3D-Aware Face Swapping

Yixuan Li Chao Ma* Yichao Yan* Wenhan Zhu Xiaokang Yang
MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
{lyx0208, chaoma, yanyichao, zhuwenhan823, xkyang}@sjtu.edu.cn

Figure 1. Demonstration of the proposed 3dSwap. Given single-view source and target images, our method synthesizes high-fidelity and multi-view-consistent images of the swapped faces and the corresponding geometries. More results can be found on our project page.

Abstract

Face swapping is an important research topic in computer vision with wide applications in entertainment and privacy protection. Existing methods directly learn to swap 2D facial images, taking no account of the geometric information of human faces. In the presence of large pose variance between the source and the target faces, there always exist undesirable artifacts on the swapped face. In this paper, we present a novel 3D-aware face swapping method that generates high-fidelity and multi-view-consistent swapped faces from single-view source and target images. To achieve this, we take advantage of the strong geometry and texture prior of 3D human faces, where the 2D faces are projected into the latent space of a 3D generative model. By disentangling the identity and attribute features in the latent space, we succeed in swapping faces in a 3D-aware manner, being robust to pose variations while transferring fine-grained facial details. Extensive experiments demonstrate the superiority of our 3D-aware face swapping framework in terms of visual quality, identity similarity, and multi-view consistency. Code is available at https://lyx0208.github.io/3dSwap.

* Corresponding authors.

1. Introduction

Face swapping aims to transfer the identity of a person in the source image to another person in the target image while preserving other attributes like head pose, expression, illumination, background, *etc*. It has attracted extensive attention recently in the academic and industrial world for its potential wide applications in entertainment [14, 30, 38] and privacy protection [7, 37, 48].

The key of face swapping is to transfer the geometric shape of the facial region (*i.e*., eyes, nose, mouth) and detailed texture information (such as the color of eyes) from the source image to the target image while preserving both geometry and texture of non-facial regions (*i.e*., hair, background, *etc*). Currently, some 3D-based methods consider geometry prior of human faces by fitting the input image to 3D face models such as 3D Morphable Model (3DMM) [8] to overcome the differences of face orientation and expression between sources and targets [7, 15, 34, 43]. However, these parametric face models only produce coarse frontal faces without fine-grained details, leading to low-resolution and fuzzy swapping results. On the other hand, following Generative Adversarial Network [24], GAN-based [6, 23, 32, 39, 40, 42] or GAN-inversion-based [44, 55, 57, 60] approaches adopt the ad-

versarial training strategy to learn texture information from inputs. Despite the demonstrated photorealistic and high-resolution images, the swapped faces via 2D GANs sustain undesirable artifacts when two input faces undergo large pose variation since the strong 3D geometry prior of human faces is ignored. Moreover, learning to swap faces in 2D images makes little use of the shaped details from sources, leading to poorer performance on identity transferring.

Motivated by the recent advances of 3D generative models [12, 13, 20, 25, 45] in synthesizing multi-view consistent images and high-quality 3D shapes, it naturally raises a question: can we perform face swapping in a 3D-aware manner to exploit the strong geometry and texture priors? To answer this question, two challenges arise. *First*, how to infer 3D prior directly from 3D-GAN models still remains open. Current 3D-aware generative models synthesize their results from a random Gaussian noise z, so that their output images are not controllable. This increases the complexity of inferring the required prior from arbitrary input. *Second*, the inferred prior corresponding to input images is in the form of a high-dimension feature vector in the latent space of 3D GANs. Simply synthesizing multi-view target images referring to the prior and applying 2D face swapping to them produces not only inconsistent artifacts but also a heavy computational load.

To address these challenges, we systematically investigate the geometry and texture prior of these 3D generative models and propose a novel 3D-aware face swapping framework 3dSwap. We introduce a 3D GAN inversion framework to project the 2D inputs into the 3D latent space, motivated by recent GAN inversion approaches [46, 47, 51]. Specifically, we design a learning-based inversion algorithm that trains an encoding network to efficiently and robustly project input images into the latent space of EG3D [12]. However, directly borrowing the architecture from 2D approaches is not yet enough since a single-view input provides limited information about the whole human face. To further improve the multi-view consistency of latent code projection, we design a pseudo-multi-view training strategy. This design effectively bridges the domain gap between 2D and 3D. To tackle the second problem, we design a face swapping algorithm based on the 3D latent codes and directly synthesize the swapped faces with the 3D-aware generator. In this way, we achieve 3D GAN-inversion-based face swapping by a latent code manipulating algorithm consisting of style-mixing and interpolation, where latent code interpolation is responsible for identity transferring while style-mixing helps to preserve attributes.

In summary, our contributions are threefold:

- To the best of our knowledge, we first address the 3D-aware face swapping task. The proposed 3dSwap method sets a strong baseline and we hope this work will foster future research into this task.
- We design a learning-based 3D GAN inversion with the pseudo-multi-view training strategy to extract geometry and texture prior from arbitrary input images. We further utilize these strong prior by designing a latent code manipulating algorithm, with which we directly synthesize the final results with the pretrained generator.
- Extensive experiments on benchmark datasets demonstrate the superiority of the proposed 3dSwap over state-of-the-art 2D face swapping approaches in identity transferring. Our reconstruction module for 3D-GAN inversion performs favorably over the state-of-the-art methods as well.

2. Related Work

Face Swapping. Face swapping has emerged as a popular research topic in the field of computer vision in recent years. Currently, it can be classified into two categories: 3D-based and GAN-based methods. Specifically, 3D-based methods [7, 15, 34, 43] fit input images into 3D parametric face models (*i.e*. 3DMM [8]) to overcome the problems of posture or perspective difference between input images. However, the performance of such methods is usually limited by the reconstruction results. GAN-based methods [6, 18, 23, 31, 32, 39, 40, 42] adopt the adversarial training strategy to generate photorealistic fake faces.

Early GAN-based face swapping methods are subject-specific, *i.e*. DeepFake [18] and Korshunova *et al*. [31] are required to train different models for different inputs. The subject-specific approaches have limited real applications since face swapping is required to be applicable to any unseen pair of input images, and such limitation is addressed in latter subject-agnostic face swapping approaches [6, 23, 32, 39, 40, 42]. To increase the resolution of generated images, MegaFS [60] firstly proposes a GAN-inversion-based face swapping method, utilizing StyleGAN [28] to synthesize megapixel-level swapping faces. Xu *et al*. [56] and StyleSwap [57] integrate the StyleGAN2 [29] generator to their face swapping pipeline, applying its strong prior to generate high-resolution swapped faces. Following these approaches, we furtherly extend the face swapping task into 3D latent space to capture fine-grained details of face shape and strengthen the robustness under large pose variance.

3D-Aware Generative Models. The 3D-aware generative models are aimed to synthesize 3D-aware (*i.e*., can be explicitly controlled by the camera pose) images from 2D image collections. HoloGAN [41] firstly proposes a 3D-aware generative model through learning the voxel features, whereas it only generates low-resolution results due to the limitation of computational cost. Recently, several works utilize the NeRF [36] representation [12, 20, 25, 45, 50]. GRAF [50] adopts the approach of patch sampling to elim-

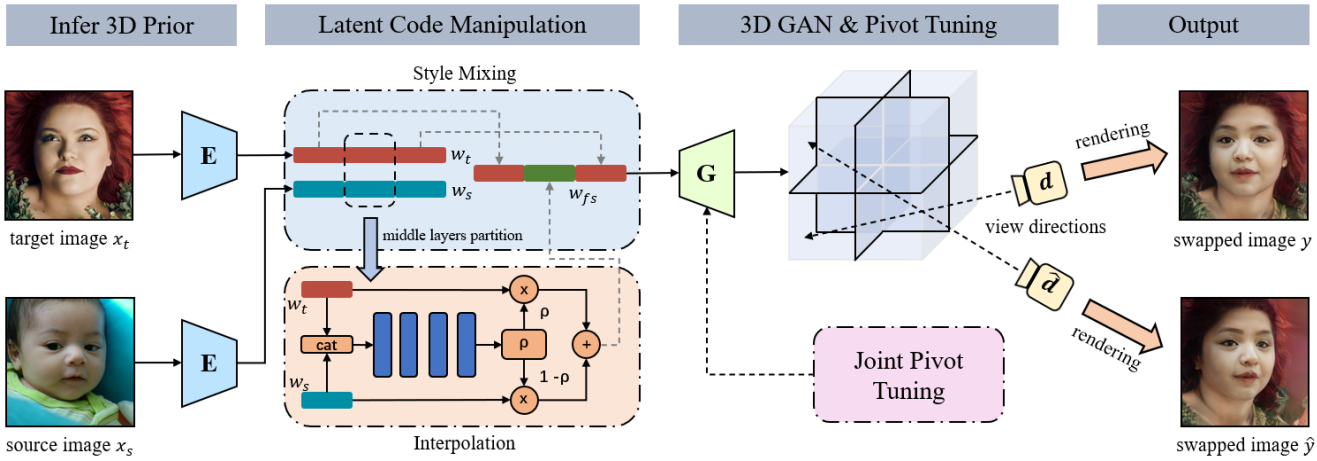

Figure 2. The pipeline of our 3D-aware face swapping method, 3dSwap. In the first stage, we infer 3D geometry and texture prior of both source and target images with an encoder. We then design a latent code manipulation algorithm consisting of style mixing and interpolation to conduct face swapping based on these priors. Finally, swapped faces in any view direction can be synthesized by 3dSwap after fine-tuning the parameters of the generator following the joint pivot tuning optimization.

inate computational costs during training. GRAM [20] estimates radiance manifolds to produce realistic images with fine details and strong 3D consistency. StyleNeRF [25] integrates NeRF with style-based generators and proposes a better up-sampler and a new regularization loss to mitigate inconsistencies. StyleSDF [45] presents a Signed Distance Field (SDF) based on 3D modeling that defines detailed 3D surfaces. EG3D [12] raises a novel tri-plane representation for efficient 3D-aware image generation. Due to the strong generative capability of these 3D-aware generative models, we leverage them to infer fine 3D prior from 2D images for our 3D-aware face swapping framework.

GAN Inversion. Since Generative Adversarial Network [24], numerous generative models reflect great abilities in synthesizing high-quality images [9,12,25,28,29,45]. To fully leverage these well-trained GANs, the task of GAN inversion emerges recently. In particular, GAN inversion is aimed to project a given image back to a vector w in the latent space of a pretrained GAN model so that this image can be faithfully reconstructed from w by the generator.

Early works invert images into Gaussian noise $z \in R^{1\times512}$ or semantic latent space $\mathcal{W} \in R^{1\times512}$ [1,16,17,59]. Abdal *et al*. [2] firstly extend latent space to $\mathcal{W}+ \in R^{18\times512}$ for more accurate reconstruction. To predict the latent code, learning-based methods [3, 26, 46, 51, 52] train an encoder for latent projection, while optimization-based methods [1,2,16,17] directly find the optimal code step-by-step from noise. Hybrid methods [4, 47, 59] combine both to optimize latent codes initialized by encoders.

In addition, there are a few inversion works for 3D generative models. Pix2NeRF [10] is proposed to generate Neural Radiance Fields (NeRF) [36] of an object applying a single input image based on a pretrained π-GAN [13]. Connor *et al*. [33] leverage EG3D [12] and a pretrained 3DMM predictor [22] to reconstruct a 3D human face, which could be further animated or edited. Our reconstruction model is also in this catalog, while the adopted learning-based algorithm is more robust and efficient compared with them.

3. Method

3.1. Overview

Given single-view source and target images, we aim to synthesize multi-view-consistent face images with identity from source image x_s and other attributes from target image x_t. Fig. 2 demonstrates the overall pipeline and notations of the proposed 3dSwap. First, to extract accurate geometry and texture prior from 2D images, we conduct a learning-based 3D GAN inversion, training an encoding network to project the inputs into the latent space of a 3D-aware generative model. Specifically, we design a pseudo-multi-view optimization strategy to train the encoder with a feature pyramid architecture from pSp [46], empowering the latent code projection with the 3D consistency of the state-of-the-art 3D GAN, *i.e*. EG3D [12] (Sec. 3.2). Then, to disentangle identity from attributes in the latent space, we design a latent code manipulation algorithm consisting of style mixing and interpolation (Sec. 3.3). Finally, for the purpose of improving the overall quality of our results, bridging the gap between 2D image generating and 3D rendering, we implement a joint pivot tuning on parameters of the pretrained EG3D generator (Sec. 3.4). The networks are trained with a set of well-designed loss functions to enforce identity transferring and attribute preserving (Sec. 3.5).

3.2. Inferring 3D Prior from 2D Images

To infer geometry and texture prior from a 2D image, we leverage the state-of-the-art 3D-aware generative model, *i.e*. EG3D [12] by projecting the inputs into its latent space. Since the optimization-based algorithm [47] is inefficient and less robust to non-front faces, we propose a learning-based inversion algorithm where an encoding network is trained to project the single-view inputs into the 3D latent space. Different from 2D StyleGAN-like models which totally rely on the latent code w to generate the corresponding output: $y = \mathcal{G}(w)$, the 3D-aware generative model has an extra input d which controls the pose of synthesized image: $y = \mathcal{G}(w, d)$. This indicates that latent codes and generated images are not bijections for 3D GANs since multi-view images of the same person can be synthesized using the same w but different d. Taking this property into account, we design a pseudo-multi-view training strategy, using a generated image in a different view from the source image to improve the consistency of latent code projection. Fig. 3 illustrates the pipeline of our design.

Specifically, we first use an encoder to project the input image x into the latent space $\mathcal{W}$ and get a high-dimension intermediate latent vector $w_x = \mathcal{E}_\theta(x)$, where $\mathcal{E}_\theta(\cdot)$ is the pSp encoder with parameters θ. Then, with the pretrained EG3D generator $\mathcal{G}(\cdot,\cdot)$ and input direction d estimated by Deep3d Face Reconstruction [21], we synthesize the reconstructed result $x' = \mathcal{G}(w_x, d)$. For a 2D GAN inversion approach, this ground-truth and reconstructed image pair (x, x') is enough, but it is inadequate for 3D GANs due to the non-bijective property.

Ideally, this issue can be addressed by feeding multi-view images of a person into the encoder and minimizing the distance between their output vectors. However, it is difficult to obtain large-scale multi-view data, and we usually only have single-view images of a person in the training dataset. To this end, we additionally sample a random direction $\hat{d}$ and use the generator to synthesize $\hat{x} = \mathcal{G}(w_x, \hat{d})$ with the same latent code. This output image $\hat{x}$, which is called a pseudo-input since it is generated by the 3D GAN, is again fed into the encoder-decoder structure to get $w_{\hat{x}} = \mathcal{E}_\theta(\hat{x})$ and $\hat{x}' = \mathcal{G}(w_{\hat{x}}, d)$.

Now, we can define our optimization objectives. Following the usual inversion approaches, we apply some pixel-wise loss functions between the input x and its reconstruction x'. Under the setting of our pseudo-multi-view input, we add constraints between the two latent codes w_x and $w_{\hat{x}}$ for the purpose of maintaining 3D consistency. We further restrain pixel-level distance between the second-order output $\hat{x}'$ synthesized with $w_{\hat{x}}$ and the origin input x to reinforce such constraint. In summary, this three-termed optimization can be written as:

$$\min_{\theta}\{\mathcal{L}(x, x') + \eta\mathcal{L}(x, \hat{x}') + \mathcal{L}(w_x, w_{\hat{x}})\}, \qquad (1)$$

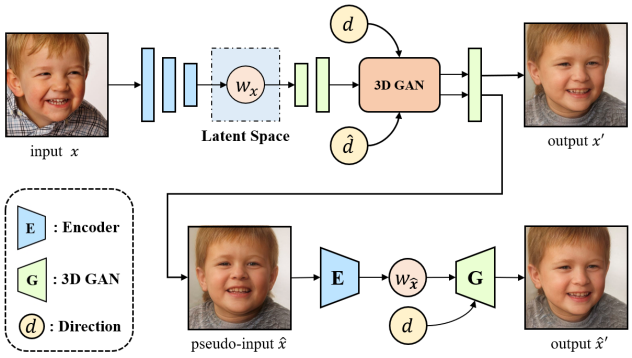

Figure 3. The pipeline of our pseudo-multi-view training strategy.

where θ is the parameter of encoder, η is a trade-off parameter and $\mathcal{L}(\cdot,\cdot)$ denotes the loss functions which will be further discussed in Sec. 3.5. After optimizing the parameters of the encoding network with this strategy, we can obtain rather accurate 3D prior w_x from any given input x.

3.3. Face Swapping via Latent Code Manipulation

To take full advantage of the prior extracted from the 3D GAN model, we calculate the latent code for the swapped face based on latent codes $w_s = \mathcal{E}_\theta(x_s)$ of the source image x_s and $w_t = \mathcal{E}_\theta(x_t)$ of the target image x_t. Before that, we step back and think about what these latent codes represent.

A face image usually contains different attributes such as face shape, hairstyle, skin color, *etc*. With the encoder discussed in Sec. 3.2, we embed all these attributes in the high-dimension latent vectors. However, identity features depending on the geometry of facial region (*i.e*., eyes, nose, mouth, cheek, and so on) also implicitly lie in such latent codes. For the task of face swapping, it is desirable if identity features can be disentangled from attribute features in the latent code. Afterward, we can simply exchange the identity part of the latent codes to achieve face swapping.

Since such identity and attributes are typically entangled in the latent codes, we design an interpolation strategy between the source and target latent codes with learnable coefficients. Here, the source latent code w_s plays a leading role in the identity part while w_t dominates the others. To obtain these coefficients, we concatenate w_s and w_t to form a 1×1024 vector and feed it into a four-layer Multilayer Perceptron whose output ρ is the interpolation coefficient.

Moreover, StyleGAN-like [28,29] models share the style mixing property of latent codes, which means that different layers of latent codes control different parts of attributes. For example, coarse spatial resolutions control high-level aspects like face shape and orientation while fine resolution latent control details like hair color. Motivated by this, we also investigate the layer-wise attributes in EG3D and observed similar properties. This allows us to generate more desirable swapping results by only performing interpolation

on part of the latent codes.

In summary, the latent code of swapped face w_{fs} can be obtained by:

$$w_{fs}^{(i)} = \begin{cases} \rho^{(i)} \times w_t^{(i)} + (1-\rho^{(i)}) \times w_s^{(i)} & i \in [5,9], \\ w_t^{(i)} & otherwise, \end{cases} \tag{2}$$

where the superscript i denotes the layer-wise expression of w_{fs} and the choice of layer, from layer 5 to layer 9, follows the definition of "middle" from StyleGAN [28], while a slight modification is made since the dimension of EG3D latent space is lower (*i.e.* $\mathcal{W} \in R^{14\times512}$). To better disentangle identity and attributes, we apply a Sigmoid-shaped activation function with a factor $\lambda = 100$ to the ρ generated by MLPs, enforcing the coefficients to be closer to 0 or 1:

$$\rho_{new}^{(i)} = (1 + e^{-\lambda \rho_{old}^{(i)}})^{-1}. \tag{3}$$

3.4. Joint Pivot Tuning

With the encoding network trained by the well-designed optimization strategy in Sec. 3.2, we can project an input image into a code in the 3D latent space. However, the inevitable reconstruction error will degrade the performance of face swapping, which is a downstream task of 3D GAN inversion. Also, we observe that directly swap faces via latent manipulation leads to slight artifacts in the non-facial region. Motivated by PTI [47], we adopt pivot tuning on the parameters of the pretrained EG3D generator using a fixed latent code w_{fs} from Sec. 3.3, but in an optimizing direction considering both reconstruction quality and face swapping performance. The process of this "joint" pivot tuning is:

$$\begin{aligned} \min_{\theta^*} \{ & \mathcal{L}(x_{s/t}\ ,\ \mathcal{G}_{\theta^*}(w_{s/t}, d_{s/t})) + \\ & \mathcal{L}(x_t \cdot M_f\ ,\ \mathcal{G}_{\theta^*}(w_{fs}, d_t) \cdot M_f)\}, \end{aligned} \tag{4}$$

where θ^* is the parameter of EG3D generator, d_s is the direction of the source image, M_f is a binary mask that shields facial region and $\mathcal{L}(\cdot,\cdot)$ is the optimization constraint including MSE, LPIPS [58] and ID [19] losses.

Finally, with this finetuned generator and the latent code calculated by Eq. 2, we can synthesize the swapped face y in any direction d by:

$$y = \mathcal{G}_{\theta^*}(w_{fs}, d). \tag{5}$$

3.5. Objective Functions

GAN Inversion Losses. In Eq. 1, we generally use $\mathcal{L}(\cdot,\cdot)$ to denote the loss function of our pseudo-multi-view training strategy. Here, we give its detailed form. Following the previous work [46], we use three different objectives for supervising a pair of input image x and reconstruction x' (and the same for $\hat{x}'$), including pixel-wise $\mathcal{L}_1$ loss, Learned Perceptual Image Path Similarity [58] loss $\mathcal{L}_{LPIPS}$, and identity similarity loss $\mathcal{L}_{id}$ maximizing the cosine similarity between two identity embeddings estimated by ArcFace [19]. The total reconstruction loss between x and x' is:

$$\begin{aligned} \mathcal{L}_{rec}(x,x') = \lambda_1 \mathcal{L}_1(x,x')\ +\ & \lambda_2 \mathcal{L}_{LPISP}(x,x') \\ & + \lambda_3 \mathcal{L}_{id}(x,x'), \end{aligned} \tag{6}$$

where λ_1, λ_2 and λ_3 are loss weights.

For the constraint between two latent codes, we adopt a cosine similarity:

$$\mathcal{L}_{lat}(w_x, w_{\hat{x}}) = 1 - cos(w_x, w_{\hat{x}}). \tag{7}$$

Besides, we adopt the latent code regularization loss from pSp [46], which constrains the generated latent vector in a region to be close to the average latent vector:

$$\mathcal{L}_{reg}(x) = \|\mathcal{E}_\theta(x) - \bar{x}\|_2, \tag{8}$$

where $\bar{x}$ is the average of 10000 randomly sampled latent codes of EG3D generator. The overall loss function for 3D GAN inversion is:

$$\begin{aligned} \mathcal{L}_{inv} = \mathcal{L}_{rec}(x,x') + \eta \mathcal{L}_{rec}(x,\hat{x}') + \mathcal{L}_{lat}(w_x, w_{\hat{x}}) \\ + \mathcal{L}_{reg}(x). \end{aligned} \tag{9}$$

Face Swapping Losses. For training our face swapping module, we first design a masked pixel-wise $\mathcal{L}_2$ loss for the face irrelevant region:

$$\mathcal{L}_2(x_t, y) = \|x_t \cdot M_f - y \cdot M_f\|_2, \tag{10}$$

where M_f is the binary mask same as in Sec. 3.4. We generate this mask according to the face segmentation labels of FFHQ [28] datasets. For 3D GAN inversion, we adopt the LPIPS [58] loss $\mathcal{L}_{LPIPS}(x_t, y)$ to learn the perceptual similarities and increase the quality of the generated images, and the binary mask is also added before feeding the image into the perceptual feature extractor.

For 3D-aware face swapping, we additionally synthesize the swapped face $\hat{y}$ in the view of the source image, calculating both $\mathcal{L}_{id}(x_s, y)$ and $\mathcal{L}_{id}(x_s, \hat{y})$ for better identity transferring.

Besides, $\mathcal{L}_{color}$ is designed to maintain the skin color of swapped faces:

$$\mathcal{L}_{color}(x_s, y) = \|\bar{\mathcal{C}}(x_s \cdot (1 - M_f)) - \bar{\mathcal{C}}(y \cdot (1 - M_f))\|_2, \tag{11}$$

where $\bar{\mathcal{C}}(\cdot)$ denotes an average RGB value of the masked region.

The overall loss function for training the face swapping module is:

$$\begin{aligned} \mathcal{L}_{fs} = \mathcal{L}_2(x_t, y) + \mathcal{L}_{LPIPS}(x_t, y) + \mathcal{L}_{id}(x_s, y) \\ + \mathcal{L}_{id}(x_s, \hat{y}) + \mathcal{L}_{color}(x_s, y). \end{aligned} \tag{12}$$

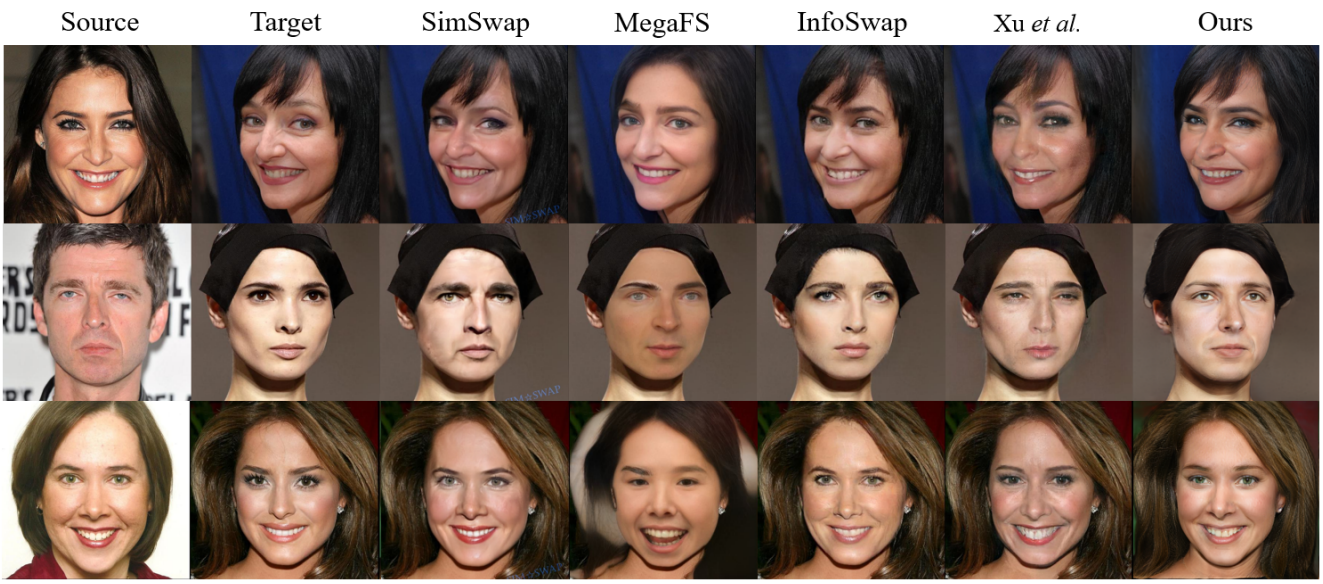

Figure 4. Qualitative comparison of face swapping on CelebA-HQ dataset. Compared with all these 2D approaches, our method extracts facial shapes more accurately and transfers identity better. Moreover, since we conduct face swapping in latent space and a well-trained 3D GAN directly synthesizes the results, there are no obvious artifacts in the facial region.

4. Experiments

In this section, we first compare the proposed 3dSwap with some state-of-the-art 2D-images-based face swapping approaches. Furthermore, face swapping in a 3D-aware manner and extra evaluation metrics designed for 3D face swapping are analyzed. We finally carry out ablation studies to evaluate the effectiveness of our major design.

4.1. Implementation Details

In all experiments, Ranger optimizer [54] is applied to train our networks with a learning rate of 1×10^{-4}. Hyper-parameters are set as $\lambda_1 = \lambda_3 = 1, \lambda_2 = 0.8$ in Eq. 6 and $\eta = 0.25$ in Eq. 9. For training time, the inversion module is trained for 1,000,000 steps on 4 NVIDIA RTX3090 GPUs for about 3 days while the face swapping module is trained for 500,000 steps also on 4 GPUs for about 2 days. The pivot tuning optimization during inference time takes about 8 minutes on a single GPU.

4.2. Datasets

We conduct experiments on two datasets: 1) The FFHQ [28] dataset contains 70,000 high-quality images of human faces crawled from Flicker with considerable variation in age, ethnicity, and background. All images of this dataset are in a resolution of 1024×1024. 2). The CelebA-HQ [27] dataset is the high-quality version of the large-scale face attributes dataset CelebA [35] which contains 30,000 images in 1024×1024. Specifically, we train our model on FFHQ, while comparison experiments are executed on CelebA-HQ. We follow the data preprocessing way of EG3D to crop images according to facial landmarks and resize them into a resolution of 512×512. Due to the relatively expensive inference cost of 3dSwap mentioned in Sec. 4.1, we operate the following comparison experiments on 1000 source-target image pairs.

4.3. Comparison with 2D Face Swapping Methods

In this section, we compare the proposed 3dSwap with four 2D swapping methods: SimSwap [14], MegaFS [60], Infoswap [23] and Xu *et al.* [56]. These four methods are representative GAN-based [14, 23] and GAN-inversion-based [56, 60] approaches in recent years with state-of-the-art performance. Moreover, their official source codes are publicly available for us to make fair comparisons.

Qualitative Comparison. The qualitative comparison results are shown in Fig. 4. Compared with all these 2D face swapping approaches, our methods transfer more accurate geometry features (*i.e.*, facial contour) and detailed texture features like eye color to targets, reflecting better identity-transferring performance. Also, since we directly synthesize our final results with a well-trained generator with a properly calculated latent code, the swapped face we generate is more realistic without obvious artifacts in the facial region. More qualitative results on CelebA-HQ are provided in the supplementary material.

Quantitative Comparison. We adopt several evaluation metrics in our quantitative experiments to show the effectiveness of our model in Table 1. Following MegaFS [60],

Method	ID ↑	Pose ↓	Exp. ↓
SimSwap [14]	0.57	**1.49**	**10.48**
MegaFS [60]	0.48	3.95	14.08
InfoSwap [23]	<u>0.61</u>	2.50	<u>10.63</u>
Xu *et al*. [56]	0.54	2.66	12.94
Ours	**0.72**	<u>1.68</u>	13.76

Table 1. **Quantitative Results.** We compare our model with four competing methods in ID Similarity for identity transferring and Pose & Expression Error for attribute preserving.

we measure the ID similarity by calculating the cosine similarity between face embeddings of the source and swapped faces that are estimated by a pretrained face recognition network [19]. Meanwhile, pose error computes the $\mathcal{L}_2$ distance between the estimated Euler Angle [49] of the target and swapped images. For expression error, we calculate an average distance among estimated facial landmarks [5].

For cosine similarity of identity, which is a crucial indicator for face swapping since it evaluates the quality of identity transferring, we significantly outperform all these 2D approaches. Such results and the visual effects in Fig. 4 together show that our method transfers identity better due to the application of 3D prior. For attribute preserving, our method which can be explicitly controlled by a camera pose performs rather well in pose error since it is only slightly weaker than SimSwap [14] but it reflects a poorer performance compared with 2D approaches in expression error. However, we can still claim that the proposed 3dSwap is superior to 2D methods in identity transferring and performs close to them in attribute preserving after considering all three quantitative comparison results.

4.4. Further Analysis on 3D-Aware Face Swapping

As the first 3D-aware face swapping method, the proposed 3dSwap is specialized in synthesizing multi-view-consistent results. In this section, we conduct more experiments in this track, showing some visualized comparisons on 3D consistency and raising brand-new criteria for 3D-aware face swapping.

Visualization on Multi-View Images. To compare with 2D face swapping approaches in fairness, we first synthesize multi-view target images by using our reconstruction module and then apply SimSwap [14] and InfoSwap [23] to them. The visualized results are shown in Fig. 5, where results under different views are not as consistent as ours (*i.e*. shape of nose, mouth, and eyebrows changes) for the 2D face swapping method. More artifacts can be discovered when the target images are sideward. Please refer to the video in the supplementary material for more intuitional comparisons.

Figure 5. Visualized comparison on Multi-view results among Infoswap [23], Simswap [14] and Ours.

Criteria for 3D-Aware Face Swapping. In Sec. 4.3, the performance of identity transferring is evaluated based on the face embedding estimated by pretrained face recognition networks [19]. However, such networks are not enough robust to pose variance so it could be an unfair criterion for face swapping. For 3D-aware face swapping, we can simply synthesize a swapped face in the view of the source image. In this way, the "Aligned Identity Similarity" can be a reasonable standard to evaluate 3D-aware face swapping models. Moreover, inspired by human's ability to recognize a familiarized person from any direction, we synthesize the swapped face into 9 different fixed poses and calculate an average identity similarity together with images in source and target views. We report our results of these two evaluation metrics in Table 2 and images under these fixed poses are shown in the supplementary material.

Metric	Aligned ID Sim.↑	Average ID Sim. ↑
Ours	0.85	0.42

Table 2. **Quantitative Results of New Metrics.** We test the proposed 3dSwap under the two new evaluation metrics.

4.5. Ablation Studies

In this section, we conduct ablation experiments on the CelebA-HQ dataset to evaluate the effectiveness of the major design of the proposed 3dSwap.

Effectiveness of 3D GAN Inversion. Since previous works [12, 33] do not release the code of their 3D GAN-inversion part, we follow the paper of EG3D to reproduce a pivot tuning inversion [47] to the generator with the same hyperparameters. In this section, we mainly compare our design with the optimization-based latent code projection of PTI on EG3D to show the effectiveness of the learning-based inversion algorithm we use. For the sake of fairness,

Figure 6. Qualitative Comparison on 3D GAN inversion. Comparing to the directly application of pivot tuning inversion, our design reconstruct details (*i.e*. shape and color of eyes, glasses *etc*.) better.

Figure 7. Visualization of face swapping results with and without style mixing.

both models are tested on the same 2000 images in CelebA-HQ and adopt a parameter tuning of the pretrained generator for 500 steps.

We show the qualitative comparison results in Fig. 6. Our design performs better in details reconstruction (*i.e*., eye shape, glasses, *etc*.) despite the optimization-based approach still recovers accurate face shape, hair color, *etc*.

For 3D GAN Inversion, we adopt the same metrics as 2D GAN inversion: $\mathcal{L}_2$ distance (or MSE loss) to calculate the pixel-wise similarity, LPIPS [58] distance to evaluate the perceptual similarity and MS-SSIM [53] to show the structural similarity. Additionally, we calculate ID similarity to ensure the accuracy of the reconstruction, and the results are reported in Table 3. Our design outperforms the optimization-based approaches in all of the four criteria.

Method	MSE ↓	LPIPS ↓	SSIM ↑	ID Sim.↑
EG3D with Opt.	0.0896	0.2761	0.6197	0.7318
Ours	**0.0168**	**0.1049**	**0.7348**	**0.8616**

Table 3. **Quantitative Results on 3D GAN inversion.** We compare our 3D GAN inversion module with an optimization-based inversion on EG3D under four common evaluation metrics in the 2D GAN inversion task.

Effectiveness of Style Mixing. As mentioned in Sec. 3.3, we adopt style mixing and latent code interpolation for face swapping. Here, we briefly show the effectiveness of style mixing. A comparison of our model with and without style mixing can be seen in Fig. 7. Identity can be ideally transferred between sources and targets under both settings, however, attributes including skin color, background, *etc*. would be prominently affected if we interpolate in all layers of latent codes as shown in the third column.

5. Conclusion

We propose a novel 3D-aware face swapping method 3dSwap that generates high-fidelity and multi-view-consistent swapped faces. To leverage both geometry and texture prior of the 3D human face, we project the input images into the latent space of the 3D-aware generative model by introducing a learning-based inversion. A latent code manipulation algorithm, consisting of style mixing and latent code interpolation, is then designed to achieve 3D GAN-inversion-based face swapping. We further bridge the image quality between 2D generating and 3D rendering by applying a joint pivot tuning. To the best of our knowledge, 3dSwap is the first 3D-aware face swapping method, thus it sets a strong baseline for future research on 3D forgery detection and face swapping.

Limitations. Since we need to project input images into the latent space of a 3D GAN which contains far more information than that of 2D GANs, we tune the parameters of the pretrained generator during testing, leading to a rather long inference time. Moreover, since the final results are rendered by a 3D generator, our method fails to accurately reconstruct clothing, backgrounds, *etc* in the image limited by the current development of 3D-aware generative models.

Broader Impacts. Although not the purpose of this work, photorealistic swapped faces may potentially be abused. On the other hand, our model can be used to generate high-quality and multi-viewed examples to facilitate face forgery detection [11].

Acknowledgements. This work was supported by NSFC (62201342), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and the Fundamental Research Funds for the Central Universities.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *ICCV*, pages 4431–4440, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *CVPR*, pages 8293–8302, 2020.

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *ICCV*, pages 6691–6700, 2021.

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *CVPR*, pages 18511–18521, 2022.

[5] Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016.

[6] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *CVPR*, pages 6713–6722, 2018.

[7] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. *Comput. Graph. Forum*, 23(3):669–676, 2004.

[8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, pages 187–194, 1999.

[9] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[10] Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. Pix2nerf: Unsupervised conditional π-gan for single image to neural radiance fields translation. In *CVPR*, pages 3971–3980, 2022.

[11] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pages 4103–4112, 2022.

[12] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, pages 16123–16133, 2022.

[13] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021.

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

[15] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. 3d-model-based face replacement in video. In *SIGGRAPH*, 2009.

[16] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of gans. In *CVPR*, pages 5770–5779, 2020.

[17] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE Trans. Neural Networks Learn. Syst.*, 30(7):1967–1974, 2019.

[18] DeepFakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes. Accessed: 2022-10-18.

[19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019.

[20] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: generative radiance manifolds for 3d-aware image generation. In *CVPR*, pages 10663–10673, 2022.

[21] 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*, pages 285–295, 2019.

[22] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Trans. Graph.*, 40(4):88:1–88:13, 2021.

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

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020.

[25] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *ICLR*, 2022.

[26] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *CoRR*, abs/2007.01758, 2020.

[27] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

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

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8107–8116, 2020.

[30] Ira Kemelmacher-Shlizerman. Transfiguring portraits. *ACM Trans. Graph.*, 35(4):94:1–94:8, 2016.

[31] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *ICCV*, pages 3697–3705, 2017.

[32] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *CoRR*, abs/1912.13457, 2019.

[33] Connor Z. Lin, David B. Lindell, Eric R. Chan, and Gordon Wetzstein. 3d GAN inversion for controllable portrait image animation. *CoRR*, abs/2203.13441, 2022.

[34] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3d model based approach. In *ICME*, pages 333–338, 2012.

[35] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020.

[37] Saleh Mosaddegh, Loïc Simon, and Frédéric Jurie. Photorealistic face de-identification by aggregating donors' face components. In *ACCV*, pages 159–174, 2014.

[38] Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M. Weber. High-resolution neural face swapping for visual effects. *Comput. Graph. Forum*, 39(4):173–184, 2020.

[39] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *ACCV*, pages 117–132, 2018.

[40] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. RSGAN: face swapping and editing using face and hair representation in latent spaces. In *SIGGRAPH*, pages 69:1–69:2, 2018.

[41] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *ICCV*, pages 7587–7596, 2019.

[42] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: subject agnostic face swapping and reenactment. In *ICCV*, pages 7183–7192, 2019.

[43] Yuval Nirkin, Iacopo Masi, Anh Tuan Tran, Tal Hassner, and Gérard G. Medioni. On face segmentation, face swapping, and face perception. In *AFGR*, 2018.

[44] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *ACM Trans. Graph.*, 39(6):225:1–225:14, 2020.

[45] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *CVPR*, pages 13493–13503, 2022.

[46] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *CVPR*, pages 2287–2296, 2021.

[47] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *TOG*, pages 1–13, 2022.

[48] Arun Ross and Asem A. Othman. Visual cryptography for biometric privacy. *IEEE Trans. Inf. Forensics Secur.*, 6(1):70–81, 2011.

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

[50] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020.

[51] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.*, 40(4):133:1–133:14, 2021.

[52] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *CVPR*, pages 11369–11378, 2022.

[53] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *ACSSC*, 2003.

[54] Less Wright. Ranger - a synergistic optimizer. https://github.com/lessw2020/Ranger-Deep-Learning-Optimizer. Accessed: 2022-9-18.

[55] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *CVPR*, pages 7632–7641, 2022.

[56] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *CVPR*, pages 7632–7641, 2022.

[57] 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, 2022.

[58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[59] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, pages 597–613, 2016.

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