DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, Jiwen Lu[†]
Tsinghua University

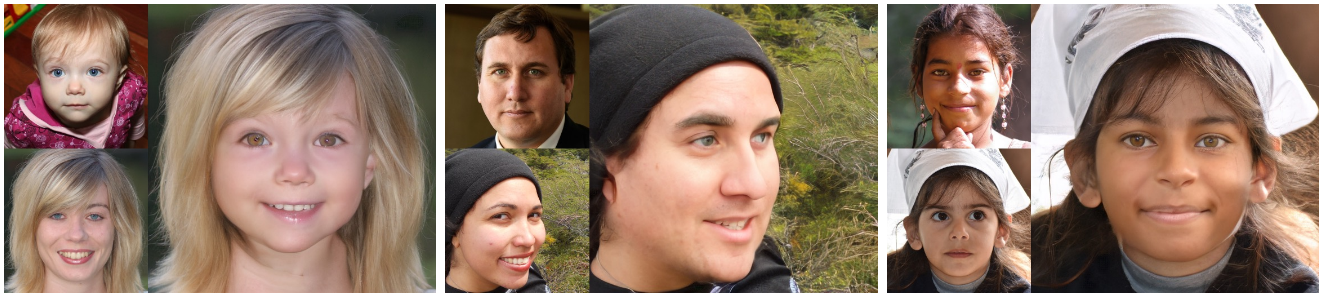

Figure 1. Face swapping results of DiffSwap at 512 × 512 resolution. In each group, we show the swapped face (right) generated by replacing the target face (bottom-left) with the source face (top-left). Benefiting from the generative power and controllability, our model can produce high-fidelity swapped faces that are robust to the differences in face shape, face pose and skin color between the source face and target face. Best viewed in color.

Abstract

In this paper, we propose DiffSwap, a diffusion model based framework for high-fidelity and controllable face swapping. Unlike previous work that relies on carefully designed network architectures and loss functions to fuse the information from the source and target faces, we reformulate the face swapping as a conditional inpainting task, performed by a powerful diffusion model guided by the desired face attributes (e.g., identity and landmarks). An important issue that makes it nontrivial to apply diffusion models to face swapping is that we cannot perform the time-consuming multi-step sampling to obtain the generated image during training. To overcome this, we propose a midpoint estimation method to efficiently recover a reasonable diffusion result of the swapped face with only 2 steps, which enables us to introduce identity constraints to improve the face swapping quality. Our framework enjoys several favorable properties more appealing than prior arts: 1) Controllable. Our method is based on conditional masked diffusion on the latent space, where the mask and the conditions can be fully controlled and customized. 2) High-fidelity. The formulation of conditional inpainting can fully exploit the generative ability of diffusion models and can preserve the background of target images with minimal artifacts. 3) Shape-preserving. The controllability of our method enables us to use 3D-aware landmarks as the condition during generation to preserve the shape of the source face. Extensive experiments on both FF++ and FFHQ demonstrate that our method can achieve state-of-the-art face swapping results both qualitatively and quantitatively.

1. Introduction

There has been growing interest in face swapping technology from both vision and graphics communities [1, 2, 6, 21, 36, 39] because of its broad applications in creating digital twins, making films, protecting privacy, *etc*. The goal of face swapping is to transfer the identity of the source face to a target image or a video frame while keeping the attributes (*e.g.*, pose, expression, background) unchanged.

There are two essential steps in realizing high-quality face swapping: encoding the identity information of the source face effectively and blending identity and attributes from different images seamlessly. Early work [4, 24] on face swapping adopts 3D models [5] to represent the source face and directly replace the reconstructed faces in the target image based on 3D structural priors, leading to recognizable artifacts. The development of generative adversarial networks (GAN) [12] provides a strong tool to generate photo-realistic face images. Many recent methods [2, 6, 21, 39] perform face swapping by extracting the identity feature

[†] Corresponding author

from the source image and then injecting it into the generative models powered by adversarial training. However, these methods tend to make minor modifications to the target image, which may fail in totally transferring the identity information when the face shapes of the source and the target face largely differ.

Very recently, diffusion-based models (DM) [22, 25, 26] have exhibited high customizability for various conditions and impressive power in generating images with high resolution and complex scenes. It is natural to ask: *whether the strong generation ability of diffusion models can benefit face swapping?* However, we find it is nontrivial to apply diffusion models to the task of face swapping. Since there is no ground-truth data pair for face swapping, face swapping models are usually trained in a weakly-supervised manner, where several losses about image fidelity, identity, and facial attributes are imposed to guide the training. These supervisory signals can be easily added to GAN-based models but it is difficult for DMs. Different from previous generative models like GANs [12, 16] and VAEs [14, 19], DMs learn a denoising autoencoder to gradually recover the data density step-by-step. Although the autoencoder can be efficiently learned by performing score matching [15] at an arbitrary step during training, image generation using an already trained DM requires executing the autoencoder sequentially for a large number of steps (typically, 200 steps), which is computationally expensive.

To tackle these challenges, we propose the first diffusion model based face swapping framework, which can produce high-fidelity results faces with high controllability. Figure 2 shows the overview of our method. Different from existing methods [1,6,21,36] that modify the target face to match the identity of the source face, we reformulate face swapping as a conditional inpainting task guided by the identity feature and facial landmarks. Our diffusion model is learned to generate a face that shares the same identity as the source face and is spatially aligned with the target face. In order to introduce identity constraints during training, we propose a midpoint estimation method that can efficiently generate swapped faces with only 2 steps. Our framework is by design highly controllable, where both the conditioned landmark and the inpainting mask can be customized during inference. Thanks to this property, we propose the 3D-aware masked diffusion where we perform the inpainting inside the 3D-aware mask conditioned on the 3D-aware landmark that explicitly enforces the shape consistency between the source face and the swapped face.

We conducted extensive experiments on FaceForensics++ [27] and FFHQ [16] to verify the effectiveness of our model both and quantitatively. On FF++ dataset, our method outperforms previous methods in both ID retrieval (98.54%) and FID (2.16), while achieving comparable results on pose error and expression error. Qualitative results show that our method can generate high-fidelity swapped faces that can better preserve the source face shape than the previous method. Besides, we also demonstrate the scalability and controllability of our method. Our model can be easily extended to higher-resolution such as 512×512 with affordable extra computational costs and allows region-customizable face swapping by controlling the inpainting mask. Our results demonstrate that DiffSwap is a very promising face swapping framework that is distinct from the existing methods and enjoys high fidelity, controllability, and scalability.

2. Related Work

Face Swapping. Existing face swapping methods can be roughly categorized into 3D-based methods and GAN-based methods. The 3D-base methods [4, 24] usually leverage the 3DMM [5] to introduce structural priors. However, these methods are involved human intervention or produce recognizable artifacts. The GAN-based methods [6,17,21,23,29,36,38] are mostly target-oriented, which fuses the identity information from the source face to the target features and uses GAN to ensure the fidelity of the swapped face. However, these methods always contain multiple loss functions and balancing them requires careful tuning of the hyper-parameters. Besides, these methods tend to make minor modifications on the target face and thus cannot deal with the cases where the shapes between the source face and the target face largely differ. Although some existing works using the 3DMM features [20, 36] to guide the swapping, we find this implicit incorporation of 3D information still cannot ensure the shape consistency. Different from previous works, our method train a diffusion model conditioned on the identity feature and facial landmark, which enables us to delicately control the facial shape using 3D priors during inference.

Diffusion Models. Diffusion models, emerging as another family of generative models, have achieved state-of-the-art results [9] recently. Different from previous generative models like GANs [12, 16] and VAEs [14, 19, 32] that often suffer from instability during training, the optimization of diffusion models is equivalent to score matching [3] and can be implemented using a simple MSE loss [15]. The stable training of the diffusion model also makes it more flexible to capture the conditional data density [3, 26]. Therefore, it is a promising direction to further exploit the controllability and high-fidelity of the diffusion models. However, the application of the diffusion model to the face swapping task is nontrivial because 1) we do not have the ground truth of the swapped face, thus the original objectives in DMs can not help to perform face swapping. 2) the image generation of DMs requires multiple steps of model evaluation which is prohibited during training, leading to the difficulty to in-

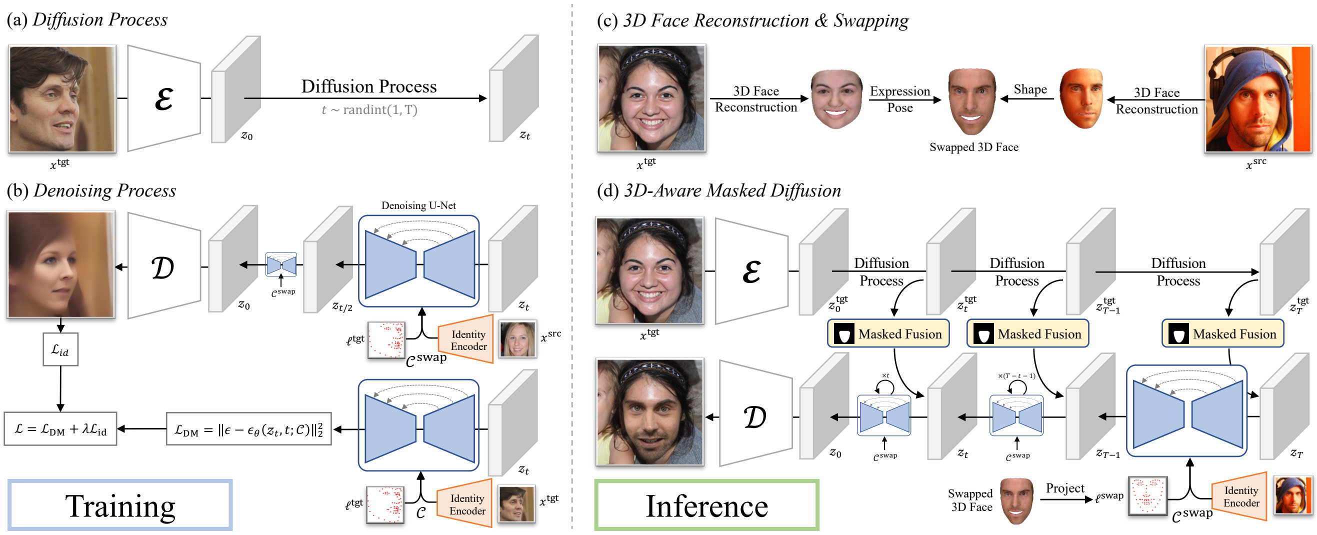

Figure 2. **Overview of *DiffSwap*.** DiffSwap is a diffusion model based framework for face swapping. During training (left), we train a conditional denoising U-Net ϵ_θ to predict the gradient of the data density ϵ. We also devise a midpoint estimation method to obtain a swapped face to enable explicit identity constraint. During inference (right), we leverage 3D face reconstruction to obtain the swapped 3D face and use the projected landmark to perform 3D-Aware masked diffusion to produce shape-preserving face swapping results.

clude identity constraints for face swapping. In this work, we solve the above issues by reformulating face swapping as conditional inpainting to make use of the original DM objective and by adopting a new midpoint estimation to recover the swapped face during training to enable the computation of identity loss. Benefiting from the success of DMs, our framework can generate high-fidelity face swapping results with high controllability.

3. Method

In this section, we will describe our method DiffSwap, the first diffusion model based face swapping framework in detail. An overview of our method is illustrated in Figure 2.

3.1. Preliminaries: Diffusion Models

Diffusion Models [30] are a family of generative models that can recover the data distribution from a Gaussian noise by learning the reverse process of a Markov Chain. Let $\boldsymbol{z}_t$ be the random variable at t-th timestep, the Markov Chain is defined as

$$\boldsymbol{z}_t \sim \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{z}_{t-1}, (1-\alpha_t)\boldsymbol{I}), \tag{1}$$

where $\{\alpha_t\}$ is a predefined sequence of coefficients controlling the variance schedule. The close form of the distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_0)$ of can be easily derived from the above formulation:

$$\begin{aligned}\boldsymbol{z}_t &= \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon},\\ \bar{\alpha}_t &= \prod_{s=1}^{t}\alpha_s, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),\end{aligned} \tag{2}$$

which makes it possible to efficiently sample an arbitrary $\boldsymbol{z}_t$ during training. By minimizing the ELBO of the reverse process, the training objective of diffusion models can be decomposed into a summation of step-wise KL-divergence between the predicted distribution of a reverse step and the corresponding posterior of the forward process, which can be further simplified into the following form through reparameterization:

$$L_{\text{DM}} = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t(\boldsymbol{z}_0,\boldsymbol{\epsilon}),t)\|_2^2\right], \tag{3}$$

where $\boldsymbol{z}_t$ is obtained using Equation (2) and $\boldsymbol{\epsilon}_\theta$ denotes the denoising autoencoder learned to predict the $\boldsymbol{\epsilon}$, which can be viewed as the gradient of the log data density. In visual generative tasks, the denoising autoencoder is usually implemented as an U-Net. During inference, diffusion models gradually predict the $\boldsymbol{\epsilon}_\theta(x_t, t)$ and recover the $\boldsymbol{z}_0$.

However, the vanilla diffusion models suffer from heavy computational costs in both training and inference, since the diffusion process is directly operated on the pixel space ($\boldsymbol{z}_t \in \mathbb{R}^{3\times H\times W}$). To address this issue, [26] find it helpful to decompose the whole generative procedure into semantic and perceptual levels. They propose the latent diffusion model (LDM) where the image is first compressed into a latent space (*e.g.* 64×64) through a pre-trained VQGAN [10], and a diffusion model is trained on that latent space instead of the original pixel space. Following LDM [26], our method also performs face swapping on the latent space for efficient training and inference.

3.2. Face Swapping as Conditional Inpainting

The goal of face swapping is to transfer the identity of the source face to a target image while keeping the attributes (pose, expression, background, *etc.*) unchanged. Most existing face swapping methods [1,6,21] adopt a pipeline similar to face editing, where the model gradually injects the identity information into the features from the target image. However, these methods tend to make *small* modifications to the target image, thus often failing to preserve the shape of the source face. Although some recent work [36] adopts the shape-identity feature to guide the face swapping, we find that this implicit injection of the shape information cannot produce satisfactory results when the shapes of the source and target faces largely differs (*e.g.*, Figure 4).

To address the above issues, we reformulate face swapping as conditional inpainting, where both the identity and the attributes of the generated faces are controllable through the conditioning vectors. Specifically, we first train a VQ-GAN that can transform the input image $\boldsymbol{x}_0 \in \mathbb{R}^{3\times H\times W}$ to the latent space as $\boldsymbol{z}_0 \in \mathbb{R}^{C\times H'\times W'}$. We then train a conditional diffusion model on that latent space with a specific emphasis on identity consistency. The diffusion model can be written as $\epsilon_\theta(z_t, t; \mathcal{C})$, which performs denoising given the conditioning set $\mathcal{C}$ and the timestep t. During inference, we construct a mask $\mathcal{M} \in \{0,1\}^{H'\times W'}$ on the latent space and use it to control the inpainting region.

Design of Conditioning Inputs. From a generative perspective, we aim to generate a new face that shares the identity with the source face and spatially aligns well with the target face. To extract the identity feature, we use a pre-trained face recognition model [7] $\mathcal{E}_{\text{id}}$. We then use a two-layer MLP to project the identity feature to a pre-defined dimension D to obtain the identity condition:

$$\boldsymbol{c}_{\text{id}} = \text{MLP}(\mathcal{E}_{\text{id}}(\boldsymbol{x})) \in \mathbb{R}^D. \tag{4}$$

Another important condition input is the facial landmark $\boldsymbol{\ell} \in \mathbb{R}^{68\times 2}$, since it can control both the pose and the expression of the generated face. Similarly, we can use an MLP to extract landmark features $\boldsymbol{c}_{\text{lmk}} \in \mathbb{R}^D$.

A diffusion model trained with the identity feature and facial landmark as the conditioning inputs can already achieve face swapping by using the identity feature from the source image and the landmark feature from the target image. Inspired by [37], we also include the region features (*e.g.*, eyes, nose, mouth) as another conditioning inputs to further improve the similarity between the swapped face and the source face. Specifically, we consider three regions including the eyes, nose, and mouth. For the sake of simplicity, we use the facial landmark to get the region masks $\mathcal{M}_{\text{eyes}}, \mathcal{M}_{\text{nose}}, \mathcal{M}_{\text{mouth}}$, and apply them on $\boldsymbol{z}_0$ to extract region features. Similarly, the region features are projected into a D dimension feature $\boldsymbol{c}_{\text{region}} \in \mathbb{R}^{3\times D}$ using an MLP. We then utilize the multi-head self-attention mechanism [33] (MHSA) to capture the interactions among different regions:

$$\boldsymbol{c}_{\text{region}} \leftarrow \boldsymbol{c}_{\text{region}} + \text{MHSA}(\text{LayerNorm}(\boldsymbol{c}_{\text{region}})), \tag{5}$$

where the region feature is updated via a residual connection. By combining $\boldsymbol{c}_{\text{id}}$ and $\boldsymbol{c}_{\text{region}}$, we are able to better leverage the identity information from both global and local levels. The final conditioning features that are fed into our network are $\mathcal{C} = \{\boldsymbol{c}_{\text{id}}, \boldsymbol{c}_{\text{lmk}}, \boldsymbol{c}_{\text{region}}\}$.

Training Objectives. The optimization of our DiffSwap is similar to the diffusion model, where we use the conditioning features $\mathcal{C}$ to guide the denoising in each timestep. The loss function for the diffusion model is defined as

$$L_{\text{DM}} = \mathbb{E}_{\boldsymbol{z}_0=\mathcal{E}(\boldsymbol{x}),\mathcal{C},\boldsymbol{\epsilon}\in\mathcal{N}(0,1),t}\left[\|\boldsymbol{\epsilon}, \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t; \mathcal{C})\|_2^2\right]. \tag{6}$$

From the theoretical analysis of the underlying mechanism of diffusion models [15], the ϵ_θ can be viewed as the learned conditional score $\nabla \log p(\boldsymbol{z}|\mathcal{C})$ [3]. Therefore, optimizing Equation 6 helps the model to learn how to recover the conditional distribution $p(\boldsymbol{z}|\mathcal{C})$.

In the face swapping task, it is crucial to make sure the swapped face can preserve the identity of the source face. Previous methods often explicitly introduce an identity loss that aims to maximize the cosine similarity between the swapped face and the source face in the feature space. However, since the image generation of diffusion models requires multiple model evaluations on different timesteps (*e.g.*, 200 steps), obtaining such an image is time-consuming during training. Therefore, it is nontrivial to add identity loss to our framework. A naive solution is to recover the $\boldsymbol{z}_0$ directly from the $\boldsymbol{z}_t$ and the learned conditional score $\boldsymbol{\epsilon}_\theta$. Considering the reparameterized forward process (Equation (2)), given the feature $\boldsymbol{z}_t$ and the learned conditional score $\epsilon_\theta(\boldsymbol{z}_t, t; \mathcal{C})$, we have:

$$\hat{\boldsymbol{z}}_0^{\text{vanilla}} = \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta}{\sqrt{\bar{\alpha}_t}}. \tag{7}$$

However, the above prediction of $\boldsymbol{z}_0$ is inaccurate because the gradient of the data density is computed at $\boldsymbol{z}_t$. To obtain a better estimation of $\boldsymbol{z}_0$ with minimal extra computational costs, we propose a midpoint estimation method that only requires two times of model evaluations and can produce a reasonably swapped face for the computation of identity loss. Specifically, we first divide the t by half to get $t_1 = \lfloor t/2 \rfloor$. From the forward process, we have

$$\begin{aligned} \boldsymbol{z}_t &= \sqrt{\prod_{\tau=t_1+1}^{t} \alpha_\tau}\,\boldsymbol{z}_{t_1} + \sqrt{1 - \prod_{\tau=t_1+1}^{t} \alpha_\tau}\,\boldsymbol{\epsilon} \\ &= \sqrt{\bar{\alpha}_t/\bar{\alpha}_{t_1}}\,\boldsymbol{z}_{t_1} + \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t_1}}\,\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \in \mathcal{N}(0,1). \end{aligned} \tag{8}$$

We can then first estimate the $\boldsymbol{z}_{t_1}$ given the predicted score at $\boldsymbol{z}_t$:

$$\hat{\boldsymbol{z}}_{t_1} = \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t_1}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t; \mathcal{C})}{\sqrt{\bar{\alpha}_t/\bar{\alpha}_{t_1}}}. \tag{9}$$

Once we have the predicted $\hat{\boldsymbol{z}}_{t_1}$, we can then apply Equation (7) to estimate $\boldsymbol{z}_0$ by substitute $\boldsymbol{z}_t$ to $\hat{\boldsymbol{z}}_{t_1}$:

$$\hat{\boldsymbol{z}}_0^{\text{midpoint}} = \frac{\hat{\boldsymbol{z}}_{t_1} - \sqrt{1-\bar{\alpha}_{t_1}}\boldsymbol{\epsilon}_\theta}{\sqrt{\bar{\alpha}_{t_1}}}. \tag{10}$$

The midpoint estimation only requires one extra evaluation of the model $\boldsymbol{\epsilon}_\theta$. To validate the effectiveness of the midpoint estimation, we provide a visualization in Figure 3, where we compare the recovered swapped face using vanilla estimation (VE) and midpoint estimation (ME). We also provide the final swapped face (sampled during inference for 200 steps) in the last column for reference. We show that the vanilla estimation of $\boldsymbol{z}_0$ is inaccurate with few sampling steps. On the other hand, the proposed midpoint estimation can produce the swapped face that is much closer to the final result than the vanilla estimation in only 2 steps.

Equipped with the midpoint sampling, we are able to effectively obtain a reasonably swapped face during training, by simply modifying the condition $\mathcal{C}$ as

$$\mathcal{C}^{\text{swap}} = \{\boldsymbol{c}_{\text{id}}^{\text{src}}, \boldsymbol{c}_{\text{lmk}}^{\text{tgt}}, \boldsymbol{c}_{\text{region}}^{\text{src}}\}, \tag{11}$$

we can then compute the identity loss by

$$\mathcal{L}_{\text{id}} = 1 - \text{CosSim}(\mathcal{E}_{\text{id}}(\boldsymbol{x}^{\text{src}}), \mathcal{E}_{\text{id}}(\mathcal{D}(\hat{\boldsymbol{z}}_0(\boldsymbol{z}_t^{\text{tgt}}; \mathcal{C}^{\text{swap}})))). \tag{12}$$

The final objective can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{LDM}} + \lambda\mathcal{L}_{\text{id}}, \tag{13}$$

where $\lambda = 0.1$ is a hyper-parameter to balance the numerical scale of the two terms.

3D-Aware Masked Diffusion We now describe how to perform face swapping during inference in our framework. Given the target image $\boldsymbol{x}^{\text{tgt}}$, we first use the encoder $\mathcal{E}$ to embed it into the latent space as $\boldsymbol{z}_0^{\text{tgt}}$. Secondly, we construct a mask $\mathcal{M}$ to specify the area to perform face swapping. We then perform the conditional inpainting through masked diffusion:

$$\begin{aligned}
&\boldsymbol{z}_t^{\text{tgt}} \leftarrow \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0^{\text{tgt}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t \\
&\boldsymbol{z}_t \leftarrow \mathcal{M} \odot \boldsymbol{z}_t + (1-\mathcal{M}) \odot \boldsymbol{z}_t^{\text{tgt}} \\
&\boldsymbol{z}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t; \mathcal{C}^{\text{swap}})\right) + \sigma_t\boldsymbol{n}_t \\
&\boldsymbol{z}_T, \boldsymbol{\epsilon}_t, \boldsymbol{n}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \quad t = T, T-1, \ldots, 1.
\end{aligned} \tag{14}$$

Note that we follow the reverse sampling method in [15] to sample $\boldsymbol{z}_{t-1}$ from $p_\theta(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t; \mathcal{C}^{\text{swap}})$. Finally, we use the decoder to transform $\boldsymbol{z}_0$ back to the image space and the $\mathcal{D}(\boldsymbol{z}_0)$ is the swapping result.

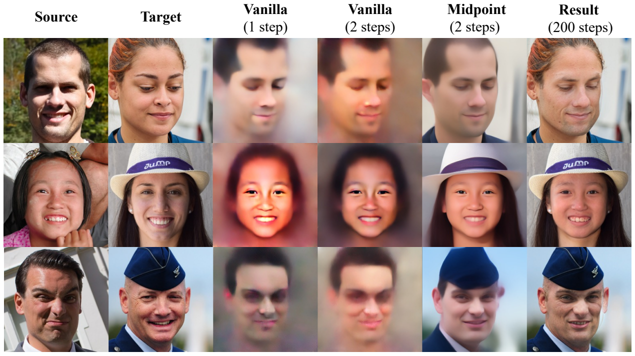

Figure 3. **Comparison between the vanilla estimation (VE) and the midpoint estimation (ME).** Given the source image and the target image, we visualize the estimated swapped face using VE (Equ. (7)) and ME (Equ. (10)) during training, as well as the final swapped face during inference. We show that compared with the VE, ME can generate plausible results that are more close to the final swapped face with only 2 steps of sampling.

Our framework is by design highly controllable because we can change both the mask $\mathcal{M}$ and the conditioning inputs $\mathcal{C}^{\text{swap}}$ during inference. Therefore, it is possible to perform careful designs on $\mathcal{M}$ or $\mathcal{C}^{\text{swap}}$ to produce more plausible face swapping results. Apart from the identity feature and the region features that come from the source face, we aim to seek a better choice for the landmark feature $\boldsymbol{c}_{\text{lmk}}$. Conditioning on a better input landmark, our model has the potential to solve the corner case where the shapes between the source face and the target face largely differ. To obtain a 3D-aware facial landmark that shares the pose and expressions with the target face and preserve the shape of the source face, we adopt a 3D face reconstruction library [8] to extract the 3D information of both the source face and target face. The 3D face reconstruction results consist of several parameters describing the shape, expression, pose, texture, *etc*., thus we can simply replace the shape of the reconstructed target face with that of the source face. We then reconstruct a new face using the swapped parameters and obtain the corresponding 2D facial landmark $\boldsymbol{\ell}^{\text{swap}}$, which can be further fed into our model for face swapping. To deal with the misalignment between the shape of the source face and the target face, the mask $\mathcal{M}$ must cover both the original landmark of the target face $\boldsymbol{\ell}^{\text{tgt}}$ and the 3D-aware landmark $\boldsymbol{\ell}^{\text{swap}}$. In our implementation, we simply compute a convex hull of the concatenation of the two sets of landmarks $[\boldsymbol{\ell}^{\text{tgt}}, \boldsymbol{\ell}^{\text{swap}}]$ to obtain the 3D-aware mask $\mathcal{M}^{\text{swap}}$.

Discussion. The idea of using inpainting to solve face swapping is investigated in some previous works like FaceInpainter [20]. However, our framework is distinct in two aspects: 1) FaceInpainter computes the swapped face

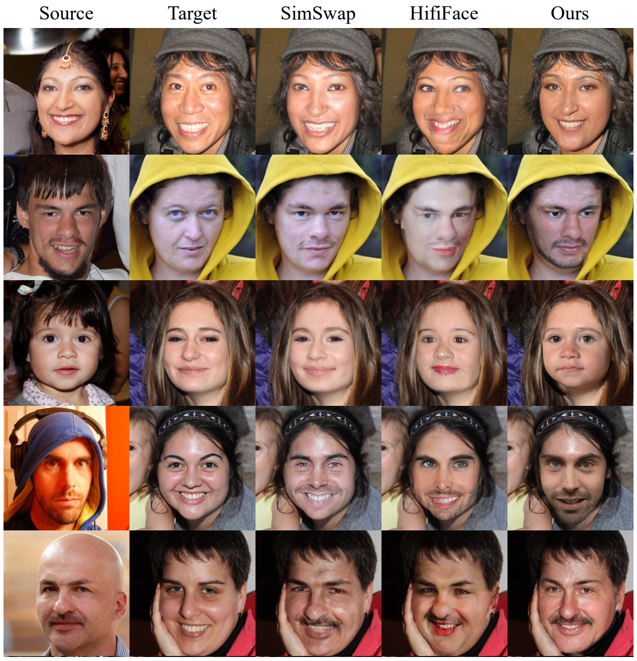

Figure 4. **Qualitative comparisons on FFHQ [16].** Our method can produce high-fidelity results that preserve both the identity and the shape of the source face.

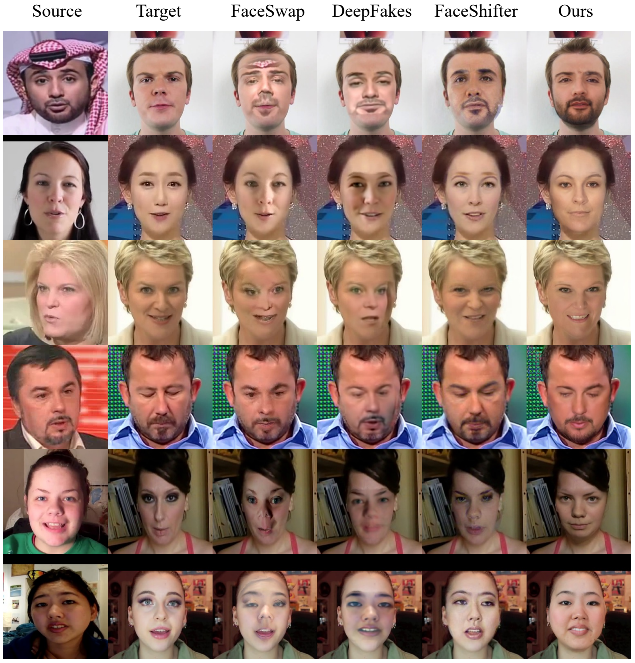

Figure 5. **Qualitative comparisons on FF++ [27].** Our method generalizes well to unseen data distribution and can also better preserve both the identity and the face shape.

Table 1. **Quantitative Comparisons on FF++ [27].** We report the ID retrieval, pose error, expression error, and the Fréchet inception distance and show that our method achieves very competitive results compared with existing methods.

Method	ID Retrieval ↑	Pose ↓	Expression ↓	FID↓
Deepfakes [1]	86.43	3.96	8.98	4.07
FaceShiter [21]	90.04	2.19	6.77	3.50
SimSwap [6]	93.07	**1.36**	**5.07**	3.04
MegaFS [40]	89.12	3.69	10.12	4.62
InfoSwap [11]	95.82	2.54	6.99	4.74
StyleSwap [39]	97.87	1.51	5.27	2.58
DiffSwap (Ours)	**98.54**	2.45	5.35	**2.16**

by directly combining the generated face and the original mask using the facial parsing mask on the image space, which might introduce artifacts and is prune to the parsing result. Our method, however, is based on the masked diffusion on the latent space which will smooth the masking boundary during the gradually denosing. 2) The mask in FaceInpainter is fixed to be the face parsing result, while our framework allows arbitrary input of the inpainting mask and thus is more controllable.

4. Experiments

Datasets We train our model on the FFHQ [16] dataset. FFHQ contains 70,000 high-quality face images that are crawled from the web and is widely used in the training of generative models [16, 26]. The original resolution of FFHQ is 1024×1024 and we use the resized images of 256×256 and 512×512 in our experiments. Following common practice, we also evaluate our method on FaceForensics++(FF++) dataset [27] which contains 1,000 videos, as well as the face swapping results of some previous methods.

Implementation Details. In all of our experiments, we use a latent space with $3 \times 64 \times 64$, which makes our diffusion process computationally efficient. Following previous works [15, 26], we use a U-Net architecture for the ϵ_θ network, where the conditioning features are injected using the cross-attention mechanism [33]. We train our diffusion model with a global batch size of 32 on 8 NVIDIA Tesla A100 GPU for 100K iterations. We use the Adam [18] optimizer with the base learning rate of 2e-6 and the linear scaling rule [8]. For the training of the first stage autoencoder $\mathcal{E}, \mathcal{D}$, we adopt the VQ-regularization [26] and the global batch size is set as 64 for 256×256 and 32 for 512×512 resolutions. During inference, we use the DDIM [31] sampler with 200 steps following [26]. For more implementation details, please refer to the supplementary materials.

4.1. Comparisons with Existing Methods

In this section, we will evaluate the effectiveness of our method both quantitatively and qualitatively on FF++ [27] and FFHQ [16] datasets.

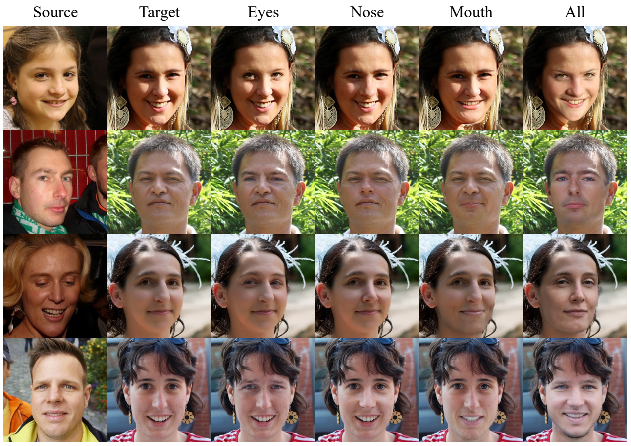

Figure 6. **Region-controllable face swapping.** By constructing masks covering different regions, our method can control which region to be swapped.

Quantitative Comparison. We conduct experiments on FF++ [27] dataset and compare our result with previous methods in Table 1. Following common practice, we sample 10 images from each video to perform face swapping and compute the metrics including ID retrieval, pose error, expression error, and the Fréchet inception distance(FID). To compute ID Retrieval, we first extract the identity feature using a different face recognition model [35]. For each swapped face, we compute the nearest face from all the frames in FF++ using the cosine similarity and check whether it is from the source video. The pose error is computed by the L-2 distance between the results' and the targets' poses estimated by a pose estimator [28]. The expression error is the L-2 distance between the expression embeddings extracted by [34] of the swapped face and the target face. The results show that our method outperforms previous methods in ID Retrieval and FID, indicating that we can generate high-fidelity swapped faces and can better preserve the source identity. Meanwhile, we also achieve comparable results on pose and expression, demonstrating that our method can also keep the target attributes.

Qualitative Comparisons. We perform qualitative comparisons on both FFHQ [16] and FF++ [27]. For FFHQ, we compared our method with two open-source methods SimSwap [6] and HifiFace [36] and the results are shown in Figure 4. We demonstrate that our method can yield high-fidelity face swapping results, especially on face shapes and local characteristics (eyes, nose, mouth). Specifically, in the third row of Figure 4, our model can successfully transfer the face shape of the children benefiting from the 3D-aware masked diffusion while other methods tend to keep the shape of the target face. These results indicate that explicitly controlling the landmark of the swapped face is more useful to preserving the source shape than implicitly injecting a 3D-related feature like HifiFace [36]. We also perform the qualitative evaluation on FF++ dataset, where we use the swapping results contained in the FF++ dataset including FaceSwap [2], DeepFakes [1] and FaceShifter [21]. From the results in Figure 5, we show that our method can also better preserve both the identity and the face shape of the source image, indicating that our method generalizes well to unseen data distribution.

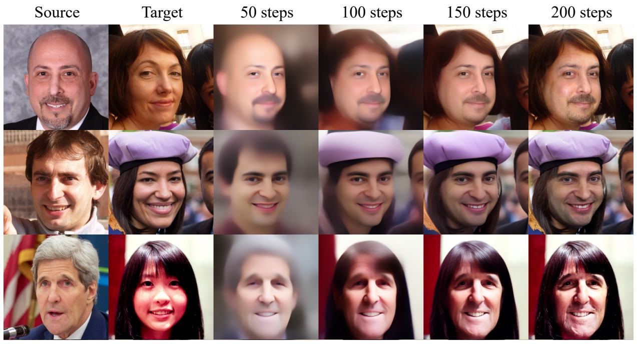

Figure 7. **Visualization of the diffusion process.** Our model start by generating a face aligned with the target face and gradually add details to make it similar to the source face.

4.2. Analysis

Region-Controllable Face Swapping. Unlike previous methods that swap the whole identity information to the target face, our method is more controllable to allow specifying which region to be swapped. This can be easily achieved by changing the masks during inference. To demonstrate the controllability of our framework, we construct three masks that cover the eyes, nose, and mouth, respectively. We then perform the masked diffusion inside those masks to achieve region-controllable face swapping, as shown in Figure 6. We also include the swapping results of the whole face in the last column. For the region swapping, we use the landmark of the target face as the condition instead of the 3D-aware landmark. We show that our method can swap a specific region, leaving the unmasked part unchanged. These results also demonstrate that our model can capture the low-level attributes of the regions, which are crucial to recognizing the identity of a person.

Visualization of Diffusion Process. To better understand how our model performs face swapping, we provide a visualization of the intermediate output of the diffusion steps in Figure 7. Specifically, we use Equation (7) to predict the $\hat{z}_0$ given the intermediate latent features z_t, and decode it back to image space $\mathcal{D}(\hat{z}_0)$ for visualization. We find that at very early steps (*e.g.* 50 steps), our model can already generate a blurred face that shares the same pose with the target face. Afterward, our model gradually refines the face to match the given conditioning landmark and adds details to ensure identity consistency with the source face. We also find that our model can deduce the lighting from the background by

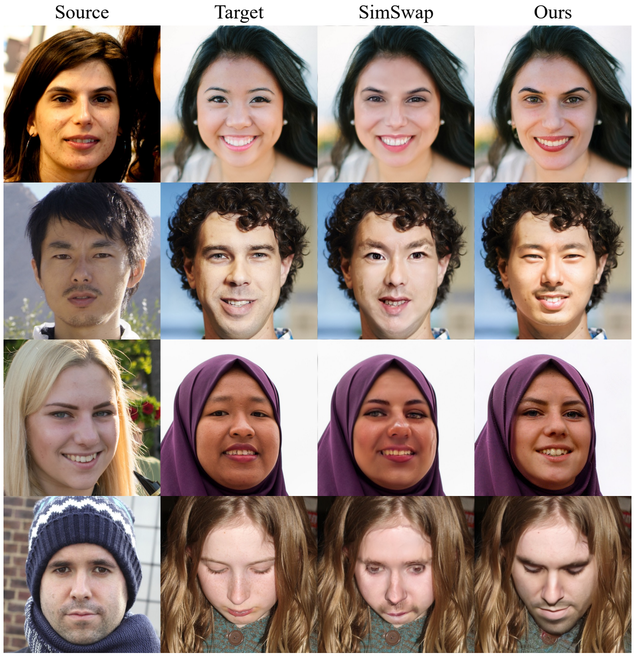

Figure 8. **Qualitative comparisons on FFHQ [16] at** 512×512 **resolution.** Our method remains robust to different poses and shapes between the source and target faces at higher resolution.

progressively performing conditional inpainting.

Face Swapping at 512×512**.** To further demonstrate the scalability of our method, we perform experiments on 512×512 resolution. To achieve this, we train another VQGAN that takes as input a 512×512 image while the latent space is still 64×64. The 512×512 VQGAN is constructed by adding extra layers to the original 256×256 one, thus we can use the pre-trained weights of the 256×256 VQGAN for fast adaptation. We then fine-tune our diffusion model on the new latent space for 10K iterations. Figure 8 compares the face swapping results of SimSwap [6] and our method at 512×512 resolution. We show that our model can still generate high-fidelity swapping results at a higher resolution, and is robust to different poses and face shapes between the source and target faces.

Pose Controlling via DiffSwap. Apart from the competitive performance of our method in the face swapping problem which aims to integrate the identity into a target pose, we now demonstrate another usage of our method, *i.e.*, to control the pose of a specific person. To achieve this, we first extract the 3D parameters of the input face using the 3D face reconstruction library [13]. We then rotate the 3D face to some specific poses and render the corresponding landmarks in the 2D space. The images with various poses can then be generated by using the 2D landmarks at different poses and the identity feature as the conditioning inputs. As is shown in Figure 9, we visualize the source face at 5 different poses. No mask is used due to the large variance of poses. The results demonstrate that our model can disentangle the identity and landmark features and can successfully model the conditional data distribution.

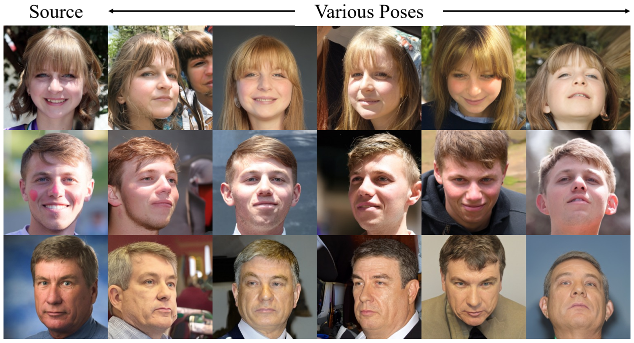

Figure 9. **Pose controlling via DiffSwap.** As another application, DiffSwap can also control the pose of a specific person. We show that the identity and the pose are well disentangled.

Limitations. Despite the effectiveness of DiffSwap, we find that there still exist some disadvantages of our method. Firstly, since we formulate the face swapping as conditional inpainting, some detailed attributes of the target face can not be fully preserved. Secondly, our method is nondeterministic due to the generative formulation and thus sometimes suffers from instability. Thirdly, our method cannot deal with occlusion. We will improve our method from the above aspects in future work.

5. Conclusion

We have presented a new framework named DiffSwap which leveraged the powerful diffusion model by reformulating face swapping as conditional inpainting. Several efforts have been taken to adapt the diffusion model to face swapping, including the designs on conditioning inputs and the midpoint estimation during training. We have developed a 3D-aware masked diffusion to explicitly ensured the consistency of face shape for the first time. Extensive experiments demonstrate our framework can achieve favorable results compared to previous methods and enjoys better controllability and scalability. We hope our attempt can inspire future work to further explore the formulation and implementation of face-swapping to achieve better results.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 62125603, and in part by a grant from the Beijing Academy of Artificial Intelligence (BAAI).

References

[1] Deepfakes: Faceswap. https://github.com/deepfakes/faceswap. Accessed: 2022-9-1. 1, 2, 4, 6, 7

[2] Faceswap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. Accessed: 2022-10-15. 1, 7

[3] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 2, 4

[4] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004. 1, 2

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 1, 2

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *ACMMM*, pages 2003–2011, 2020. 1, 2, 4, 6, 7, 8

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 4

[8] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *CVPRW*, 2019. 5, 6

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 2

[10] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 3

[11] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *CVPR*, pages 3404–3413, 2021. 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 1, 2

[13] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 8

[14] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016. 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 4, 5, 6

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2, 6, 7, 8

[17] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-swap: A simple enhancement for face-swapping with smoothness. In *CVPR*, pages 10779–10788, 2022. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[20] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. Faceinpainter: High fidelity face adaptation to heterogeneous domains. In *CVPR*, pages 5089–5098, 2021. 2, 5

[21] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. In *CVPR*, 2020. 1, 2, 4, 6, 7

[22] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

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

[24] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *FG 2018*, pages 98–105. IEEE, 2018. 1, 2

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 6

[27] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pages 1–11, 2019. 2, 6, 7

[28] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *CVPRW*, pages 2074–2083, 2018. 7

[29] Changyong Shu, Hemao Wu, Hang Zhou, Jiaming Liu, Zhibin Hong, Changxing Ding, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Few-shot head swapping in the wild. In *CVPR*, pages 10789–10798, 2022. 2

[30] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 3

[31] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. 6

[32] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 4, 6

[34] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In *CVPR*, pages 5683–5692, 2019. 7

[35] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, pages 5265–5274, 2018. 7

[36] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 1, 2, 4, 7

[37] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *CVPR*, pages 7632–7641, 2022. 4

[38] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *CVPR*, pages 7642–7651, 2022. 2

[39] Zhiliang Xu, Hang Zhou, Zhibin Hong, Ziwei Liu, Jiaming Liu, Zhizhi Guo, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Styleswap: Style-based generator empowers robust face swapping. In *ECCV*, pages 661–677. Springer, 2022. 1, 6

[40] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *CVPR*, pages 4834–4844, 2021. 6