MetaSwap: Towards Elegant One-Shot Face Swapping Framework

Project of AI2612 Machine Learning Project, 2023 Fall, SJTU

Xiangyuan Xue 521030910387

Shengmin Yang 521030910430

Yi Ai 521030910429

Figure 1. Examples of our face swapping results. The first row consists of the source images, and the second row consists of the target images. The third row shows the results of our face swapping framework, which preserves the identity of the source faces and the attribute of the target faces. Our framework shows strong generalization ability and high fidelity in multiple scenarios.

Abstract

In this project, we investigate recent advances in one-shot face swapping frameworks. First, we describe the development of face swapping technology and present the related works. Then we propose a novel one-shot face swapping framework based on FaceShifter and illustrate the detailed methods. We implement all the codes using the latest deep learning framework and conduct qualitative and quantitative experiments. In addition, we use SimSwap as the baseline and compare the qualitative results to demonstrate the effectiveness of our framework. Finally, we present our discussion about the limitations of these one-shot face swapping frameworks and the possible future directions. We believe that face swapping is a promising research topic and will be widely applied in the future. Our code is available at `https://github.com/xxyQwQ/metaswap`.

1. Introduction

In the field of computer vision and artificial intelligence, face swapping stands as a captivating technology with the ability to replace a target face in an image or video with a source face while maintaining the congruity and expression dynamics of the original content. To be specific, it describes the process of replacing one face in an image, i.e., the source face, with another face from a different image, i.e., the target face, which preserves the identity information of the source face and the attribute information of the target face. This sophisticated process opens up myriad applications, revolutionizing sectors ranging from the entertainment industry, where it is used in live streaming and the film sector for enhancing production quality, to amusement-oriented applications that allow users to alter their images for engaging digital interactions.

Despite its varied applications, face swapping is not universally celebrated and has been the subject of intense debate. The ethical considerations, potential for criminal misuse, and legal ramifications of impersonation have ignited serious concerns about its unchecked propagation. The ease with which individuals' identities can be forged poses profound questions about privacy and the integrity of digital content in our increasingly interconnected world.

Addressing the technical front, there are two main paradigms in face swapping technologies: dedicated models and one-shot models. The dedicated models typically

require extensive data for each specific individual for training, which brings better fidelity but cannot generalize well to unseen faces. The one-shot models seek to universally apply face swapping irrespective of the facial variation presented. With a given source face, the one-shot model can replace the face in any target image, which is once and for all but much more challenging. Here we mainly focus on the context of the latter, i.e. the one-shot models, which can serve as a more general and practical solution, especially in the film and live streaming industry.

The challenges in developing a successful one-shot face swapping framework are manifold. After investiating the literature, we summarize them into four main aspects:

- Since no ground truth is available for the face swapping task, it is difficult to train the model in a supervised manner. Instead, unsupervised learning is required.
- The one-shot paradigm requires the model to generalize well to any source and target face, but out-of-distribution problems have always been a challenge in deep learning.
- The resulting images should be free from tell-tale signs such as stitching marks and visual artifacts, which requires a strong ability to synthesize realistic images.
- Real-world scenarios are much more complicated than expected. The model should be robust to various facial accessories, e.g. glasses and jewelry.

Faced with these challenges, many researchers have proposed various solutions, among which the GAN-based methods have achieved impressive results, including SimSwap [6], FSGAN [25], and FaceShifter [21]. Our novel framework draws inspiration from FaceShifter [21], which is an elegant two-stage approach and achieves state-of-the-art performance. In this project, our contributions in the realm of one-shot face swapping are substantial and diverse:

- We meticulously collect and clean a segment of training data, which is critical for the training of our model.
- Instead of directly copying the implementation from SimSwap [6], we design and implement an elegant one-shot face swapping framework inspired by FaceShifter [21], which is more versatile and flexible. Since the author has never released the code, we implement the framework by ourselves with the latest deep learning framework.
- We implement the face swapping framework as a whole process, boasting end-to-end training and inference capabilities that cater to both image and video content.
- We develop an exclusive dataset and an automatic evaluation module to assess the quality of the face swapping model. Our empirical results showcase a remarkable leap in one-shot face swapping performance.

2. Related Work

Early works [2, 32] about face swapping focus on directly replacing the pixel values in the inner regions of the face, which makes them highly susceptible to variations in pose and perspective. Such primitive methods are still widely used in video editing software, which demonstrates their simplicity and efficiency to some extent.

An improved way is to utilize 3D models [4, 22] to transform the face, which attempts to address these discrepancies by employing 3D modeling to account for differences in pose or perspective. However, it is difficult to reconstruct a 3D face model from a single image, and the time complexity of 3D modeling is extremely high. Therefore, the accuracy and robustness of 3D face reconstruction are still not satisfactory with the current proceeding.

More recently, deep learning methods have demonstrated impressive results in generating face images. Early in 2017, conditional variational autoencoder (CVAE) [1] is proposed to generate face images. Convolutional neural networks (CNNs) are also used in face swapping [20]. However, the generated images are blurry and lack high-frequency details. This problem is solved by the introduction of generative adversarial networks (GANs) [12], which is applied in multiple face swapping methods [23, 24] and largely improves the quality of the generated images.

Nowadays, researchers are exploring even more advanced methods for one-shot face swapping, which can to the largest extent preserve the identity of the source face while maintaining the realism of the target face. More powerful GANs [6, 21, 25] are proposed to achieve this goal, bringing the face swapping technology to a new level. FSGAN [25] introduces reactment and segmentation modules to extract the required information from the source face and the target face, and then uses inpainting and blending modules to generate the final image, thus splitting the face swapping process into several subtasks. FaceShifter [21] proposes a two-stage framework, which first uses a GAN-based network to generate a rough face swapping result, and then uses a refinement network to enhance the details of the result, thus significantly simplifying the face swapping process. SimSwap [6] further simplifies the face swapping process by introducing an end-to-end framework without the need for two-stage training.

With the advent of modern deep learning architectures, face swapping can be empowered with more advanced techniques. As transformers [31] become the mainstream in natural language processing (NLP), they are recently applied in face swapping [8], where the attention mechanism is fully exploited to capture the details for face swapping. As a core method in AI generated content (AIGC), diffusion models [16] can generate diverse and high-quality images. Previous works [19] also explores how to adapt them to face swapping, where the process of face swapping is naturally formulated as a diffusion process and a denoising process. Face swapping keeps evolving with the development of deep learning, pacing towards better performance

and higher efficiency. We believe that more and more advanced methods will be proposed by AIGC researchers in the future.

3. Method

Suppose the dataset $\mathcal{D}$ consists of face images from m different identities, and the j-th identity provides n_j different images. Let x_i^j denote the i-th face image from the j-th identity, where $i \in \{1, 2, \ldots, n_j\}$ and $j \in \{1, 2, \ldots, m\}$. Given a source face image $x_s \in \mathcal{D}$ and a target face image $x_t \in \mathcal{D}$, the face swapping process aims to generate a result face image x_r that preserves the identity of x_s and the attribute of x_t. To be exact, the generated result x_r should be the same face as x_s, but presents the same facial expression and pose as x_t. Let G denote the proposed generator. We expect it can generalize well to the unseen face images.

We utilize the adversarial learning paradigm to train the injective generator G. Specifically, we introduce a multi-scale discriminator D to distinguish the generated faces $x_r = G(x_s, x_t)$ from the real faces x_s, x_t. We expect that the discriminator D can encourage the generator G to learn how to generate plausible faces, which follows the common practice of GANs [17, 18, 33]. The detailed network architecture will be introduced in the following parts.

3.1. Injective Generator

We empirically follow the classic FaceShifter [21] framework to construct the generator G. Note that the generator in FaceShifter is composed of two networks, i.e. the AEI-Net and the HEAR-Net, but much engineering practice has shown that the HEAR-Net is not necessary for the face swapping task. Therefore, we mainly adopt the former network as our injective generator, which can be divided into three modules: an attribute encoder E_α, an identity encoder E_β, and a denormalization decoder D_γ. The attribute encoder E_α extracts the attribute information $\alpha_1, \alpha_2, \ldots, \alpha_8$ from the target face image x_t. The identity encoder E_β extracts the identity information β from the source face image x_s. Then the denormalization decoder D_γ combines the attribute and identity information to generate the result face image x_r. The overview of the proposed injective generator is shown in Figure 2.

Attribute Encoder. The attribute encoder E_α applies the U-Net [29] architecture, which includes 7 down-sampling blocks and 7 up-sampling blocks, extracting 8 feature maps $\alpha_1, \alpha_2, \ldots, \alpha_8$ with different scales from the target face image x_t, as shown in the blue part of Figure 2.

$$E_\alpha(x_t) = \{\alpha_1, \alpha_2, \ldots, \alpha_8\}. \tag{1}$$

Each block is composes of a 4×4 convolution layer with stride 2, a normalization layer, and an activation layer. There are skip connections between feature maps with the same scale. The attribute encoder E_α is trained in a self-supervised manner, which expects the attribute of the result face image $x_r = G(x_s, x_t)$ to be the same with that of the target face image x_t.

Identity Encoder. Facial recognition has long been a tough problem in computer vision. Luckily, the recent advances in face recognition have changed the situation. Based on the powerful pretrained face recognition models, we can easily extract reliable identity representations. In this project, we adopt the Arcface [9] model as the identity encoder E_β, which utilizes a ResNet-50 backbone [15] and contrastive learning methods to learn discriminative identity representations. The identity encoder E_β extracts a 512-dimensional feature vector β from the source face image x_s as the identity information, as shown in the orange part of Figure 2.

$$E_\beta(x_s) = \beta. \tag{2}$$

Similar to the attribute encoder E_α, the identity encoder E_β is trained in a self-supervised manner, which expects the identity of the result face image $x_r = G(x_s, x_t)$ to be the same with that of the source face image x_s.

Denormalization Decoder. Both the attribute information $\alpha_1, \alpha_2, \ldots, \alpha_8$ and the identity information β will be fed into the denormalization decoder D_γ, which is composed of 8 cascaded denormalization blocks $D_\gamma^1, D_\gamma^2, \ldots, D_\gamma^8$, as shown in the green part of Figure 2.

$$D_\gamma(\{\alpha_i\}_{i=1}^8, \beta) = x_r. \tag{3}$$

The output of the denormalization decoder D_γ is the result face image $x_r = G(x_s, x_t)$, which can be supervised by the target face image x_t when the source face image x_s and the target face image x_t are the same face image.

The detailed structure of the denormalization block is shown in Figure 3. The denormalization block is composed of two identical layers, each of which contains a denormalization layer, an activation layer and a 3×3 convolution layer. There is also a skip connection between the input and output of the denormalization block, which follows the classical residual learning strategy [15].

$$D_\gamma^i(\alpha_i, \beta, \gamma_{i-1}) = \gamma_i, \tag{4}$$

where $i \in \{1, 2, \ldots, 8\}$. Specifically, we state that $\gamma_0 = \beta$ and $\gamma_8 = x_r = G(x_s, x_t)$ for simplicity.

The detailed structure of the denormalization layer is also shown in Figure 3. The denormalization layer follows the design of Adaptive Instance Normalization (AdaIN) [17, 18], which is composed of a normalization step and a denormalization step. Inspired by FaceShifter [21], we also introduce an attention mechanism to fuse different feature maps softly. To be exact, the input feature map γ is first normalized in the channel dimension.

$$\bar{\gamma} = \frac{\gamma - \mu_\gamma}{\sigma_\gamma}, \tag{5}$$

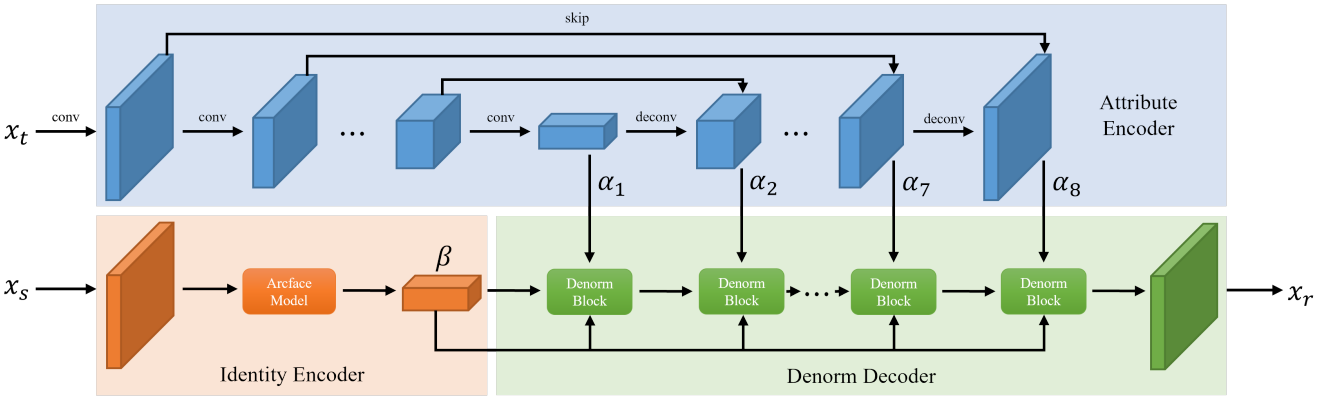

Figure 2. The overview of the proposed injective generator G. The generator G mainly contains three modules: an attribute encoder E_α, an identity encoder E_β, and a denormalization decoder D_γ. The attribute encoder E_α applies the U-Net architecture. The identity encoder E_β applies a pretrained Arcface model. The denormalization decoder D_γ is composed of 8 cascaded denormalization blocks.

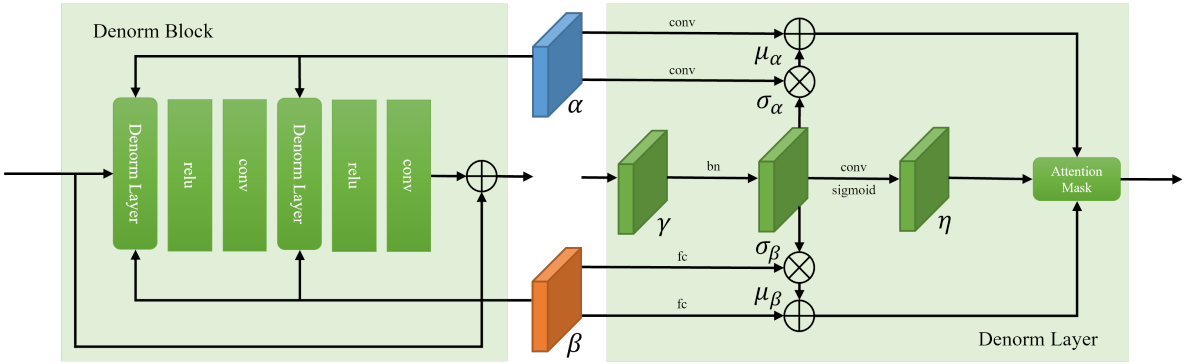

Figure 3. The detailed structure of the denormalization block and the denormalization layer. The denormalization layer is composed of a normalization step, a denormalization step and an attention mechanism. The denormalization block is composed of a skip connection and two identical layers, each of which contains a denormalization layer, an activation layer and a convolution layer.

where μ_γ and σ_γ are the mean and standard deviation of the input feature map γ respectively.

Then the attribute feature map α will be fed into a 1×1 convolution layer to predict the mean μ_α and standard deviation σ_α of the normalized feature map $\bar{\gamma}$ for warping.

$$\hat{\alpha} = \sigma_\alpha \times \bar{\gamma} + \mu_\alpha, \tag{6}$$

where $\hat{\alpha}$ is the predicted attribute feature map.

The identity feature vector β will be fed into a linear layer to predict the mean μ_β and standard deviation σ_β of the normalized feature map $\bar{\gamma}$ for warping.

$$\hat{\beta} = \sigma_\beta \times \bar{\gamma} + \mu_\beta, \tag{7}$$

where $\hat{\beta}$ is the predicted identity feature map.

In addition, the normalized feature map $\bar{\gamma}$ will be fed into a 1×1 convolution layer and a sigmoid activation layer to form the attention map η, which is used as a weighted mask to fuse the feature maps $\hat{\alpha}$ and $\hat{\beta}$.

$$\hat{\gamma} = (1 - \eta) \times \hat{\alpha} + \eta \times \hat{\beta}, \tag{8}$$

where $\hat{\gamma}$ is the output feature map. Finally, the denormalization layer completes the entire process.

The key idea of the denormalization layer is to adaptively adjust the effective regions of the attribute feature map and the identity feature map, so that the result face image x_r can inherit the attribute of the target face image x_t and the identity of the source face image x_s. Compared with the AdaIN [17, 18], the denormalization layer can fuse the feature maps of arbitrary styles instead of pairwise exchanging, which shows better flexibility in face swapping.

3.2. Multi-scale Discriminator

The discriminator block D^i follows the traditional convolutional neuron network paradigm, which is composed of 5 down-sampling blocks and a classification head, as shown in Figure 4. Each down-sampling block is composed of a 4×4 convolution layer with stride 2, a normalization layer, and an activation layer, which is exactly the same as that of the attribute encoder E_α. The classification head contains a single linear layer to predict the probability distribution

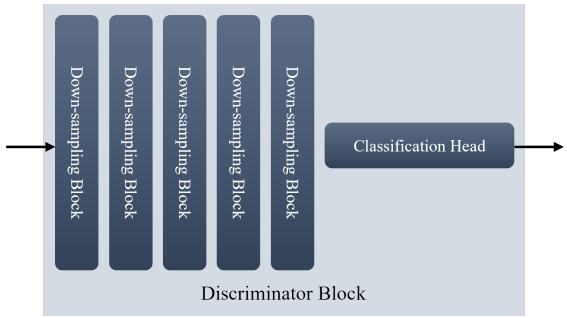

Figure 4. The structure of the discriminator block. The discriminator block is composed of 5 down-sampling blocks and a classification head as a typical convolutional neuron network.

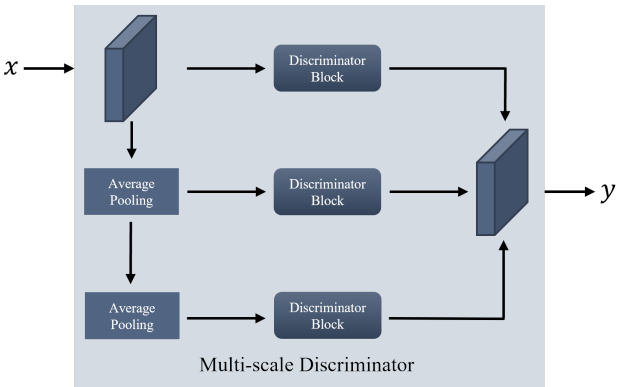

Figure 5. The overview of the multi-scale discriminator D. The discriminator D is composed of 2 average pooling layers and 3 discriminator blocks D^1, D^2 and D^3. The input image x will be down-sampled by the average pooling layers to form 3 different scales and fed into the corresponding discriminator blocks.

from the extracted feature map.

$$D^i(x) = \hat{y}^i, \tag{9}$$

where D^i represents the i-th discriminator block, $\hat{y}^i_\phi$ represents the predicted authenticity and $\hat{y}^i$ represents the predicted authenticity. We expect that the discriminator block can extract useful features to distinguish the generated face images from the real face images.

Inspired by [33], we apply a multi-scale discriminator D to enhance the performance of the discriminator, as shown in Figure 5. The multi-scale discriminator D is composed of 2 average pooling layers and 3 discriminator blocks D^1, D^2 and D^3. The input image x will be down-sampled by the average pooling layers to form 3 different scales, so the corresponding discriminator blocks can evaluate the input image x from different perspectives and perform better supervision. Previous works [21] have shown that the multi-scale discriminator can effectively improve the quality of the generated results, especially in the high-resolution tasks.

$$D(x) = \{D^1(x), D^2(x'), D^3(x'')\}, \tag{10}$$

where x' and x'' represent the input image x down-sampled once and twice respectively.

3.3. Training Objective

We utilize adversarial learning to train the injective generator G and the multi-scale discriminator D and introduce 4 kinds of loss functions: adversarial loss $\mathcal{L}_{adv}$, attribute loss $\mathcal{L}_{att}$, identity loss $\mathcal{L}_{ide}$ and reconstruction loss $\mathcal{L}_{rec}$. The overall loss function is a weighted sum over them, which is formally defined as follows.

$$\begin{aligned} \mathcal{L}^G_{all} &= \lambda^G_{adv}\mathcal{L}^G_{adv} + \lambda^G_{att}\mathcal{L}^G_{att} \\ &\quad + \lambda^G_{ide}\mathcal{L}^G_{ide} + \lambda^G_{rec}\mathcal{L}^G_{rec}, \\ \mathcal{L}^D_{all} &= \lambda^D_{adv}\mathcal{L}^D_{adv}, \end{aligned} \tag{11}$$

where λ^G_{adv}, λ^G_{att}, λ^G_{ide}, λ^G_{rec} and λ^D_{adv} are the hyperparameters to balance the loss functions.

Adversarial Loss. The adversarial loss is to train the discriminator D to distinguish the generated face images $x_r = G(x_s, x_t)$ from the real face images x_s, x_t, which indirectly encourages the generator G to generate more plausible face images. Hinge loss is applied as the adversarial loss, which is formally defined as follows.

$$\begin{aligned} \mathcal{L}^G_{adv} &= -\sum_{s=1}^{3} \max\{0, 1 + D^s(G(x_s, x_t))\} \\ \mathcal{L}^D_{adv} &= \sum_{s=1}^{3} \max\{0, 1 - D^s(x_s)\} \\ &\quad + \sum_{s=1}^{3} \max\{0, 1 + D^s(G(x_s, x_t))\} \end{aligned} \tag{12}$$

Attribute Loss. Intuitively, we expect that the result face images $x_r = G(x_s, x_t)$ can inherit the attribute of the target face images x_t. Therefore, The feature maps $\alpha_1, \alpha_2, \ldots, \alpha_8$ extracted by the attribute encoder E_α should be invariant. The attribute loss not only encourages the generator G to generate plausible face images, but also encourages the attribute encoder E_α to extract valid attribute features. The attribute loss is formally defined as follows.

$$\mathcal{L}^G_{att} = \frac{1}{2}\|E_\alpha(G(x_s, x_t)) - E_\alpha(x_t)\|^2_2. \tag{13}$$

Identity Loss. We also expect that the result face images $x_r = G(x_s, x_t)$ can follow the identity of the source face images x_s. Therefore, the feature vector β extracted by the identity encoder E_β should be invariant. The identity loss encourages the generator G to maintain the exact identity

under the guide of the pretrained identity encoder E_β. Cosine similarity is applied as the identity loss, which is formally defined as follows.

$$\mathcal{L}_{ide}^{G} = 1 - \frac{E_\beta(G(x_s, x_t)) \cdot E_\beta(x_s)}{\|E_\beta(G(x_s, x_t))\|_2 \cdot \|E_\beta(x_s)\|_2}. \tag{14}$$

Note that Euclidean distance is not suitable for the identity loss because it distorts in the high-dimensional space, whiie cosine distance caters to the Arcface [9] representation.

Reconstruction Loss. The reconstruction loss represents the pixel-wise difference between the result face images $x_r = G(x_s, x_t)$ and the target face images x_t, which is a direct supervision to the generator G. This term takes effect when the source face images x_s and the target face images x_t are the same. Mean squared error (MSE) is applied as the reconstruction loss, which is formally defined as follows.

$$\mathcal{L}_{rec}^{G} = \begin{cases} \frac{1}{2}\|G(x_s, x_t) - x_t\|_2^2, & \textit{if}\, x_s = x_t \\ 0, & \textit{otherwise} \end{cases}. \tag{15}$$

To fully exploit the reconstruction loss, we should intentionally sample the same face images x_s and x_t with a certain probability during the training process. Such a sampling strategy can to the largest extent ensure the invariance of the generator G and lead to better performance.

3.4. Model Inference

Until now, we have introduced the training process of the proposed face swapping framework, but there are still some details to be discussed when it comes to inference.

Image Inference. Different from the training process where the face images are well aligned, the face images in the wild can be various. Hence, the first step of the inference process is to align the face images. We adopt the MTCNN [35] model to detect the facial landmarks and crop the face images. Then the cropped face images will be fed into the injective generator G to generate the result face images $x_r = G(x_s, x_t)$. With the inverse transformation, we can obtain the final images with the same size as the input images. To make the fusion more natural and smooth, we introduce a post-processing step based on Laplacian pyramid blending [11] to fuse the result image with the original image. This is a classic image fusion method in traditional computer vision, which can effectively maintain the details and avoid the artifacts, as shown in Figure 6.

Video Inference. With the image inference process implemented, video inference becomes a trivial task. We can simply apply the image inference process to each frame of the video and merge the result frames to form the final video. However, the result video may contain some artifacts due to the misalignment of the face images. Such problems can be addressed with a temporal smoothing strategy, which is not discussed in this project. Experiment results show that the proposed face swapping framework has outstanding stability and robustness, especially in video inference.

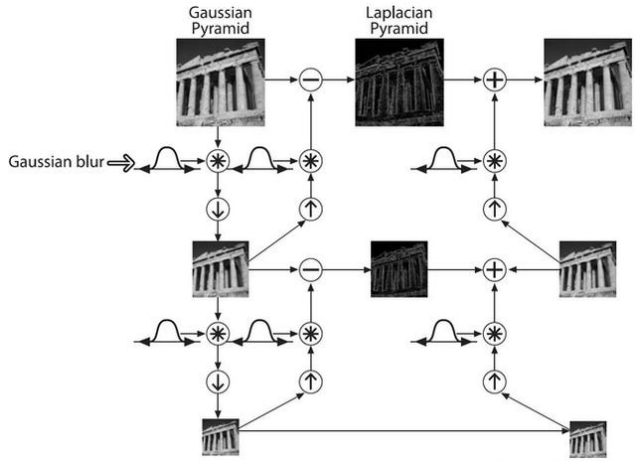

Figure 6. Illustration of the Laplacian pyramid blending. The algorithm fuses the images in multiple scales to reconstruct the image, which effectively maintains the details and avoids the artifacts.

4. Experiments

4.1. Experiment Setup

Dataset. According to the requirement, we mainly use the VGGFace2 dataset [5] to train the model. VGGFace2 is a common face recognition dataset, which contains 3.31M images of 9131 identities. The images are collected from the Internet and are labeled with the identity. We align and crop the face images into 224×224 resolution to match our model input. In addition, we participate in the construction of a new dataset, which contains face images captured from high definition movies. An automatic script is utilized to roughly capture the face images from the movies. Then we use K-means algorithm [13] to cluster the face images into different identities. Finally, we manually clean the collected data by checking the clustering results and reserving the valid images. These images are also resized into 224×224 resolution and appended to the main dataset for training. The final dataset contains 682k images of 9118 identities, which makes the training process stable while keeping relatively acceptable computational cost.

Implementation details. The model is trained for 500k iterations with a batch size of 16 on a single NVIDIA GeForce RTX 3090 GPU, which takes about 120 hours. We use the Adam optimizer for training. The learning rate is set to 0.0004 for both the generator and the discriminator. The weights of the loss function are $\lambda_{\text{adv}}^{G} = 1$, $\lambda_{\text{att}}^{G} = 10$, $\lambda_{\text{ide}}^{G} = 5$, $\lambda_{\text{rec}}^{G} = 10$ and $\lambda_{\text{adv}}^{D} = 1$, which is exactly the same as the paper [21]. As is mentioned in Section 3, we use the same source and target face image with a probability of

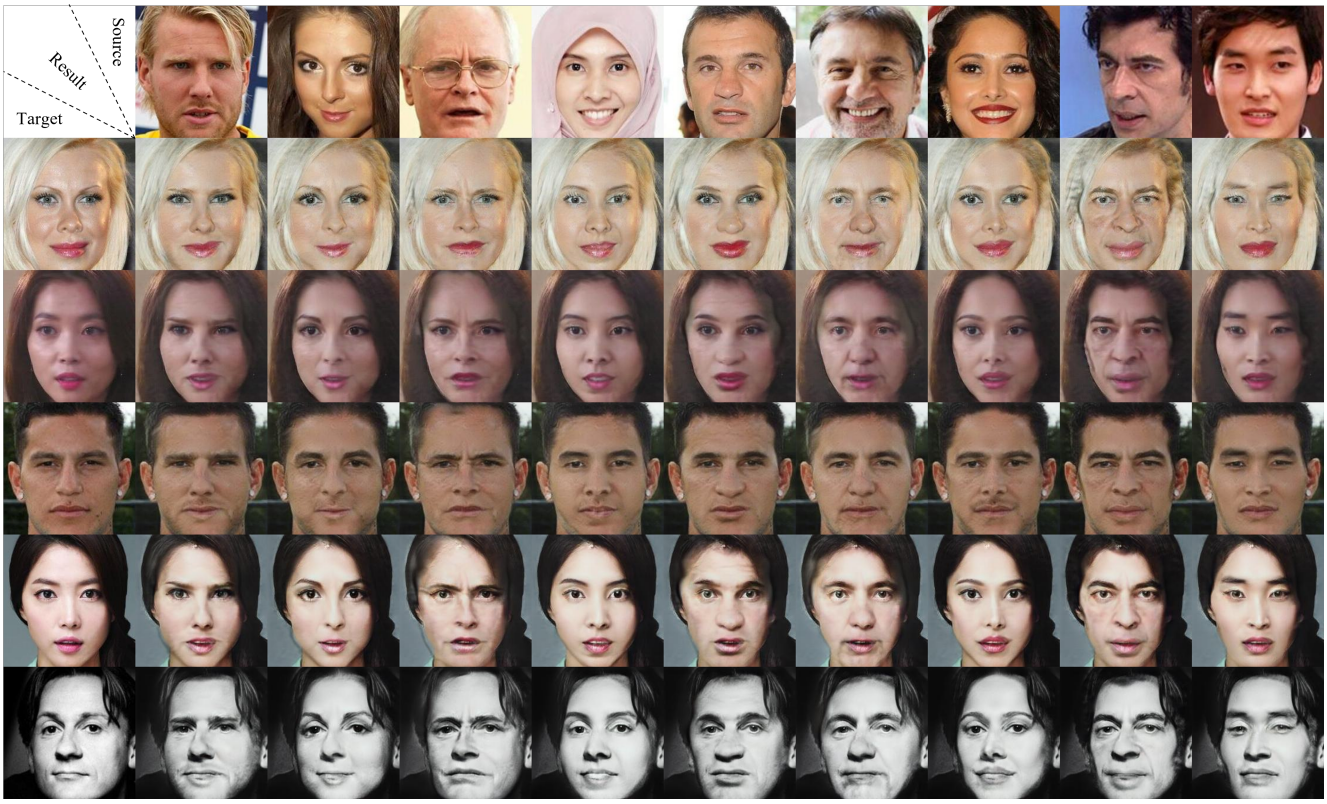

Figure 7. Some results generated by our face swapping framework. The first row consists of the source images, and the first column consists of the target images. Other images in the matrix are generated by our face swapping framework. We can see that the result face images preserve the attributes of the target images and the identities of the source face images.

20% to exploit the reconstruction loss.

Evaluation metrics. As an unsupervised task in computer vision, there are no standard evaluation metrics for face swapping. In this project, we follow FaceShifter [21] and SimSwap [6] to utilize ID retrieval and posture as the evaluation metrics. ID retrieval is first proposed in FaceShifter [21], which searches for the nearest face image in the test set for each generated image and computes the mean accuracy, thus measuring the identity preservation ability of the model. Posture is widely used in face swapping tasks, which measures the similarity of the head pose between the target image and the result image, thus measuring the attribute preservation ability of the model. We use a pretrained Arcface model [9] and a pretrained HopeNet model [30] to extract the identity feature and the posture vector respectively. We manually select 100 images from the FFHQ dataset [17] as the test set, which has never been seen by the model during the training process. According to the requirement, we also compute the cosine distance of the identity feature between the source image and the result image. Note that this metric is computed on a separate test set, which is different from the previous experiment settings.

4.2. Qualitative Results

We present some qualitative results of our face swapping framework in Figure 7. The first row consists of the source images and the first column consists of the target images. Other images in the matrix are generated by our face swapping framework. We can see that the result face images preserve the attributes of the target images and the identities of the source face images. The generated images are also visually realistic and natural. We owe the success to the strong warping ability of the injective generator together with the powerful attribute and identity encoder.

Table 1. Qualitative results of our face swapping framework. We use SimSwap as the baseline and apply three evaluation metrics, including ID retrieval, posture and cosine distance.

Method	Retrieval↑	Posture↓	Distance↓
SimSwap	71.84%	**2.59**	0.23
MetaSwap	**99.66%**	4.80	**0.17**

The metrics of our face swapping framework are shown in Table 1. We use SimSwap as the baseline, which is

trained on the same dataset for the same number of iterations. The computation details of the metrics have been described previously. We can see that our face swapping framework outperforms SimSwap in ID retrieval and cosine distance, which demonstrates the effectiveness of our framework in preserving the identity of the source face images. However, our framework performs worse than SimSwap in posture, which indicates that the framework is not so good as SimSwap at preserving the attribute of the target face images due to the lack of supervision on semantic features. We should note that preserving the identity and the attribute is a trade-off, which will be further discussed in Section 5. Such results are consistent with previous works [6, 21], where FaceShifter shows better performance in ID retrieval while SimSwap shows better performance in posture. Anyway, we claim that our face swapping framework is effective in preserving the identity of the source face image and the attribute of the target face image.

5. Discussion

Both FaceShifter and SimSwap are classic face swapping frameworks, which can achieve high-quality face swapping results in various scenarios, as is shown in Section 4. However, there are still some limitations in these GAN-based frameworks. In this section, we will discuss the limitations of these frameworks and present some possible solutions.

Preserving attribute or identity is a trade-off. As is illustrated in Section 3, the proposed generator G is devoted to injecting the identity of the source face into the target face, using sophisticated denormalization operation and attention mechanism. However, such implementation will inevitably lead to the loss of the attribute of the target face, which is reflected in Table 1 where the posture is much higher compared to the baseline. To address this problem, FaceShifter introduces another network for refinement, i.e. the HEAR-Net shown in Figure 8. However, this method breaks the end-to-end pipeline and cannot solve the attribute loss problem fundamentally.

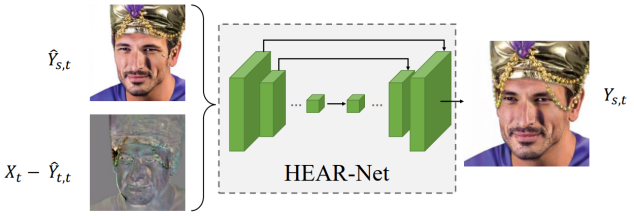

Figure 8. HEAR-Net proposed in FaceShifter. HEAR-Net applies the U-Net structure to refine the swapped face, restoring the details lost in AEI-Net, thus preserving the attribute of the target face.

In contrast, SimSwap modifies the discriminator D and introduces the weak feature matching loss, expecting to provide a better supervision signal, as is shown in Figure 9. The weak feature matching loss is derived from the common attribute loss, which is defined as

$$\mathcal{L}_{wfm}^{G} = \sum_{i=m}^{M} \|D_i(G(x_s, x_t)) - D_i(x_t)\|_1, \tag{16}$$

where D_i denotes the i-th layer of the discriminator and M is the number of layers. The intuition is that the shallow layers focus on the texture information, while the deep layers can reflect the attribute information. Therefore, the weak feature matching loss encourages the generator to preserve the attribute of the target face and avoids harming the identity injection. However, this design is weird in that the discriminator itself is not trained to extract the attribute information, but to distinguish the fake images from the real ones. Therefore, multiple optimization objectives are mixed in the discriminator, thus reducing the interpretability of the whole framework.

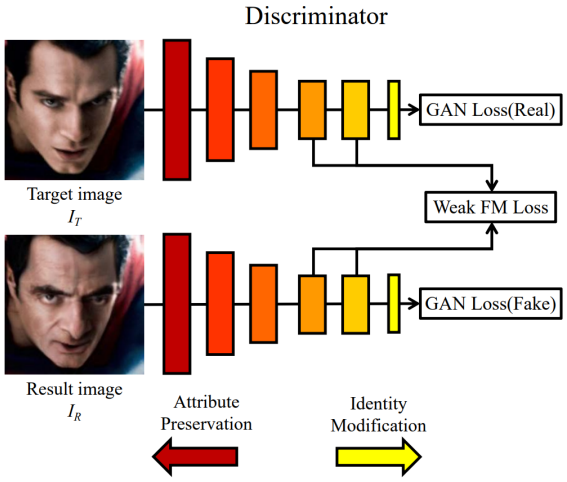

Figure 9. Discriminator proposed in SimSwap. The weak feature matching loss is introduced to preserve the attribute of the target face while avoiding harming the identity injection.

We should highlight that preserving attribute or identity is naturally a trade-off. There is no explicit boundary between the attribute and identity information. Even the same facial expression can be presented differently by different people. We might as well consider the extreme cases. If the generator directly outputs the same image as the target face, the attribute loss will be zero. On the contrary, if the generator outputs the same image as the source face, the identity loss will be zero. Therefore, what we can do is to find a balance between the attribute and identity loss, instead of being obsessed with finding a perfect solution.

Monocular vision is inadequate for perspective transformation. Theoretical research [14] has proved that a single camera is not enough to reconstruct the 3D structure of the scene in that the depth information is lost in the projection

process. This may cause severe problems in face swapping tasks, especially when the target face is not frontal. If the generator is fed with images of side faces, it tends to perform poorly. The model may mistake the angle of the target face. The pretrained MTCNN model may even fail to detect the face in the image, as is shown in Figure 10.

Figure 10. Some failure cases of our face swapping framework. The first and second row consist of the source and target images respectively. The third row shows the results of our face swapping framework, which performs poorly with side faces as the targets.

We owe this problem to two reasons. First, the model suffers from the long tailed problem because the training set is mainly composed of frontal faces and contains few side faces. Second but more importantly, the monocular vision is inadequate for perspective transformation. In other words, it is impossible for the model to reconstruct all the spatial information from a single image, which is the common problem of the one-shot face swapping frameworks, where the depth information is invalid.

Stereo vision [14] is a possible solution in traditional computer vision, which utilizes two cameras to capture the scene from different perspectives. This method is based on a mathematical model named epipolar geometry shown in Figure 11, which is of vital importance in the field of binocular vision. With the help of epipolar geometry, we can reconstruct the 3D structure of the scene from at least two images. This hints that the face swapping quality can be significantly improved if we can provide more perspectives of the target face, pushing the one-shot paradigm to few-shot.

With the advent of deep learning, the 3D reconstruction model can be parameterized as a neural network. Early in 2016, researchers have proposed a deep learning based method [10] for stereo vision. In recent years, more powerful models for 3D face reconstruction have been proposed [3], which significantly empowers the traditional face swapping frameworks. A typical example is HifiFace [34], a state-of-the-art face swapping framework based on 3D shape-aware identity extractor, as is shown in Figure 12. This framework fully exploits the spatial information, which brings huge advantages in the face swapping task.

GANs fail to provide sufficient scalability. Admittedly,

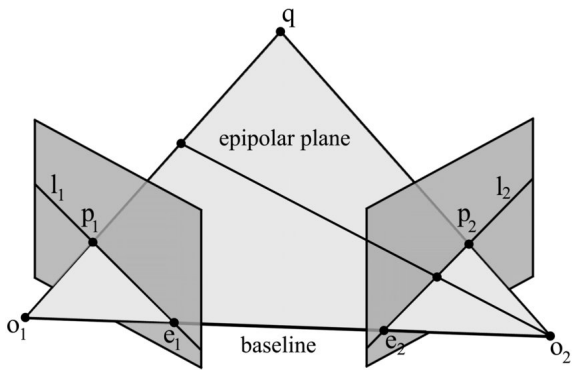

Figure 11. Illustration of epipolar geometry. The stereo vision system consists of two cameras. The epipolar plane is the plane that contains the baseline and the optical axes of the two cameras. The epipolar line is the intersection of the epipolar plane and the image plane. The epipolar constraint states that the corresponding points of the same 3D point must lie on the same epipolar line.

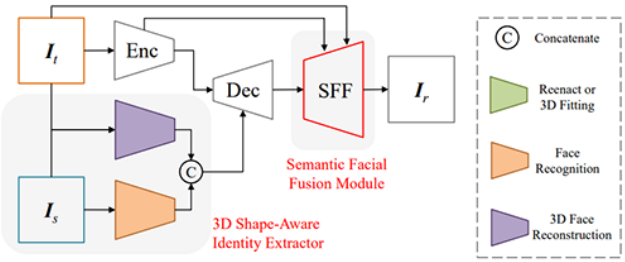

Figure 12. The overview of the HifiFace framework. A 3D shape-aware identity extractor and a semantic facial fusion module are proposed to improve quality and fidelity of the swapped faces.

GANs [12] have achieved great success in the generation tasks with high quality and low cost. However, when it comes to large-scale generation tasks, GANs cannot provide sufficient scalability. On one hand, training GANs is a tricky task with many possible pitfalls such as gradient vanishing and mode collapse. On the other hand, the representation ability of GANs is limited by the convolutional architecture. Therefore, generalizing to any face is extremely challenging for GAN-based face swapping frameworks.

This problem is being overcome with the emergence of diffusion models [16], which describes the generation process as a forward diffusion process and a backward denoising process. Diffusion models have achieved great success in text-to-image and image-to-image generation tasks, typically exemplified by Stable Diffusion [28] and DALL-E [27]. Previous works have proved that diffusion models are extremely scalable and can fit diverse data distributions. Recently proposed diffusion policy [7] in reinforcement learning (RL) also presents the potential of diffusion models.

Based on the powerful diffusion models, the face swapping framework can be entirely overturned. Recently pro-

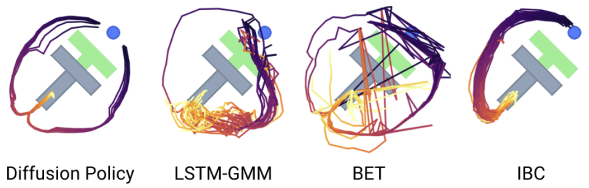

Figure 13. Intuitive illustration of the diversity of diffusion models. Diffusion models can accurately fit diverse data distributions compared to traditional deep learning models.

posed DiffFace [19] is a novel diffusion-based face swapping framework as is shown in Figure 14, where adversarial training is replaced by facial guidance, which is completely different from the traditional GAN-based frameworks. Empowered by diffusion models, DiffFace can naturally generate faces with high fidelity and easily learn from large-scale datasets. Stable Diffusion also provides a plugin for face swapping, which has acceptable performance for users with limited computing resources. However, there are still some problems to be solved in the diffusion-based face swapping frameworks, especially the temporal stability and noise robustness of the generated face images.

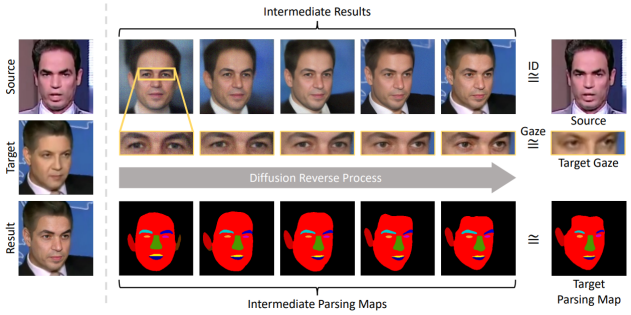

Figure 14. The overview of the DiffFace framework. DiffFace is a novel diffusion-based face swapping framework, which is completely different from the traditional GAN-based frameworks.

One-shot face swapping is at an awkward position in the film industry. As one of the most promising application scenario for face swapping technology, film industry lays high requirements on the quality and fidelity of the generated results. However, the one-shot face swapping frameworks fail to show explicit advantages compared to the dedicated face swapping models in quality. Since the training expense is completely affordable for the film makers, the price advantage of the one-shot face swapping frameworks loses its significance. Therefore, the one-shot face swapping frameworks are at an awkward position in the film industry. Luckily, the emergence of large models brings a new opportunity, where the base models with large capacity and pretrained on large-scale datasets are extremely powerful with their strong generalization ability. With some specific finetuning, they can achieve much better performance compared to the dedicated models, which is exactly the reproduction of large language models (LLMs) [26].

Face swapping technology is faced with ethical and privacy issues. The identity changing characteristic of face swapping determine the ethical and privacy problems in practical application. Intentionally using AI face swapping may lead to the transmission of false information, infringement of privacy, and other consequences. To avoid this technology being used for illegal activities such as political propaganda and privacy theft, we can take measures to make our facial changes more noticeably. This is not to decrease the performance of the model, but to add watermark or perform some unique encryption like modifying in frequency domain, which can be detected easily. Anyway, we should admit that no matter what effective encryption method we use, we cannot completely eliminate the ethical impact it brings. This problem is not just about face swapping since the ethical issues of AI applications have always been a long-standing problem. Researchers and legislators should cooperate to face the problem and explore how to address it, instead of leaving it in a grey zone.

6. Conclusion

In this project, we investigate recent advances in one-shot face swapping frameworks. We describe the development of face swapping technology and present the related works in Section 1 and 2. We propose a novel one-shot face swapping framework based on FaceShifter and illustrate the detailed methods in Section 3. We implement all the codes using the latest deep learning framework and conduct qualitative and quantitative experiments in Section 4. We use SimSwap as the baseline and compare the qualitative results to demonstrate the effectiveness of our framework. We find that FaceShifter is better at preserving the identity of the source face, while SimSwap is better at preserving the attribute of the target face, which is consistent with the results in the original paper. We present our discussion in Section 5 about the limitations of these one-shot face swapping frameworks and the possible future directions. We believe that preserving attribute or identity is inevitably a trade-off and monocular vision is inadequate for perspective transformation in face swapping. We also believe that diffusion models and other large pretrained models are promising for future research in the field of face swapping.

A. Member Contribution

Xiangyuan Xue. Investigate the paper, design the framework, implement the training code, run the experiments and write the majority of the report.

Shengmin Yang. Implement the evaluation code, write the test script and write the introduction of the report.

Yi Ai. Implement the inference code and write a part of the discussion in the report.

References

[1] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 2

[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: Automatically replacing faces in photographs. *ACM Transactions on Graphics*, 27(3):251–258, 2008. 2

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. *acm siggraph*, pages 187–194, 2002. 9

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

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 6

[6] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 2, 7, 8

[7] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023. 9

[8] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. Face transformer: Towards high fidelity and accurate face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 668–677, 2023. 2

[9] 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. 3, 6, 7

[10] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5515–5524, 2016. 9

[11] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987. 6

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 9

[13] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. 6

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 8, 9

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 9

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 4, 7

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3, 4

[19] Kihong Kim, Yunho Kim, Seokju Cho, Junyoung Seo, Jisu Nam, Kychul Lee, Seungryong Kim, and KwangHee Lee. Difface: Diffusion-based face swapping with facial guidance. *arXiv preprint arXiv:2212.13344*, 2022. 2, 10

[20] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2

[21] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 3, 5, 6, 7, 8

[22] Yuan Lin, Shengjin Wang, Qian Lin, and Feng Tang. Face swapping under large pose variations: A 3d model based approach. In *2012 IEEE International Conference on Multimedia and Expo*, pages 333–338. IEEE, 2012. 2

[23] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018. 2

[24] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI 14*, pages 117–132. Springer, 2019. 2

[25] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 2

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback.

Advances in Neural Information Processing Systems, 35: 27730–27744, 2022. 10

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 9

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 9

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 3

[30] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 7

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[32] Hong-Xia Wang, Chunhong Pan, Haifeng Gong, and Huai-Yu Wu. Facial image composition based on active appearance model. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 893–896. IEEE, 2008. 2

[33] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3, 5

[34] 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. 9

[35] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23 (10):1499–1503, 2016. 6