT module that performs identity interaction more granularly to control the identity-related feature fusion.

More importantly, we joint AttrT and IdT in a unified framework according to the intrinsic relationship of two tasks, termed *Feature Transfer*, which could further tackle the problems of identity inconsistency in isolated face reenactment and the dilemma of attributes preservation in isolated face swapping. Specifically, the IdT in face reenactment serves as an identity enhancement module, which allows the ambiguity warped feature to learn detailed texture information from the source face, resulting in higher identity consistency between the transformed and source faces. For face swapping, we apply AttrT on source features to align their attributes with the target face before sending to IdT. The aligned source face makes our method preserve the attributes of the target face excellently while not harming the identity modification performance, as depicted in Fig. 1.

In summary, we make the following three contributions:

- We propose a novel end-to-end unified framework to achieve both face reenactment and swapping. Our method does not rely on prior knowledge during inference stage to disentangle identity and attribute representations.
- We thoroughly exploit the intrinsic relationship between face reenactment and swapping, and design the novel joint-learned AttrT and IdT that transfer the attributes and maintain identity consistency for reenactment, simultaneously facilitate attributes preservation and integrate identity for swapping.
- Abundant experiments qualitatively and quantitatively demonstrate the superiority of our method to generate high-fidelity faces, *i.e.*, more attributes-alike identity-preserving reenacted faces and identity-consistent attributes-preserving swapped faces than both SOTA unified and isolated methods.

## 2 Related Work

**Unsupervised Disentanglement.** Unsupervised separation of face attributes is very common in face manipulation. Zheng *et al*. [49] disentangle facial semantics from StyleGAN without external supervision. Liu *et al*. [24] adopt structure-texture independent network to achieve attribute disentanglement. Some methods [40, 44] of face reenactment extract structure cues from drive face by imposing reconstruction loss. Similar to these works, we design a feature disentanglement module to decouple identity and attribute features unsupervisedly.

**Face Reenactment.** Face reenactment aims to animate the source face into pose and expression of the target face, which could be roughly divided into two categories: instruction-based and warping-based. Instruction-based methods animate the source face instructed by the target structure. Works [41, 15, 46, 13, 6] adopt landmarks and segmentation maps to indicate the facial attribute. Recently, 3DMMs have been proven to be very effective for modeling faces. Some works [19, 21] fit 3DMMs by pose and expression from the target face, and identity from the source face to achieve reenactment. Moreover, some AdaIN-based [16] methods [44, 43] encode target attributes in the latent vector space, which is later injected into the source face.

However, the above methods could not explicitly indicate the transformations between the source and target faces. Subsequently, warping-based methods have grown to be popular. X2Face [40] learns flow fields from the target face, which are used to warp the embedded source face at the pixel level, but it suffers unnatural head deformations. In the follow-up work, MonkeyNet [33] uses sparse key points to predict flow fields for source appearance driving. FOMM [34] utilizes relative key-point locations to further improve identity preservation. Recently, several image-level warping-based works [31, 48, 11] separate motion estimation and warped source face refinement into two stages. However, when under extreme conditions, structural priors are not reliable, leading to identity degradation. Unlike the above, our method is independent of facial structure, and the identity transfer further boosts the model's robustness to extreme conditions.

**Face Swapping.** Face swapping aims to change the facial identity but keep other face attributes constant. Early efforts [37, 36] rely on 3D geometrical operations to transfer identity. However, these approaches fail to produce high-fidelity images since their performance depends on the accuracy of the non-trainable external 3D models. Recently, with the development of GANs [12], learning-based methods have made significant progress. IPGAN [1] recombines learned disentangled identity and attribute embeddings to generate swapped faces. But vectorized representations inevitably bring information loss. Consequently, FaceShifter [23] integrates the identity and attributes by a carefully designed fusion module. SimSwap [5] proposes a weak feature matching loss to improve target attributes preservation. HifiFace [39] uses an extra 3D shape model to pay more attention to shape change. MegaFS [50] follows the GAN-Inversion [18] paradigm to

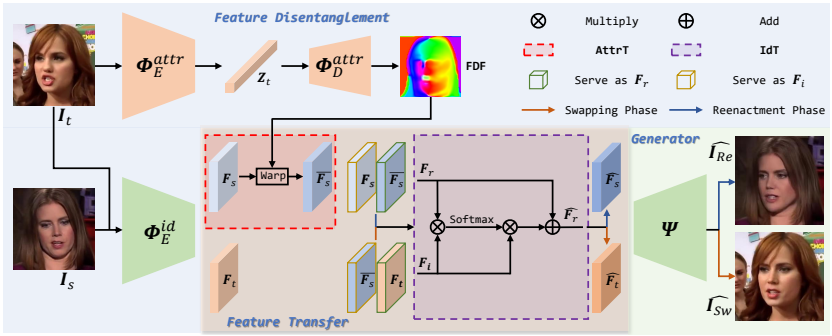


**Fig. 2. Overview of the proposed method.** Given a source face $\boldsymbol{I}_s$ and target face $\boldsymbol{I}_t$, *Feature Disentanglement* first extracts the disentangled representations, obtaining accurate *Feature Displacement Fields* (FDF) generated from attributes vector $\boldsymbol{Z}_t$, and identity features $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$, which are then sent to *Feature Transfer*. Specifically, *Attribute Transfer* (AttrT) uses accurate FDF to puppeteer $\boldsymbol{F}_s$ into the aligned $\bar{\boldsymbol{F}}_s$ with the desired attributes. *Identity Transfer* (IdT) receives the $\boldsymbol{F}_t$ (serve as $\boldsymbol{F}_r$) and $\bar{\boldsymbol{F}}_s$ (serve as $\boldsymbol{F}_i$) to transfer the identity information from source to the target if in swapping task, otherwise fed with the $\bar{\boldsymbol{F}}_s$ (serve as $\boldsymbol{F}_r$) and $\boldsymbol{F}_s$ (serve as $\boldsymbol{F}_i$) to preserve the identity details after global transformation in reenactment task. Subsequently, the reenacted features $\hat{\boldsymbol{F}}_s$ and swapped features $\hat{\boldsymbol{F}}_t$ go through a powerful *Generator* $\boldsymbol{\Psi}$ to synthesize $\hat{\boldsymbol{I}}_{Re}$ and $\hat{\boldsymbol{I}}_{Sw}$, respectively.

generate high-resolution swapped faces. However, the above methods use direct concatenation, or AdaIN [16], which fails to model identity-related feature interaction and is prone to introduce identity-unrelated cues, resulting in poor identity consistency and attribute preservation. To alleviate these problems, we design identity transfer based on self-attention to finely integrate identity information and further plug with attribute transfer to help maintain attributes.

## 3 Method

In this paper, we propose a novel efficient paradigm to complete face reenactment and swapping in an end-to-end unified framework. As shown in Fig. 2, given a target face $\boldsymbol{I}_t$ that provides attribute cues and a source face $\boldsymbol{I}_s$ that provides identity cues, the model learns to animate the source $\boldsymbol{I}_s$ guided by the FDF extracted from $\boldsymbol{I}_t$ to generate reenacted face $\hat{\boldsymbol{I}}_{Re}$, while integrate the identity information from $\boldsymbol{I}_s$ into the target $\boldsymbol{I}_t$ to generate swapped face $\hat{\boldsymbol{I}}_{Sw}$.

### 3.1 Architecture

**Feature Disentanglement.** The previous methods [26, 4, 29] that use identity and structure priors to extract corresponding disentangled representations, which are fixed as the information guidance during training. However, such representations would be inaccurate under challenging conditions, leading to per-

formance degradation. Here we introduce *Feature Disentanglement* module consisting of two embedders to decouple identity and attribute features. Specifically, Identity Embedder employs an encoder $\boldsymbol{\Phi}_E^{id}$ that embeds $\boldsymbol{I}_t$ and $\boldsymbol{I}_s \in \mathbb{R}^{3\times H\times W}$ to low-resolution semantic feature maps, obtaining $\boldsymbol{F}_t$ and $\boldsymbol{F}_s \in \mathbb{R}^{C\times H/8\times W/8}$, which contain more identity-related spatial details than that in vector space.

$$\boldsymbol{F}_t, \boldsymbol{F}_s = \boldsymbol{\Phi}_E^{id}(\boldsymbol{I}_t), \boldsymbol{\Phi}_E^{id}(\boldsymbol{I}_s). \tag{1}$$

Attribute Embedder is designed in an Encoder-Decoder architecture, the encoder $\boldsymbol{\Phi}_E^{attr}$ embeds $\boldsymbol{I}_t$ to disentangled attribute vector $\boldsymbol{Z}_t \in \mathbb{R}^{512}$, which is sent to the followed $\boldsymbol{\Phi}_D^{attr}$ to estimate *Feature Displacement Fields* $\boldsymbol{W}_{fdf} \in \mathbb{R}^{2\times H/4\times H/4}$.

$$\boldsymbol{W}_{fdf} = \boldsymbol{\Phi}_D^{attr}(\boldsymbol{\Phi}_E^{attr}(\boldsymbol{I}_t)). \tag{2}$$

To guide these embedders to learn the desired disentangled descriptors, we impose several reconstruction losses during the reenactment training phase to achieve disentanglement in a fully unsupervised manner.

**Attribute Transfer.** Current image-level warping-based methods usually cause severe artifacts, which require further refinement in the two-stage paradigm. To align the source face with the desired attributes and preserve visual textures more efficiently, we employ *Attribute Transfer* module that adopts the warping operation on source identity features. Specifically, we first downsample FDF from the Attribute Embedder to match the resolution of the source features. Such design is resolution-free for different identity features, and a higher resolution of FDF contains fine-grained attribute details. As shown in Fig. 2, FDF clearly displays the approximate location and movement of each facial region, *e.g.*, eyes, mouth, left face, and right face, indicating the coordinate offsets specifying which position in the source feature maps could be sampled to generate the targets. Then the warped source features $\bar{\boldsymbol{F}}_s$ could be calculated by the equation:

$$\bar{\boldsymbol{F}}_s = \mathrm{AttrT}(\boldsymbol{F}_s, \mathrm{RS}(\boldsymbol{W}_{fdf})), \tag{3}$$

where RS is resize operation, $\bar{\boldsymbol{F}}_s$ is the coarse feature maps with desired pose and expression like target face while keeping the same identity of the source face.

**Identity Transfer.** To effectively model the identity-related feature interaction and finely aggregate identity information between identity and reference faces, we propose a novel *Identity Transfer* module, which is well designed based on the self-attention mechanism. As shown in Fig. 2, given identity features $\boldsymbol{F}_i$ and reference features $\boldsymbol{F}_r$ with the same dimensions, the query is extracted by one convolution from $\boldsymbol{F}_r$, and the key and value are extracted from $\boldsymbol{F}_i$ in the same way, obtaining $\boldsymbol{Q}_r, \boldsymbol{K}_i, \boldsymbol{V}_i \in \mathbb{R}^{C/4\times H/8\times W/8}$ with reduced channel numbers. Then $\boldsymbol{Q}_r$ and $\boldsymbol{K}_i$ are employed to calculate the correlation matrix $\boldsymbol{M}$, which further multiplies $\boldsymbol{V}_i$ to obtain $\boldsymbol{F}_{i\to r}$. A zero-initialized learned scale parameter $\alpha$ is applied on $\boldsymbol{F}_{i\to r}$ to control the identity transfer flow when added to the $\boldsymbol{F}_r$:

$$\boldsymbol{F}_{i\to r} = \mathrm{softmax}(\boldsymbol{Q}_r(\boldsymbol{K}_i)^T)\boldsymbol{V}_i = \boldsymbol{M}\boldsymbol{V}_i, \tag{4}$$

$$\hat{\boldsymbol{F}}_r = \alpha \boldsymbol{F}_{i \to r} + \boldsymbol{F}_r. \tag{5}$$

Benefiting from representing identity with feature maps, IdT is allowed to learn the explicit correlation between identity and reference faces on identity-related regions and transfer the identity information to the reference face adaptively.

**Generator.** To generate authentic faces, we adopt the stylemap resizer along with the synthesis network of the StyleMapGAN [20] as our *Generator* $\boldsymbol{\Psi}$. Specifically, the stylemap resizer is a common decoder with the corresponding feature size to the encoder $\boldsymbol{\Phi}_E^{id}$. We further add a skip connection that brings the texture details from the $\boldsymbol{\Phi}_E^{id}$ to keep the facial contents and background. The synthesis network could learn more sufficient facial prior than the typical decoder, enabling us to generate more realistic faces. Such a powerful generator with skip connection successfully embeds the transformed features to the high-fidelity faces:

$$\hat{\boldsymbol{I}_{Re}}, \hat{\boldsymbol{I}_{Sw}} = \boldsymbol{\Psi}(\hat{\boldsymbol{F}}_s, \boldsymbol{F}_s^i), \boldsymbol{\Psi}(\hat{\boldsymbol{F}}_t, \boldsymbol{F}_t^i), \tag{6}$$

where $i$ indicates the feature maps of $i$-th layer in $\boldsymbol{\Phi}_E^{id}$.

### 3.2 Face Reenactment and Swapping

The isolated AttrT is designed for face reenactment and IdT is for face swapping, we joint two modules in a unified framework, termed *Feature Transfer*, which enables each task to capture complementary information from the other. Specifically, as shown in Fig. 2, for face reenactment, we send $\boldsymbol{I}_t$ to attribute embedder and $\boldsymbol{I}_s$ to identity embedder, AttrT first spatially transforms source feature $\boldsymbol{F}_s$ by FDF. Subsequently, the warped source feature $\bar{\boldsymbol{F}}_s$ is in the same pose and expression as the target face. A followed IdT borrows the source identity cues from $\boldsymbol{F}_s$ (serve as $\boldsymbol{F}_i$ in Sec. 3.1) to $\bar{\boldsymbol{F}}_s$ (serve as $\boldsymbol{F}_r$) for better identity preservation. For face swapping, the Identity Embedder is given $\boldsymbol{I}_t$ and $\boldsymbol{I}_s$. AttrT first aligns the $\boldsymbol{F}_s$ with the attributes of $\boldsymbol{I}_t$, obtaining $\bar{\boldsymbol{F}}_s$ (serve as $\boldsymbol{F}_i$), which along with $\boldsymbol{F}_t$ (serve as $\boldsymbol{F}_r$) are sent into IdT for identity transfer. The aligned source features help preserve the attributes when performing identity-related feature interaction. By fully exploiting the intrinsic relationship of two tasks, the joint AttrT and IdT effectively achieve both tasks, while solving the challenges of identity inconsistency in reenactment and attribute preservation in swapping.

### 3.3 Objective Functions

We use several loss terms to train our unified framework: reconstruction loss $\mathcal{L}_{rec}$, perceptual loss $\mathcal{L}_p$, and adversarial loss $\mathcal{L}_{adv}$ for face reenactment. Based on these three losses, identity loss $\mathcal{L}_{id}$ and contextual loss $\mathcal{L}_{ctx}$ are adopted for face swapping. Thus, the total loss for two tasks are defined as follow:

$$\begin{aligned} \mathcal{L}_{all}^{Re} &= \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_p\mathcal{L}_p, \\ \mathcal{L}_{all}^{Sw} &= \lambda_{adv}^{'}\mathcal{L}_{adv} + \lambda_{rec}^{'}\mathcal{L}_{rec} + \lambda_p^{'}\mathcal{L}_p + \lambda_{id}^{'}\mathcal{L}_{id} + \lambda_{ctx}^{'}\mathcal{L}_{ctx}. \end{aligned} \tag{7}$$

**Reconstruction Loss.** When the source and target faces are from the same identity, we define a reconstruction loss to calculate pixel-level errors:

$$\begin{aligned} &\mathcal{L}_{rec} = \left\| \hat{\boldsymbol{I}} - \boldsymbol{I}_t \right\|_2 \textit{ if } \boldsymbol{I}_s, \boldsymbol{I}_t \textit{ in same identity}, \\ &\text{where } \hat{\boldsymbol{I}} \in \{\hat{\boldsymbol{I}_{Re}}, \hat{\boldsymbol{I}_{Sw}}\}. \end{aligned} \tag{8}$$

**Perceptual Loss.** Besides measuring the difference between two faces at the pixel level, we adopt LPIPS [47] loss to calculate the semantic errors:

$$\mathcal{L}_p = \left\| \phi_{vgg}(\hat{\boldsymbol{I}}) - \phi_{vgg}(\boldsymbol{I}_t) \right\|_2, \tag{9}$$

where $\phi_{vgg}(\cdot)$ represents the pre-trained VGG16 [35] network.

**Identity Loss.** We calculate the cosine similarity to estimate the identity consistent between swapped and source faces:

$$\mathcal{L}_{id} = 1 - \cos(R(\boldsymbol{I}_s), R(\hat{\boldsymbol{I}_{Sw}})), \tag{10}$$

where $R(\cdot)$ is a pre-trained ArcFace [8] network.

**Contextual Loss.** We utilize contextual loss [25] to mitigate the effect of excessive information on the swapped faces:

$$\mathcal{L}_{ctx} = -\log(\mathrm{CX}(\phi_{vgg}(\hat{\boldsymbol{I}}_t), \phi_{vgg}(\boldsymbol{I}_{Sw}))). \tag{11}$$

**Adversarial Loss.** We adopt adversarial training to ensure the authenticity of the transformed faces:

$$\mathcal{L}_{adv} = \mathbb{E}_{x \sim p_x} \left[ \log D\left(x\right) \right] + \mathbb{E}_{\tilde{x} \sim p_{\tilde{x}}} \left[ \log\left(1 - D\left(\tilde{x}\right)\right) \right], \tag{12}$$

where $p_x$ ans $p_{\tilde{x}}$ are real and generated image distributions.

## 4 Experiments

### 4.1 Datasets and Implementations Details

**Datasets.** For face reenactment, we leverage the VoxCeleb2 [7] dataset, which contains over 1 million utterances for over 6,000 celebrities. Following the pre-processing method in FOMM [34], we crop faces from the original videos and resize them to $256 \times 256$ for training and testing. For face swapping, we adopt the CelebA-HQ [17] dataset, which consists of 28,000 faces for training and 2,000 for testing. We use the FaceForensics++ [32] dataset for further evaluation.

**Evaluation Metric.** For face reenactment, we use FID [14] to evaluate the realism of the generated images. The Average Pose Errors (APE) and Average Expression Errors (AEE) are used to estimate the motion accuracy, and Cosine SIMilarity (CSIM) is used to measure identity preservation. D3DFR [10] is used to extract the pose and expression coefficients. CosFace [38] is used to extract identity embedding. We additionally use LPIPS [47] to evaluate reconstruction quality. For face swapping, we evaluate the performance by IDentity Retrieval (IDRet), APE, and AEE. IDRet retrieves the closest face to evaluate identity modification while APE and AEE evaluate attributes preservation.


**Fig. 3.** Qualitative comparison with four SOTA methods on VoxCeleb2 [7] test set. The top shows the results of the reconstruction task and the bottom of the reenactment.

**Implementation Details.** For face reenactment, we perform reconstruction during training, *i.e.*, randomly sample the source and target faces from the same video. The values of the loss weights are set to $\lambda_{adv} = 1$, $\lambda_{rec} = 5$, $\lambda_p = 5$. For face swapping, we train on the CelebA-HQ dataset with resized $256 \times 256$ faces. The values of the loss weights are set to $\lambda^{'}_{adv} = 1$, $\lambda^{'}_{rec} = 1$, $\lambda^{'}_{p} = 1$, $\lambda^{'}_{id} = 2.5$, $\lambda^{'}_{ctx} = 0.5$. The ratio of the training data with $\boldsymbol{I}_t = \boldsymbol{I}_s$ and $\boldsymbol{I}_t \neq \boldsymbol{I}_s$ is set to $1:4$. We train two tasks in stages. First, face reenactment is trained for $500K$ iterations from scratch. Then we only load the trained attribute embedder weights for initialization, and then train the face swapping stage with $500K$ iterations. Both tasks adopt Adam optimizer with $\beta_1 = 0$, $\beta_2 = 0.99$, $lr = 1e-4$ for updating whole model weights, using 2 Tesla V100 GPUs and 8 batch size.

### 4.2 Comparison with SOTAs

**Qualitative Results.** For face reenactment, we conduct two tasks: *Reconstruction* where the source and target faces are of the same identity, and *Reenactment* where the source face is driven to mimic the motions of another cross-identity individual. We perform qualitative comparisons with X2Face [40], Bi-layer [42], FOMM [34], and PIRenderer [31]. As shown in Fig. 3, we sample two pairs for reconstruction and four pairs for reenactment from VoxCeleb2 test set. It can be seen that X2Face suffers from severe warping artifacts. Bi-layer over-smooths the facial details and could not generate authentic background. Recent FOMM and PIRenderer could generate high-quality results, but they fail to handle extreme

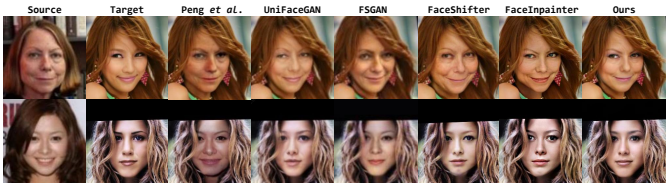


**Fig. 4.** Qualitative comparison with three SOTA unified methods and two SOTA isolated methods. Images are from official attached results in the corresponding paper.

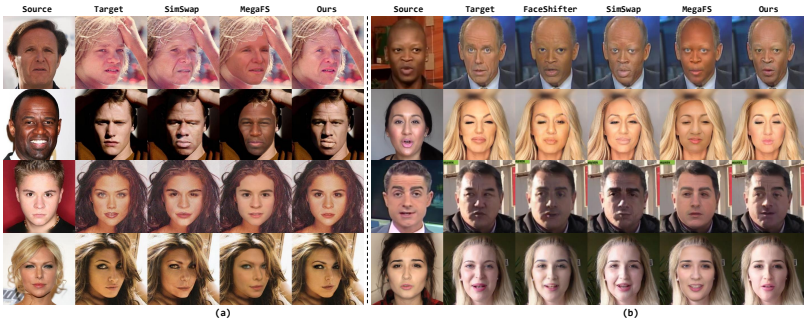


**Fig. 5.** Qualitative comparison with SOTA methods. (a) The results on CelebA-HQ [17] test set. We not compare with FaceShifter since its official codes are not available. (b) The results on FaceForensics++ [32]. FaceShifter attaches official results in this dataset.

cases. As shown in row 3, they are sensitive to the lighting in the target face due to their dependence on structural information. In rows 4 and 6, the source identity could not be well-preserved when the source face shape is very different from the target. Besides, under the large-pose conditions in rows 5 and 6, they fail to maintain source identity and are less realistic. In contrast, our method successfully generates more realistic results with accurate pose and expression while still preserving the source identity in various conditions.

For face swapping, we first compare our method with three SOTA unified works (Peng *et al.* [29], UniFaceGAN [4], and FSGAN [27]) and two SOTA isolated methods (FaceShifter [23] and FaceInpainter [22]). Obviously, recent unified works fail to preserve the attributes of the target face (*e.g.*, skin color and gaze) and maintain the low identity consistency. Our method exhibits a strong identity modification ability as SOTA FaceInpainter while achieving more high-fidelity swapped faces with authentic skin textures. We further conduct a series of qualitative experiments to compare with FaceShifter, SimSwap [5], and MegaFS [50] on CelebA-HQ and FaceForensics++ dataset. As shown in Fig. 5, we sample eight pairs of significant gaps between pose, expression, skin color, and lighting. Notably, FaceShifter could produce highly identifiable faces but unable to keep the details of the attributes, *e.g.*, closed eyes in row 2 of Fig. 5 (b), while SimSwap improves the attributes preservation but in poor identity consistency, *e.g.*, pupil color in row 1 and small mouth in row 2 of Fig. 5 (b).

**Table 1.** Quantitative results on the tasks of reconstruction for VoxCeleb2 [7] test set. **Bold** and underline represent optimal and suboptimal results. The up arrow indicates that the larger the value, the better the model performance, and vice versa.

| Method | FID↓ | LPIPS↓ | AEE↓ | APE↓ | CSIM↑ | Auth.↑ |
|---|---|---|---|---|---|---|
| X2Face [40] | 93.99 | 0.3612 | 2.76 | 0.194 | 0.6122 | 0.12 |
| Bi-layer [42] | 149.14 | 0.5443 | 1.92 | 0.055 | 0.7002 | - |
| FOMM [34] | 48.96 | **0.1817** | **1.55** | 0.044 | 0.8462 | 0.30 |
| PIRenderer [31] | 55.57 | 0.2634 | 2.09 | 0.059 | 0.8186 | 0.25 |
| Ours | **45.91** | 0.1907 | 1.70 | **0.042** | **0.8607** | **0.33** |

**Table 2.** Quantitative results on the tasks of reenactment for VoxCeleb2 [7] test set.

| Method | FID↓ | AEE↓ | APE↓ | CSIM↑ | Auth.↑ |
|---|---|---|---|---|---|
| X2Face [40] | 119.32 | 3.99 | 0.274 | 0.4329 | 0.09 |
| Bi-layer [42] | 195.67 | 3.18 | 0.073 | 0.5353 | - |
| FOMM [34] | 72.20 | 3.25 | **0.067** | 0.5365 | 0.26 |
| PIRenderer [31] | 73.73 | **3.08** | 0.079 | 0.4737 | 0.21 |
| Ours | **54.34** | 3.15 | 0.070 | **0.5703** | **0.44** |

**Table 3.** Quantitative results on the tasks of face swapping for FaceForensics++ [32].

| Method | IDRet↑ | AEE↓ | APE↓ | Id-Attr.↑ | Auth.↑ |
|---|---|---|---|---|---|
| DeepFakes [30] | 79.84 | 3.75 | 0.260 | 0.04 | 0.02 |
| FaceShifter [23] | 92.59 | 3.47 | 0.206 | 0.11 | 0.13 |
| SimSwap [5] | 90.02 | **3.13** | **0.039** | 0.13 | 0.08 |
| MegaFS [50] | 93.87 | 3.45 | 0.202 | 0.13 | 0.10 |
| Ours | **99.45** | 3.21 | 0.052 | **0.59** | **0.67** |

Comparing with SOTA unified and isolated methods shows the superiority of our method to generate both identity-consistent and attributes-preserving faces.

**Quantitative Results.** We first evaluate the effectiveness of the face reenactment with different SOTA methods on VoxCeleb2. We randomly sample 1,000 videos from the test set and set 6 random seeds to generate 6K pairs in total. The reconstruction results are summarized in Tab. 1 and the reenactment ones are in Tab. 2. It can be seen that FOMM and PIRenderer achieve impressive results on AEE and APE due to the accurate structure prior in most conditions. However, as facial structures involve identity information, they would inevitably cause poor identity preservation when guiding the reenactment, which can be inferred from the low CSIM. Besides, these methods struggle to handle challenging conditions and generate unrealistic facial textures, resulting in low FID. The above observations are consistent with the qualitative results in Fig. 3.

For face swapping, we conduct quantitative comparisons on FaceForensics++ and follow the settings in MegaFS, which carefully check the aligned faces and manually categorize all videos into 885 identities. As depicted in Tab. 3, SimSwap preserves attributes of the target face better but a relatively poor performance


**Fig. 6.** Qualitative ablation study for contextual loss of face swapping. We zoom in the red dotted rectangles of the forehead area for more clear comparison.

on IDRet. Our method achieves comparable results in attribute preservation and maintains the highest IDRet, outperforming MegaFS with a large margin. Results show that our method is better considering both identity consistency with the source and attribute preservation with the target. Our method is not compared with works [29, 4] quantitatively due to unavailability of their codes.

**Human Study.** We conduct the human study to evaluate the performance of each method in two tasks. Concretely, we randomly sample 200 pairs from the corresponding test set. Each pair is compared 5 times by different volunteers. For face reenactment, the volunteers are asked to choose the most realistic image among the generated results of all methods. Similarly, for face swapping, the volunteers are invited to select the one that most resembles the source and shares similar attributes with the target, as well as the one that looks most realistic. The results are shown in Tab. 1, Tab. 2, and Tab. 3 of the Id-Attr. (abbreviated from Identity-consistent and Attribute-preserving) and Auth. (abbreviated from Authentic), our generated faces are preferred by volunteers, meaning that we can generate higher fidelity images than SOTA methods on both tasks.

### 4.3 Ablation Study and Further Analysis

**Loss Functions.** To evaluate the effectiveness of contextual loss, we report the ablation study in Fig. 6. We sample two pairs of significant identity differences in the forehead area. FaceShifter tends to show excessive source identity and cause apparent artifacts. Comparing the results of columns 3 and 4, our method without contextual loss can already alleviate texture distortions due to the effectively semantic interactions. The last column exhibiting a better performance illustrates that contextual loss further helps preserve the target textures. We further evaluate the effectiveness of identity and adversarial losses on FaceForensicss++. IDRet suffers a sharp decline without identity loss, from 99.45 to 0.08, while FID shows higher value without adversarial loss, from 3.29 to 4.01.

**Network Components.** We perform qualitative and quantitative experiments to evaluate the effectiveness of AttrT and IdT. As shown in Fig. 7 (a), IdT is critical for identity transfer since our model without it fails to generate identity-consistent faces. From columns 1 and 5, AttrT aligns the source face with the


**Fig. 7.** Qualitative ablation study of our method with different components. (a) The results of face swapping, please attention to the mouth of row 1 and eyes of row 2. (b) The results of face reenactment, we tag CSIM on reenacted faces for clear comparison.

**Table 4.** Quantitative ablation study of our approach with different components. We mark the optimal and suboptimal results without considering meaningless values.

| Method | Face Swapping | | | Face Reenactment | | | |
|---|---|---|---|---|---|---|---|
| | IDRet↑ | AEE↓ | APE↓ | FID↓ | AEE↓ | APE↓ | CSIM↑ |
| w/o IdT | 0.05 | 0.02 | 0.000 | <u>56.28</u> | <u>3.16</u> | <u>0.071</u> | <u>0.5570</u> |
| w/o AttrT | <u>99.43</u> | <u>3.30</u> | <u>0.058</u> | 44.45 | 4.34 | 0.424 | 0.9884 |
| Ours | **99.45** | **3.21** | **0.052** | **54.34** | **3.15** | **0.070** | **0.5703** |

target face in terms of attributes, *e.g.*, closed mouth and closed eyes. Comparing the results of columns 4 and 6, we can observe that the aligned source features help achieve more attributes-preserving results. Similarly, we can draw the conclusions from Fig. 7 (b) that AttrT successfully transfers the accurate motions from the target, and IdT further boosts the identity consistency with the source face on the reenactment task, as depicted in columns 4 and 5. Moreover, the above observations could also be summarized from Tab. 4, the sharp decreasing of IDRet, and rapidly rising of AEE and APE illustrate that IdT and AttrT are indispensable to the corresponding tasks. In the meantime, AttrT significantly improves the attributes preservation for face swapping with a smaller AEE and APE, and IdT keeps better identity preservation for face reenactment with the improvement of CSIM.

**Interpretability of AttrT.** Since AttrT receives the FDF from the attribute embedder, we first perform a qualitative experiment to demonstrate the disentanglement of learned attribute representations. This experiment is based on the intuition that a well-disentangled attribute descriptor contains almost no identity information and can be used to purely measure the similarity of pose and expression between different identity individuals. Specifically, we randomly sample three images from the VoxCeleb2 test set and select their most consistent faces with all test images, according to the cosine similarity between two attribute embeddings. The results are shown in Fig. 8 (a), where the retrieved images have similar poses and expressions to those of the query images. Besides, we visualize the feature maps and FDF in Fig. 9. As expected, FDF explicitly models the absolute pose and expression of the target face, which can adaptively animate the source to desired attributes.


**Fig. 8.** (a) Image retrieval using the disentangled attribute embeddings. The Top-4 retrieved images have similar poses and expressions but different identity from the query. (b) Attention visualization of IdT. The color bars indicate activation values. The points in the target face could correctly match similar semantic areas in the source.

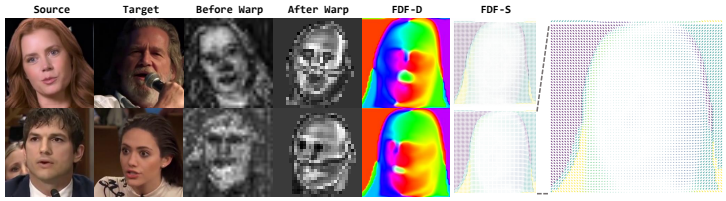


**Fig. 9.** The visualization of feature maps before and after warping, and FDF represented in dense flow (FDF-D) and sparse displacement (FDF-S). These two pairs are sampled from Fig. 3. We enlarge one FDF-S to show better motion details.

**Interpretability of IdT.** To better understand the effect of the IdT, we visualize the attention maps in Fig. 8 (b). Specifically, we select four points from different regions in the target face, *i.e.*, forehead, eyes, face, and background. The visualized attention maps indicate that each location pays more attention to semantically similar areas, allowing the explicitly identity-related semantic interaction to achieve the generation of highly identity-consistent.

## 5 Conclusions

In this paper, we propose a novel end-to-end unified paradigm to complete face reenactment and swapping. Our method shows several appealing properties: 1) We extract identity and attributes during inference stage without any prior knowledge, which is more robust under some challenging conditions. 2) To allow sufficient feature interaction, we design a novel AttrT to transfer attributes for reenactment and IdT to integrate identity for swapping. 3) We effectively joint AttrT and IdT by exploiting their underlying similarity to achieve better identity consistency in reenactment and better attribute preservation in swapping.

**Acknowledgments** This work is supported by the Key R&D Program Project of Zhejiang Province (2021C01035).

## References


1. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6713–6722 (2018)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques. pp. 187–194 (1999)
3. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13786–13795 (2020)
4. Cao, M., Huang, H., Wang, H., Wang, X., Shen, L., Wang, S., Bao, L., Li, Z., Luo, J.: Unifacegan: A unified framework for temporally consistent facial video editing. IEEE Transactions on Image Processing **30**, 6107–6116 (2021)
5. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2003–2011 (2020)
6. Chen, Z., Wang, C., Yuan, B., Tao, D.: Puppeteergan: Arbitrary portrait animation with semantic-aware appearance transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13518–13527 (2020)
7. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
8. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
9. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5154–5163 (2020)
10. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
11. Doukas, M.C., Zafeiriou, S., Sharmanska, V.: Headgan: One-shot neural head synthesis and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14398–14407 (2021)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Ha, S., Kersner, M., Kim, B., Seo, S., Kim, D.: Marionette: Few-shot face reenactment preserving identity of unseen targets. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10893–10900 (2020)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
15. Huang, P.H., Yang, F.E., Wang, Y.C.F.: Learning identity-invariant motion representations for cross-id face reenactment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7084–7092 (2020)
16. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)

17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
19. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) **37**(4), 1–14 (2018)
20. Kim, H., Choi, Y., Kim, J., Yoo, S., Uh, Y.: Exploiting spatial dimensions of latent in gan for real-time image editing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 852–861 (2021)
21. Koujan, M.R., Doukas, M.C., Roussos, A., Zafeiriou, S.: Head2head: Video-based neural head synthesis. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 16–23. IEEE (2020)
22. Li, J., Li, Z., Cao, J., Song, X., He, R.: Faceinpainter: High fidelity face adaptation to heterogeneous domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5089–5098 (2021)
23. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
24. Liu, K., Cao, G., Zhou, F., Liu, B., Duan, J., Qiu, G.: Towards disentangling latent space for unsupervised semantic face editing. IEEE Transactions on Image Processing **31**, 1475–1489 (2022)
25. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 768–783 (2018)
26. Ngo, L.M., Karaoglu, S., Gevers, T., et al.: Unified application of style transfer for face swapping and reenactment. In: Proceedings of the Asian Conference on Computer Vision (2020)
27. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7184–7193 (2019)
28. Nitzan, Y., Bermano, A., Li, Y., Cohen-Or, D.: Face identity disentanglement via latent space mapping. arXiv preprint arXiv:2005.07728 (2020)
29. Peng, B., Fan, H., Wang, W., Dong, J., Lyu, S.: A unified framework for high fidelity face swap and expression reenactment. IEEE Transactions on Circuits and Systems for Video Technology (2021)
30. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J., et al.: Deepfacelab: A simple, flexible and extensible face swapping framework. arXiv preprint arXiv:2005.05535 (2020)
31. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
32. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
33. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: Animating arbitrary objects via deep motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2377–2386 (2019)
34. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019)

35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
36. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
37. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)
38. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)
39. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hififace: 3d shape and semantic prior guided high fidelity face swapping. arXiv preprint arXiv:2106.09965 (2021)
40. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–686 (2018)
41. Wu, W., Zhang, Y., Li, C., Qian, C., Loy, C.C.: Reenactgan: Learning to reenact faces via boundary transfer. In: Proceedings of the European conference on computer vision (ECCV). pp. 603–619 (2018)
42. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: European Conference on Computer Vision. pp. 524–540. Springer (2020)
43. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9459–9468 (2019)
44. Zeng, X., Pan, Y., Wang, M., Zhang, J., Liu, Y.: Realistic face reenactment via self-supervised disentangling of identity and pose. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12757–12764 (2020)
45. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
46. Zhang, J., Zeng, X., Wang, M., Pan, Y., Liu, L., Liu, Y., Ding, Y., Fan, C.: Freenet: Multi-identity face reenactment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5326–5335 (2020)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
48. Zhang, Z., Li, L., Ding, Y., Fan, C.: Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3661–3670 (2021)
49. Zheng, Y., Huang, Y.K., Tao, R., Shen, Z., Savvides, M.: Unsupervised disentanglement of linear-encoded facial semantics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3917–3926 (2021)
50. Zhu, Y., Li, Q., Wang, J., Xu, C.Z., Sun, Z.: One shot face swapping on megapixels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4834–4844 (2021)