DOI: 10.1111/cgf.14949

Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

Semantics-guided generative diffusion model with a 3DMM model condition for face swapping

Xiyao Liu[1], Yang Liu[1], Yuhao Zheng[1],Ting Yang[1],Jian Zhang[1]*,Victoria Wang[2] and Hui Fang[3]*

[1]School of Computer Science and Engineering, Central South University, Changsha, 410083, China,
[2] School of Criminology and Criminal Justice, Faculty of Humanities and Social Sciences, University of Portsmouth, PO12HY, U.K. †
[3] Department of Computer Science, Loughborough University, Loughborough, LE113TU, U.K.

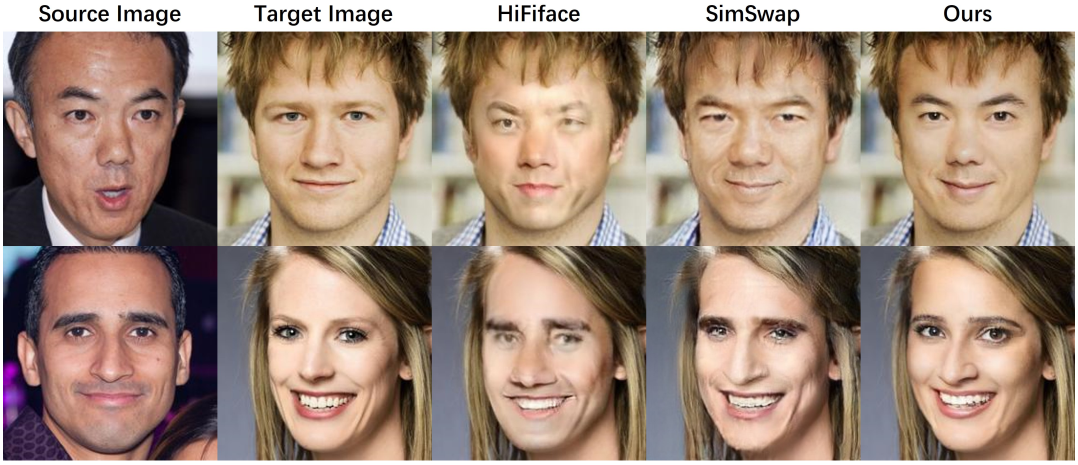

Figure 1: Face swapping results generated by our approach. *Compared to existing GAN-based face swapping approaches, our method effectively embeds the identity information from the source image, maintains the attributes from the target image, and generates images with enhanced realism.*

Abstract

Face swapping is a technique that replaces a face in a target media with another face of a different identity from a source face image. Currently, research on the effective utilisation of prior knowledge and semantic guidance for photo-realistic face swapping remains limited, despite the impressive synthesis quality achieved by recent generative models. In this paper, we propose a novel conditional Denoising Diffusion Probabilistic Model (DDPM) enforced by a two-level face prior guidance. Specifically, it includes (i) an image-level condition generated by a 3D Morphable Model (3DMM), and (ii) a high-semantic level guidance driven by information extracted from several pre-trained attribute classifiers, for high-quality face image synthesis. Although swapped face image from 3DMM does not achieve photo-realistic quality on its own, it provides a strong image-level prior, in parallel with high-level face semantics, to guide the DDPM for high fidelity image generation. The experimental results demonstrate that our method outperforms state-of-the-art face swapping methods on benchmark datasets in terms of its synthesis quality, and capability to preserve the target face attributes and swap the source face identity.

CCS Concepts

• ***Computing methodologies*** → *Computer graphics; Image manipulation; Computational photography;*

1. Introduction

Face swapping is a technique that transfers associated facial identity from a source image to a target media (which can be an image frame or a video clip); while preserving facial attributes of

† *Corresponding authors: Jian Zhang and Hui Fang

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

delivered by EUROGRAPHICS DIGITAL LIBRARY
www.eg.org diglib.eg.org

the target media, including pose, expressions, lighting condition, and background [NNN*22]. Since the last decade, face swapping has attracted significant research attention [LBY*19, ZFW*20, XYH*21, LPG*23], due to its various applications in film-making [Ver20], virtual human creation [ML21], and privacy protection [NNN*22], etc.

Recently, many methods have been proposed for high-fidelity face swapping, including 3D Morphable Model (3DMM) fitting [NMT*18, NKH19, JLW*20] and Generative adversarial networks (GANs) [CCNG20, WCZ*21]. 3DMM [BV03], one of the most traditional face representations, has been applied to disentangle and parameterise face identity and other attributes of two facial images, as a means to blend source identity into a target image. However, the quality of the synthesised face is far from desirable due to its linear approximation assumption of the 3DMM [BV03]. In contrast, GANs are more capable of generating high-quality swapped face images, as the result of the constant adversarial competition between a generator network and a discriminator network. However, GANs often suffer from training instabilities, which leads to oscillations during optimization and convergence issues [BSSE21].

Denoising Diffusion Probabilistic Models (DDPMs) [DN21, ND21, PCWS22, RBL*22] have emerged as an alternative to GANs [CWD*18]. As evidenced in [BSSE21], on the one hand, in terms of high synthesis quality and diversity, DDPMs outperform GANs in various applications. On the other hand, synthesis diversity hinders the embedding of desired high semantics into their synthesised images. To tackle this issue, recently, conditional denoising diffusion probabilistic models utilise classifier guidance [DN21] or text-embeddings [SCS*22], in order to provide prior knowledge as a constraint to facilitate the preservation of high semantics in those synthesised images. Nevertheless, research on the effective utilisation of semantic guidance in actual applications remains limited.

In this paper, we propose a novel conditional DDPM-based face swapping approach to achieve high-fidelity face synthesis and; at the same time, to embed extra face semantics, e.g., facial identity, into the synthesised image by using multiple prior guidance. Specifically, at the model training stage, we utilise a 3DMM to generate a swapped face image as an image-level condition to facilitate the effective convergence of our DDPM model. At the model inference stage, besides the image-level condition, we have used high-level face semantics extracted from several pre-trained attribute classifiers, as a means to provide further guidance to enhance identity embedding and attribute preservation in the diffusion process. As illustrated in Figure 1, compared to state-of-the-art (SoTA) methods, especially GAN-based models, our approach is capable of producing photo-realistic face images with embedded identity information from the source images. Our contributions can be summarised as follows:

- To our best knowledge, we are the first to introduce a conditional DDPM approach enforced by multiple semantic guidance for face swapping. Although we are aware of the fact that some previous work either used image condition in DDPM or used semantic guidance to achieve generic synthesis, we have not discovered any other work that use these two components simultaneously in this field.
- We exploit an image-level condition generated by a 3DMM to exploit prior knowledge for effective face swapping.
- We make use of a high-semantic level guidance driven by information extracted from several pre-trained attribute classifiers to achieve high-quality face image synthesis.

The remainder of the paper is organised as follows. Next, related work is discussed in Section 2. In Section 3, we explain the details of our proposed approach. Section 4 presents the experimental results to demonstrate that the proposed method outperforms SoTA benchmarks. Finally, Section 5 concludes the paper.

2. Related Work

2.1. Face Swapping

Face swapping aims to change the facial identity of a target image but to keep its other facial attributes. Early attempts require manual interactions with 3D-based models to synthesise a new face image [BKD*08, BSVS04, CTL*09]. To address this limitation, Face2Face [TZS*16] fits a 3DMM to both source and target faces, which enables automatic face swapping. Further, Nirkin et al. [NMT*18] have combined 3DMM and face segmentation model to achieve a robust face swapping approach under uncontrolled conditions. However, due to inherent linear approximation characteristics, 3DMM fitting approaches cannot produce photo-realistic synthesis.

Face swapping approaches have been dominated by GAN-based models [BCW*18, LBY*19, ZLW*21]. IPGAN [BCW*18] applies two encoders to disentangle identity from the source face and attributes from the target face, and recombine these two vectors as an input of a generator to swap faces. Similarly, SimSwap [CCNG20] presents an ID injection module as a conditional input of a generator to embed identity into the synthesised face. While in FS-GAN [NKH19], two generators, named a reenactment generator and a segmentation generator, are used to produce the reenacted source face and the background target image. They are further combined by using an inpainting generator and a blending generator, to synthesise new faces. To handle facial occlusions, FaceShifter [LBY*19] designs a second stage to refine occluded face regions by identifying them in a self-supervised manner.

Sharing the most similarity to our work, HifiFace [WCZ*21] extracts conditional facial representations from a 3DMM [BV99] as an extra input of a generator of GANs to produce 3D shape-aware identity face synthesis. Our method distinguishes from it in two aspects: (i) we utilise a DDPM to generate more realistic face swapping images, and (ii) we apply an image-level condition generated from 3DMM for finer synthesis guidance of our DDPM.

2.2. Diffusion Model

Diffusion model, as an emerging generative model, is able to gradually denoise and produce high-quality synthesised images through learning a reversed diffusion process [SDWMG15, HJA20]. It exploits a deep neural network to learn the reverse trajectory in order to recover high-quality data samples from noisy images, assuming a data distribution can be converted into a simple distribution, e.g., Gaussian distribution, by adding noises step by step,

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

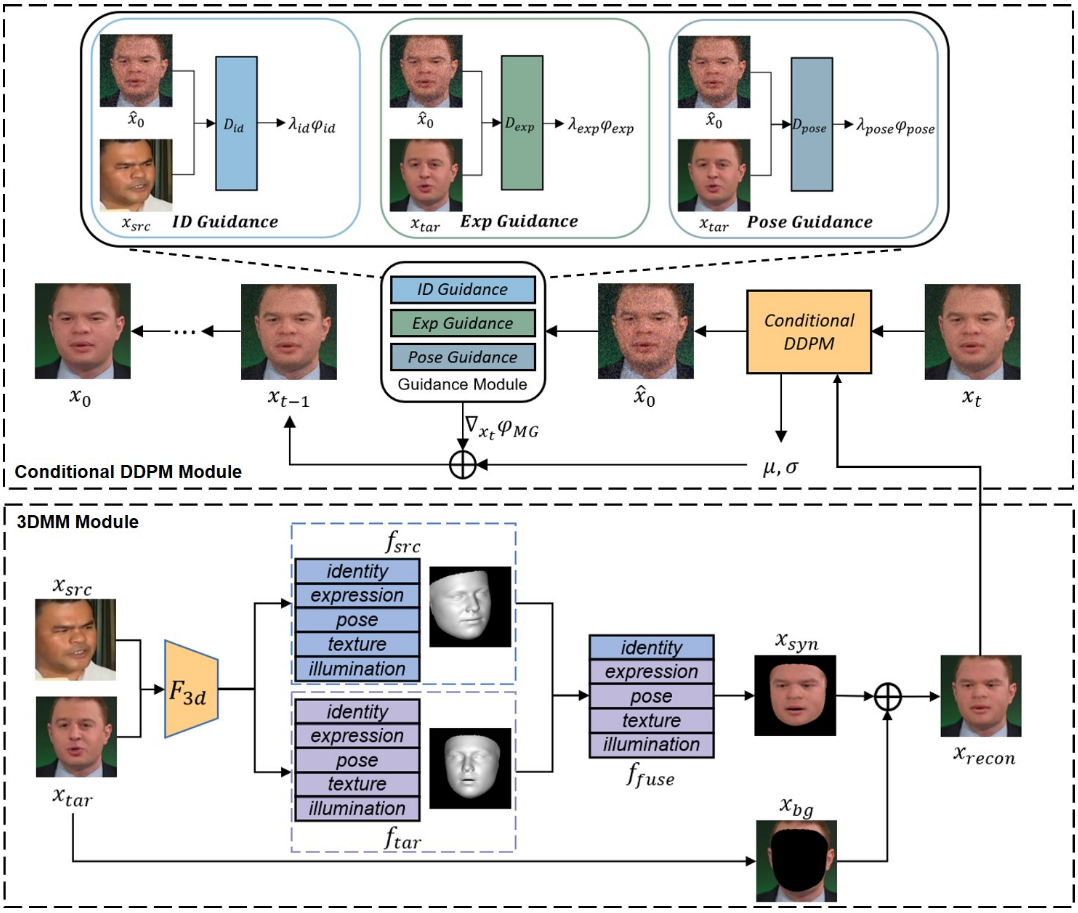

Figure 2: The overview of our approach. *We use 3DMM to extract coefficients of the source and target faces to obtain f_{src} and f_{tar}. We then recombine the identity parameters from f_{src} and the other parameters from f_{tar} to achieve f_{fuse}. Subsequently, we reconstruct and render the face image x_{recon} using f_{fuse}. In the sampling process, x_{recon} serves as an image-level condition input for our conditional DDPM, while multiple additional facial attribute classifiers are deployed to guide the generation of the conditional DDPM.*

i.e., diffusion process. Upon demonstrating its potential to generate photo-realistic image samples [HJA20], DDPM has been applied in various image processing tasks, including image super-resolution [SHC*22], text-to-image generation [NDR*21], and image editing [MHS*21], etc.

Different models of conditional DDPMs are proposed to enforce conditions to preserve high semantics in targeted synthesised images. For instance, In [DN21], classifier guidance is proposed to use gradients to guide diffusion sampling toward a random class. This strategy has further improved the FID metric for image synthesis. In a super-resolution task, Saharia et al. [SHC*22] utilise low-resolution images as the condition of a DDPM to generate their corresponding high-resolution images. Furthermore, conditional DDPMs have also been widely applied to text-to-image generation [NDR*21, RBL*22] and text-guided image editing [ALF22] by injecting text representations. Our model of conditional DDPM investigates different types of conditions, including image-level conditions and high-semantic conditions, to apply the conditional DDPM to a specific application, i.e., face swapping, for the first time.

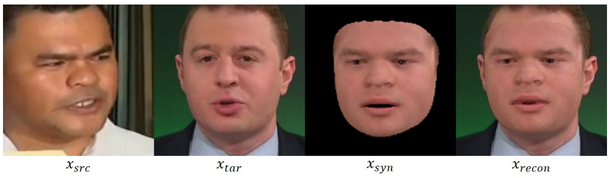

Figure 3: The process of using 3DMM to obtain the reconstructed image x_{recon}.

3. Conditional DDPM for face swapping

As illustrated in Figure 2, our approach utilises 3DMM model fitting to generate an initialised swapped face image (Subsection. 3.1) as an image-level condition to guide the diffusion process. In addition to the image-level condition, after training the conditional DDPM, three high-semantic classifiers, including identity, expression and pose extractors, are used to provide our model of conditional DDPM with further guidance. The actual training process of the conditional DDPM is explained in Subsection 3.2. In Subsection 3.3, we detail the high-semantics guidance at the inference stage.

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

3.1. Image-level condition generation by 3DMM

3DMM is a linear 3D-aware facial model that disentangles and parameterised a 2D face image into a list of parameters, which can be used to reconstruct the image. In our work, we deploy a pre-trained 3DMM model from [DYX*19]. Given an input face image, the 3DMM regresses a vector v = (α, β, δ, γ, p) $\in R^{257}$, where α $\in R^{80}$, β $\in R^{64}$, δ $\in R^{80}$, γ $\in R^{27}$ and $p \in R^6$ to represent the identity, the expression, the texture, the illumination and the pose, respectively. Using these parameters, the 3D face shape (S) and its texture (T) can be represented through two linear equations:

$$\mathbf{S} = \mathbf{S}(\alpha,\beta) = \overline{\mathbf{S}} + \mathbf{B}_{id}\alpha + \mathbf{B}_{exp}\beta \tag{1}$$

$$\mathbf{T} = \mathbf{T}(\delta) = \overline{\mathbf{T}} + \mathbf{B}_t\delta \tag{2}$$

where (i) $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ are the average face shape and texture; (ii) $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_t$, scaled with standard deviations, are the PCA bases of identity, expression, and texture, respectively; and (iii) α, β, and δ are the corresponding coefficient vectors for generating a 3D reconstructed face.

To generate the image-level condition from 3DMM, as illustrated in Figure 2, we fit both the source image and the target image to the 3DMM model, so as to extract two sets of parameters, f_{src} and f_{tar}. After generating f_{fuse} via recombining identity parameters from the source image and other attributes from the target image, the conditional image of our DDPM x_{recon} is synthesised by combining the rendered face region x_{syn} and the target background image x_{bg}. The conditional image generation can be formulated as:

$$x_{recon} = P\left(x_{bg}, RR\left(Swap\left(F_{3d}\left(x_{src}\right), F_{3d}\left(x_{tar}\right)\right)\right)\right) \tag{3}$$

where (i) $P(\cdot)$ is linear fusion function; (ii) x_{src} is the source image; (iii) x_{tar} is the target image; (iv) x_{bg} is the background image for s_{tar}; (v) $RR(\cdot)$ refers to 3D reconstruction and rendering; (vi) $Swap(\cdot)$ is a parameter swap function; and (vii) the $F_{3d}(\cdot)$ is the 3DMM. An example of the conditional image is illustrated in Figure 3. Although photo-realistic quality has not been achieved, the reconstructed face offers a reasonable condition to guide the DDPM model, as a means to generate the swapped face with all desired attributes.

3.2. Training process of the conditional DDPM

Diffusion models generate a realistic image from a standard Gaussian distribution by reversing a recurrent noising process. We formulate our problem statement as learning a parametric approximation to $p(x|x_{recon})$ through a stochastic iterative refinement that maps a 3D reconstructed image x_{recon} to a photo-realistic image $x \in R^d$. Subsequently, the problem is approached by adapting the DDPM in [HJA20] to a conditional image generation model. Following the work in [HJA20], we first define a forward Markovian diffusion process q that gradually adds Gaussian noise to image x_0 over T iterations as shown in the following equation:

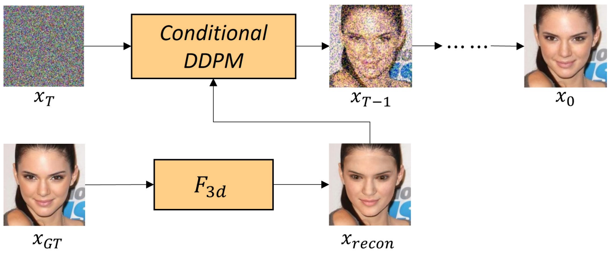

Figure 4: Training process of conditional DDPM. *The ground truth image x_{GT} is passed through the 3D reconstruction model to obtain the reconstructed image x_{recon}, x_{GT} and x_{recon} are paired into the conditional DDPM for training, where x_{recon} is the condition.*

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t \mid \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}\right) \tag{4}$$

where the scalar parameters α_t are hyper-parameters, subject to $0 < \alpha_t < 1$, which determines the variance of the noise added at each iteration. Inference under our model is defined as a reverse Markovian process, which is expressed as:

$$p_\theta\left(x_{t-1} \mid x_t, x_{recon}\right) = \mathcal{N}\left(x_{t-1} \mid \mu_\theta\left(x_t, x_{recon}, t\right), \sigma_t^2\mathbf{I}\right) \tag{5}$$

where x_{recon} is the conditional image, $\mu_\theta(\cdot)$ and $\sigma_t(\cdot)$ represent the mean and variance of the distribution, which can be parameterized by using deep neural networks. In practice, it is well known that the use of noise approximation model f_θ works best instead of using $\mu_\theta(\cdot)$ [HJA20]. Thus, $\mu_\theta(x_t, x_{\text{recon}})$ can be expressed as:

$$\mu_\theta\left(x_t, x_{recon}, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}f_\theta\left(x_t, x_{recon}, t\right)\right) \tag{6}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$ and we set the variance of $p_\theta\left(x_{t-1} \mid x_t, x_{recon}\right)$ to $(1-\alpha_t)$, a default given by the variance of the forward process in [PPSRHR16]. With these definitions, each iteration of the iterative refinement under our model is expressed as:

$$x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}f_\theta\left(x_t, x_{recon}, t\right)\right) + \sqrt{1-\alpha_t}\varepsilon_t \tag{7}$$

where $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Given x_t, the reverse process of the diffusion model outputs x_{t-1}. After iterations, a high-quality synthesised image $\hat{x}_0$ is obtained as the final output.

The training process is illustrated in Figure 4. After generating the condition image from 3DMM described in the previous subsection, it is applied as a condition of the DDPM to learn the denoising neural network. Notably, at the training stage, different from the inference stage, we adopt a self-supervised technique to train the conditional DDPM. In doing so, we generate the conditional image by synthesising the face patch via the 3DMM, and subsequently render it with the same image background. The denoising model

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

Algorithm 1: Training the Conditional DDPM

```
Data: Groundtruth image $x_{GT}$
Begin
$x_{bg} \leftarrow x_{GT}$
$x_{recon} = P\left(x_{bg}, RR\left(F_{3d}\left(x_{GT}\right)\right)\right)$
repeat
$\left(x_{recon}, x_0\right) \sim p(x_{recon}, x)$
$\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Take a gradient descent step on
$\nabla_\theta \left\|f_\theta\left(x_t, x_{recon}, t\right) - \varepsilon\right\|_2^2$
until converged
End
```

$f_\theta\left(x_t, x_{recon}, t\right)$ takes the conditional image x_{recon} and the noise image x_t as input, and is trained to predict the noise vector for iterative recovery. The proposed loss function for training f_θ is:

$$L = \left\|f_\theta\left(x_t, x_{recon}, t\right) - \varepsilon\right\|_2^2 \tag{8}$$

where x_t is a noisy version of input image x_{GT} at timestep t using Equation 4. The pseudocode for training the conditional DDPM is shown in Algorithm 1.

3.3. High-semantic Guidance at the inference stage

The use of guidance has become a popular technique to embed further semantics into a diffusion model. To embed one high semantic attribute y, a classifier $p(y|x_t, t)$ is used in noised images to compute the probability of y [DN21]. Subsequently, its gradient is derived to guide the diffusion model for sampling in the next iteration. This process can be formulated as:

$$p_\theta\left(x_{t-1} \mid x_t, y\right) = \mathcal{N}\left(\mu + s\sigma\nabla_{x_t}\log p_\theta\left(y \mid x_t\right), \boldsymbol{\sigma}\mathbf{I}\right) \tag{9}$$

where s is a constant to represent the guidance scale, y is class label, and μ and σ are mean and variance of the data distribution. Since the guidance at the early iterations of the diffusion process is weak, we deploy the method in [SHC*22] to estimate $\hat{x}_0$ and use the estimated output to compute the guidance gradients. This strategy has made the diffusion model converge more efficiently. The estimated $\hat{x}_0$ is computed using the following equation:

$$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t} f_\theta\left(x_t, x_{recon}, t\right)\right) \tag{10}$$

where $f_\theta\left(x_t, x_{recon}, t\right)$ represents the denoising model. To enhance the effect of the face swapping, we utilise features, including identity embedding subspace from Arcface [DGXZ19], pose estimator from [DGV*20], and expression extractors from [NHSC19], as the semantic guidance during the diffusion sampling process. Features from these pre-trained models are used instead of 3DMM parameters as guidance because (i) they are more distinctive semantic features compared to the 3DMM parameters, and (ii) the two-level conditions could complement each other by reducing redundancy, which could be introduced if the 3DMM parameters are used.

Algorithm 2: High-semantics Guidance for Conditional DDPM

```
Data: source image $x_{src}$, target image $x_{tar}$
Result: output $x_0$ includes the identity information of the
        $x_{src}$ and the attribute information of the $x_{tar}$
Begin
$x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$x_{bg} \leftarrow x_{tar}$
$x_{recon} = P\left(x_{bg}, RR\left(Swap\left(F_{3d}\left(x_{src}\right),\ F_{3d}\left(x_{tar}\right)\right)\right)\right)$
for $t = T, \cdots, 1$ do
$\hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(x_t - \sqrt{1-\bar{\alpha}_t} f_\theta\left(x_t, x_{recon}, t\right)\right)$
$\varphi_{id} \leftarrow cos\left(D_{id}\left(\hat{x}_0\right), D_{id}\left(x_{src}\right)\right)$
$\varphi_{exp} \leftarrow ||F_N(D_{exp}(\hat{x}_0)) - F_N(D_{exp}(x_{tar}))||_2^2$
$\varphi_{pose} \leftarrow ||F_N(D_{pose}(\hat{x}_0)) - F_N(D_{pose}(x_{tar}))||_2^2$
$\varphi_{MG} \leftarrow \lambda_{id}\varphi_{id} - \lambda_{exp}\varphi_{exp} - \lambda_{pose}\varphi_{pose}$
$\hat{x}_{t-1}$ $\mathcal{N}(\mu + \sigma\nabla_{x_t}\varphi_{MG}, \sigma)$
End for
Return $x_0$
End
```

Identity Guidance. To enhance the source identity information in the generated images, we ensure that the identity vectors extracted from both source and generated images are close to each other in the Arcface embedding subspace. The identity similarity we use as guidance is:

$$\varphi_{id} = \cos(D_{id}(x_{src}),\ D_{id}(\hat{x}_0)) \tag{11}$$

where D_{id} is the identity extractor [DGXZ19], x_{src} is the source image and $\hat{x}_0$ is obtained by the Equation 10.

Expression Guidance. To align the facial expression of the generated image to the one in the target image, we use the MLCR model [NHSC19] and compute the l2 distance between the expression vectors from the paired images to guide the conditional DDPM sampling process:

$$\varphi_{exp} = ||F(D_{exp}(x_{tar})) - F(D_{exp}(\hat{x}_0))||_2^2 \tag{12}$$

where D_{exp} is the expression extractor [NHSC19], and $F(\cdot)$ is the normalization function.

Pose Guidance. Pose plays a crucial role in face swapping, and most existing face swapping methods cannot preserve the pose of the target face very well. Thus, some deepfake detection methods use head pose as a cue for the detection of face swapping [YLL19]. Bearing this in the mind, we incorporate pose guidance into the sampling process of the conditional DDPM to ensure head pose preservation in the target image. In specific, we deploy the pose estimator proposed in [DGV*20] to extract pose vectors of the target face x_{tar} and the denoised image by localising their key fiducial

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

Method \ Metric	ID Retrieval ↑	ID Cos ↑	Exp. Error ↓	Pose. Error↓
FaceSwap	54.19%	0.330 ± 0.16	0.14± 0.06	2.48 ± 1.92
Deepfakes	81.96%	0.437 ± 0.13	0.19 ± 0.06	4.24 ± 2.60
HiFiface	92.17%	0.565 ± 0.09	0.30 ± 0.06	2.94 ± 1.64
Simswap	93.07%	0.597 ± 0.08	**0.12 ± 0.04**	2.40 ± 1.40
FSGANv2	94.60%	0.589 ± 0.10	0.13 ± 0.05	2.40 ± 2.24
Ours	**96.01%**	**0.614± 0.08**	**0.12 ± 0.04**	**2.24 ± 1.13**

Table 1: Quantitative comparison on FF++ dataset. We compare out approach with other SOTA methods based on ID retrieval, ID cosine similarity, expression error and pose error , respectively. Here, ↑: the higher the better; ↓: the lower the better.The expression error and pose error are evaluated based on the L2 distances generated by expression feature extractor [VA19] and pose estimator [RCR18]. Note that the model used for testing is different from the model used in the semantic guidance in our approach.

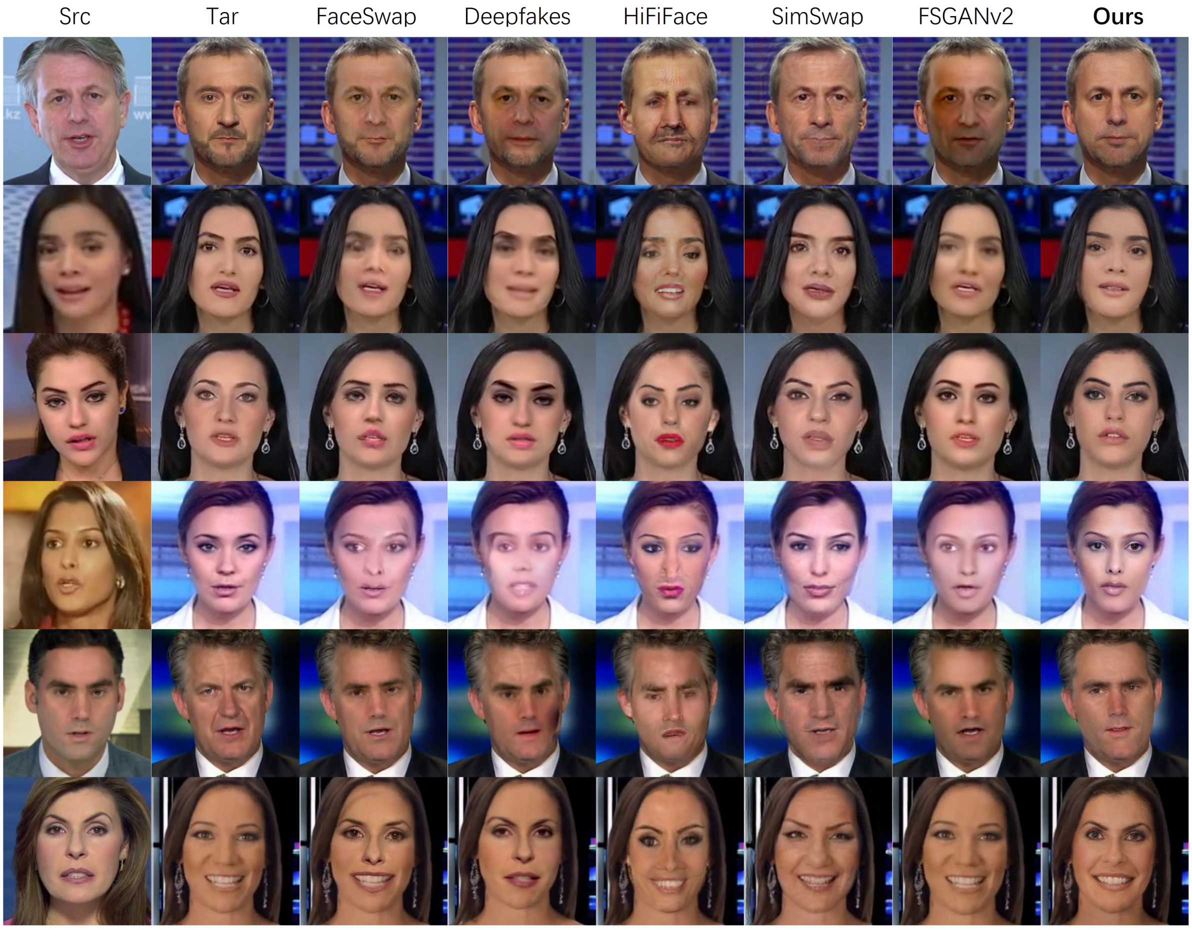

Figure 5: Comparison to other SoTA methods. *The methods we compare include: FaceSwap, Deepfakes, HiFiFace, SimSwap, FSGANv2. Here, Src represents the source image, Tar represents the target image. From the generated results, our method can preserve the identity information of the source image and the attribute information of the target image, and generate higher-quality images at the same time.*

points. We then compute the l2 distance between x_{tar} and the pose vector of x_t to enhance their consistency:

$$\varphi_{pose} = \|F(D_{pose}(x_{tar})) - F(D_{pose}(\hat{x}_0))\|_2^2 \tag{13}$$

where D_{pose} is the key points extractor implemented in [DGV*20]. To guide the diffusion sampling process towards the desired images, we integrate gradients from these multiple guidance modules formulated as follows:

$$\begin{gathered} x_{t-1} \sim \mathcal{N}(\mu + \sigma\nabla_{x_t}\varphi_{MG}, \sigma) \\ \mu = \mu_\theta(x_t, x_{recon}, t), \sigma = (1-\alpha_t) \\ \varphi_{MG} = \lambda_{id}\varphi_{id} - \lambda_{exp}\varphi_{exp} - \lambda_{pose}\varphi_{pose} \end{gathered} \tag{14}$$

where $\mu_\theta(\cdot)$ and σ represent the mean and variance of $p_\theta(x_t, x_{recon}, t)$, λ_s are weights to balance these items and they are set heuristically as: λ_{id}=1000, λ_{exp}=200, λ_{pose}=100. The pseudo-

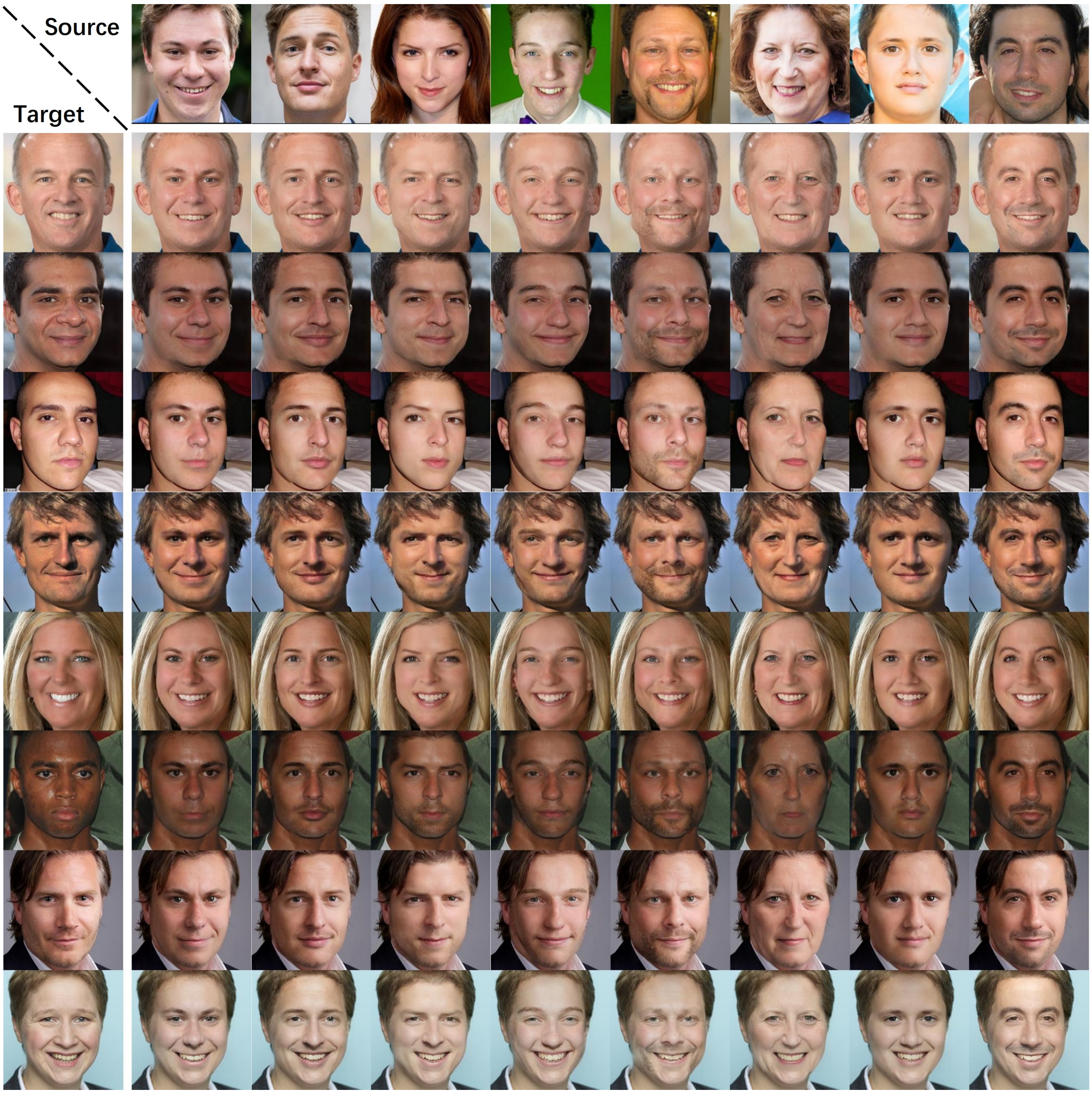

Figure 6: Face matrix generated by our approach. *These images are randomly picked from the Internet. All the target and source images are excluded from the training set. The results show that our method can embed the identity of the source face while preserving the attributes of the target face.*

code of conditional DDPM with multiple facial guidance is summarized in Algorithm 2.

4. Experiment

4.1. Implementation Details

To optimize the conditional DDPM for our task, we first construct a new dataset derived from two commonly-used face datasets, which are CelebaHQ [KALL17] and VGGFace2 [CSX*18]. Secondly, we train the conditional DDPM using the new dataset, after aligning and cropping faces to 224 * 224. More specifically, we execute a 3D reconstruction and rendering process on the images from CelebaHQ and VGGFace2 via F_{3d} to obtain reconstructed corresponding images. During the training process, we fed both the original and reconstructed images into the conditional DDPM. The new dataset comprises 30k pairs of original and reconstructed images, each with a resolution of 224*224. We implement our network using PyTorch [PGM*19]. Adam optimizer [KB14] is used for training, and the learning rate is set to 0.0001. Our model is trained at 2000 epochs.

Competing Methods. We compare our model with state-of-the-art face swap methods. These include FaceSwap [Fac12] and Deepfakes [Dee12], which are famous open-sourced tools; and SimSwap [CCNG20], HiFiFace [WCZ*21], and FSGANv2 [NKH22], which are all face swapping methods based on GANs. We use the officially released FaceSwap and Deepfakes results from the

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

FF++ [RCV*19] dataset, and the officially released codes and models for producing swapped results of other methods.

4.2. Quantitative Comparisons

We conduct quantitative experiments on the FF++ dataset [RCV*19]. Following the approach outlined in [LBY*19], we uniformly sample 10 frames from the 280 videos to obtain 2800 faces for evaluation. We utilise four commonly used metrics: the identity (ID) cosine similarity and ID retrieval scores are measured between the swapped results, and the source uses an independent pre-trained identity-recognition network [WWZ*18]. Specifically, the computation process of the ID retrieval score is identical to that in [LBY*19]. To evaluate the extent to which the generated results preserve other attributes of the target image, we calculate the expression error and pose error based on the L2 distances generated by expression [VA19] and pose extractors [RCR18]. To ensure evaluation fairness, the pose and expression extractor models for evaluation are different and independent to the ones used for guidance in our approach.

As shown in Table 1, our method achieved the best performance when compared to other SoTA methods. In terms of source ID embedding, the ID retrieval of our method reaches 96.01% which is 1.41% better than the runner-up. Similarly, the average ID cosine similarity between our synthesised faces and source faces reaches 0.614, which is the best performance in all the competing methods. For the preservation of expression from target faces, ours is equivalent to SimSwap but better than any other method. In terms of pose preservation, the average error between the target faces and synthesised faces is 2.24, which is also the best among all the methods.

4.3. Qualitative Evaluation

The qualitative results are illustrated in Figure 5 and Figure 6. As shown in Figure 5, we compare our results with FaceSwap [Fac12], Deepfakes [Dee12], SimSwap [CCNG20], HiFiFace [WCZ*21] and the latest work FSGANv2 [NKH22]. The comparisons are based on the test data provided by FF++. As can be observed, the image generated by SimSwap effectively retains the identity information of the source image. However, it falls short of adequately preserving attribute information, such as expressions, from the target image. Similarly, the image created by FSGANv2 faces the same challenges, including inconsistencies in the positioning of the eyes in the generated image relative to the target. Moreover, the unnatural lighting detracts from the realism of the generated results. In contrast, our method can achieve a better balance between identity information and attribute information retention. The generated image not only embeds the identity information of the source image; but also preserves the attributes of the target image, e.g., expression, pose and background. Furthermore, our method is able to generate more realistic results when compared to other face swapping methods.

We further produce a visualisation matrix to illustrate the synthesis quality by swapping a set of source faces with another set of target faces in Figure 6. Here, we randomly pick eight source images and eight target images to avoid cherry-picking illustrations. This matrix demonstrates that our approach has successfully swapped faces with a generalised capability.

Method	ID Similarity↑	Att.Preservation ↑	Naturalness↑
3DMM	2.67 ± 0.56	2.56 ± 0.57	2.29 ± 0.55
FaceSwap	2.73 ± 0.55	2.67 ± 0.56	2.67 ± 0.57
Deepfakes	2.84 ± 0.50	2.76 ± 0.55	2.62 ± 0.56
HiFiface	3.13 ± 0.52	3.18 ± 0.54	3.33 ± 0.60
SimSwap	3.58 ± 0.61	3.36 ± 0.58	3.62 ± 0.57
FSGANv2	3.24 ± 0.58	3.38 ± 0.58	3.67 ± 0.54
Ours	**3.67 ± 0.56**	**3.51 ± 0.58**	**3.89 ± 0.43**

Table 2: User Study's Ranking Scores. Higher scores indicate more realistic results generated by the method

4.4. User Evaluation

We conducted a user study on the FF++ dataset with 45 participants with normal vision (25 male, 20 female; aged 20 to 52, average age 26.91 ± 6.36) to further compare the synthesised image quality of the face swapping methods. Participants were informed about the purpose of the study but not about any of our hypotheses. Participants were given the dataset of source images, target images, and the face swapping results generated by these methods. Then, they were asked to rank the quality of the fake images based on the following criteria: (i) identity similarity with the source images; (ii) attribute preservation, which includes expressions and backgrounds consistent with the target images; and (iii) naturalness, which indicates whether there are visible artefacts on the synthesised face and whether the figures resemble real faces. Prior to the evaluation, detailed instructions were provided to the participants. We devised a scoring system with the highest score being 5 and the lowest being 1 for each case.

The results reported in Table 2 are based on the responses of the 45 participants. We can observe that GAN-based methods outperform traditional methods, such as FaceSwap and Deepfakes, with large margins. Regarding the GAN-based methods, the identity preservation score of Simswap is higher than that of HiFiface and FSGANv2, but SimSwap does not perform very well on attribute preservation of target images. Although the Identity Similarity score of FSGANv2 is lower than SimSwap, the score of attribute preservation is higher than SimSwap. In contrast, our method achieved the best results in all the evaluation questions. It demonstrates that our conditional DDPM outperforms GAN-based methods to generate higher quality images, and the guidance strategy we proposed during the sampling process of the diffusion model does work effectively. In addition, we can also find the images directly generated by 3DMM is far from photo-realistic due to its dimension reduction and linear approximation characteristics [EST*20, ZGC*21]. Note that the detailed user-study comparison results for specific individual persons are provided in our Supplemental Material.

4.5. Ablation Study and Analysis

In this section, ablation studies are carried out qualitatively and quantitatively to demonstrate the superiority of our method by using the two-stage conditional guided DDPM framework. In Figure 7, some qualitative examples are illustrated (with more qualitative results provided in our Supplemental Material). Further, quantitative results are presented in Table 3.

Conditional DDPM without Guidance. By comparing the third

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

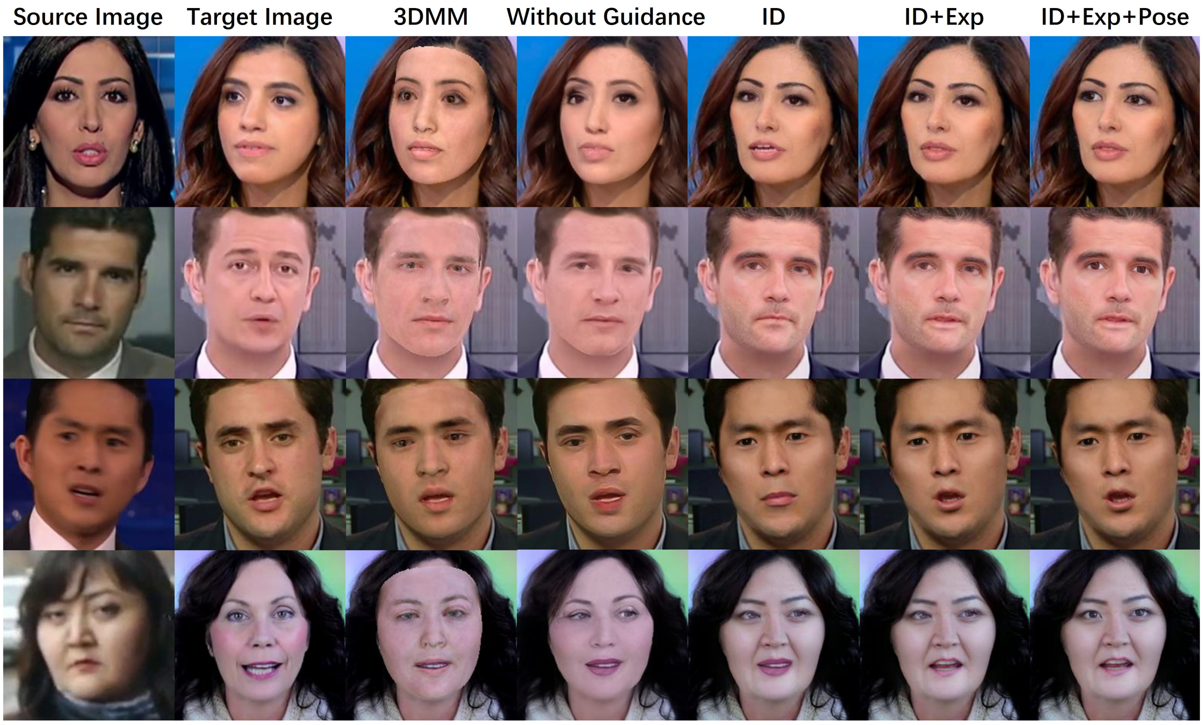

Figure 7: Ablation study for high-semantics guidance. *The third column represents the results constructed by 3DMM; the fourth column represents the results generated without guidance; the fifth column represents the results guided by identity only; the sixth column represents the results guided by both identity and expression; and the seventh column represents the results guided by identity, expression, and pose. The guidance effectively embeds identity information of source faces and preserves other attributes of target faces in the face swapping results.*

and fourth columns in Figure 7, we can observe that the use of conditional DDPM produces synthesised faces with much higher visual quality. This is also confirmed by the results in Table 2.

Identity Guidance. In the sampling process of conditional DDPM, identity guidance is the most important factor to embed the identity of a source image. To verify the effectiveness of identity guidance, we conduct an ablation study by using no guidance and only the identity guidance to generate synthesised results in the conditional DDPM sampling process. As shown in Figure 7, by comparing the faces of the fourth and fifth columns, i.e. without and with ID guidance, we can see that the use of source identity to guide the conditional DDPM produces synthesised faces with visually similar featured characteristics of the source faces. Furthermore, it is observed from Table 3, when identity guidance is not added, the identity cosine similarity between the generated image and the source image is only 0.424, but the identity cosine similarity between the generated image and the source image is improved to 0.603 after we add the identity guidance. Obviously, identity guidance is essential in the denoising process of DDPM for face swapping.

Expression Guidance. To verify the effectiveness of expression guidance, we make another experiment to show the results by using combined identity and expression guidance. As illustrated in row 3 of Figure 7, the addition of the expression guidance leads to a better preservation of the attribute from the target images when compared to the results with identity guidance only. This is also evidenced in Table 3. After adding the identity guidance only, the ID Cos becomes better, but the expression error becomes larger. In contrast, the use of both guidance terms can improve the preservation of expressions of target images without compromising the performance of identity embedding.

Guidence \ Metric	ID Cos ↑	Exp. Error ↓	Pose. Error↓
w/o	0.424	0.16 ± 0.06	2.40 ± 1.41
ID	0.603	0.18 ± 0.06	2.36 ± 1.35
ID + Exp	0.604	0.11 ± 0.07	2.35 ± 1.42
ID + Exp + Pose	0.614	0.12 ± 0.04	2.24 ± 1.31

Table 3: Quantitative ablation study. 'w/o' signifies that no guidance is implemented during the sampling process of the conditional DDPM.

Pose Guidance. Once we have the pose guidance added to our approach, it can be observed that other attributes in the target images, such as eye gaze, can be preserved better, as illustrated in row 4 of Figure 7. The reason for this result is because fiducial points, e.g., eye centres, are used in the pose guidance. In Table 3, the quantitative results confirm that the pose error is further improved. The ablation study has proved the effectiveness of the high-semantics guidance in our conditional DDPM face swapping model.

5. Failure cases and limitation

Although our method can generate high quality face-swapping results, it still has limitations and could generate some failure cases, as illustrated in Figure 8. Since many facial attributes, e.g., wrinkles

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

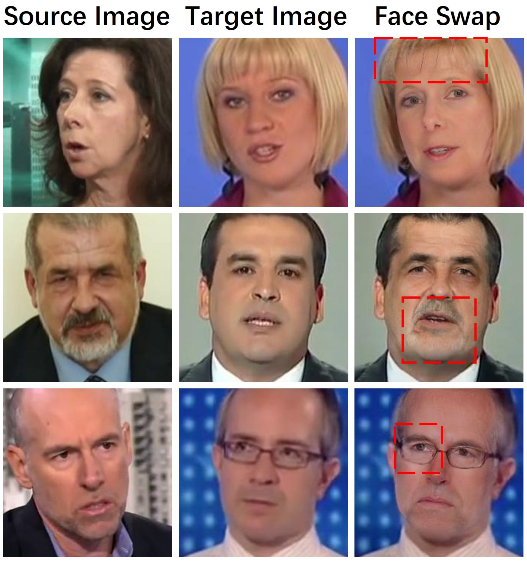

Figure 8: Failure cases of our method. *The first row illustrates the uncompleted swapping at hair region, second row with partial beard transfer, and third row with faded eye glasses.*

or facial hairing, could reflect identity information, our holistic facial model generated uncompleted swapping effects. Additionally, the failure cases may also be due to the fact that 3DMM is a model constructed based on a limited 3D facial dataset. Thus, our model could be further enhanced to improve local facial attribute swapping and synthesis diversity. Last but not the least, our model is adapted from a standard diffusion model, which suffers from the use of huge computational resources and long training time, which makes its fine-tuning difficult.

6. Conclusion

In this paper, we have proposed a novel conditional DDPM for face swapping. The model is implemented by two-level face prior guidance to achieve photo-realistic synthesised face images. We use an image-level condition reconstructed from a 3DMM to ensure the effective convergence of our DDPM. Further, high-level face semantics guidance is applied at the model inference stage to ensure the identity embedding from a source image; and the preservation of other semantic attributes of a target image. Experimental results convincingly demonstrate that our method exhibits superior performance compared to other SoTA methods. This is particularly noticeable in the quality of image synthesis, as well as the faithful preservation and swapping of face attributes and identity, respectively. Nevertheless, despite these substantial improvements, we must recognise the potential for further improvement and expansion of our model. Our future work will focus on refining the high-level semantic guidance and exploring methods to augment the image-level condition more effectively. We will also investigate ways to make the model training more efficient to further improve the performance of our model.

Acknowledgments

This research is supported by the Science and Technology Innovation Program of Hunan Province (2022GK5002), the Special Foundation for Distinguished Young Scientists of Changsha (kq2209003), 111 Project (No. D23006), Foreign Expert Project of China (G2023041039L) and in part by the High Performance Computing Center of Central South University.

References

[ALF22] Avrahami O., Lischinski D., Fried O.: Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18208–18218. 3

[BCW*18] Bao J., Chen D., Wen F., Li H., Hua G.: Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 6713–6722. 2

[BKD*08] Bitouk D., Kumar N., Dhillon S., Belhumeur P., Nayar S. K.: Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*. 2008, pp. 1–8. 2

[BSSE21] Batzolis G., Stanczuk J., Schönlieb C.-B., Etmann C.: Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606* (2021). 2

[BSVS04] Blanz V., Scherbaum K., Vetter T., Seidel H.-P.: Exchanging faces in images. In *Computer Graphics Forum* (2004), vol. 23, Wiley Online Library, pp. 669–676. 2

[BV99] Blanz V., Vetter T.: A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (1999), pp. 187–194. 2

[BV03] Blanz V., Vetter T.: Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence 25*, 9 (2003), 1063–1074. 2

[CCNG20] Chen R., Chen X., Ni B., Ge Y.: Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), pp. 2003–2011. 2, 7, 8

[CSX*18] Cao Q., Shen L., Xie W., Parkhi O. M., Zisserman A.: Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (2018), IEEE, pp. 67–74. 7

[CTL*09] Cheng Y.-T., Tzeng V., Liang Y., Wang C.-C., Chen B.-Y., Chuang Y.-Y., Ouhyoung M.: 3d-model-based face replacement in video. In *SIGGRAPH'09: Posters*. 2009, pp. 1–1. 2

[CWD*18] Creswell A., White T., Dumoulin V., Arulkumaran K., Sengupta B., Bharath A. A.: Generative adversarial networks: An overview. *IEEE signal processing magazine 35*, 1 (2018), 53–65. 2

[Dee12] Deepfakes. *Deepfakes. https://github.com/deepfakes/faceswap* (2018-12). 7, 8

[DGV*20] Deng J., Guo J., Ververas E., Kotsia I., Zafeiriou S.: Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 5203–5212. 5, 6

[DGXZ19] Deng J., Guo J., Xue N., Zafeiriou S.: Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), pp. 4690–4699. 5

[DN21] Dhariwal P., Nichol A.: Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems 34* (2021), 8780–8794. 2, 3, 5

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

[DYX*19] DENG Y., YANG J., XU S., CHEN D., JIA Y., TONG X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2019), pp. 0–0. 4

[EST*20] EGGER B., SMITH W. A., TEWARI A., WUHRER S., ZOLLHOEFER M., BEELER T., BERNARD F., BOLKART T., KORTYLEWSKI A., ROMDHANI S., ET AL.: 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (ToG) 39*, 5 (2020), 1–38. 8

[Fac12] Faceswap. *FaceSwap. https://github.com/MarekKowalski/FaceSwap.* (2016-12). 7, 8

[HJA20] HO J., JAIN A., ABBEEL P.: Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems 33* (2020), 6840–6851. 2, 3, 4

[JLW*20] JIANG L., LI R., WU W., QIAN C., LOY C. C.: Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 2889–2898. 2

[KALL17] KARRAS T., AILA T., LAINE S., LEHTINEN J.: Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017). 7

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 7

[LBY*19] LI L., BAO J., YANG H., CHEN D., WEN F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457* (2019). 2, 8

[LPG*23] LIU K., PEROV I., GAO D., CHERVONIY N., ZHOU W., ZHANG W.: Deepfacelab: Integrated, flexible and extensible face-swapping framework. *Pattern Recognition 141* (2023), 109628. 2

[MHS*21] MENG C., HE Y., SONG Y., SONG J., WU J., ZHU J.-Y., ERMON S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations* (2021). 3

[ML21] MIRSKY Y., LEE W.: The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR) 54*, 1 (2021), 1–41. 2

[ND21] NICHOL A. Q., DHARIWAL P.: Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning* (2021), PMLR, pp. 8162–8171. 2

[NDR*21] NICHOL A., DHARIWAL P., RAMESH A., SHYAM P., MISHKIN P., MCGREW B., SUTSKEVER I., CHEN M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021). 3

[NHSC19] NIU X., HAN H., SHAN S., CHEN X.: Multi-label co-regularization for semi-supervised facial action unit recognition. *Advances in neural information processing systems 32* (2019). 5

[NKH19] NIRKIN Y., KELLER Y., HASSNER T.: Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 7184–7193. 2

[NKH22] NIRKIN Y., KELLER Y., HASSNER T.: Fsganv2: Improved subject agnostic face swapping and reenactment. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*, 1 (2022), 560–575. 7, 8

[NMT*18] NIRKIN Y., MASI I., TUAN A. T., HASSNER T., MEDIONI G.: On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (2018), IEEE, pp. 98–105. 2

[NNN*22] NGUYEN T. T., NGUYEN Q. V. H., NGUYEN D. T., NGUYEN D. T., HUYNH-THE T., NAHAVANDI S., NGUYEN T. T., PHAM Q.-V., NGUYEN C. M.: Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding 223* (2022), 103525. 2

[PCWS22] PREECHAKUL K., CHATTHEE N., WIZADWONGSA S., SUWAJANAKORN S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10619–10629. 2

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., ET AL.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (2019). 7

[PPSRHR16] PÉREZ-PELLITERO E., SALVADOR J., RUIZ-HIDALGO J., ROSENHAHN B.: Psyco: Manifold span reduction for super resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 1837–1845. 4

[RBL*22] ROMBACH R., BLATTMANN A., LORENZ D., ESSER P., OMMER B.: High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695. 2, 3

[RCR18] RUIZ N., CHONG E., REHG J. M.: Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018), pp. 2074–2083. 6, 8

[RCV*19] ROSSLER A., COZZOLINO D., VERDOLIVA L., RIESS C., THIES J., NIESSNER M.: Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (2019), pp. 1–11. 8

[SCS*22] SAHARIA C., CHAN W., SAXENA S., LI L., WHANG J., DENTON E. L., GHASEMIPOUR K., GONTIJO LOPES R., KARAGOL AYAN B., SALIMANS T., ET AL.: Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems 35* (2022), 36479–36494. 2

[SDWMG15] SOHL-DICKSTEIN J., WEISS E., MAHESWARANATHAN N., GANGULI S.: Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning* (2015), PMLR, pp. 2256–2265. 2

[SHC*22] SAHARIA C., HO J., CHAN W., SALIMANS T., FLEET D. J., NOROUZI M.: Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). 3, 5

[TZS*16] THIES J., ZOLLHOFER M., STAMMINGER M., THEOBALT C., NIESSNER M.: Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 2387–2395. 2

[VA19] VEMULAPALLI R., AGARWALA A.: A compact embedding for facial expression similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5683–5692. 6, 8

[Ver20] VERDOLIVA L.: Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing 14*, 5 (2020), 910–932. 2

[WCZ*21] WANG Y., CHEN X., ZHU J., CHU W., TAI Y., WANG C., LI J., WU Y., HUANG F., JI R.: Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965* (2021). 2, 7, 8

[WWZ*18] WANG H., WANG Y., ZHOU Z., JI X., GONG D., ZHOU J., LI Z., LIU W.: Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 5265–5274. 8

[XYH*21] XU Z., YU X., HONG Z., ZHU Z., HAN J., LIU J., DING E., BAI X.: Facecontroller: Controllable attribute editing for face in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2021), vol. 35, pp. 3083–3091. 2

[YLL19] YANG X., LI Y., LYU S.: Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019), IEEE, pp. 8261–8265. 5

[ZFW*20] ZHU H., FU C., WU Q., WU W., QIAN C., HE R.: Aot: Appearance optimal transport based identity swapping for forgery detection. *Advances in Neural Information Processing Systems 33* (2020), 21699–21712. 2

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

[ZGC*21] ZHANG Z., GE Y., CHEN R., TAI Y., YAN Y., YANG J., WANG C., LI J., HUANG F.: Learning to aggregate and personalize 3d face from in-the-wild photo collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 14214–14224. 8

[ZLW*21] ZHU Y., LI Q., WANG J., XU C.-Z., SUN Z.: One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 4834–4844. 2

© 2023 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.