International Journal of Computer Vision (2025) 133:4538–4554
https://doi.org/10.1007/s11263-025-02395-6

# UniFace++: Revisiting a Unified Framework for Face Reenactment and Swapping via 3D Priors

Chao Xu[1] · Yijie Qian[1] · Shaoting Zhu[1] · Baigui Sun[2] · Jian Zhao[3,4] · Yong Liu[1] · Xuelong Li[3,4]

Received: 4 July 2024 / Accepted: 8 February 2025 / Published online: 11 March 2025


**Abstract**

Face reenactment and swapping share a similar pattern of identity and attribute manipulation. Our previous work UniFace has preliminarily explored establishing a unification between the two at the feature level, but it heavily relies on the accuracy of feature disentanglement, and GANs are also unstable during training. In this work, we delve into the intrinsic connections between the two from a more general training paradigm perspective, introducing a novel diffusion-based unified method UniFace++. Specifically, this work combines the advantages of each, *i.e.*, stability of reconstruction training from reenactment, simplicity and effectiveness of the target-oriented processing from swapping, and redefining both as target-oriented reconstruction tasks. In this way, face reenactment avoids complex source feature deformation and face swapping mitigates the unstable seesaw-style optimization. The core of our approach is the rendered face obtained from reassembled 3D facial priors serving as the target pivot, which contains precise geometry and coarse identity textures. We further incorporate it with the proposed Texture-Geometry-aware Diffusion Model (TGDM) to perform texture transfer under the reconstruction supervision for high-fidelity face synthesis. Extensive quantitative and qualitative experiments demonstrate the superiority of our method for both tasks.

**Keywords** Face reenactment · Face swapping · Unified model · Diffusion models · 3D priors

## 1 Introduction

The latest research has observed notable progress in face reenactment and swapping technologies due to their wide-ranging applications within the metaverse. The goal of face reenactment is to migrate facial attributes, including pose and expressions, from a target face to a source face, without altering the source face's identity. On the other hand, face swapping is aimed at transferring the identity of a source face onto a target face, while maintaining the target face's original attributes intact. Although these two tasks share the same pattern, current methods (Zhao et al., 2023; Jiang et al., 2023; Gao et al., 2023; Zhang et al., 2023; Xu et al., 2023) seldom adopt a unified framework to address these two tasks. In this paper, we focus on delving into the similarities between them and exploring the potential enhancements for each in a unified view.

Communicated by Svetlana Lazebnik.

Chao Xu and Yijie Qian contributed equally to this work.

✉ Chao Xu
21832066@zju.edu.cn

Yijie Qian
22332148@zju.edu.cn

Shaoting Zhu
zhust@zju.edu.cn

Baigui Sun
sunbaigui85@gmail.com

Jian Zhao
zhaoj90@chinatelecom.cn

Yong Liu
yongliu@iipc.zju.edu.cn

Xuelong Li
xuelong_li@chinatelecom.cn

[1] State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China

[2] Walf 1069B Lab, Sany Group, Guangzhou, China

[3] The Institute of AI (TeleAI), China Telecom, Beijing, China

[4] School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China

Our previous conference work, Xu et al. (2022a), has made preliminary explorations into this issue. In that work, we focus on how to reuse identity and attribute feature extraction modules, as well as their transfer networks, to enhance the robustness of each task. In this work, we shift our focus away from feature-level considerations for this issue, instead aiming to unify the two tasks from a holistic framework perspective based on our two observations. First, in both tasks, the target supplies the geometry while the source contributes the appearance. Face swapping often does a good job of preserving the structure information of the target face, thanks to the explicit provision of these information in the target, necessitating only minor adjustments of the facial identity. However, as shown in Fig. 1a, reenactment usually involves significantly reassembling the source features to align with the target's structural information, which poses a greater challenge. Thus it invites us to ponder, *whether we can modify the source-oriented approach in face reenactment to a target-oriented one to assist in achieving better structural alignment*. Second, face reenactment typically maintains a relatively stable high level of identity consistency under the same-identity setting, which is attributed to its learning under reconstruction supervision. In contrast, face swapping lacks ground truth and needs to balance the influence of coarse-grained identity loss and attribute loss, as illustrated in Fig. 1c. Here we pose another question, *whether we can achieve stable training under solely reconstruction loss supervision in face swapping to ensure high identity consistency*. Additionally, UniFace employs Goodfellow et al. (2020), whose unstable adversarial min-max objective training process can lead to quality degradation. Now, diffusion models (Ho et al., 2020a; Nichol & Dhariwal, 2021) have become the mainstream framework for generation.

In this work, we propose a novel framework, named UniFace++, a diffusion-based, target-oriented reconstruction framework that unifies both tasks. Specifically, we frame face reenactment as a target-oriented texture transfer, replacing the conventional source-oriented feature rearrangement, as shown in Fig. 1b, and adopt a multi-conditional diffusion model to avoid unstable training of GANs, termed Texture-Geometry-aware Diffusion Model (TGDM). In particular, benefiting from the explainable and disentangled parameter space of 3DMMs (Deng et al., 2019b), we combine the texture-related coefficients from the source face with the geometry-related ones from the driving conditions to construct 3D descriptors, which are projected to the image domain and serve as the target pivot. Besides, vectorized 3DMMs coefficients are also injected to the main dataflow by AdaIN (Huang & Belongie, 2017). To further supplement source texture to rendered face, we employ cross-attention that accurately models the correspondences between the source and target appearance. To this end, TGDM is dedicated to transferring the source texture to the target rendered face, which preserves explicit structural information but avoids complex texture deformations. Furthermore, we connect TGDM to face swapping task. Unlike current diffusion-based methods (Kim et al., 2022; Zhao et al., 2023) that still require balancing identity transfer and attribute preservation, we derive an entirely new reconstruction-based training framework with a *single* face input, as shown in Fig. 1d, with no extra tricks for sampling either. Echoing the reenactment process, we regard the rendered meshes of the recombined source and target codes as the pivotal, with the target image where the face region is masked serving as the source appearance, and incorporate supplementary global identity cues to facilitate the reconstruction of the face.

In summary, we make the following four contributions:

- We examine the similarities between face reenactment and face swapping from a holistic paradigm perspective and, based on this, reconstruct the frameworks for both tasks to share a target-oriented reconstruction-based framework, termed UniFace++, which facilitates a stable training process and high-performance outcomes.
- We propose a novel TGDM pipeline based on the multi-conditional diffusion model to afford complex texture transfer and maintain overall facial geometry.
- Abundant experiments are conducted qualitatively and quantitatively to demonstrate the superiority of UniFace++ for both tasks over SOTA methods.

## 2 Related Works

### 2.1 Face Reenactment

Face Reenactment (Wu et al., 2018; Chen et al., 2020b; Ren et al., 2023) involves taking the source face and replicating its pose and expression as the target. Previous efforts (Huang et al., 2020a; Zhang et al., 2020; Ha et al., 2020) directly combine target landmarks and source face for training. Then, with the success of AdaIN (Huang & Belongie, 2017), subsequent works (Zeng et al., 2020; Zakharov et al., 2019) encode the target attributes in vectorized information and then inject them into the source face. Among them, Bounareli et al. (2023) leverages a superior pre-trained StyleGAN2 generator (Karras et al., 2019, 2020) and introduces hypernetworks to animate the source with target expressions. However, the above methods fail to explicitly indicate the movements between the source and target faces. Subsequently, warping-based methods learn to warp and synthesize the target faces based on estimated motion fields. These methods (Wiles et al., 2018; Siarohin et al., 2019a) usually separate motion estimation and warped source face refinement into two stages. The most representative work is FOMM (Siarohin et al.,

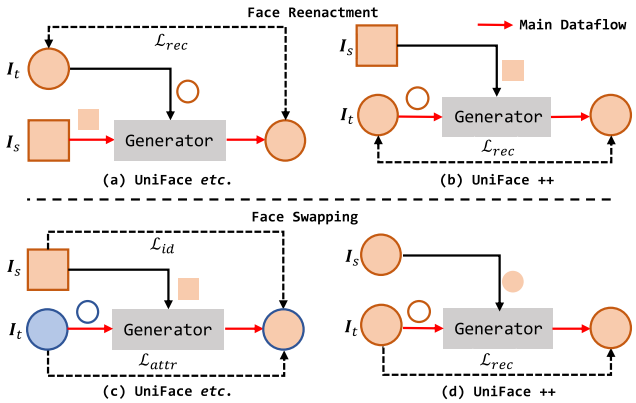


**Fig. 1** An illustrative comparison of training phase among UniFace, other mainstream methods, and proposed UniFace++. We can initially summarize that for both tasks, the source provides identity information (represented by *solid* shapes), while the target provides structural information (represented by *hollow* shapes). For face reenactment (top part), **a** UniFace *etc*are source-oriented rearranging operation, *i.e.*, reshaping the solid square to a circular form, while **b** UniFace++ serve as the target-oriented transfer task, *i.e.*, filling the hollow circle with the texture of the square. For face swapping (bottom part), **c** UniFace *etc*usually suffer from unstable training introduced by seesaw-style optimizing of identity and attribute losses across different identities (*orange* and *blue*). By contrast, **d** UniFace++ with simple reconstruction loss for stable training while still maintaining comparable results. Besides, it can be observed from the **b** and **d** that the proposed UniFace++ effectively unifies the two tasks within a single framework (Color figure online)

2019b), which uses relative key-point locations to predict flow fields to drive the appearance of the source. Other follow-up works (Tao et al., 2022; Zhao & Zhang, 2022; Zhang et al., 2023) focus on improving motion flows and warping operation accuracy. Some studies further introduce 3D cues as structural guidance to refine flow field, such as face mesh (Ren et al., 2021; Zhang et al., 2021; Doukas et al., 2021; Gao et al., 2023) and depth information (Hong et al., 2022, 2023).

More recent work, Rochow et al. (2024), samples each target pixel with a transformer-based decoder conditioned on keypoints and an expression vector that are extracted from the driving frame. Wei et al. (2024) introduces a cross-attention mechanism to guide the source animation.

The descriptions above all indicate that current methods are source-oriented, requiring the resampling of the source to align with the target's attribute, which involves significant deformation and increases the complexity of training, as shown in Fig. 1a. In contrast, face swapping only requires minor adjustments to the target face's identity while preserving its inherent structural information. Consequently, we *reframe face reenactment into a target-oriented paradigm*, using the desired textual 3D face mesh as a pivot for further texture and identity enhancement, as shown in Fig. 1b.

## 2.2 Face Swapping

Face Swapping aims to change the target identity according to the given source but keep other facial attributes constant. Early face swap works (Blanz et al., 2004; Bitouk et al., 2008; Cheng et al., 2009; Lin et al., 2012) mainly focus on 3D-based methods but suffer from poor visual quality. Then, GAN-based (Goodfellow et al., 2020) methods (Perov et al., 2020; Natsume et al., 2018; Bao et al., 2018) have made significant progress. Specifically, Li et al. (2020) adaptively integrates identity and attribute for face synthesis. Chen et al. (2020a) introduces a feature matching loss hoping to preserve more attribute embeddings. Wang et al. (2021b) and Li et al. (2021) introduce 3D face descriptor for better geometry structure of swapped results. Gao et al. (2021a) decouples identity and attribute information to better balance the two aspects. With the success of Karras et al. (2019, 2020), many works have emerged as a solution for high-resolution generation. Specifically, Zhu et al. (2021) is a pioneering work based on pSp (Richardson et al., 2021) and subsequent works (Xu et al., 2022d; Rosberg et al., 2023) further design fusion strategies for better attribute preservation. E4S (Liu et al., 2023b) and RAFSwap (Xu et al., 2022b) explicitly encode facial components related to each identity to enhance ID consistency. However, they lack flexibility in application due to the fixed generator. Consequently, Luo et al. (2022) redesigns the StyleGAN2 module and opens parameters for training. Concurrent Xu et al. (2022e) introduces a mask branch and an ID inversion strategy to empower high-fidelity and robust face swapping.

As the diffusion model shows excellent performance in many fields, Kim et al. (2022) makes the first attempt that uses facial guidance during denoising sampling. Zhao et al. (2023) is closely related to our work, which resorts to 3D information for explicit semantic and geometrical control.

Despite the impressive progress achieved by the above methods, it is still remains a challenge to fully transfer the face identity from the source face while preserving identity-unrelated attributes of the target images due to seesaw-style training losses, *i.e.*, improving one aspect at the expense of another, as shown in Fig. 1c. We observe that face reenactment yields relatively stable high-fidelity results in the same-identity setting, attributing this to its strong reconstruction supervision. Inspired by this, we *reformulate the training of face swapping as a single-image reconstruction task*, achieving stable optimization without the need for paired ground truth data, as shown in Fig. 1d.

## 2.3 Unified Framework

Unified Framework has been preliminarily explored in our conference work Xu et al. (2022a), where we harmonizes the two tasks from the feature processing level. In this work,

we adopt a holistic framework perspective to more seamlessly integrate the advantages of both tasks based on the above discussions, resulting in a more stable and superior performance. Besides, the GAN employed in UniFace has an unstable adversarial min-max optimization objective that struggles to stably converge to satisfactory results. The diffusion models have overcome these issues and demonstrated powerful capabilities in the field of generation (Avrahami et al., 2022; Harvey et al., 2022; Ho et al., 2022b; Fan et al., 2022). Thus, we build UniFace++ based on this foundation.

### 2.4 Diffusion Model

Diffusion models (Ho et al., 2020a; Nichol & Dhariwal, 2021) are recently proposed generative models that can synthesize high-quality images, which are trained without discriminators, so they are more reliable and robust during training compared to GANs. Additionally, they do not suffer from common issues such as mode collapse or vanishing gradients, which are inevitable in the training process of GANs. After achieving great success in the unconditional generation, diffusion models are adapted to enable conditional generation. Dhariwal and Nichol (2021) introduce classifier-guided diffusion, which forces the produced noise to approach the desired condition. Ho *et al.*further (Ho & Salimans, 2022) develop a Classifier-Free Guidance approach that allows conditional editing without having to pretrain classifiers. Despite these advantages, diffusion models are hindered by their slow sampling speed due to the thousands of times on one sample for complete pixel space-based denoising. To address this issue, Song et al. (2020) propose DDIM reduce sample time, and Rombach et al. (2022) propose the Latent Diffusion Models (LDMs), which transfer the training and inference processes to a compressed lower-dimension latent space for more efficient computing.

Application fields of the diffusion model vary from image generation (Avrahami et al., 2022; Fan et al., 2022; Ruiz et al., 2022; Saharia et al., 2022), video generation (Ho et al., 2022c, a; Wu et al., 2022; Molad et al., 2023; Zhou et al., 2020), audio generation (Huang et al., 2023; Liu et al., 2023a), 3D representation generation (Poole et al., 2022; Xu et al., 2022c; Li et al., 2022), and many others.

## 3 Preliminaries

### 3.1 Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs follow the idea of latent variable models that consist of a forward diffusion process and a reverse diffusion process. Specifically, a diffusion process gradually adds noise to the data sampled from the target distribution $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$ as a Markov chain. Each step $q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ (for $t \in \{0, \ldots, T\}$) is defined as a Gaussian distribution with a fixed or learned variance schedule $\beta_t \in (0, 1)$:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) := \mathcal{N}\left(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right). \tag{1}$$

By the Bayes' rules and Markov property, the latent variable $\boldsymbol{x}_t$ can be expressed as:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, and $\alpha_t = 1 - \beta_t$. Then, the reverse process $q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ is parametrized by another Gaussian transition:

$$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_\theta(\boldsymbol{x}_t, t)), \tag{3}$$

where $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ are predicted by the trained deep neural networks $\boldsymbol{\epsilon}_\theta$, which is optimized under the objective $\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right]$. Thus, given $\boldsymbol{x}_t$, $\boldsymbol{x}_{t-1}$ can be sampled by using:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\right) + \sigma_t \boldsymbol{z}, \tag{4}$$

where $\boldsymbol{z} \in \mathcal{N}(0, \mathbf{I})$. Furthermore, according to Song et al. (2020), $\boldsymbol{x}_0$ can be approximate derived by $x_t$ and $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$:

$$\hat{\boldsymbol{x}_0} := \frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{x}_t, t)}{\sqrt{\alpha_t}}. \tag{5}$$

This facilitates the use of pixel-level supervision and the perceptual losses during the training stage in Sect. 4.2.

### 3.2 3D Morphable Models (3DMMs)

Recent methods estimate the 3D face descriptors of 2D images by optimizing a neural network to extract 3D parameters from a face image. Thus we follow the previous work D3DFR (Deng et al., 2019b) that adopts ResNet50 as the backbone to predict 3DMM coefficients, which consist of identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, expression $\boldsymbol{\beta} \in \mathbb{R}^{64}$, texture $\boldsymbol{\delta} \in \mathbb{R}^{80}$, illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$, and pose $\boldsymbol{p} \in \mathbb{R}^{6}$. Note that the original 3DMM fails to control the gaze direction, we explicitly model the gaze like Park et al. (2018), providing the normalized direction vector from the center of the eye to the pupil in four dimensions $\boldsymbol{\omega} \in \mathbb{R}^4$. Therefore, given an input face $\boldsymbol{I}$, the output coefficients $\boldsymbol{\rho} \in \mathbb{R}^{261}$:

$$\boldsymbol{\rho} = \mathcal{D}(\boldsymbol{I}) = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{p}, \boldsymbol{\omega}\}. \tag{6}$$

With 3DMM, the 3D shape $\mathbf{S}$ and albedo texture $\mathbf{T}$ could be parameterized as:

$$\begin{aligned} \mathbf{S} &= \bar{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta}, \\ \mathbf{T} &= \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta}, \end{aligned} \tag{7}$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ denote the mean face shape and albedo texture. $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_{t}$ are the bases of identity, expression, and texture computed via PCA. We project the reconstructed 3D face onto the 2D image plane with a differentiable renderer $\boldsymbol{R}$ according to its illumination $\boldsymbol{\gamma}$ and pose $\boldsymbol{p}$:

$$\boldsymbol{I}_{3d} = \mathcal{R}(\mathbf{S}, \mathbf{T}, \boldsymbol{\gamma}, \boldsymbol{p}). \tag{8}$$

We naturally choose the rendered image $\boldsymbol{I}_{3d}$ as the intermediate geometry condition due to its several appealing properties: 1) Compared with other structural representations, *e.g.*, landmarks and segmentation maps, 3DMMs provide an explainable and disentangled parameter space, which enables direct recombining of corresponding factors when conducting the specific face manipulation task. Besides, mapping other cues onto 3DMMs is much easier since no additional spatial information is required. 2) Rendered face images provide more detailed semantic and explicit geometry than vectorized parameters, thus reducing the training difficulty of our framework for both face reenactment and swapping.

## 4 Method

### 4.1 Overview

Our previous work delves deeply into the intrinsic connections between face reenactment and face swapping, proposing UniFace to unify the two. However, it primarily focuses on the feature level and relies heavily on the feature decoupling capability of reenactment. In this work, we revisit the paradigm of both tasks from a more general viewpoint, proposing an enhanced framework known as UniFace++. whose critical component termed Texture-Geometry-aware Diffusion Model (TGDM). It is built upon the multi-conditional diffusion model, allowing complex texture transfer and overall facial geometry preservation, stable and effective training either. In Sects. 4.2 and 4.3, we respectively describe how TGDM can be simultaneously adapted for face reenactment and swapping. We will supply more details in the following.

### 4.2 TGDM for Face Reenactment

Most recent face reenactment methods are source-oriented that model the deformation to animate the source into the driving pose and expression. However, it is still quite challenging to achieve the accurate desired geometry and maintain the complex identity appearance under various conditions, yielding noticeable artifacts and degradation problems. Inspired by the target-oriented swapping framework that the desired facial structure is inherently provided by the target itself and the remaining task is to transfer the global identity onto this target structure, we reconstruct the source-oriented reenactment framework into a target-oriented one. For this purpose, we design the Texture-Geometry-aware Diffusion Model (TGDM), which focuses on transferring the source texture to the rendered geometry face. In this part, we give the descriptions of the network structure and the training details.

*Feature Extraction* As shown in Fig. 2, we combine the appearance-related 3DMM coefficients (identity, texture, and illumination) from the source image $\boldsymbol{I}_s$ with the motion-related coefficients (expression, pose, and gaze) from the driving image $\boldsymbol{I}_d$ to construct the desired 3D face descriptors $\hat{\boldsymbol{\rho}} = \{\boldsymbol{\alpha}_s, \boldsymbol{\beta}_d, \boldsymbol{\delta}_s, \boldsymbol{\gamma}_s, \boldsymbol{p}_d, \boldsymbol{\omega}_d\}$, along with its rendered face $\boldsymbol{I}_{3d}$ as the geometry conditions, which explicitly represent accurate facial structure and coarse identity semantics.

*Architecture.* Following Ho et al. (2020b), our conditional denoising model $\boldsymbol{\epsilon}_\theta$ is designed by the UNet-based backbone, consisting of the encoder $\boldsymbol{\Psi}_E$ and decoder $\boldsymbol{\Psi}_D$. As shown in Fig. 2, TGDM is conditioned on three external inputs. First, the spatially aligned rendered face $\boldsymbol{I}_{3d}$ is concatenated channel-wise with the noisy face $\boldsymbol{Z}_T$, which is obtained by adding noise to $\boldsymbol{I}_d$ according to Eq. 2. They are fed to the first layer of the network to guide the denoising process, ensuring the intermediate noise and the output face follow the given facial geometry. But $\boldsymbol{I}_{3d}$ struggles with precise identity control due to its coarse semantics and unrealistic textures. Consequently, the texture encoder $\boldsymbol{\Phi}_E$ provides the multiscale features $\boldsymbol{F}_s = \{\boldsymbol{F}_s^0, \ldots, \boldsymbol{F}_s^k\}$ to provide the desired identity texture patterns, where $k$ is 1, *i.e.*, we adopt two resolution texture features in $16 \times 16$ and $32 \times 32$. To mix the source texture within the noise prediction branch and eliminate the effects of misalignment, we design the Texture Attention-based (TexAtt) module that employs the cross-attention mechanism for harmonious integration. Concretely, as shown in Fig. 2, each TexAtt receives the source texture feature $\boldsymbol{F}_s^i$ and the noise feature $\boldsymbol{F}_d^i$, the query is extracted by one convolution from $\boldsymbol{F}_d^i$, and the key and value are extracted from $\boldsymbol{F}_s^i$ in the same way, obtaining $\boldsymbol{Q}_d, \boldsymbol{K}_s, \boldsymbol{V}_s \in \mathbb{R}^{C_i/4 \times H_i \times W_i}$, which have reduced channel numbers. Then $\boldsymbol{Q}_d$ and $\boldsymbol{K}_s$ are used to calculate the correlation matrix $\boldsymbol{M}$, which further multiplies $\boldsymbol{V}_s$ to obtain $\boldsymbol{F}_{s\to d}^i$. A zero-initialized learned scale parameter $\tau$ is applied on $\boldsymbol{F}_{s\to d}^i$ to control the source texture transfer flow when added to the $\boldsymbol{F}_d^i$:

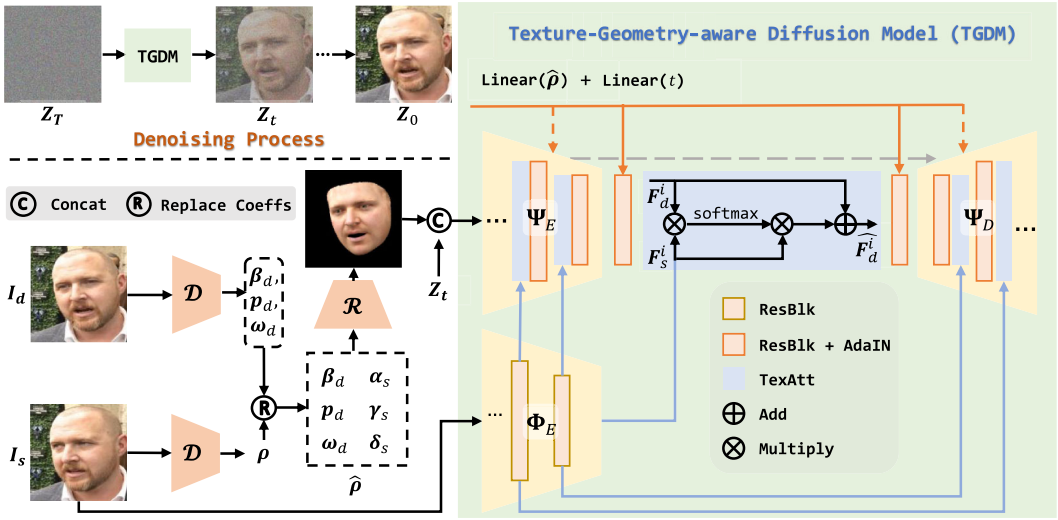


**Fig. 2** Overview of the paradigm of face reenactment. Given the source and target faces, we first represent them in geometry-related 3DMM coefficients and then project the recombined ones to the rendered face $\boldsymbol{I}_{3d}$, which serves as the intermediate structural representation. To obtain realistic faces, we develop a multi-conditional diffusion model, termed TGDM, that learns geometry prior from $\boldsymbol{I}_{3d}$ by simply concatenated input, transfers source appearance from $\boldsymbol{I}_s$ by cross attention, and supplements implicit identity and geometry information by AdaIN

$$\boldsymbol{F}^i_{s\to d} = \text{softmax}(\boldsymbol{Q}_d(\boldsymbol{K}_s)^T)\boldsymbol{V}_s = \boldsymbol{M}\boldsymbol{V}_s, \tag{9}$$

$$\hat{\boldsymbol{F}}^i_d = \tau \boldsymbol{F}^i_{s\to d} + \boldsymbol{F}^i_d. \tag{10}$$

In addition to the explicit conditions, the modified coefficients $\hat{\boldsymbol{\rho}}$ further supplement the implicit geometry cues, especially the gaze direction not included in the rendered face. It added with embedded time, forming the last condition $\boldsymbol{C} = \text{Linear}(\hat{\boldsymbol{\rho}}) + \text{Linear}(t)$, which is injected into the noise predictor via the adaptive instance normalization (AdaIN) (Huang & Belongie, 2017):

$$\text{AdaIN}(\boldsymbol{F}^i_d, \boldsymbol{C}) = \sigma_c(\boldsymbol{C})\frac{\boldsymbol{F}^i_d - \mu(\boldsymbol{F}^i_d)}{\sigma(\boldsymbol{F}^i_d)} + \mu_c(\boldsymbol{C}), \tag{11}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ is the average and variance operation of the input feature $\boldsymbol{F}^i_d$ respectively. $\mu_c(\cdot)$ and $\sigma_c(\cdot)$ are used to estimate the adapted mean and bias according to the given condition. To this end, all condition information is properly integrated into the network $\boldsymbol{\epsilon}_\theta(\boldsymbol{Z}_t, \boldsymbol{F}_s, \boldsymbol{I}_{3d}, \hat{\boldsymbol{\rho}}, t)$ to predict the noise for face reenactment.

*Objectives.* We first adopt the regular simple *Denoising Loss*:

$$\mathcal{L}_{simple} = \left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{Z}_t, \boldsymbol{F}_s, \boldsymbol{I}_{3d}, \hat{\boldsymbol{\rho}}, t)\right\|_2, \tag{12}$$

where $\boldsymbol{\epsilon}$ is an added noise on $\boldsymbol{I}_d$. Besides, we estimate the fully denoised face $\hat{\boldsymbol{Z}}_0$ according to the Eq. 5, which enables further constraints on the image level. Concretely, we measure the difference between $\hat{\boldsymbol{Z}}_0$ and $\boldsymbol{I}_d$ at the pixel and perceptual level by a *Reconstruction Loss* $\mathcal{L}_{rec}$ as $\mathcal{L}_2$ distance and a *Perceptual Loss* as the LPIPS loss (Zhang et al., 2018):

$$\mathcal{L}_{rec} = \left\|\hat{\boldsymbol{Z}}_0 - \boldsymbol{I}_d\right\|_2, \tag{13}$$

$$\mathcal{L}_p = \left\|\phi_{vgg}(\hat{\boldsymbol{Z}}_0) - \phi_{vgg}(\boldsymbol{I}_d)\right\|_2, \tag{14}$$

where $\phi_{vgg}(\cdot)$ represents the pre-trained VGG16 (Simonyan & Zisserman, 2014) network. Thus, the total loss is defined as follows:

$$\mathcal{L} = \lambda_{simple}\mathcal{L}_{simple} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_p\mathcal{L}_p, \tag{15}$$

where $\lambda_{simple} = 10$, $\lambda_{rec} = 1$, and $\lambda_p = 1$.

### 4.3 TGDM for Face Swapping

Despite the impressive progress of recent methods, GAN- and diffusion-based face swapping methods still suffer from the dilemma that the improvement of source face identity consistency at the expense of sacrificing target attribute preservation. For example, Kim et al. (2022) employs identity and attribute expert models to guide the noise prediction, and the balance between them is critical to producing high-quality swapped faces. However, it is complex and needs many experimental attempts. We attribute this phenomenon to the training phase of playing the seesaw-style game, which struggles to balance all identity-unrelated attributes preservation and the source identity fusion. Inspired by the stable and appealing performance brought about by reconstruction loss in same-identity face reenactment, we attempt to reframe the

**Table 1** Quantitative results on the tasks of same-identity and cross-identity setting on VoxCeleb1

| Method | Same-Identity | | | | | | Cross-Identity | | | | | PP↑ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | LPIPS ↓ | Exp ↓ | Angle ↓ | Gaze ↓ | ID-C ↑ | FID ↓ | Exp ↓ | Angle ↓ | Gaze ↓ | ID-C ↑ | FID ↓ | |
| FOMM (Siarohin et al., 2019b) | 16.32 | 0.3459 | 5.52 | 0.0474 | 0.0749 | 0.6552 | 27.56 | 7.14 | 0.0613 | 0.0961 | 0.5445 | 41.77 | 0.04 |
| (Ren et al., 2021) | 16.72 | 0.3549 | 5.41 | 0.0546 | 0.0773 | 0.6576 | 28.90 | 6.90 | 0.0673 | 0.0971 | 0.5503 | 37.95 | 0.06 |
| NTHS (Wang et al., 2021a) | 18.13 | 0.3588 | 5.98 | 0.0625 | 0.0903 | 0.7091 | 27.07 | 7.77 | 0.0814 | 0.1166 | 0.6159 | 38.48 | 0.02 |
| HifiHead (Zhu et al., 2022) | 15.72 | 0.3678 | 5.39 | 0.0693 | 0.0625 | **0.8722** | **21.53** | 6.80 | 0.0871 | 0.0746 | **0.8394** | **33.77** | 0.04 |
| TPSM (Zhao & Zhang, 2022) | **19.29** | 0.3366 | 5.28 | 0.0412 | 0.0660 | 0.6918 | 25.57 | 6.88 | 0.0536 | 0.0853 | 0.5917 | 39.28 | 0.10 |
| DAM (Tao et al., 2022) | 18.05 | 0.3440 | 5.46 | 0.0484 | 0.0737 | 0.6535 | 28.11 | 7.08 | 0.0626 | 0.0949 | 0.5415 | 44.09 | 0.11 |
| DaGAN++ (Hong et al., 2023) | 17.89 | 0.3565 | 5.42 | 0.0502 | 0.0751 | 0.6771 | 27.89 | 7.02 | 0.0648 | 0.0925 | 0.5553 | 42.19 | 0.08 |
| MCNet (Hong & Xu, 2023) | 18.29 | 0.3576 | 5.18 | 0.0398 | 0.0653 | 0.7127 | 26.72 | 6.67 | 0.0509 | 0.0872 | 0.5845 | 37.23 | 0.10 |
| HyperReenact (Bounareli et al., 2023) | 13.38 | 0.3609 | 5.23 | 0.0484 | 0.0793 | 0.6429 | 28.35 | 6.76 | 0.0567 | 0.0986 | 0.5498 | 38.94 | 0.05 |
| AniPortrait (Wei et al., 2024) | 15.50 | 0.3563 | 6.12 | 0.0683 | 0.0944 | 0.6853 | 27.19 | 7.24 | 0.0608 | 0.1012 | 0.6111 | 41.28 | 0.03 |
| FSRT (Rochow et al., 2024) | 16.55 | 0.3450 | 5.36 | 0.0463 | 0.0758 | 0.6787 | 27.09 | 6.94 | 0.0581 | 0.0969 | 0.5655 | 39.91 | 0.14 |
| UniFace (Xu et al., 2022a) | 16.18 | 0.3541 | 5.81 | 0.0554 | 0.0792 | 0.6603 | 27.17 | 7.35 | 0.0609 | 0.1006 | 0.5694 | 40.07 | 0.05 |
| Ours(UniFace++) | 18.55 | **0.3346** | **5.09** | **0.0315** | **0.0554** | 0.7718 | 25.51 | **5.82** | **0.0349** | **0.0596** | 0.7017 | 35.16 | **0.18** |

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. PP means user preference percentage

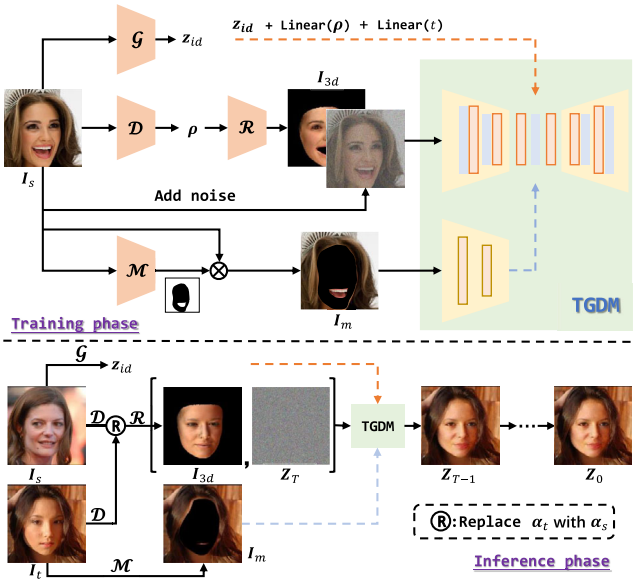


**Fig. 3** The new paradigm of face swapping is built upon the TGDM. During the training phase, we focus on reconstructing the input face $\boldsymbol{I}_s$ from given identity- and attribute-related conditions, *i.e.*, identity embedding $z_{id}$, 3D face descriptor $\boldsymbol{\rho}$, rendered face $\boldsymbol{I}_{3d}$, and masked face $\boldsymbol{I}_m$. During inference, we first render the recombined coefficients to capture the desired geometry prior and coarse identity texture, which incorporates with other conditions to generate final swapped results

face swapping as a single-image reconstruction task, which is also built upon the TGDM.

Specifically, as shown at the top of Fig. 3, there are two modifications. First, we completely mask the face region of the source texture image with the help of the mask predictor $\mathcal{M}$ (Yu et al., 2018) to ensure that the ground truth identity information is not visible to the network, obtaining $\boldsymbol{I}_m$:

$$\boldsymbol{I}_m = \boldsymbol{I}_s \times \mathcal{M}(\boldsymbol{I}_s). \tag{16}$$

Second, because of the low-dimensional linear representation of 3DMMs, the rendered images often lack photo-realism and fine texture details like wrinkles. We further supplement the identity embedding from the expert identity model $\mathcal{G}$ (Huang et al., 2020b). In this way, the renderer image $\boldsymbol{I}_{3d}$, identity embedding $z_{id}$, and Linear($\boldsymbol{\rho}$) focused on affording identity cues and identity-unrelated attributes of the face region, while $\boldsymbol{I}_m$ makes up for the absence of hair and background. Notably, the mouth areas are also served as the background, which is discussed in Sect. 5.3. During training, as Eq. 15, our scheme does not require complex losses. Instead, the reconstruction loss is sufficient. The hyperparameter setting is the same as Eq. 15 either. For inference, given the source $\boldsymbol{I}_s$ and the target $\boldsymbol{I}_t$, we first render the $\boldsymbol{I}_{3d}$ with the identity factor of the source and the remaining parameters of the target. As shown in the bottom of Fig. 3, $\boldsymbol{I}_{3d}$ is sensitive to the geometric structure, exhibiting the exact desired face shape,

and $z_{id}$ contains source identity semantics. Combining both of them guarantees identity similarity. To this end, following the standard denoising process, our method successfully transfers semantics and textures of the source identity to the target, while fully keeping the identity-unrelated attributes without any complex sampling tricks.

### 4.4 Temporally Coherent Video Generation

The aforementioned methods primarily focus on the image level. In order to output temporally consistent face swapping and driving results, we make minor modifications to the TGDM. Inspired by Stypulkowski et al. (2023), we introduce two temporally consecutive frames to provide motion information, which are concatenated with the source input ($\boldsymbol{I}_s$ in reenactment paradigm Fig. 2 and $\boldsymbol{I}_m$ in swapping paradigm Fig. 3) and injected into the textual encoder $\boldsymbol{\Phi}_E$ through attention to incorporate temporal cues. During training, given that our ground truth frame $\boldsymbol{I}_d$ is randomly selected from a sequence at the timestamp $k$, three distinct cases arise: Initially, for $k >= 2$, there is an ample supply of preceding frames to act as motion frames. In the case of $k = 1$, the frame at $k = 0$ is paired with a single-channel black image to constitute the motion frame. Finally, for $k = 0$, the motion frame is composed of two single-channel black images. Then, the sampling procedure is consistent, and it should be noted that the motion frames utilized in the inference phase are derived from the frames predicted earlier, since we do not have access to the ground truth frames. In this way, we generate temporally coherent videos frame by frame.

## 5 Experiment

### 5.1 Datasets and Implementation Details

*Datasets.* For face reenactment, we leverage the VoxCeleb1 (Nagrani et al., 2017) dataset for training. Among them, we select the high-resolution (720P) ones and follow the preprocessing method in FOMM Siarohin et al. (2019b) to crop the videos and resize them to $256 \times 256$. We artificially construct various challenging pairs as a test set to evaluate the model's overall performance. Besides VoxCeleb1, we randomly select 1,000 images from the VGGFace2-HQ (Cao et al., 2018)[1] dataset to serve as the source images, and correspondingly sample target faces from (Nagrani et al., 2017), to compare generalization capability with other methods. For face swapping, we utilize the high-quality (Lee et al., 2020) dataset for training, which has 30,000 images with fine-grained mask annotation. Rossler et al. (2019) is used for testing, which is a forensics dataset consisting of 1,000 videos. Besides, we further randomly sample 1,000 identity pairs from the test set of CelebA-HQ datasets for further evaluation.

*Metrics.* We evaluate the reenactment performance from three aspects: **Quality** is measured using PSNR, LPIPS (Zhang et al., 2018), and FID (Heusel et al., 2017), which are commonly used in most available models (Ren et al., 2021; Wang et al., 2021a; Zhu et al., 2022; Rochow et al., 2024; Xu et al., 2022a; Hong et al., 2023, 2022; Hong & Xu, 2023; Bounareli et al., 2023); **Attribute Consistency** is assessed using Exp, Angle, and Gaze, following Zhu et al. (2022), which calculates the average Euclidean distances of expression, pose, and gaze coefficients extracted by 3D face reconstruction model (Deng et al., 2019b) between the generated and target images; **Identity Consistency** is also measured like most competitors (Wang et al., 2021a; Zhu et al., 2022; Xu et al., 2022a; Hong et al., 2023, 2022; Bounareli et al., 2023), by calculating the identity cosine similarity in the feature space, termed ID-C, where C indicates that ID embeddings are extracted by Huang et al. (2020b). Similarly, we evaluate the swapping performance from three aspects: **Quality** is judged using FID (Zhao et al., 2023; Xu et al., 2022d; Zhu et al., 2021; Liu et al., 2023b; Rosberg et al., 2023). **Identity Similarity** is estimated using ID-A (Zhu et al., 2021; Kim et al., 2022; Shiohara et al., 2023), where A means (Deng et al., 2019a) extractor. **Attribution Preservation** is assessed using Exp and Angle (Kim et al., 2022; Shiohara et al., 2023).

However, we observe that the reliability of the discriminative model decreases under challenging conditions. Inspired by Kim et al. (2022), we introduce a relative distances metric to measure not only how close synthesis is to positive pairs, but also how far synthesis and negative pairs, as formalized in Eq. 17. This approach better reflects how humans perceive facial changes, especially in challenging cases.

$$R - \mathcal{D} := \frac{\mathcal{D}\left(\mathbf{Z}_{syn}, \mathbf{Z}_p\right)}{\mathcal{D}\left(\mathbf{Z}_{syn}, \mathbf{Z}_p\right) + \mathcal{D}\left(\mathbf{Z}_{syn}, \mathbf{Z}_n\right)}, \tag{17}$$

where $\mathcal{D}$ can be any distance metric, $\mathbf{Z}_{syn}$ is the generated face, $\mathbf{Z}p$ has the desired characteristic, while $\mathbf{Z}_n$ has the undesired ones.

Consequently, we supplement our evaluation with additional challenging metrics to better reflect identity consistency and attribute preservation under various conditions, termed R-ID-A, R-ID-C, R-Exp, R-Angle, and R-Gaze, which are the relative distance versions of the original metrics.

*Implementation Details.* For face reenactment, we randomly sample the source and target faces from the same video in VoxCeleb1 for training. It takes about 4 days by using 4 V100 GPUs with 8 batch sizes and a 0.0002 learning rate for 200K iterations. For face swapping, we train its model as the aforementioned setting for approximately 3 days. For

[1] https://github.com/NNNNAI/VGGFace2-HQ.


**Fig. 4** Qualitative comparison with SOTA methods on VoxCeleb1 test set. We present various challenging cases which show the significant difference between the source and target of the pose, occlusion, and face size. Please pay attention to the area indicated by the red arrow

**Table 2** Quantitative results on the tasks of cross-identity setting when the source faces are sampled from VGGFace2-HQ

| Method | R-Exp ↓ | R-Angle ↓ | R-Gaze ↓ | R-ID-C ↑ | FID ↓ |
|---|---|---|---|---|---|
| DaGAN++ (Hong et al., 2023) | 6.88 | 0.0692 | 0.1054 | 0.5134 | 68.09 |
| MCNet (Hong & Xu, 2023) | 6.64 | <u>0.0562</u> | <u>0.0955</u> | 0.5433 | 54.73 |
| HyperReenact (Bounareli et al., 2023) | 6.73 | 0.0629 | 0.1128 | 0.4982 | <u>51.60</u> |
| AniPortrait (Wei et al., 2024) | 7.45 | 0.0720 | 0.1321 | 0.5077 | 60.15 |
| FSRT (Rochow et al., 2024) | <u>6.51</u> | 0.0597 | 0.1052 | <u>0.5569</u> | 53.62 |
| Ours | **5.76** | **0.0401** | **0.0656** | **0.6294** | **45.83** |

We use relative distances metric in this experiment
Bold and underline values represent optimal and suboptimal results


**Fig. 5** Qualitative comparison with recent SOTA methods on VGGFace2-HQ dataset

the diffusion model, the length of the denoising step $T$ is set to 1000, and a linear noise schedule is adopted for both the training and inference process. Notably, to stale the training procedure, only MSE loss of noise is used at the beginning of the training, when it has been decreased below 0.05, MSE loss of image, and LIPIS loss then start to work. Besides, the UNet of TGDM receives 256 × 256 resolution images and performs 16 down-sample ratios.

## 5.2 Face Reenactment

### 5.2.1 Comparison with Baselines

*Qualitative Results.* We perform qualitative comparisons with Siarohin et al. (2019b); Ren et al. (2021), NTHS (Wang et al., 2021a; Zhu et al., 2022), TPSM (Zhao & Zhang, 2022), DAM (Tao et al., 2022; Hong et al., 2023), MCNet (Hong & Xu, 2023; Bounareli et al., 2023; Rochow et al., 2024), recent diffusion-based method Wei et al. (2024), and our previous version Xu et al. (2022a) in the *Cross-Identity* setting, where the source and the target are of different identities. As shown in Fig. 4, we sample nine pairs from VoxCeleb1 for visualization. First, the top three pairs have a significant difference in *face size*. It can be seen that FOMM-based methods, *e.g.*, TPSM, DAM, and MCNet, produce over-smooth facial textures and suffer from noticeable warping artifacts. HifiHead could generate realistic faces, but their poses are inconsistent with the target. Another stylegan-based method, HyperReenact, exhibits the opposite phenomenon, where attribute similarity is high, but image distortion is severe, particularly in non-facial areas. By contrast, the results of our method are of high quality and with the desired attributes. Second, the target faces of the middle ones show *rich micro-expressions*. Recent methods just imitate mouth shape and head direction, and they ignore the emotion embodied in the target. For example, the target of the fourth row is surprised, and the sixth is contempt. For comparison, our results exhibit accurate emotion styles, *i.e.*, surprised forehead lines, delighted mouth corners, and disdainful eyes. Finally, the bottom pairs suffer from *occlusions* in the source or the target. It is difficult for FOMM-based methods to estimate the precise key points, even with the introduction of depth cues in DaGAN++ to enhance accuracy. Thus they usually suffer from extremely distorted facial shapes (the head area of row 7). Other methods also struggle to animate the occluded objects to fit the desired pose, which is attributes to source-oriented methods struggle to reassemble unseen parts into a coherent and reasonable result.

In contrast, our method designed in a target-oriented manner thus avoids these problems, which is not sensitive to occlusion and reasonably preserves the non-facial parts in the generated results (the headphones of row 8 and the hat of row 9). Moreover, these cases are all under large-pose conditions. Aniportrait, as the latest method, although it has fully integrated the facial structure information with the source face, it still produces noticeable artifacts in these challenging scenarios. Our previous work UniFace is a GAN-based source-oriented framework, so it is not surprising that it could not produce competitive results. Note that we introduce the gaze information in 3D facial prior, and it can be seen from rows 5 and 6 that it does work. Thus these visualizations convincingly demonstrate the superiority of our target-oriented method that successfully transfers the source texture to the target rendered image, providing more realistic results with accurate pose and detailed expression while preserving the source identity.

To fully assess the generalization and adaptability of the proposed method, we provide a qualitative comparison with SOTA methods on VGGFace2-HQ dataset. We choose competitors from publications dating from 2023 onwards, *i.e.*, DaGAN++, MCNet, FSRT, HyperReenact, and AniPortrait. As shown in Fig. 5, our results are conditioned on explicitly decoupled 3DMM coefficients and rendered 3D faces, which are not very sensitive to the source and driving subjects. Therefore, despite the different cropping and alignment methods between VGGFace2-HQ and VoxCeleb1, our method can still effectively handle these source faces and map them to the desired attributes under various driven expressions and poses. In contrast, comparative methods exhibit noticeable texture degradation and inconsistencies in identity preservation.

*Quantitative Results.* We quantitatively compare the proposed method with several aforementioned SOTA methods both in *Same-Identity* and *Cross-Identity* settings. We randomly sample 200 identities from the test set and set 5 random seeds to generate 1K pairs in total. The results are summarized in Table 1. Benefiting from the explicit facial representation contained in the target rendered face, our method achieves an impressive performance of facial attributes, *i.e.*, far ahead in metrics Exp, Angle, and Gaze, indicating that our model can animate the source face that is highly faithful to the given structure cues. Beyond that, our approach was nearly the best in overall quality and similarity to the identity. HifiHead obtains the lowest FID and shows the best identity consistent, yet suffers from severe pose error, which can be concluded from rows 2 and 8 of Fig. 4 either.

Besides, we report the quantitative results on VGGFace2-HQ in Table 2. We employ more challenging metrics, *i.e.*, relative distance metrics, to assess the performance of our method on this unseen and cross-domain dataset. Consistently, our method achieves superior performance on all these metrics.

Finally, we attach the overall user preference percentage results in the rightmost column in Table 1. Concretely, we randomly sample 200 pairs from the corresponding test set. Each pair is compared 5 times by different volunteers, who are asked to choose the preferred one in terms of three metrics: realism, identity-consistency, and attribute-alignment. The results show that our method outperforms other competing methods.

### 5.2.2 Ablation Study and Analysis

We perform qualitative and quantitative ablation studies to validate the merits of the proposed designs. For a fair com-


**Fig. 6** Qualitative ablation study of our method with different variations on VoxCeleb1 dataset, including loss functions, method components, TextAtt structures

parison, we train our method and all these baselines with the *same* setting, *e.g.*, same batch sizes and training iterations.

*Ablating Loss Functions.* we first present ablation experiments to explore the impact of various loss functions. As shown in Fig. 6, using only a simple loss ($\lambda_{simple} = 1$, $\lambda_{rec} = 0, \lambda_p = 0$, column 3, Exp1) can produce structurally consistent driving results, but the texture and color differ significantly from the source. Introducing a reconstruction loss ($\lambda_{simple} = 1, \lambda_{rec} = 1, \lambda_p = 0$, column 4, Exp2) further leads to degradation in the results, possibly because the estimated fully denoised face $\hat{\mathbf{Z}}_0$ according to the Eq. 5 is inaccurate. Directly applying a strong pixel-level loss can easily cause oscillations, while a perceptual loss constrains at the feature level and can produce relatively better results ($\lambda_{simple} = 1, \lambda_{rec} = 0, \lambda_p = 1$, column 5, Exp3). Consistent conclusions can also be drawn from the Table 3. Since Exp2 exhibits severe degradation, we do not calculate the metrics for it.

Furthermore, we perform a sensitive analysis on these three loss functions to obtain an optimal combination of weights. As shown in Fig. 6, we introduce two variants, $\lambda_{simple} = 1, \lambda_{rec} = 1, \lambda_p = 1$ (column 6, Exp4), $\lambda_{simple} = 1, \lambda_{rec} = 10, \lambda_p = 10$ (column 7, Exp5), and the setting $\lambda_{simple} = 10, \lambda_{rec} = 1, \lambda_p = 1$ we used is shown in column 13. From these visualizations, it can be observed that the Exp5 has an excessively large weight on the pixel-level reconstruction loss and perceptual loss, leading to an absence of denoising capabilities, while the output of the Exp4 is also unsatisfactory, exhibiting noticeable degradation. Consistently, Exp4 shows a noticeable decline in all metrics, as shown in Table 3. The results of the Exp5 are pure noise, so we do not calculate metrics for it.

*Ablating Method Component.* As shown in Fig. 6, when the 3D face descriptor is absent (column 10, Exp8), the overall quality of the generated face does not significantly deteriorate; however, details such as eye gazes, which are challenging to convey in the rendered face, are inaccurately represented, as can be seen in the rows 1, 3, and 4. This demonstrates that the 3D face descriptor is primarily responsible for providing detailed information about the eye area. This is also evident from the comparison between the rows 8

**Table 3** Quantitative ablation study of our approach with different module on VoxCeleb1

| Method | Exp ↓ | Angle ↓ | Gaze ↓ | ID-C ↑ | FID ↓ |
|---|---|---|---|---|---|
| Exp1 | 6.34 | 0.0389 | 0.1037 | 0.5544 | 57.91 |
| Exp2 | n/a | n/a | n/a | n/a | n/a |
| Exp3 | 6.10 | 0.0372 | 0.0789 | 0.6525 | 42.36 |
| Exp4 | 7.22 | 0.0470 | 0.1448 | 0.4981 | 54.17 |
| Exp5 | n/a | n/a | n/a | n/a | n/a |
| Exp6 | 6.06 | 0.0353 | 0.1252 | 0.6323 | 46.63 |
| Exp7 | 6.69 | 0.0443 | 0.1346 | 0.5204 | 52.96 |
| Exp8 | 5.98 | 0.0350 | 0.0948 | 0.6855 | 41.83 |
| Exp9 | n/a | n/a | n/a | n/a | n/a |
| Exp10 | 9.10 | 0.4376 | 0.1845 | 0.2233 | 70.49 |
| Ours | **5.82** | **0.0349** | **0.0596** | **0.7017** | **35.16** |

Bold values represent optimal and suboptimal results

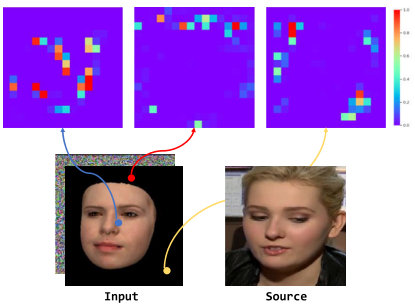


**Fig. 7** Attention visualization of TextAtt. The color bars indicate activation values. The points in the input rendered face could correctly match similar semantic and geometrical areas in the source


**Fig. 8** Qualitative comparison of face swapping results with other SOTA models on FaceForensics++. The results of our model better reflect the source identity, especially the face shape and local characteristics. Additionally, they are more faithful to the target image for non-identity-related attributes. The results of other methods are from the Kim et al. (2022) main paper and officially released codes

and 11 of Table 3, where there is a noticeable decline in the gaze metric, *i.e.*, from 0.0596 to 0.0948.

Then, when there are no source features (column 11, Exp9), only the facial textures remain. This proves that cross-attention effectively supplements the textures missing from the rendered face, such as hair and background, and further refines the facial textures. We do not include this variant in the Table 3 because it is unable to output a complete facial image.

Finally, when there is no rendered face $\boldsymbol{I}_{3d}$ (column 12, Exp10), the 3D face descriptor must solely provide facial geometric information. Comparing the results of columns 12 and 13, we can observe that the implicit representation is insufficient to effectively support geometry alignment, which underscores the importance of the rendered face. This conclusion is further supported by the sharp performance decline in the row 10 of the Table 3.

*Ablating Feature Fusion Method of TextAtt.* We design two variations to evaluate the effectiveness of TextAtt. As shown in Fig. 6, we adopt the image-level (column 8, Exp6) and feature-level (column 9, Exp7) concatenation for feature injection as two baselines, these two baselines are able to generate the desired pose and expression, but they have a limited ability to retain the source appearance, exhibiting severe color jitting and artifacts, especially the feature-level concatenation. Contrary to the above competitors, our results using cross attention show higher quality, which illustrates the effectiveness of cross attention as the feature transfer module, reducing the difficulty of training and speeding up the convergence of the model. Besides, the above observations could also be summarized from Table 3, our proposed method improves all metrics by a large margin.

*Interpretability of TextAtt.* To better understand the cross-attention mechanism, we visualize the attention maps of the TextAtt in UNet middle block, which is $16 \times 16$ resolution. As shown in Fig. 7, we select three points from different regions in the noise feature, *i.e.*, head, face, and background. The visualized attention maps indicate that each location pays more attention to the geometrically and semantically similar areas, *e.g.*, the red point is sampled from the head region, which has a higher response with the corresponding region of the source feature. Consequently, such attention-based design allows sufficient texture transfer to achieve photo-realistic and identity-consistent face generation.

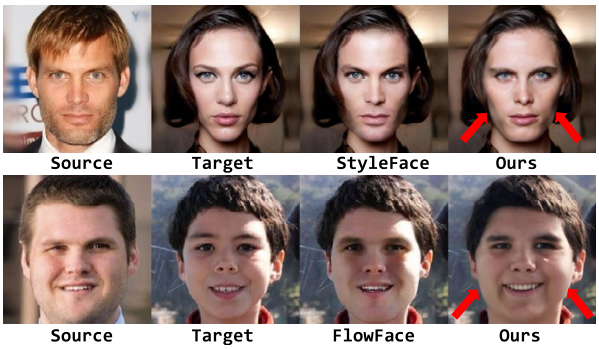


**Fig. 9** Qualitative comparison of face swapping results with other SOTA models without officially released codes. The inferred images are directly copied from their papers

## 5.3 Face Swapping

### 5.3.1 Comparison with Baselines

**Qualitative Results.** We first conduct qualitative experiments to compare our method with diffusion-based methods:

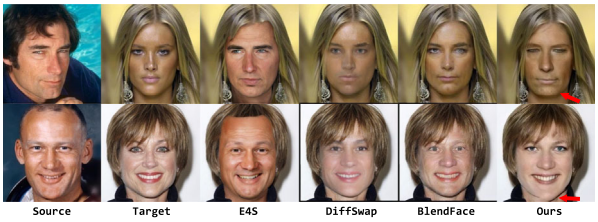


**Fig. 10** Qualitative comparison of face swapping results with recent SOTA models on CelebA-HQ dataset

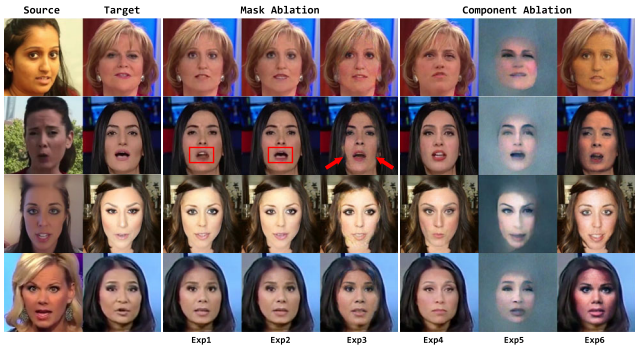


**Fig. 11** Qualitative ablation study of our method with different variations on FaceForensics++ dataset

ROOP,[2] Kim et al. (2022); Zhao et al. (2023), GAN-based methods: Li et al. (2020); Wang et al. (2021b); Chen et al. (2020a); Zhu et al. (2021), InfoSwap (Gao et al., 2021a), High-Res (Xu et al., 2022d), E4S (Liu et al., 2023b; Shiohara et al., 2023), and our previous version Xu et al. (2022a) on the Rossler et al. (2019) dataset. As shown in Fig. 8, our model outperforms other models in obtaining high-fidelity face swapping results, especially on face shapes and local characteristics (eyes, nose, mouth), and preserving non-identity-related attributes such as hair and background. Specifically, in the third row, our model can successfully transfer a wide face into a slim face benefiting from the explicit geometry guidance while other methods tend to keep the shape of the target face. DiffSwap and DiffFace, while possessing some ability to alter facial shapes, still exhibit unstable identity and attribute adversarial processes, as seen in the columns 4 and 13, which fail to produce consistent results. Similarly, HiFace employs implicit 3D-aware features, and struggles to effectively preserve facial shape. Besides, in the first row, our result has larger eyes and is more consistent with the source, while other methods lack this feature, resulting in low identity consistency.

Moreover, Fig. 9 presents more qualitative comparisons with other SOTA methods that are without officially released codes, e.g., Luo et al. (2022), and Zeng et al. (2022) when there are pronounced differences in facial contours. Figure 10 presents the comparative results on the CelebA-HQ dataset to further verify the generalization of our method. We select DiffSwap, E4S, and BlendFace for comparison, which are published in and after 2023. It can be observed that the results of E4S fail to preserve the same skin tone as the target, while BlendFace and DiffSwap exhibit low identity consistency. In contrast, our method effectively maintains target attributes while possessing a high degree of identity similarity, *i.e.*, learning the flat chin characteristic of the first case and the long face feature of the second case. Please pay attention to the area indicated by the red arrow.

*Quantitative Results.* We further report quantitative results compared to some of the above method with officially released codes in Table 4. Benefiting from the explicit structural information, our method has a large lead in the Exp and Pose metrics. At the same time, due to the sufficient identity-related texture and semantic transfer, the ID-A metric is also the best. We observe that ROOP slightly leads our method in terms of FID. This is attributes to the fact that ROOP inherits the generative priors that Stable Diffusion has learned from large-scale data, thus ensuring the overall quality of the generated results. Consequently, our UniFace++ surpasses that of other methods in almost all aspects, including the user preference percentage in the rightmost column in Table 4. We also supplement the quantitative results on the CelebA-HQ dataset in Table 5. Obviously, our method is leading in all aspects of the challenging metrics. Qualitative and quantitative experiments both prove that our method is better considering both identity consistency with the source and attribute preservation with the target.

### 5.3.2 Ablation Study and Analysis

We perform qualitative ablation studies to validate the merits of the proposed designs. For a fair comparison, we train our method and all these baselines with the *same* setting, *e.g.*, same batch sizes and training iterations.

*Ablating Mask Form.* The critical operation of our reconstruction-based face swapping paradigm is to mask the source face to avoid identity information leaking. Thus we report a visualization to explore the effect of the mask area. As depicted in Fig. 11, we design three variations, *i.e.*, the Normal mask (column 3, Exp1) covers the all face area, the Small mask (column 4, Exp2) treats the mouth area as the background, and the Dilated mask (column 5, Exp3) dilates the Normal mask to cover more areas. There is no apparent difference between the Normal and Small types in terms of identity and attributes by comparing columns 3 and 4, but the Small obtains the more realistic mouth area since it can learn information from the Small masked source. Please pay attention to the red rectangle of row 2. The results of Dilated show the artifacts around the face contour and lead to image degra-

[2] https://github.com/s0md3v/roop.

**Table 4** Quantitative comparison of face swapping on FaceForensics++ dataset

| Method | ID-A ↑ | Exp ↓ | Angle ↓ | FID ↓ | PP ↑ |
|---|---|---|---|---|---|
| FaceShifter (Li et al., 2020) | 0.5283 | 2.54 | 0.3001 | 17.82 | 0.08 |
| HifiFace (Wang et al., 2021b) | 0.5792 | 2.56 | 0.3116 | 18.91 | 0.11 |
| MegaFS (Zhu et al., 2021) | 0.3409 | 3.08 | 0.3385 | 21.68 | 0.02 |
| InfoSwap (Gao et al., 2021b) | 0.5914 | 2.93 | 0.2874 | 21.23 | 0.06 |
| High-res (Xu et al., 2022d) | 0.3182 | 2.92 | 0.2288 | 21.79 | 0.04 |
| E4S (Liu et al., 2023b) | 0.5621 | 2.66 | 0.2138 | 16.67 | 0.07 |
| BlendFace (Shiohara et al., 2023) | 0.5333 | 2.47 | 0.2051 | 19.23 | 0.07 |
| DiffSwap (Zhao et al., 2023) | 0.4022 | 3.03 | 0.2867 | 20.50 | 0.05 |
| ROOP | 0.5861 | 2.20 | 0.1877 | **15.56** | 0.17 |
| UniFace (Xu et al., 2022a) | 0.5835 | 2.67 | 0.2790 | 16.72 | 0.12 |
| Ours(UniFace++) | **0.6121** | **1.94** | **0.1122** | 15.87 | **0.21** |

Bold and underline values represent optimal and suboptimal results

**Table 5** Quantitative comparison of face swapping on CelebA-HQ test set

| Method | R-ID-A ↑ | R-Exp ↓ | R-Angle ↓ | FID ↓ |
|---|---|---|---|---|
| E4S (Liu et al., 2023b) | 0.4735 | 2.44 | 0.2427 | 15.62 |
| BlendFace (Shiohara et al., 2023) | 0.4368 | 2.32 | 0.2040 | 16.54 |
| DiffSwap (Zhao et al., 2023) | 0.3628 | 2.91 | 0.3070 | 20.36 |
| Ours(UniFace++) | **0.4948** | **1.81** | **0.1359** | **14.35** |

We use relative distances metric in this experiment
Bold and underline values represent optimal and suboptimal results

dation. On the basis of these phenomena, we choose Small masks experimentally, as depicted in $\boldsymbol{I}_m$ of Fig. 3.

*Ablating Method Component.* To verify that the components in our method are indispensable, we present qualitative ablation studies in Fig. 11. The vectorized features contain crucial source identity embeddings $\boldsymbol{z}_{id}$, and the absence of these features (column 6, Exp4) leads to a decrease in identity consistency, which can be observed by comparing the columns 4 and 6. Then, the absence of the source feature (column 7, Exp5) results in the same phenomena described in the reenactment task, *i.e.*, there is no texture outside of the facial area, and the texture within the facial area is also not realistic. Finally, we observe that without the rendered face $\boldsymbol{I}_{3d}$ (column 8, Exp6), the color of the swapped results is prone to be similar to the source rather than the target, and fails to preserve the facial expression of the target (rows 2 and 3), which further demonstrates the necessity of the explicit facial geometry as the condition.

## 6 Conclusion

In this paper, we revisit a unified framework for face reenactment and swapping, constructing a target-oriented reconstruction paradigm, termed UniFace++, which shows several appealing properties: (1) We reframe the face reenactment as a target-oriented texture transfer, instead of the source-oriented feature rearrange, to avoid complex source texture deformation. (2) We reframe the face swapping as a single image reconstruction task, which mitigates the challenge of balancing identity transfer and attribute preservation. (3) Our proposed Texture-Geometry-aware Diffusion Model (TGDM) decomposes the complex transfer problem into a multi-conditional denoising process, where a Texture Attention-based module accurately models the correspondences between appearance and geometry cues contained in source and target conditions, and incorporates extra implicit information for high-fidelity face generation. (4) Our extensive results demonstrate the superiority of the proposed framework for both face reenactment and swapping.

*Limitations and Future Works.* Constrained by computational resources, we utilize the Denoising Diffusion Probabilistic Model (DDPM) to validate the effectiveness of our UniFace++, which currently only supports a resolution of 256, and takes about 45 ms on one V100 GPU to generate a single face under the $T = 1000$ DDPM setting. In subsequent work, we plan to introduce more advanced techniques such as Stable Diffusion (SD) to strengthen our framework in terms of effectiveness and efficiency.[3,4,5,6]

[3] https://www.robots.ox.ac.uk/vgg/data/voxceleb/vox1.html.

[4] https://github.com/NNNNAI/VGGFace2-HQ.

[5] https://mmlab.ie.cuhk.edu.hk/projects/CelebA/CelebAMask_HQ.html.

[6] https://github.com/ondyari/FaceForensics.

**Acknowledgements** This work was supported by the Key R&D Project of Zhejiang Province under Grant 2024C01172 and National Natural Science Foundation of China (62476224).

**Data Availability** Our method utilizes the open-source datasets, *i.e.*, VoxCeleb1 (Nagrani et al., 2017) and VGGFace2-HQ (Cao et al., 2018) for face reenactment, CelebA-HQ (Lee et al., 2020) and FaceForensics++ (Rossler et al., 2019) for face swapping. We will release the code upon acceptance for reproduction.

## References


Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18208–18218).

Bao, J., Chen, D., Wen, F., Li, H., & Hua, G. (2018). Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6713–6722).

Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. (2008). Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH* (pp. 1–8).

Blanz, V., Scherbaum, K., Vetter, T., & Seidel, H. P. (2004). Exchanging faces in images. *Computer Graphics Forum, Wiley Online Library, 23*, 669–676.

Bounareli, S., Tzelepis, C., Argyriou, V., Patras, I., & Tzimiropoulos, G. (2023). Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7149–7159).

Cao, Q., Shen, L., Xie, W., Parkhi, O.M., & Zisserman, A. (2018) Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67–74). IEEE.

Chen, R., Chen, X., Ni, B., & Ge, Y. (2020a) Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2003–2011).

Chen, Z., Wang, C., Yuan, B., & Tao, D. (2020b). 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).

Cheng, Y.T., Tzeng, V., Liang, Y., Wang, C.C., Chen, B.Y., Chuang, Y.Y., & Ouhyoung, M. (2009) 3d-model-based face replacement in video. In *SIGGRAPH'09: Posters* (pp. 1–1).

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019a). 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).

Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., & Tong, X. (2019b) 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*.

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems, 34*, 8780–8794.

Doukas, M. C., Zafeiriou, S., & Sharmanska, V. (2021). Headgan: One-shot neural head synthesis and editing. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14398–14407).

Fan, W.C., Chen, Y.C., Chen, D., Cheng, Y., Yuan, L., & Wang, Y.C.F. (2022) Frido: Feature pyramid diffusion for complex scene image synthesis. arXiv preprint arXiv:2208.13753

Gao, G., Huang, H., Fu, C., Li, Z., & He, R. (2021a). Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3404–3413).

Gao, G., Huang, H., Fu, C., Li, Z., & He, R. (2021b). Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3404–3413).

Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., & Lu, Y. (2023). High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5609–5619).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144.

Ha, S., Kersner, M., Kim, B., Seo, S., & Kim, D. (2020). Marionette: Few-shot face reenactment preserving identity of unseen targets. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 10893–10900.

Harvey, W., Naderiparizi, S., Masrani, V., Weilbach, C., & Wood, F. (2022) Flexible diffusion modeling of long videos. arXiv preprint arXiv:2205.11495

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, *30*

Ho, J., & Salimans, T. (2022) Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems, 33*, 6840–6851.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., & Fleet, D.J., et al. (2022a) Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D.J. (2022b) Video diffusion models. arXiv preprint arXiv:2204.03458

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., & Fleet, D. J. (2022c). *Video diffusion models.*, arXiv preprint arXiv:2204.03458

Hong, F. T., & Xu, D. (2023). Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 23062–23072).

Hong, F. T., Zhang, L., Shen, L., & Xu, D. (2022). Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3397–3406).

Hong, F. T., Shen, L., & Xu D (2023) Dagan++: Depth-aware generative adversarial network for talking head video generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, P. H., Yang, F. E., & Wang, Y. C. F. (2020a). 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).

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., & Zhao, Z. (2023) Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. arXiv preprint arXiv:2301.12661

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision* (pp. 1501–1510).

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., & Huang, F. (2020b). Curricularface: Adaptive curriculum learning loss for

deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5901–5910).
Jiang, D., Song, D., Tong, R., & Tang, M. (2023). Styleipsb Identity-preserving semantic basis of stylegan for high fidelity face swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 352–361).
Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4401–4410).
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110–8119).
Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., Kim, S., & Lee, K. (2022) Diffface: Diffusion-based face swapping with facial guidance. arXiv preprint arXiv:2212.13344
Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan Towards diverse and interactive facial image manipulation. In *IEEE conference on computer vision and pattern recognition (CVPR)*.
Li, J., Li, Z., Cao, J., Song, X., & He, R. (2021). Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5089–5098).
Li, L., Bao, J., Yang, H., Chen, D., & Wen, F. (2020). Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5074–5083).
Li, M., Duan, Y., Zhou, J., Lu, J. (2022) Diffusion-sdf: Text-to-shape via voxelized diffusion. arXiv preprint arXiv:2212.03293
Lin, Y., Wang, S., Lin, Q., & Tang, F. (2012) Face swapping under large pose variations: A 3d model based approach. In *2012 IEEE international conference on multimedia and expo* (pp. 333–338). IEEE
Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., & Plumbley, M.D. (2023a) Audioldm: Text-to-audio generation with latent diffusion models. arXiv preprint arXiv:2301.12503
Liu, Z., Li, M., Zhang, Y., Wang, C., Zhang, Q., Wang, J., & Nie, Y. (2023b). Fine-grained face swapping via regional gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8578–8587).
Luo, Y., Zhu, J., He, K., Chu, W., Tai, Y., Wang, C., & Yan, J. (2022). Styleface: Towards identity-disentangled face generation on megapixels. In X. V. I. Part (Ed.), *Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings* (pp. 297–312). Springer.
Molad, E., Horwitz, E., Valevski, D., Acha, A.R., Matias, Y., Pritch, Y., Leviathan, Y., & Hoshen, Y. (2023) Dreamix: Video diffusion models are general video editors. arXiv preprint arXiv:2302.01329
Nagrani, A., Chung, J.S., & Zisserman, A. (2017) Voxceleb: a large-scale speaker identification dataset. arXiv preprint arXiv:1706.08612
Natsume, R., Yatagawa, T., & Morishima, S. (2018) Rsgan: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:1804.03447
Nichol, A. Q., & Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning* (pp. 8162–8171). PMLR
Park, S., Zhang, X., Bulling, A., & Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications* (pp. 1–10).
Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., & Jiang, J., et al. (2020) Deepfacelab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535
Poole, B., Jain, A., Barron, J.T., & Mildenhall, B. (2022) Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988
Ren, Q., Lu, Z., Wu, H., Zhang, J., & Dong, Z. (2023). Hr-net: a landmark based high realistic face reenactment network. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(11), 6347–6359.
Ren, Y., Li, G., Chen, Y., Li, T. H., & Liu, S. (2021). Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 13759–13768).
Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., & Cohen-Or, D. (2021). Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2287–2296).
Rochow, A., Schwarz, M., & Behnke, S. (2024). Fsrt: Facial scene representation transformer for face reenactment from factorized appearance head-pose and facial expression features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7716–7726).
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684–10695).
Rosberg, F., Aksoy, E.E., Alonso-Fernandez, F., & Englund, C. (2023) Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3454–3463)
Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J.,& Nießner, M. (2019) Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*.
Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2022) Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. arXiv preprint arXiv:2208.12242
Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., & Lopes, R. G., et al. (2022) Photorealistic text-to-image diffusion models with deep language understanding. arXiv preprint arXiv:2205.11487
Shiohara, K., Yang, X., & Taketomi, T. (2023). Blendface: Redesigning identity encoders for face-swapping. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 7634–7644).
Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019a). Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2377–2386).
Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019b) First order motion model for image animation. *Advances in Neural Information Processing Systems*, *32*.
Simonyan, K., & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
Song, J., Meng, C., & Ermon, S. (2020) Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502
Stypulkowski, M., Vougioukas, K., He, S., Zieba, M., Petridis, S., & Pantic, M. (2023). Diffused heads: Diffusion models beat gans on talking-face generation. arXiv preprint arXiv:2301.03396
Tao, J., Wang, B., Xu, B., Ge, T., Jiang, Y., Li, W., & Duan, L. (2022). Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3637–3646).
Wang, T. C., Mallya, A., & Liu, M. Y. (2021a). One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10039–10049).

Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., & Ji, R. (2021b) Hififace: 3d shape and semantic prior guided high fidelity face swapping. arXiv preprint arXiv:2106.09965

Wei, H., Yang, Z., & Wang, Z. (2024) Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv preprint arXiv:2403.17694

Wiles, O., Koepke, A., Zisserman, A. (2018) 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).

Wu, J.Z., Ge, Y., Wang, X., Lei, W., Gu, Y., Hsu, W., Shan, Y., Qie, X., & Shou, M.Z. (2022) Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. arXiv preprint arXiv:2212.11565

Wu, W., Zhang, Y., Li, C., Qian, C., & Loy, C. C. (2018). Reenactgan: Learning to reenact faces via boundary transfer. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 603–619).

Xu, C., Zhang, J., Han, Y., Tian, G., Zeng, X., Tai, Y., Wang, Y., Wang, C., & Liu, Y. (2022). Designing one unified framework for high-fidelity face reenactment and swapping. In X. V. Part (Ed.), *Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings* (pp. 54–71). Springer.

Xu, C., Zhang, J., Hua, M., He, Q., Yi, Z.,& Liu, Y. (2022b) Region-aware face swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7632–7641).

Xu, C., Zhu, J., Zhang, J., Han, Y., Chu, W., Tai, Y., Wang, C., Xie, Z., & Liu, Y. (2023) High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6609–6619).

Xu, J., Wang, X., Cheng, W., Cao, Y.P., Shan, Y., Qie, X., & Gao, S. (2022c) Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models. arXiv preprint arXiv:2212.14704

Xu, Y., Deng, B., Wang, J., Jing, Y., Pan, J., & He, S. (2022d). High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7642–7651).

Xu, Z., Zhou, H., Hong, Z., Liu, Z., Liu, J., Guo, Z., Han, J., Liu, J., Ding, E., & Wang, J. (2022). Styleswap: Style-based generator empowers robust face swapping. In X. I. V. Part (Ed.), *Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings* (pp. 661–677). Springer.

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 325–341).

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9459–9468).

Zeng, H., Zhang, W., Fan, C., Lv, T., Wang, S., Zhang, Z., Ma, B., Li, L., Ding, Y., & Yu, X. (2022) Flowface: Semantic flow-guided shape-aware face swapping. arXiv preprint arXiv:2212.02797

Zeng, X., Pan, Y., Wang, M., Zhang, J., & Liu, Y. (2020). Realistic face reenactment via self-supervised disentangling of identity and pose. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*, 12757–12764.

Zhang, B., Qi, C., Zhang, P., Zhang, B., Wu, H., Chen, D., Chen, Q., Wang, Y., & Wen, F. (2023). Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 22096–22105).

Zhang, J., Zeng, X., Wang, M., Pan, Y., Liu, L., Liu, Y., Ding, Y., & Fan, C. (2020). Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5326–5335).

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018) 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).

Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021). 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).

Zhao, J., & Zhang, H. (2022). Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3657–3666).

Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., & Lu, J. (2023) Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8568–8577).

Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. (2020). Makelttalk: Speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG), 39*(6), 1–15.

Zhu, F., Zhu, J., Chu, W., Tai, Y., Xie, Z., Huang, X., & Wang, C. (2022) Hifihead: One-shot high fidelity neural head synthesis with 3d control. In *IJCAI* (pp. 1750–1756).

Zhu, Y., Li, Q., Wang, J., Xu, C. Z., & Sun, Z. (2021). One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4834–4844).