Article

Not peer-reviewed version

# ControlFace: Feature Disentangling for Controllable Face Swapping

Xuehai Zhang , Wenbo Zhou * , Kunlin Liu , Hao Tang , Zhenyu Zhang , Weiming Zhang , Nenghai Yu

Posted Date: 1 December 2023

doi: 10.20944/preprints202312.0066.v1

Keywords: face swapping; feature disentanglement; semantic hierarchy-based feature fusion; controllable identity feature transfer

*Article*

# ControlFace: Feature Disentangling for Controllable Face Swapping

**Xuehai Zhang [1], Wenbo Zhou[1,*], Kunlin Liu [2], Hao Tang[3], Zhenyu Zhang[4], Weiming Zhang[1] and Nenghai Yu [1]**

[1] Department of Cyber Science and Technology, University of Science and Technology of China, Hefei, 230026, Anhui, China; zxh141613683@mail.ustc.edu.cn
[2] ZTE corporation; liu.kunlin@zte.com.cn
[3] Carnegie Mellon University; bjdxtanghao@gmail.com
[4] Nanjing University, Suzhou Campus; zhangjesse@foxmail.com
* Correspondence: welbeckz@ustc.edu.cn

**Abstract:** Face swapping is an intriguing and intricate task in the field of computer vision. Currently, most mainstream face swapping methods employ face recognition models to extract identity features and inject them into the generation process. Nonetheless, such methods often struggle to effectively transfer identity information, result in generated results failing to achieve a high identity similarity with the source face. Furthermore, if we can accurately disentangle identity information, we can achieve controllable face swapping, thereby providing more choices to users. In pursuit of this goal, we propose a new face swapping framework (ControlFace) based on the disentanglement of identity information. We disentangle the structure and texture of the source face, encoding and characterizing them in the form of feature embeddings separately. According to the semantic level of each feature representation, we inject them into the corresponding feature mapper and fuse them adequately in the latent space of StyleGAN. Owing to such disentanglement of structure and texture, we are able to controllablely transfer parts of the identity features. Extensive experiments and comparisons with state-of-the-art face swapping methods demonstrate the superiority of our face swapping framework in terms of transferring identity information, producing high-quality face images and controllable face swapping.

**Keywords:** face swapping; feature disentanglement; semantic hierarchy-based feature fusion; controllable identity feature transfer

## 1. Introduction

Face swapping is a technique that transfers the identity information from a source face to a target face while preserving the identity-independent attributes (e.g., pose, expression, lighting, and background) of the target face. It has a wide range of applications in game production and film creation for creating fictional characters as well as protecting personal privacy. This technique has attracted considerable attention from researchers in the field of computer vision.

The key problem of face swapping technology is identity transfer, *i.e.*, how to precisely and adequately transfer the identity-relevant facial features, comprising both structure and texture, to the target face. Most current methods [1–5] use a pre-trained 2D face recognition network [6] to extract identity features and inject them into a generator to achieve identity transfer. However, due to the difference between face generation and face recognition tasks, the identity information extracted by this network may miss many important facial structure details, like face contour, which can result in swapped faces with structure between the source and target faces. More importantly, previous methods rarely consider the transfer of texture features of the face, such as skin color. We believe that texture, as well as structure, is an important component of face identity information. Such an identity transfer method can lead to huge identity differences between the swapped face and the source face in human visual perception, which can be seen in Figure 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



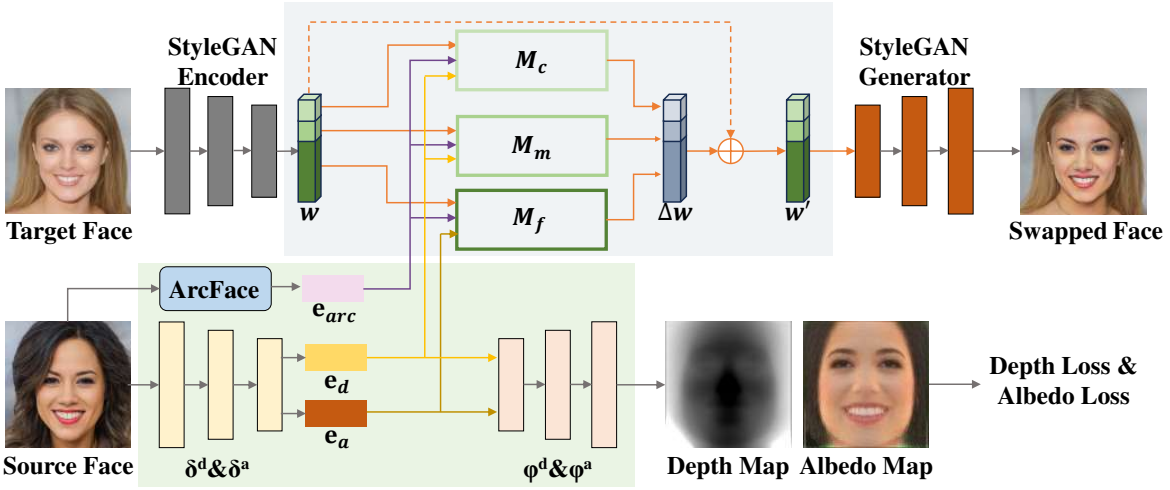


**Figure 1.** Overview of the proposed ControlFace. We extract the identity information of the source face using two 3D autoencoders and a 2D face recognition model and represent it as three identity embeddings. Meanwhile, we obtain the latent code of the target face in the $\mathcal{W}+$ space by the StyleGAN encoder. We inject the identity embeddings into the $\mathcal{W}+$ space according to their semantic levels with the face feature fusion network. Finally, we use the StyleGAN generator to get the swapped face.

Another significant challenge pertains to the development of a more versatile and user-friendly face swapping methodology. We think that it is essential to provide users with the capability to controllably transfer specific facial attributes according to their preferences. For instance, users should have the option to transfer solely the texture of the source face while preserving the structure of the target face, and vice versa. However, since current face swapping methods can not disentangle the structure and texture of the face, controllable face swapping cannot be achieved.

To effectively address the above problems, we propose a novel face-swapping network based on the disentanglement of face features. Inspired by works on face reconstruction [7,8], our method uses two 3D autoencoders to disentangle the facial structure and texture, characterizing them as depth embedding and albedo embedding respectively. Additionally, we complement them with Arcface embedding extracted by the 2D face recognition network [6] for more information about the internal structure of the face. The combination of these three feature embeddings collectively constitutes the identity representation extracted from the source face. Leveraging this disentanglement approach, Leveraging this disentanglement approach, we enable controllable face swapping by extracting a portion of the identity embeddings from the source face and another portion from the target face, thereby transferring partial identity information we choose.

Simultaneously, we encode the target face image into a latent code $w \in \mathbb{R}^{18*512}$ using the StyleGAN encoder to preserve its identity-independent information. In order to fuse the structure and texture from the source face and the identity-independent information from the target face together to control and guide the generation of face swapping, we design a face feature fusion network. We inject the extracted feature embeddings into the feature mappers according to their semantic levels and fuse them with the identity-independent information to obtain a new latent code $w'$ in the $\mathcal{W}+$ space, and then generate the swapped face results through the StyleGAN generator.

During the training process, to more effectively disentangle the structure and texture information, we design three types of training losses: 1) identity-consistent losses used to guide the transfer of identity-related information (structure and texture); 2) attribute-consistent losses used to preserve identity-independent information (expression, pose, and lighting); 3) ancillary loss used to improve the fidelity of the generated image and facilitate convergence of model training.

Overall, our contribution can be summarised in the following three points:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



- We propose a new idea for the face-swapping task, *i.e.*, that we can transfer the identity information more adequately and flexibly by identity feature disentanglement, based on which we propose a new high-quality face swapping framework (ControlFace) and achieve controllable face swapping.
- We propose a novel approach for disentangling structure and texture, and accordingly propose a semantic hierarchy-based face feature fusion module, where different semantic levels of features are fused to enable the model to efficiently learn these features and generate the swapped face. Moreover, We design some loss functions to make the disentanglement more adequate and accurate.
- Extensive experiments demonstrate the effectiveness of our approach to transfer identity information and perform controllable face swapping.

## 2. Related Work

### *2.1. GAN Inversion*

The purpose of GAN inversion is to reconstruct the input image as accurately as possible by mapping it to the corresponding latent code. In this way, we can edit the latent code in order to perform the desired image manipulation. There are two key points in this technique: the latent space and the inversion algorithm. StyleGAN [9,10] can generate high-resolution face images with realistic detail information due to its powerful representation and synthesis capabilities. Its latent space has good disentanglement properties [11–14] and is well suited for feature editing. Recent StyleGAN works [15,16] extends the latent space from $w \in \mathbb{R}^{1*512}$ to $w \in \mathbb{R}^{18*512}$, obtaining better reconstruction results. Our approach accomplishes the fusion of face features of different semantic levels based on the $\mathcal{W}+$ space of the pre-trained StyleGAN model.

### *2.2. Face Swapping*

As a research interest in computer vision, face swapping tasks have a long history. Most of the early face swapping studies [17–19] are based on 3D shape estimation for face alignment and feature transfer, which can produce obvious traces of forgery. Most of the GAN-based methods [1–4,20–25] are target-oriented methods, which use an encoder to extract the identity information of the source face and transfer it to the target face. These methods use a discriminator to improve the fidelity of the swapped images. [1,2,4] obtain the identity embedding from the face recognition model [6], which is injected into the layers of the generator network for fusion. HifiFace [3] adds landmark obtained from 3D Morphable Model (3DMM) to this identity embedding to complement the identity-related geometric information. [22,23] represent source and target faces with latent codes by a pre-trained StyleGAN encoder, and fuse them in the $\mathcal{W}+$ space according to the semantic level. These methods control the attributes of the target faces through landmarks or segmentation masks. Recently, some face swapping methods based on diffusion model [5] are proposed.

However, all of these face swapping methods above only transfer the structure of the source face, neglecting the texture. They generate swapped images with skin colors that are consistent with the target face. E4S [26] is capable of texture transfer, but its feature disentangling approach and its reliance on pre-trained face reenactment models affect its generation quality. Our approach uses a 2D face recognition model [6] and two 3D autoencoders to extract identity information, resulting in more adequate identity transfer and more high-quality controllable face swapping.

### *2.3. Feature Disentanglement*

Existing face disentangling methods can be classified into parametric and non-parametric methods. Parametric disentangling methods [27–33] separate face features such as shape, expression, texture, pose, lighting, etc. By modeling the face with a parametric 3DMM. Such methods fit the 3DMM parameters by optimization algorithms or use deep neural networks to regress the results on the input

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



images. Non-parametric methods no longer require predefined models and parameters. SFS-Net [34], Unsup3d [7] perform unsupervised training based on guessing from shading to shape. LAP [8] exploits multi-image consistency in a non-parametric paradigm to disentangle faces into global facial structure and texture features. GAN2Shape [35] and LiftedGAN [36] attempt to disentangle face facial features using 2D GAN. Unlike the above methods, NPF [37] performs 3D face modeling through a neural rendering mechanism and therefore performs better in terms of detail, resolution, and non-face objects.

## 3. Method

In this section, we will describe our method ControlFace, which is based on a StyleGAN model. After cropping and aligning the given target and source faces, we first use the StyleGAN encoder to obtain the latent code $w$ of the target face in $\mathcal{W}+$ space, while extracting the identity embeddings of the source face using two 3D autoencoders and a 2D face recognition model [6]. Then we inject the disentangled identity embeddings of different semantic levels into $\mathcal{W}+$ space with three feature mappers in the face feature fusion network, obtaining the latent code change $\Delta w$. Finally, we input the new latent code $w' = w + \Delta w$ into the pre-trained StyleGAN generator to generate the face-swapping result. The overall framework is illustrated in Figure 1, and each component will be described in detail below:

### *3.1. Disentangling of Identity Feature*

To achieve controllable face swapping, we first need to accurately and adequately disentangle the structure and texture of the source face. Inspired by work on non-parametric 3D face reconstruction [7,8], we use an autoencoder-based approach to disentangle face features. These works use four autoencoders $\phi^d$, $\phi^a$, $\phi^\omega$, $\phi^l$ to separate each face image into four parts: depth map $d \in \mathbb{R}_+$, albedo map $a \in \mathbb{R}^3$, viewpoint $\omega \in \mathbb{R}^6$, and global light direction $l \in \mathbb{S}^2$. Such disentanglement is achieved by using the UV relationship of the face features and the basic symmetry principles of structure and texture as follows:

$$\hat{\mathbf{I}} = \Pi(\Lambda(a, d, l), d, \omega), \hat{\mathbf{I}}' = \Pi\left(\Lambda\left(a', d', l\right), d', \omega\right), \tag{1}$$

where $\Pi$ and $\Lambda$ are the illumination and projection steps in the reconstruction process, respectively, and $a'$ and $d'$ are the flipped versions of $a$ and $d$. The method constrains the self-encoders according to the symmetry relationship $\mathbf{I} \approx \hat{\mathbf{I}}'$.

We employ such depth maps and albedo maps obtained from the symmetry principles as a representation of structure and texture in our face swapping method. These maps exhibit high identity consistency since facial identity information in the image possesses significantly greater symmetry compared to non-identity information. We encode the face separately using a depth autoencoder $\phi^a = (\delta^a, \varphi^a)$ as well as an albedo autoencoder $\phi^d = (\delta^d, \varphi^d)$ that are pre-trained, obtaining a depth embedding $e_d$ and an albedo embedding $e_a$ which represent the structure and the texture of the face. We inject these two embeddings into the generative network to guide the face generation. During training, we use the depth map $d_{ref}$ and albedo map $a_{ref}$ generated from the decoder $\varphi^d$, $\varphi^a$ to constrain the face-swapping results.

However, since depth maps represent the structure of a face by displaying its distance from the observer in terms of the size of each pixel's grey value, depth embedding is not sufficient for representing local structure features of the face, especially the eyes, eyebrows and other detailed parts of the face. Therefore, we still need to complement the source face structure using ArcFace [6], a 2D face recognition network. We use it to map the source face into a 512-dimensional ArcFace embedding $e_{\text{arc}}$ , which will have high cosine similarity for different photos of the same identity.

Due to the characteristics of the face recognition task, the identity-related information extracted by this face recognition network [6] contains more structured information about the interior of the face, while it is hard for this network to extract texture and face contour effectively. So in our method,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



we complementarily characterize the structure by combining depth embedding $e_d$ with ArcFace embedding $e_{arc}$ while representing the texture through albedo embedding $e_a$.

Based on our disentanglement of structure and texture, our method achieves the capability to perform controllable face swapping tasks. To this goal, we devise a multi-mode training strategy that allows the model to learn four modes of identity feature transfer during training, including Complete Identity Transfer, Structure-only Transfer, Texture-only Transfer, and Self-swapping. When Structure-only Transfer is performed, we extract $e_d$ and $e_{arc}$ from the source face and $e_a$ from the target face. In contrast, when performing Texture-only Transfer, we extract $e_d$ and $e_{arc}$ from the target face and $e_a$ from the source face. When Self-swapping, all identity embeddings are extracted from the target face, while they are all extracted from the source face when performing Complete Identity Transfer. This enables us to achieve four different types of identity feature transfer with the same model, allowing users to choose the identity transfer mode according to their needs. It also enables a small remnant of texture information in $e_{arc}$ to be eliminated during the fusion process, making the feature disentanglement more adequate.

### *3.2. Feature Fusion Based on Semantic Hierarchy*

In order to preserve the identity-independent features of the target image and generate high-quality, high-resolution face swapping results, we use the StyleGAN model with powerful representation capabilities. To optimize computational efficiency and enhance training stability, we don't train the StyleGAN model from scratch. For a given target face image, we encode it using the end-to-end StyleGAN inversion method "e4e" [38] as latent codes $w \in \mathbb{R}^{18*512}$ in the $\mathcal{W}+$ potential space. Previous face swapping methods [22,23] tend to consider feature fusion only for high-level semantics, but we believe that low-level identity information is equally important for face swapping tasks. For this reason, we design a multi-level identity injection network for feature fusion.

Many studies [9,39] have shown that the StyleGAN encoder has robust semantic disentangling capability, which means that it is able to disentangle features at different semantic levels of the face image and represent them at different layers of the $\mathcal{W}+$ space, with the more preceding network layers in the model framework corresponding to image information at higher semantic levels. Taking this characteristic of the StyleGAN model as a basis, we separate these layers into three groups (coarse, medium, and fine). Correspondingly, we classify the latent code $w$ that represents the image in the W+ space into $w_c$, $w_m$, and $w_f$, which denote high, medium, and low-level semantic features, respectively. In order to fuse identity-related information at different semantic levels in $\mathcal{W}+$ space, we devise three facial feature mappers with the same structure: $M_c$, $M_m$, and $M_f$.

Based on the semantic levels to which different identity feature embeddings belong, we direct their injection into the corresponding mappers. Specifically, the depth embedding $e_d$ represents the global structure and contour, with a higher semantic level, and we inject it into $M_c$ and $M_m$; the albedo embedding $e_a$ represents the texture, with a lower semantic level, and we inject it into $M_f$. The ArcFace embedding obtained from a 2D face recognition network [10] represents the structure detail information of the internal face, so we inject it into the three mappers at the same time. Therefore, the output of the face feature mapper can be expressed as:

$$\Delta w_c = M_c\left(w_c, \mathrm{e}_{\mathrm{arc}}, e_d\right), \tag{2}$$

$$\Delta w_m = M_m\left(w_m, \mathrm{e}_{\mathrm{arc}}, e_d\right), \tag{3}$$

$$\Delta w_f = M_f\left(w_f, \mathrm{e}_{\mathrm{arc}}, e_a\right). \tag{4}$$

Each mapper consists of 10 units. The first 5 units perform $e_{arc}$ injection and the last 5 units perform $e_d$ or $e_a$ injection. In each unit, we further extract the useful parts of the identity embedding

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



through two fully connected networks, especially separating the textures left in the $e_{arc}$. We fuse them in $\mathcal{W}+$ space according to the following formula:

$$x' = \text{LeakyRelu}\left((1 + f_\gamma(e))\ \text{LayerNorm}\ (x) + \text{f}_\beta(e)\right), \tag{5}$$

where both $f_\beta$ and $f_\gamma$ are fully connected networks.

Finally, we use the facial parsing network [40] to predict the face region of the target face and generate the mask, and then we fuse the face region of the swapped face result with the background of the target image using Poisson Fusion. In order not to leave visible artifacts at the fusion junction, we performed a soft erosion operation on the generated masks to enable gradient transformations at the blending boundaries of the fused image.

### *3.3. Loss Functions*

Our goal is to disentangle and transfer texture and structure from the source face $x_{\text{src}}$ while preserving identity-independent attribute information such as expression, pose, and lighting of the target face $x_{\text{tgt}}$. Therefore, we design several types of loss functions to constrain the generation process of swapped face $x_{\text{swap}}$. In the following, we will introduce the identity consistency loss, attribute consistency loss, and ancillary loss of our method respectively:

#### 3.3.1. Identity-consistency Loss

For the $e_{arc}$ extracted from the 2D face recognition network [6], we use cosine similarity to compute the ArcFace loss:

$$L_{\text{arc}} = 1 - \cos\left(e_{\text{arc}}, R\left(x_{\text{swap}}\right)\right), \tag{6}$$

where R refers to the 2D face recognition model ArcFace.

In order to transfer the structure and texture of the source image more efficiently, we use the autoencoder $\phi^d$ and $\phi^a$ to disentangle the source face $x_{\text{src}}$ and swapped face $x_{\text{swap}}$ into a depth map $d_{\text{ref}}$ and albedo map $a_{\text{ref}}$, then compute the depth loss and albedo loss respectively:

$$L_{\text{depth}} = \left\|\Phi^d\left(x_{\text{swap}}\right) - d_{\text{ref}}\right\|_2, \tag{7}$$

$$L_{\text{albedo}} = \left\|\Phi^a\left(x_{\text{swap}}\right) - a_{\text{ref}}\right\|_2. \tag{8}$$

Our identity consistency loss is formulated as:

$$L_{\text{id}} = \lambda_{\text{arc}} L_{\text{arc}} + \lambda_{\text{depth}} L_{\text{depth}} + \lambda_{\text{albedo}} L_{\text{albedo}}, \tag{9}$$

where $\lambda_{\text{arc}} = 1$, $\lambda_{\text{depth}} = 10$, $\lambda_{\text{albedo}} = 10$.

While performing Structure-only Transfer or Texture-only Transfer, we calculate these loss functions according to the specific structure-reference images and texture-reference images of these modes.

#### 3.3.2. Attribute-consistency Loss

In order to keep the attribute information of the swapped face consistent with the target face, we use an advanced pre-trained 3D face model [32] to parametrically encode the shape, expression, pose, texture, and lighting information of the source face $x_{src}$ and the target face $x_{tgt}$ to obtain the 3DMM coefficients $c_{src}$ and $c_{tgt}$. Then we mix the shape coefficients (Complete Identity Transfer and Structure-only Transfer) and texture coefficients (Complete Identity Transfer and Texture-only Transfer) of the source face with the other coefficients of the target face to obtain the fused 3DMM coefficients $c_{fuse}$ so that we can obtain the indicative key point coordinates $q_{fuse}$ and the color coefficient $color_{fuse}$ of the swapped face corresponding to it through the mesh renderer and its affine model respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



By this way, we can constrain the attribute information of the expression, pose, and lighting by using the landmark loss and color loss:

$$L_{\text{landmark}} = \left\| q_{\text{fuse}} - q_{\text{swap}} \right\|_1, \tag{10}$$

$$L_{\text{color}} = \| \text{ color }_{\text{fuse}} - \text{ color }_{\text{swap}} \|_1. \tag{11}$$

Moreover, in order to limit the shape change of the swapped face for better foreground-and-background fusion, we design a segmentation loss:

$$L_{\text{seg}} = \left\| M_{\text{swap}} - M_{\text{tgt}} \right\|_1, \tag{12}$$

where $M_{swap}$ and $M_{tgt}$ are the face region masks predicted by the facial parsing network [40] for the swapped face and the target face respectively.

Our attribute-consistent loss is formulated as:

$$L_{\text{attr}} = \lambda_{\text{landmark}} L_{\text{landmark}} + \lambda_{\text{color}} L_{\text{color}} + \lambda_{\text{seg}} L_{\text{seg}}, \tag{13}$$

where $\lambda_{\text{landmark}} = 0.1$, $\lambda_{\text{color}} = 5$, $\lambda_{\text{seg}} = 1$.

#### 3.3.3. Ancillary Loss

In order to make the model converge faster in the training process, we design the Self-swapping mode, which accounts for 9% of the training steps, in which mode the source face and the target face are the same face image. We calculate the reconstruction loss using the following equation:

$$L_{rec} = \left\| x_{tgt} - x_{rec} \right\|_1, \tag{14}$$

where $x_{rec}$ denotes the swapped face obtained from the Self-swapping mode.

In order to improve the fidelity of the generated images, we used the original discriminator and the adversarial loss function of the StyleGAN2 model, resizing the swapped face images to 256 to input them to the discriminator.

Our ancillary loss is formulated as:

$$L_{\text{anci}} = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}}, \tag{15}$$

where $\lambda_{\text{rec}} = 1$, $\lambda_{\text{adv}} = 0.05$. To this end, the total loss of our proposed framework has the following form:

$$L_{total} = L_{id} + L_{attr} + L_{anci}. \tag{16}$$

## 4. Results and Discussion

### *4.1. Experimental Setup*

The CelebAMask-HQ [40] dataset contains 30K high-quality 1024 × 1024 face images with great diversity in gender, skin color, and age. We divide it into 27K and 3K images, which are used as the training and testing sets respectively. We adopt the same data preprocessing approach as e4e [38], using a pre-trained 68-keypoint detection model to detect the keypoints of all face images and subsequently cropping and align them, and then resize them to a size of 256.

We trained our model on 1 A6000 GPU. During training, we set the batch size to 16 and used the Adam optimizer [41] with $\beta_1$ and $\beta_2$ of 0.9 and 0.999, and learning rates of 0.0005 and 0.00005 for the generator and discriminator, respectively. The number of training iterations is 400, 000.

We compare our approach with previous face swapping methods that have had a large impact, including FaceShifter [1], SimSwap [2], HifiFace [3], InfoSwap [42] and MegaFS [22]. Specifically, we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



apply these methods to the CelebAMask-HQ dataset on a test set of 3000 source-target pairs to generate swapped faces.

### 4.2. Qualitative Evaluation

A qualitative comparison of our method with current state-of-the-art face-swapping methods is shown in Figure 2. It can be seen that the swapped faces obtained by our method have higher identity similarity and better-processed details than others. It can be seen that FaceShifter [1] is not able to transfer the identity information sufficiently on high-resolution face images. For SimSwap [2] and HifiFace [3], the detail parts of their results are not processed well enough, especially the part of the eyes and mouth have more obvious artifacts. The quality of swapped face generated by MegaFS [22] and InfoSwap [42] are quite higher than the methods before, but compared with our method, they are still missing a lot of important identity-related information, which directly makes their results have a relatively large identity gap with the source face, especially when there are large differences in contour and skin color between the source and target faces. Among these methods, our method is the only face swapping approach that can transfer texture features and facial contours efficiently. For source and target faces with widely varied textures, we are able to transfer the facial color and other texture details of the source face to the target face accurately, especially to transfer the beard of the source face (line 3). In addition, the facial contours of our swapped face results are more consistent with the source face, especially when the source and target faces have large differences in face shapes. This makes our swapped face results have higher identity similarities with the source face.

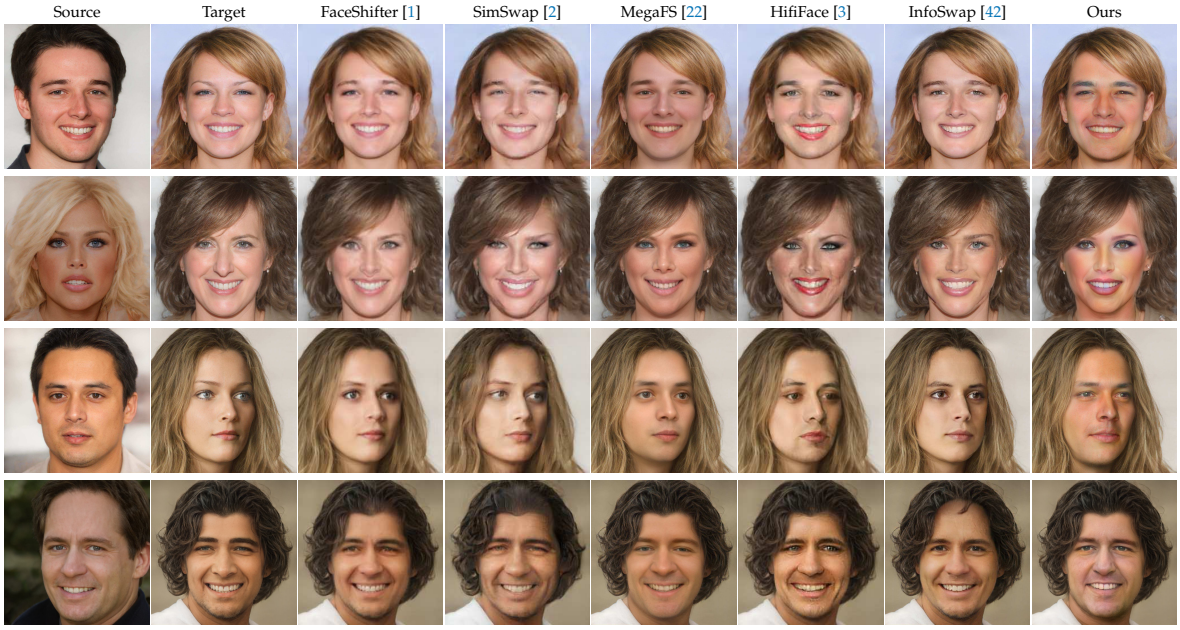


**Figure 2.** Qualitative comparison of our face swapping results with current state-of-the-art face swapping methods. Please pay attention to the texture of the face-swapped image.

### 4.3. Quantitative Evaluation

We also conduct a quantitative comparison with the leading methods to compare the ability to transfer source face identity information and preserve target face attributes. We use the face recognition model [6] to extract the $e_{arc}$ of each swapped face and its corresponding source face and calculate the cosine similarity between them to calculate the accuracy rate of identity transfer. Also, we use the depth autoencoder $\phi^d$ and the albedo autoencoder $\phi^a$ to separate the depth map and albedo map of the swapped face and the source face and compute their $l_2$ distances as an indicator of depth error and albedo error to estimate the transfer of structure and texture respectively. Moreover, in order to quantitatively calculate the preservation of each face feature, we use a 3D face model [32] to extract the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



shape, texture, expression, pose, and lighting coefficients of each swapped face and the corresponding source face and target face. We compute the $l_2$ distances of shape coefficients and texture coefficients between the swapped face and the source face, and the other coefficients between the swapped face and the target face.

As can be seen from the Table 1, our method has the highest ArcFace similarity, which indicates that the swapped faces generated by our method have a high identity transfer rate based on face recognition models. For depth error, the generated results of our method are slightly higher than other methods, which is due to the fact that we are able to transfer the source face contours better. The albedo error of our results is quite higher than other methods, which shows that our model has better results in transferring the source face texture information, while most of the other methods don't pay much attention to texture transfer.

As can be seen from the Table 2, our model achieves near-top results for each face feature. For the shape and texture coefficients by [32], our ControlFace is also more accurate than the others. For the attribution preservation, our method achieves second place for both expression and pose control behind FaceShifter [1], such result is mainly based on the landmark loss we propose. The disentanglement of texture and light is currently a big challenge in texture transfer, due to the fact that they both act together on every pixel value. Nevertheless, our model still manages to reach third in light error, while ControlFace is the only method that transfers texture in the experiment. This enables our swapped faces to have a high fidelity.

**Table 1. Quantitative Results of Identity Transfer.** We compare our model with five competing methods in ArcFace Similarity and Depth & Albedo Error for the ability of identity transfer. The best results are shown in bold. ↑ means higher is better, and ↓ means lower is better.

| Method | Arc. Simi. ↑ | Depth ↓ | Albedo ↓ |
|---|---|---|---|
| FaceShifter [1] | 49.33 | 31.68 | 49.59 |
| SimSwap [2] | 52.03 | 32.32 | 48.49 |
| MegaFS [22] | 48.49 | 33.44 | 48.18 |
| HifiFace [3] | 48.24 | 32.35 | 50.22 |
| InfoSwap [42] | 52.58 | 31.04 | 51.58 |
| **Ours** | **55.20** | **27.89** | **35.17** |

**Table 2. Quantitative Results of each face feature.** We measure the error of shape, texture, expression, pose, and lighting.

| Method | Shape ↓ | Tex. ↓ | Exp. ↓ | Pose ↓ | Light. ↓ |
|---|---|---|---|---|---|
| FaceShifter [1] | 2.07 | 5.34 | **0.74** | **0.57** | **1.08** |
| SimSwap [2] | 2.01 | 5.09 | 1.15 | 0.75 | 1.71 |
| MegaFS [22] | 2.34 | 5.25 | 1.25 | 2.77 | 3.04 |
| HifiFace [3] | 1.75 | 4.95 | 1.23 | 0.63 | 2.14 |
| InfoSwap [42] | 2.01 | 4.91 | 1.38 | 2.41 | 1.93 |
| **Ours** | **1.26** | **3.12** | 1.02 | 0.60 | 1.72 |

### *4.4. Ablation Study*

In this section, we verify the effectiveness of the identity embedding selection and feature injection of our proposed method by an ablation study. In each experiment, we only change one of the components in our framework to keep the remaining variables constant. We show the quantitative result in Table 3, where we test the identity transfer capabilities of each model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1



**Table 3. Quantitative Ablation Study.** The comparison of different strategies of identity embedding selection and feature injection.

| Method | Arc. Simi. ↑ | Depth ↓ | Albedo ↓ |
|---|---|---|---|
| **Ours** | 55.20 | 27.89 | 35.17 |
| w/o $e_{arc}$ | 49.44 | **27.80** | **34.58** |
| w/o $e_d$ | 55.97 | 28.62 | 35.07 |
| w/o $e_a$ | 56.82 | 27.86 | 38.82 |
| w/o $e_d$&$e_a$ | **57.06** | 28.73 | 39.34 |
| (a) | 53.26 | 28.43 | 36.77 |
| (b) | 51.98 | 28.71 | 35.48 |
| (c) | 51.86 | 28.46 | 36.41 |

**Choice of Identity Embeddings.** Our identity extraction network extracts a total of three identity embeddings $e_{arc}$, $e_d$ and $e_a$. To demonstrate the necessity of individual identity embeddings, we reduce 1-2 identity embeddings at a time and retrain the model. We reduced $e_{arc}$, $e_d$, $e_a$, both $e_d$ and $e_a$, respectively. When we don't inject $e_d$ or $e_a$ into the feature fusion network, we also correspondingly stopped using $L_{depth}$ or $L_{albedo}$. The experimental results show that reducing a certain embedding may lead to a better transfer of other identity features, but has a large impact on the identity information represented by that embedding.

**Feature Injection Strategy.** For feature injection, we conduct experiments with three different strategies. (a) injects the albedo embedding $e_a$ into the coarse feature mapper $M_c$ and the medium feature mapper $M_m$, and injects the depth embedding $e_d$ into the fine feature mapper $M_f$. (b) injects the depth embedding $e_d$ into the coarse feature mapper $M_c$, and injects the albeda embedding $e_a$ into the fine feature mapper $M_f$ and the medium feature mapper $M_m$. (c) injects the ArcFace embedding $e_{arc}$ into the coarse feature mapper $M_c$ and the medium feature mapper $M_m$, and no more into the fine feature mapper $M_f$. Experiments on strategies (a) and (b) show that our feature injection approach matches its semantic level. Experiments on strategy (c) show that there are a number of low-level semantic features in the ArcFace embedding $e_{arc}$, and it is necessary to inject them into the whole three mappers.

### *4.5. Controllable Face Swapping*

Distinguished from conventional face swapping methods, our model stands out by its exceptional capability to perform controllable face swapping, based upon the sufficient disentanglement of structure and texture. This is a pioneering breakthrough in the field of face swapping, as it empowers users with the freedom to choose their preferred identity transfer mode. When utilizing ControlFace for a face swapping project, users have the flexibility to decide whether to transfer the structure or texture of source faces to the target faces.

To enable a single face swapping model to seamlessly handle all of the four identity feature transfer modes, including Complete Identity Transfer, Structure-only Transfer, Texture-only Transfer, and Self-swapping, we design a probabilistic framework that governs the transfer of structural and textural information during each training step, with both probabilities set at 0.7. Consequently, the probability distribution for each of the four transfer modes during a training step is 0.49, 0.21, 0.21, and 0.09, respectively.

**Qualitative results.** We show the generation results of each identity transfer mode in Figure 3, from which we can clearly make out the significant differences between the different modes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 December 2023 doi:10.20944/preprints202312.0066.v1




**Figure 3.** Qualitative results of controllable face swapping using our method.

When Structure-only Transfer is performed, the skin color, lip color, beard, and other texture information of the swapped face remain consistent with the target image, while the structure has a high similarity with that of the source face. In the field of virtual character creation, this mode plays a pivotal role in crafting lifelike virtual personas. It facilitates the fusion of unique face structures with pre-existing character templates while retaining the technologically synthesized skin and makeup. This is instrumental in generating diverse characters for video games, augmented reality experiences, and virtual worlds, enhancing the immersive quality of these digital environments.

While Texture-only Transfer is performed, the face structure of the swapped face is basically the same as the target face, but the texture information changes considerably, which is more consistent with the source face. In the field of beauty-themed applications, such as Photoshop and beauty filters in mobile applications, Texture-only Transfer can be utilized to enhance and refine individuals' appearances in photos. Users can modify their skin texture and complexion to achieve a more desired and aesthetically pleasing look while retaining their original facial structure. This serves to meet the ever-evolving standards of beauty in the digital age, providing individuals with a means to perfect their selfies and photographs before sharing them on social media or elsewhere.

Our future research will concentrate on the further disentanglement of face structure and expression, as well as the separation of texture and lighting information. These plans have the potential to elevate the precision and interpretability of face swapping and edition techniques, which is able to give users a wider range of choices and higher quality results.

## 5. Conclusions

In this paper, we propose ControlFace, a novel framework for face swapping. This method accurately disentangles the structure and texture of a source face and extracts them in the form of identity embeddings. We inject them into the feature mapper according to their semantic level and fully fuse them with the representation $w$ of the target face in the $\mathcal{W}+$ space of StyleGAN to generate high-fidelity, high-quality swapped faces. We realize controllable face swapping by extracting some of the identity embeddings from the source face, while others from the target face. Extensive experiments

and qualitative and quantitative comparisons with current mainstream methods demonstrate the superiority of our method in identity information transfer, attribute information protection, and controllable face swapping.

**Author Contributions:** Conceptualization, X.Z., W.Z. and K.L; methodology, X.Z., W.Z, K.L, H.T. and Z.Z.; software, X.Z.; validation, X.Z.; formal analysis, X.Z., W.Z. and K.L.; investigation, X.Z. and K.L.; resources, W.Z., W.Z. and N.Y.; data curation, X.Z; writing—original draft preparation, X.Z.; writing—review and editing, W.Z., K.L., H.T. and Z.Z.; visualization, X.Z.; supervision, W.Z., W.Z. and N.Y.; project administration, W.Z., W.Z. and N.Y.; funding acquisition, W.Z., W.Z. and N.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Natural Science Foundation of China under Grant 62372423, U20B2047, 62072421, 62002334, and 62121002, Key Research and Development program of Anhui Province under Grant 2022k07020008.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Our code will be publicly available at https://github.com/ZhangXH227/ControlFace. Publicly available datasets were analyzed in this study. This data can be found here: CelebAMask-HQ: https://github.com/switchablenorms/CelebAMask-HQ.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. 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. Chen, R.; Chen, X.; Ni, B.; Ge, Y. Simswap: An efficient framework for high fidelity face swapping. In Proceedings of the Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 2003–2011.
3. 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**.
4. 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 Proceedings of the AAAI Conference on Artificial Intelligence, 2021, Vol. 35, pp. 3083–3091.
5. Zhao, W.; Rao, Y.; Shi, W.; Liu, Z.; Zhou, J.; Lu, J. DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8568–8577.
6. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4690–4699.
7. Wu, S.; Rupprecht, C.; Vedaldi, A. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1–10.
8. 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14214–14224.
9. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
10. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
11. Goetschalckx, L.; Andonian, A.; Oliva, A.; Isola, P. Ganalyze: Toward visual definitions of cognitive image properties. In Proceedings of the Proceedings of the ieee/cvf international conference on computer vision, 2019, pp. 5744–5753.

12. Jahanian, A.; Chai, L.; Isola, P. On the" steerability" of generative adversarial networks. *arXiv preprint arXiv:1907.07171* **2019**.
13. Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of gans for semantic face editing. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9243–9252.
14. Collins, E.; Bala, R.; Price, B.; Susstrunk, S. Editing in style: Uncovering the local semantics of gans. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5771–5780.
15. Abdal, R.; Qin, Y.; Wonka, P. Image2stylegan: How to embed images into the stylegan latent space? In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4432–4441.
16. Richardson, E.; Alaluf, Y.; Patashnik, O.; Nitzan, Y.; Azar, Y.; Shapiro, S.; Cohen-Or, D. Encoding in style: a stylegan encoder for image-to-image translation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2287–2296.
17. Blanz, V.; Scherbaum, K.; Vetter, T.; Seidel, H.P. Exchanging faces in images. In Proceedings of the Computer Graphics Forum. Wiley Online Library, 2004, Vol. 23, pp. 669–676.
18. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395.
19. Nirkin, Y.; Masi, I.; Tuan, A.T.; Hassner, T.; Medioni, G. On face segmentation, face swapping, and face perception. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018, pp. 98–105.
20. Natsume, R.; Yatagawa, T.; Morishima, S. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447* **2018**.
21. Nirkin, Y.; Keller, Y.; Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 7184–7193.
22. Zhu, Y.; Li, Q.; Wang, J.; Xu, C.Z.; Sun, Z. One shot face swapping on megapixels. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4834–4844.
23. Xu, Y.; Deng, B.; Wang, J.; Jing, Y.; Pan, J.; He, S. High-resolution face swapping via latent semantics disentanglement. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7642–7651.
24. Xu, Z.; Zhou, H.; Hong, Z.; Liu, Z.; Liu, J.; Guo, Z.; Han, J.; Liu, J.; Ding, E.; Wang, J. StyleSwap: Style-Based Generator Empowers Robust Face Swapping. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 661–677.
25. Luo, Y.; Zhu, J.; He, K.; Chu, W.; Tai, Y.; Wang, C.; Yan, J. StyleFace: Towards Identity-Disentangled Face Generation on Megapixels. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 297–312.
26. Liu, Z.; Li, M.; Zhang, Y.; Wang, C.; Zhang, Q.; Wang, J.; Nie, Y. Fine-Grained Face Swapping via Regional GAN Inversion. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8578–8587.
27. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face alignment across large poses: A 3d solution. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 146–155.
28. Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; Zhou, X. Joint 3d face reconstruction and dense alignment with position map regression network. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 534–551.
29. Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; Carin, L. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems* **2016**, *29*.
30. Shen, Y.; Luo, P.; Yan, J.; Wang, X.; Tang, X. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 821–830.
31. Tran, L.; Yin, X.; Liu, X. Disentangled representation learning gan for pose-invariant face recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1415–1424.

32. 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 Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2019, pp. 0–0.
33. Daněček, R.; Black, M.J.; Bolkart, T. EMOCA: Emotion driven monocular face capture and animation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 20311–20322.
34. Sengupta, S.; Kanazawa, A.; Castillo, C.D.; Jacobs, D.W. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6296–6305.
35. Pan, X.; Dai, B.; Liu, Z.; Loy, C.C.; Luo, P. Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. *arXiv preprint arXiv:2011.00844* **2020**.
36. Shi, Y.; Aggarwal, D.; Jain, A.K. Lifting 2d stylegan for 3d-aware face generation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 6258–6266.
37. Zhang, Z.; Chen, R.; Cao, W.; Tai, Y.; Wang, C. Learning Neural Proto-Face Field for Disentangled 3D Face Modeling in the Wild. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 382–393.
38. Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; Cohen-Or, D. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)* **2021**, *40*, 1–14.
39. Xia, W.; Yang, Y.; Xue, J.H.; Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2256–2265.
40. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5549–5558.
41. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
42. Gao, G.; Huang, H.; Fu, C.; Li, Z.; He, R. Information bottleneck disentanglement for identity swapping. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3404–3413.