StyleIPSB: Identity-Preserving Semantic Basis of StyleGAN for High Fidelity Face Swapping

Diqiong Jiang [1], Dan Song[2*], Ruofeng Tong[1*], Min Tang[1]
[1]Zhejiang University, China [2]Tianjin University, China

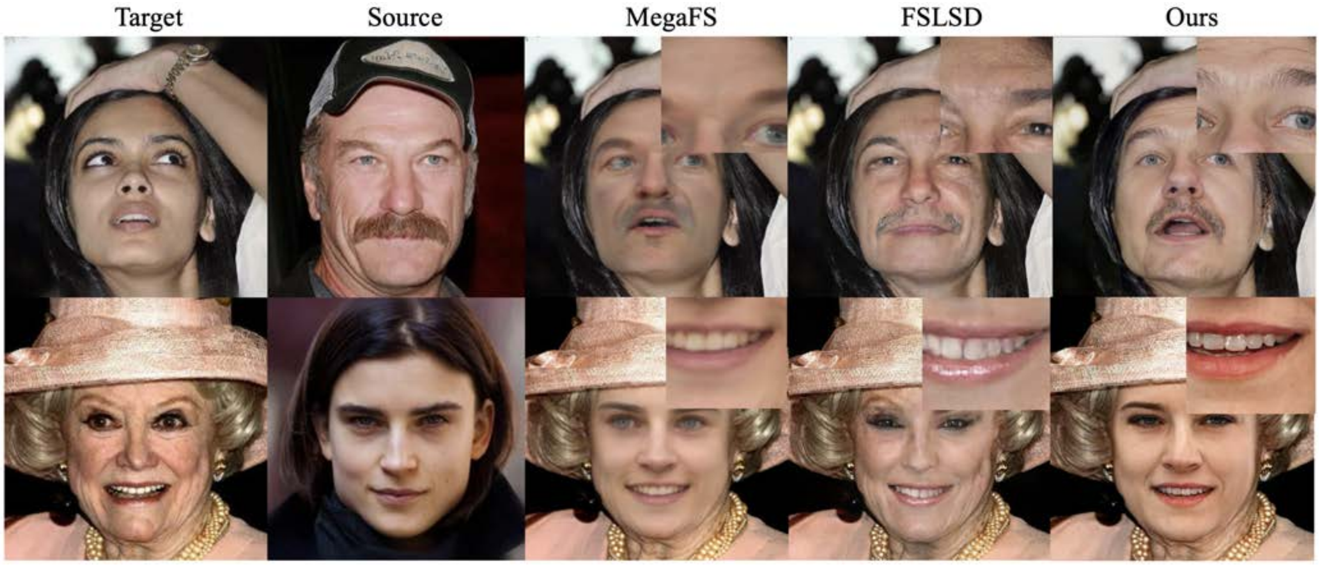

Figure 1. Compared with the existing high-fidelity face swapping methods MageFS [51] and FSLDS [47], our face swapping framework based on the identity-preserving semantic basis (StyleIPSB) can preserve the pore-level detail and source image's identity.

Abstract

Recent researches reveal that StyleGAN can generate highly realistic images, inspiring researchers to use pre-trained StyleGAN to generate high-fidelity swapped faces. However, existing methods fail to meet the expectations in two essential aspects of high-fidelity face swapping. Their results are blurry without pore-level details and fail to preserve identity for challenging cases. To overcome the above artifacts, we innovatively construct a series of identity-preserving semantic bases of StyleGAN (called StyleIPSB) in respect of pose, expression, and illumination. Each basis of StyleIPSB controls one specific semantic attribute and disentangles with the others. The StyleIPSB constrains style code in the subspace of W+ space to preserve pore-level details and gives us a novel tool for high-fidelity face swapping, and we propose a three-stage framework for face swapping with StyleIPSB. Firstly, we transform the target facial images' attributes to the source image. We learn the mapping from 3D Morphable Model (3DMM) parameters, which capture the prominent semantic variance, to the coordinates of StyleIPSB that show higher identity-preserving and fidelity. Secondly, to transform detailed attributes which 3DMM does not capture, we learn the residual attribute between the reenacted face and the target face. Finally, the face is blended into the background of the target image. Extensive results and comparisons demonstrate that StyleIPSB can effectively preserve identity and pore-level details. The results of face swapping can achieve state-of-the-art performance. We will release our code at https://github.com/a686432/StyleIPSB

1. Introduction

Facial image manipulation [36, 37, 48, 50] is a task of transforming specific attributes from the source image to the target image while persevering other attributes unchanged. Face swapping is one of the essential parts of facial im-

[*]Co-Corresponding authors

[†]The research is supported in part by the National Natural Science Foundation of China (61972342,61832016,61972341,61902277)

age manipulation, which has attracted lots of interest in the computer vision and graphics community. Face swapping aims to generate an image with the source image's identity and the target image's attributes (e.g., expression, pose, background, hair, etc.). It has wide applications in the film industry and computer games.

Current face swapping methods are mainly divided into two categories: 3D model-based methods and 2D image-based methods. 3D model-based methods [11,17,40] firstly reconstruct the 3D face models based on 3DMM from the source image and target image and transfer the non-identity parameters of the target face model to the source face model. Then they render the transferred 3D model and blend it into the target image. Although such methods can transfer coarse facial attributes such as pose, expression, and illumination, they have difficulty in generating realistic hair and teeth accessories.

With the development of generative adversarial networks, 2D image-based methods [6, 7, 22, 27, 28, 45] can synthesize photo-realistic images. The generated face images have convincing and detailed facial attributes, such as mouth, teeth, and eyebrows. Recent works [21,46,47,51] employ the pre-trained StyleGAN decoder to further improve the fidelity and synthesize pore-level details. However, as shown in Fig. 1, despite using the pre-trained StyleGAN model, their results fail to generate pore-level details and identity-preserving in challenge conditions. Overall, 2D image-based methods generate more realistic images than 3D model-based methods, but the identity-preserving and pore-level details of the images still need improvement.

To tackle the challenges of blurry images and identity-preserving in face swapping, according to our observations, the cause of the blurred images is that the regressed style code is out of W+ space. Additionally, as mentioned in [18], identity embedding is a non-smooth space, so finding the identity-preserving optimized direction is challenging. To address these problems, our method constrains the regressed style code with identity-preserving semantic bases of StyleGAN (i.e., the proposed StyleIPSB). StyleIPSB stays within the W+ space, and the identity is preserved when changing its coordinates.

The advantages of the proposed StyleIPSB are summarized as follows: (1) StyleIPSB constitutes a linear space, which is the subspace of the W+ space of StyleGAN. By ensuring the regressed style code within the W+ space of StyleGAN, we can more easily generate images with pore-level details. (2) When changing the coordinates of the StyleIPSB, the identity remains preserved as much as possible. (3) StyleIPSB can represent various poses, expressions, and illuminations. To construct the basis that satisfies the above properties, we propose a novel identity-preserving distance metric to find the orthogonal semantic directions, which are further assembled to StyleIPSB.

StyleIPSB also cooperates well with 3DMM to control facial attributes. StyleRig [39] builds the mapping of the 3DMM parameter space and W+ space of StyleGAN, which can change the facial attribute of the generated image by the 3DMM parameters. StyleRig only can manipulate images generated by StyleGAN. Pie [38] designs a non-linear optimization problem to edit the real-world image based on StyleRig, but the optimization operation is time-consuming. GIF [12] generates face images by the FLAME [23] parametric control. However, the generated images are easy to contain artifacts and change identity. Other face manipulation methods [4, 29, 37] use the network directly to find the edit direction. Still, without the guidance of 3DMM, they can only generate some basic expressions (e.g., smile) and fail to cover various expressions. In this paper, we propose the StyleGAN-3DMM mapping network, which transforms the semantic information of 3DMM parameters into StyleIPSB coordinates. The StyleGAN-3DMM mapping network reduces the gap between 3DMM and StyleIPSB. It shows that StyleIPSB is very compatible with 3DMM.

In summary, we propose a face swapping framework based on StyleIPSB and achieve state-of-the-art results. The main contributions of this paper lie in the following three aspects:

- We propose a novel method of establishing identity-preserving semantic bases of StyleGAN called StyleIPSB. The face image, generated by the linear space of StyleIPSB, remains pore-level details and identity-preserving.
- The proposed StyleGAN-3DMM mapping network serves as the bridge to narrow the gap between 3DMM and StyleIPSB, which can take advantage of the prominent semantic variance of 3DMM and the identity-preserving and high-fidelity of styleIPSB.
- We propose the face swapping framework based on StyleIPSB and StyleGAN-3DMM mapping network. Extensive results show our method outperforms others in detail-preserving and identity-preserving.

2. Related works

2.1. Image Modification Using StyleGAN

StyleGAN [14–16] is a powerful image synthesis model that can generate a wide variety of high-quality face images. Some methods [5, 31, 34, 35] find the particular attribute editing direction in W+ space using the corresponding attribute classifier. However, their attribute editing is limited by the classifier's ability and cannot control diverse expressions and illumination. To facilitate attribute manipulation in an unsupervised manner, GANSpace [13] and SeFa [36] perform decomposition to find primary directions in the latent space and explore the interpretable directions among the primary directions. CLIP2StyleGAN [4] discovery and label of StyleGAN edit-direction based on CLIP image space. However, they only find a few types of expressions and illumination and fail to discover a complete set of expressions and illumination. Some works [12,38,39] map the rigging information to face manipulation, but our StyleIPSB can generate more detail and identity-preserving results.

2.2. Face swap

In the early years, many works [24–26, 40] apply 3D face models for face swapping. Face2face [40] applies an efficient deformation transfer to track the source video facial expressions and re-render the synthesized faces with retrieved and warped mouth interiors. Ma et al. [25] reconstructed high-resolution facial geometry and appearance by capturing an individual-specific face model with fine-scale details. Those 3D model-based methods are difficult to generate high-fidelity face images, especially realistic hair, mouth, and teeth. In recent years, conditional GAN architecture has been widely used in face-swapping. Many works [7, 19, 21, 22, 28, 46, 49] use neural networks to generate high-fidelity images. SimSwap [7] and FaceShifter [22] use a face recognition network to extract the identity embedding and use a decoder to fuse the identity embedding. Recently, many methods [21, 45–47, 51] use pretrained StyleGAN as the image generator to improve the quality of generated images further. However, the images generated by their method still suffer from the lack of details and face identity shift. Our approach is better at retaining pore-level details and identify-preserving.

3. Method

This section reveals the details of our proposed method in three aspects. The first part describes the process of conducting StyleIPSB. The second part describes how to use StyleIPSB to transfer the facial attributes except for identity from the target image to the source face. The third introduces the overall framework of face swap based on StyleIPSB.

3.1. StyleIPSB Construction

This subsection aims to find a semantic basis in the W+ space of StyleGAN that satisfies the conditions mentioned in the introduction. In short, StyleIPSB should meet the following properties: a subspace of W+ space, identity-preserving, and representation ability. The previous methods fail to satisfy the above three conditions. For example, as shown in Tab. 1, GIF [12] does not meet the requirements in the W+ subspace, which would lead to generated image distortion. Some methods [12,35,36] do not concern the identity preservation and suffer the identity shift when editing. InterFaceGAN [34] cannot edit various expressions and illumination. Unlike existing methods, our method considers all three conditions when we build StyleIPSB.

Method	Subspace	ID	Representation
InterFaceGAN [34]	√	√	
GANSpace [13]	√		
SeFa [36]	√		
GIF [12]			√
Our	√	√	√

Table 1. The conditions of the basis satisfied by different methods. We compare our method with InterFaceGAN [34], GANSpace [13], GIF [12] and SeFa [36].

Compared with the previous methods, which built the basis by directly performing decomposition on eigenvalue in W+ space, we added semantic metrics and identity loss in our decomposition. Therefore, StyleIPSB is not only in the W+ subspace but can also represent the complete set of semantic information and have the property of identity preservation. So adding semantic information and maintaining identity when building StyleIPSB is the crucial point to this subsection. The following introduces the detailed process of conducting StyleIPSB.

Formulation The StyleGAN network, denoted by G, is a mapping from style code w to image I, $G : R^n \rightarrow R^{H\times W\times 3}, w \mapsto I$. The 3DMM Fitting network, denoted by M, regresses the 3DMM parameters p from image I. The parameters include pose parameter p_p, expression parameter p_e and illumination parameter p_i. In this paper, we use DECA [10] as our 3DMM Fitting network.

We first define the distance metric of the pose, expression, and illumination. The distance metrics measure the attribute difference between two images generated by style codes w_1, w_2. In addition, we add identity loss to the metric. Therefore, the distance metric increases if the difference between two images' attributes increases and the identity remains similar. Then we decompose the Hessian matrix to find the direction with the fastest distance metric change in the W+ space. Therefore, the attributes change fast but identity changes slowly along the direction we found. The distance metrics are shown as follows:

$$\begin{aligned} D_p(w_1, w_2) &= \frac{||M(G(w_1))_p - M(G(w_2))_p||^2}{L_{id}(G(w_1), G(w_2))} \\ D_e(w_1, w_2) &= \frac{||M(G(w_1))_e - M(G(w_2))_e||^2}{L_{id}(G(w_1), G(w_2))} \\ D_i(w_1, w_2) &= \frac{||M(G(w_1))_i - M(G(w_2))_i||^2}{L_{id}(G(w_1), G(w_2))} \end{aligned} \tag{1}$$

where L_{id} measures the similarity of the identities of two images. Its value represents the similarity of two identity

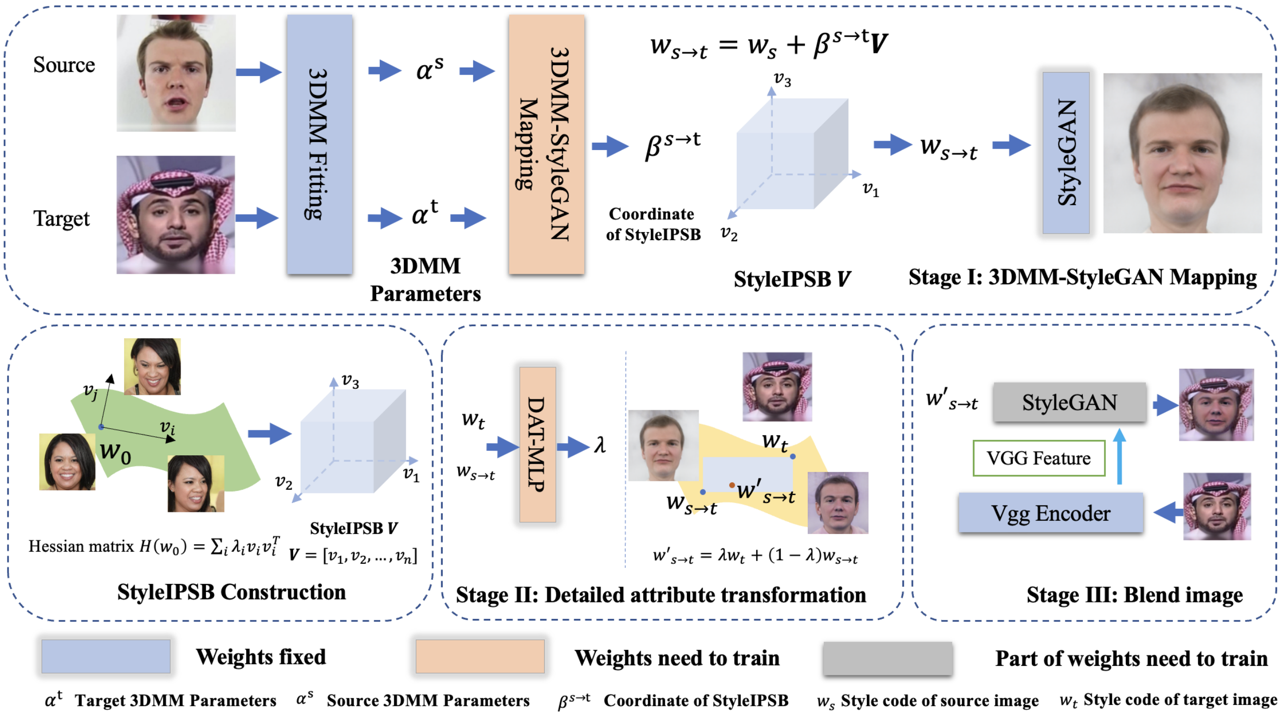
Source
Target
3DMM Fitting
α^s
α^t
3DMM Parameters
3DMM-StyleGAN Mapping
$w_{s\to t} = w_s + \beta^{s\to t} V$
$\beta^{s\to t}$
Coordinate of StyleIPSB
v_1
v_2
v_3
StyleIPSB V
$w_{s\to t}$
StyleGAN
Stage I: 3DMM-StyleGAN Mapping
v_j
v_i
w_0
StyleIPSB V
Hessian matrix $H(w_0) = \sum_i \lambda_i v_i v_i^T$
$V = [v_1, v_2, \ldots, v_n]$
StyleIPSB Construction
w_t
$w_{s\to t}$
DAT-MLP
λ
$w'_{s\to t}$
$w'_{s\to t} = \lambda w_t + (1-\lambda) w_{s\to t}$
Stage II: Detailed attribute transformation
StyleGAN
VGG Feature
Vgg Encoder
Stage III: Blend image
Weights fixed
Weights need to train
Part of weights need to train
α^t Target 3DMM Parameters
α^s Source 3DMM Parameters
$\beta^{s\to t}$ Coordinate of StyleIPSB
w_s Style code of source image
w_t Style code of target image

Figure 2. Our framework is divided into three stages. Stage one transfers the target image's expression, pose, and illumination attributes to the source image through the semantic base. Stage two transfers detailed non-identity attributes. Stage three blends the transferred image into the background of the target image.

embeddings obtained by the pre-trained face recognition network [8] from two input face images.

$$L_{id}(I_1, I_2) = 1 - \frac{ID(I_1) \cdot ID(I_2)}{||ID(I_1)||_2\,||ID(I_2)||_2} \tag{2}$$

Then, we use the Hessian matrix to encode the local distance information on the image manifold up to second-order approximation.

$$D^2(w_0, w_0 + \delta w) \approx ||\delta w||_H^2 = \delta_w^T H(w_0)\delta w \tag{3}$$

The w_0 used to construct the semantic base is randomly sampled, and $w_0 + \delta w$ is the point near w_0. We decompose the Hessian matrix $H(w_0) = \sum_i \lambda_i v_i v_i^T$ and find the first m principal components of v as our basis V. Therefore, the basis V contains the direction vectors, along which the distance metric changes first m fastest.

We use the algorithm [42] to calculate and decompose the Hessian matrix. StyleGAN has 18 different levels, and the style code of each level itself contains certain semantic features (for example, the low level includes the features of face shape and pose, and the high level has the skin color and illumination, etc.). We construct the semantic base of the pose on the first three levels of the style code and the expression on the 4th to 10th levels. The semantic base of the illumination is on the levels after the 10th. We sampled w_0 100 times and generated 100 sets of bases. And then, we average these bases and performed Schmidt orthogonalization to obtain the semantic base of the pose V_p, expression V_e, and illumination V_i, respectively. Finally, the pose, expression, and illumination base are combined as StyleIPSB $V = [V_p, V_e, V_i]$

3.2. 3DMM-StyleGAN Mapping

As shown in Fig. 2, the source and target images are fed into the 3DMM fitting network to regress the 3DMM parameters α^s, α^t. The 3DMM-StyleGAN Mapping network, which contains three six-layer multilayer perceptrons(MLPs), maps 3DMM parameters α to StyleIPSB coordinates of pose $\beta^p_{s\to t}$, expression $\beta^e_{s\to t}$ and illumination $\beta^i_{s\to t}$, respectively. The following equations calculate the transferred style code $w_{s\to t}$:

$$w_{s\to t} = w_s + \beta_{s\to t} V \tag{4}$$

where $\beta_{s\to t} = [\beta^p_{s\to t}, \beta^e_{s\to t}, \beta^i_{s\to t}]$ and w_s is the corresponding style code of the source image, which is obtained from the source image by the StyleGAN encoder [41]. Finally, $w_{s\to t}$ passes through StyleGAN to get the attribute transferred image.

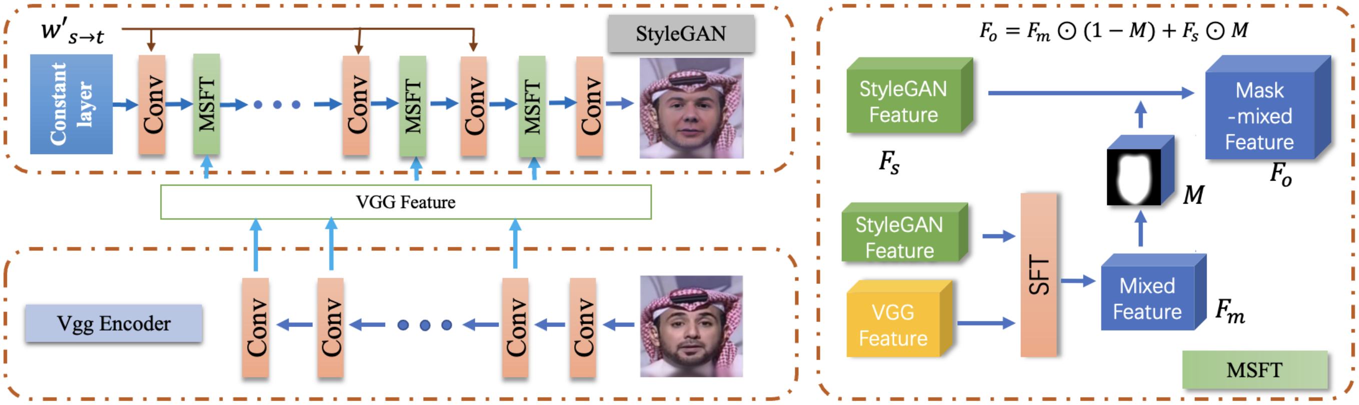

Figure 3. The network structure of the image blend. Stage three blends transformed image to the target image by mixing the VGG encoder feature with the StyleGAN feature through masked spatial feature transforms (MSFT).

3DMM-StyleGAN mapping network aims at precisely controlling the generated attribute through the 3DMM parameters while keeping the identity unchanged. Therefore, we design the attribute loss L_{attr} to make the generated face image have the same attribute as the pose, expression, and illumination as target 3DMM parameters. And we add identity loss L_{id} to preserve the identity of the source image. The total loss is shown as:

$$L = L_{id} + \varepsilon_{attr} L_{attr} \tag{5}$$

where L_{id} measure the identity similarity of two images and defined in Eq. (2). ε_{attr} is the weight of two loss functions. L_{attr} measures the difference between the attributes of the transferred face and the target face. We reconstruct the 3D face from the transferred face and the target face, and then compare their difference in 3D face geometries and rendered images:

$$\begin{aligned} L_{attr}(\alpha^s, \alpha^t, \alpha^{s\to t}) = L_{geo}(\alpha^s, \alpha^{s\to t}) \\ + L_{render}(\alpha^s, \alpha^t, \alpha^{s\to t}) \end{aligned} \tag{6}$$

where α^s,α^t is the 3DMM parameters of the source and target images. And $\alpha^{s\to t}$ is 3DMM parameters of the transferred image where $\alpha^{s\to t} = M(G(w_{s\to t}))$.

We define that $G_{3DMM}(\alpha_s, \alpha_e, \alpha_p)$ can generate 3D geometry from 3DMM parameters, and $R(\alpha_s, \alpha_e, \alpha_a, \alpha_i, \alpha_p)$ can generate rendered images from 3DMM parameters, where α_s,α_e,α_a,α_i,α_p represent shape parameters, pose parameters, expression parameters, albedo parameters, and lighting parameters, respectively. Because we only compare attribute differences between transferred and target images, we reconstruct the 3D face by the same 3DMM shape parameters.

The geometric term L_{geo} uses the L_2 loss between two face meshes:

$$\begin{aligned} & L_{geo}(\alpha^s, \alpha^{s\to t}) = \\ & \frac{1}{N}\left\|G_{3DMM}(\alpha^s_s, \alpha^t_e, \alpha^t_p) - G_{3DMM}(\alpha^{s\to t}_s, \alpha^{s\to t}_e, \alpha^{s\to t}_p)\right\|_2 \end{aligned} \tag{7}$$

where N is the number of vertices of the face mesh.

The render term L_{render} uses the L_1 loss between two rendered images:

$$\begin{aligned} & L_{render}(\alpha^s, \alpha^t, \alpha^{s\to t}) = \\ & \left\|R(\alpha^s_s, \alpha^t_e, \alpha^t_a, \alpha^t_i, \alpha^t_p) - R(\alpha^{s\to t}_s, \alpha^{s\to t}_e, \alpha^t_a, \alpha^{s\to t}_i, \alpha^{s\to t}_p)\right\|_1 \end{aligned} \tag{8}$$

3.3. Detailed Attribute Transformation

In the previous subsection, we transfer the attributes of the target face to the source face, including pose, expression, and illumination. In this subsection, to transfer attributes beyond the 3DMM expressive capabilities, we use DAT-MLP to transfer more detailed attributes.

First, we project the target images into the $W+$ space using the StyleGAN encoder to obtain the style code w_t. Then, DAT-MLP regresses λ from style codes $w_t, w_{s\to t}$ using six-layer MLP. Finally, we use the following formula to get the latent vector $w'_{s\to t}$ that contains non-identity attributes of the target image:

$$w'_{s\to t} = \lambda w_{s\to t} + (1 - \lambda) w_t \tag{9}$$

where $\lambda \in R^{18\times512}$ is constrained to be between 0 and 1 through the sigmoid activation layer. We want to transfer attributes of w_t into $w'_{s\to t}$ as many as possible. The loss function of the second stage is as follows:

$$L = L_{id} + \varepsilon_p L_p + \varepsilon_{attr} L_{attr}(\alpha^s, \alpha^t, M(G(w'_{s\to t}))) \tag{10}$$

The loss function is divided into three items. The first item is face recognition loss, which is used to constrain the identity of the generated face to be consistent with the source image. The second item L_p represents the perceptual distance between the generated image and the target image. It can help transfer residual attributes other than pose, expression, and illumination. The third term L_{attr} constrains the attributes of the generated face consistent with the target image. $\varepsilon_p, \varepsilon_{attr}$ represent the weight of the loss function.

The Eq. (9) shows that the style code $w'_{s\to t}$ is in a bounding box of two style codes $w_{s\to t}$ and w_t. We consider that most of the bounding box is still in the subspace of W+ space because the attribute of two style codes $w_{s\to t}$ and w_t is very close and generated image retains pore-level details in practice.

3.4. Blend Image

This subsection introduces how to blend the generated image into the target image. We propose Masked Spatial Feature Transform (MSFT) module to fuse the feature. Unlike traditional Spatial Feature Transform (SFT) [44], MSFT only fuses the feature of the masked regions.

The detail of this module is shown in Fig. 3. The VGGFace network [30] extracts various level features of the target image, which is injected into the StyleGAN through the MSFT module. We use Gaussian filtering to filter the masked image obtained by the face segmentation algorithm to make the boundary smoother. The StyleGAN feature is retained in the facial area, while the background area is from the mixed feature. We mix features of StyleGAN and VGGFace using SFT. The background loss and the perceptual loss make the swapped image have the same background as the target image:

$$L = L_b + \varepsilon_p L_p \tag{11}$$

L_b is the background loss used to measure the difference in the background between the generated image and the target image. L_p represents the perceptual distance between the generated image and the target image. ε_p is the weight of the loss function. The background loss is expressed as:

$$L_b = \| M \odot (I_r - I_t) \| \tag{12}$$

where M is the binarized image of the background region obtained by face segmentation algorithm [2].

4. Experiment

As mentioned above, StyleIPSB has the following advantages: (1) StyleIPSB can represent various poses, expressions, and illumination properties. (2) When the style code is modified along one base, it only changes the specific attribute while remaining identity unchanged. (3) StyleIPSB can preserve the pore-level details. (4) StyleIPSB cooperates well with 3DMM to control facial attributes. Therefore, in this section, we mainly evaluate the effectiveness of our method in the following aspects: (1) Evaluating the properties of StyleIPSB. (2) Evaluating the performance of 3DMM controlling facial attributes with StyleIPSB. (3) Comparison of face swapping results with other methods. First of all, we introduce the dataset and training detail.

4.1. Dateset and Training Detail

Dataset. We used the FFHQ database [15] to train the 3DMM StyleGAN Mapping module. Flickr-Faces-HQ (FFHQ) consists of 70,000 high-quality face images at 1024×1024 resolution. CelebAMask-HQ [20] is a large-scale face image dataset with 30,000 high-resolution face images selected from the CelebA dataset by following CelebA-HQ. FaceForensics++ [32] is a forensics dataset consisting of 1,000 original video sequences.

Training environment and hyperparameter We train the network on a GTX 3090 using the Pytorch framework. In training, our optimizer is Adam, the weight decay is 5e-5, and the batch size is 2. In order to effectively train the above three stages, we first pre-train the three stages separately and finally train them together. In the three-stage pre-training, the learning rate is 2e-4; in the final together training stage, we adjust the learning rate to 2e-5.

4.2. The Properties of StyleIPSB

The representation and the disentanglement of StyleIPSB. We conduct experiments on the representation of StyleIPSB to evaluate whether it can represent various poses, expressions, and illumination properties while identity is unchanged. Fig. 4 shows that StyleIPSB can express pitch, yaw rotations, and lighting with different colors, directions, and intensities. In terms of expressions, StyleIPSB can not only control whether the mouth is open or not, the eyes are open or closed but also control different eyeball orientations and raising eyebrows. Supplementary material reveals the effect of modifying coordinates in different base directions.

To evaluate the disentanglement of StyleIPSB, we present the quantitative results, which show the identity change when manipulating the pose. As shown in Fig. 5, the horizontal axis represents the yaw angle between the anchored face and the edited image, and the vertical axis represents identity loss as defined in Eq. (2) between the anchored face and the edited image. Our StyleIPSB outperforms Ganspace [13], which is also a basis constructed in W+ space using unsupervised learning. We achieve comparable results with the supervised method. InterfaceGAN [34]. It needs a dataset that contains left and right-facing faces. StyleIPSB uses unsupervised learning and can represent more attributes than InterfaceGAN. Fig. 5 also shows that if we introduce identity loss into the distance metric when building StyleIPSB, the identity-preserving perfor-

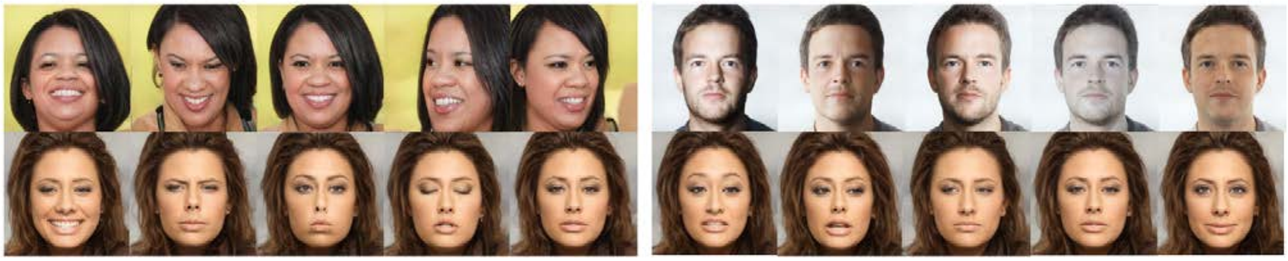

Figure 4. StyleIPSB can modify various poses, expressions, and illumination by adjusting its coordinates while keeping the identity unchanged. The first row is the results of changing coordinates of pose and illumination basis. The second row shows the results of the changing coordinate of the expression basis.

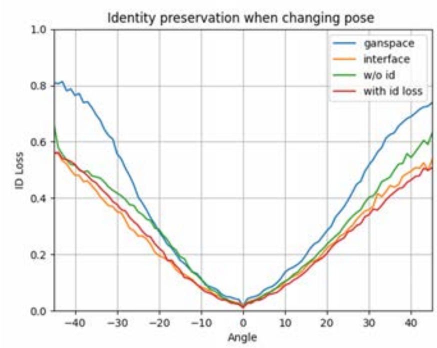

Figure 5. Effects of different methods of changing face pose attributes on face identity information

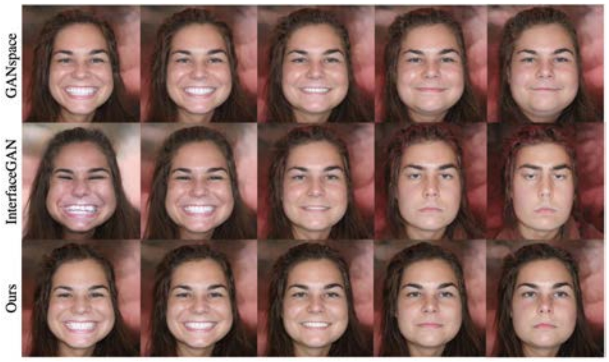

Figure 6. We compare our method with GANSpace [13] and InterfaceGAN [34] in the editing of facial expression.

mance of StyleIPSB is enhanced.

Editing results of StyleIPSB. We compare StyleIPSB with other StyleGAN-based face edit methods. Fig. 6 compares the edit results of expression (smile) with other methods. The results show that our approach has better performance in identity-preserving. StyleIPSB preserves the face shape when expression changes, while face shape is deformed in other methods.

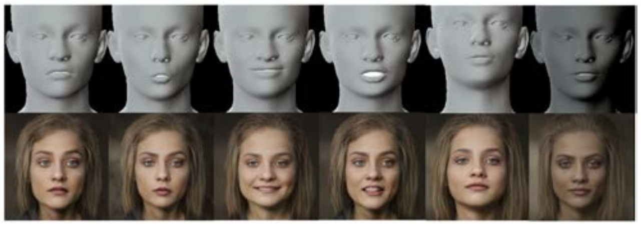

Figure 7. 3D face models control the attribute of generated images. The top is the rendered images of 3D face geometry. The bottom is the edited images controlled by the 3DMM parameters of the top 3D face geometry.

Fig. 7 shows examples of editing a face image using the 3DMM parameters of FLAME [23]. The first row is the 3D face models generated by the 3DMM parameters. The second row is the edited image controlled by those 3DMM parameters. These results show that StyleIPSB cooperates well with 3DMM to control the facial attributes of the generated image.

4.3. Result of Face Swapping

This subsection evaluates the performance of face swapping based on StyleIPSB. First, the ablation experiments evaluate the influence of StyleIPSB on attribute transfer performance. Then, we compare our face swapping with other methods.

Ablation experiment In the experiment, ”no basis” means that the 3DMM-StyleGAN mapping network directly regresses the $w_{s \to t}$ instead of $\beta_{s \to t}$. As shown in Fig. 8, using StyleIPSB can constrain the generated style code on the subspace of W+ space and preserve the pore-level details.

Tab. 2 shows the quantitative comparison results on the performance of StyleIPSB in stage one. FID (Fréchet inception distance) is used to evaluate the image quality generated by the image generation model. ”Exp” and ”Pose” represent the Euclidean distance between the target image and the transferred image expression and pose parameters, respectively. ID Similarity means the cosine similarity be-

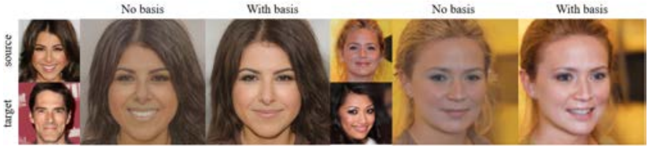

Figure 8. The qualitative ablation experiments on StyleIPSB. The results show we can generate pore-level details with StyleIPSB.

	FID ↓	Exp ↓	Pose ↓	ID similarity ↑
No Basis	26.06	3.73	**0.074**	0.65
With basis	**22.15**	**3.37**	0.078	**0.67**

Table 2. The quantitative ablation experiments on StyleIPSB.

tween the identity embedding of the source image and the transferred image. D3FR [9] is used to extract the 3D face parameters, and CosFace [43] is used to extract the identity embedding. The test set is 10,000 image pairs randomly sampled from the CelebAMask-HQ [20]. The Tab. 2 shows that using StyleIPSB can effectively reduce the value of FID because StyleIPSB is in the subspace of W+ space. In addition, the performance of transferring expressions with StyleIPSB is better in identity-preserving because the identity is preserved while changing the coordinates of StyleIPSB.

Tab. 3 shows the quantitative comparison results of different methods on the FaceForensics++ dataset. The dataset is classified into 885 identities, and the identity retrieval indicates the top-1 matching rate of swapped and the source image. We apply CosFace [43] to extract identity embedding. Expressions are measured in the same way as Tab. 2, The pose measurement method uses a pose estimator [33]. We compare our results with FaceSwap [3], DeepFakes [1], FaceShifter [22], MegaFace [51] and FSLDS [47]. As shown in the Tab. 3, our method outperforms in identity-preserving and expression transfer thanks to StyleIPSB's identity-preserving property and the ability to present various expressions. Our performance of pose transfer does not outperform other methods, which may be because StyleIPSB only contains two bases expressing poses in raw and pitch directions, which makes it hard to express very accurate poses. However, the error of the pose is only a few degrees, so our results still look accurate visually.

Fig. 9 shows the results of face swapping under challenging conditions. The results show that our method can preserve the unique noise patterns in portraiture in the red boxes. In the second row of the figure, our method successfully keeps pore-level details and identity while the source and target pose is quite different. Tab. 4 shows that our method has the better image quality and identity preservation ability than other StyleGAN-based methods. The metric is the same as in Tab. 2. We follow RAFS and compare 100k swapped faces for a fair comparison.

Method	ID Retri.(%) ↑	Exp Err. ↓	Pose Err. ↓
FaceSwap [3]	72.69	2.89	2.58
Deepfakes [1]	88.39	3.33	4.64
FaceShifter [22]	90.68	2.82	2.55
MegaFS [51]	90.83	2.92	2.64
FSLDS [47]	90.05	2.79	**2.46**
Ours	**95.05**	**2.23**	3.58

Table 3. The quantitative experiments on FaceForensics++ dataset with other methods.

	FID ↓	Exp ↓	Pose ↓	ID similarity ↑
MageFS [51]	22.03	2.85	**0.043**	0.4837
RAFS [46]	13.25	3.15	-	0.5232
FSLDS [47]	10.01	2.99	0.053	0.4761
Ours	**9.37**	**2.75**	0.078	**0.5378**

Table 4. The quantitative experiments on the CelebAHQ dataset with other methods.

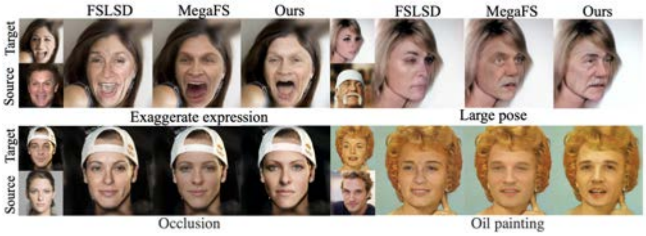

Figure 9. We compare our method with MageFS [51] and FSLDS [47] in challenging conditions.

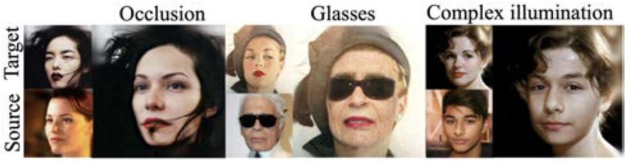

Figure 10. limitation of our method

5. Conclusion and Limitation

We have developed a new semantic basis for face swapping, called StyleIPSB, that is specifically designed to preserve identity and pore-level details. Our experiments have demonstrated that StyleIPSB outperforms other state-of-the-art methods. Despite this, there is still potential for further improvement, as shown in Fig. 10. (1) Occlusion is limited by the mask. (2) The glasses in the source image cannot be removed. (3) Light and shadow cannot be perfectly restored in the case of complex illumination.

References

[1] Deepfakes. https://github.com/ondyari/FaceForensics/tree/master/dataset/DeepFakes. 8

[2] face-parsing.pytorch. https://github.com/zllrunning/face-parsing.PyTorch. 6

[3] Faceswap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. 8

[4] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2StyleGAN: Unsupervised extraction of StyleGAN edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2, 3

[5] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics (TOG)*, 40(3):1–21, 2021. 2

[6] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018. 2

[7] 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, 3

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

[9] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 8

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

[11] Ohad Fried, Ayush Tewari, Michael Zollhöfer, Adam Finkelstein, Eli Shechtman, Dan B Goldman, Kyle Genova, Zeyu Jin, Christian Theobalt, and Maneesh Agrawala. Text-based editing of talking-head video. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2

[12] Partha Ghosh, Pravir Singh Gupta, Roy Uziel, Anurag Ranjan, Michael J Black, and Timo Bolkart. GIF: Generative interpretable faces. In *2020 International Conference on 3D Vision (3DV)*, pages 868–878. IEEE, 2020. 2, 3

[13] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANspace: Discovering interpretable GAN controls. *arXiv preprint arXiv:2004.02546*, 2020. 2, 3, 6, 7

[14] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021. 2

[15] 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. 2, 6

[16] 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. 2

[17] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[18] Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. Smooth-Swap: A simple enhancement for face-swapping with smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10779–10788, 2022. 2

[19] 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. 3

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6, 8

[21] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. FaceInpainter: High fidelity face adaptation to heterogeneous domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5089–5098, 2021. 2, 3

[22] 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, 8

[23] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 7

[24] 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. 3

[25] Luming Ma and Zhigang Deng. Real-time hierarchical facial performance capture. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 1–10, 2019. 3

[26] Luming Ma and Zhigang Deng. Real-time face video swapping from a single portrait. In *Symposium on Interactive 3D Graphics and Games*, pages 1–10, 2020. 3

[27] 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

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGANv2: Improved subject agnostic face swapping and reenactment. *arXiv preprint arXiv:2202.12972*, 2022. 2, 3

[29] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13503–13513, 2022. 2

[30] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*. British Machine Vision Association, 2015. 6

[31] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleClip: Text-driven manipulation of StyleGAN imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 2

[32] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 6

[33] 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. 8

[34] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of GANs for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020. 2, 3, 6, 7

[35] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2, 3

[36] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1532–1540, 2021. 1, 2, 3

[37] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11264, 2022. 1, 2

[38] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2, 3

[39] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. StyleRig: Rigging StyleGAN for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6142–6151, 2020. 2, 3

[40] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2, 3

[41] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 4

[42] Binxu Wang and Carlos R Ponce. The geometry of deep generative image models and its applications. *arXiv preprint arXiv:2101.06006*, 2021. 4

[43] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 8

[44] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 6

[45] 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. 2, 3

[46] Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7632–7641, 2022. 2, 3, 8

[47] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022. 1, 2, 3, 8

[48] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. TransEditor: Transformer-based dual-space GAN for highly controllable facial editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2022. 1

[49] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. MobileFaceSwap: A lightweight framework for video face swapping. *arXiv preprint arXiv:2201.03808*, 2022. 3

[50] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. StyleHeat: One-shot high-resolution editable talking face generation via pretrained StyleGAN. *arXiv preprint arXiv:2203.04036*, 2022. 1

[51] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021. 1, 2, 3, 8