# Region-Aware Face Swapping

Chao Xu[1*†] Jiangning Zhang[1†] Miao Hua[2] Qian He[2] Zili Yi[2] Yong Liu[1‡]
[1] APRIL Lab, Zhejiang University [2]Bytedance Inc.
{21832066, 186368}@zju.edu.cn, yongliu@iipc.zju.edu.cn
{huamiao, heqian, yizili}@bytedance.com

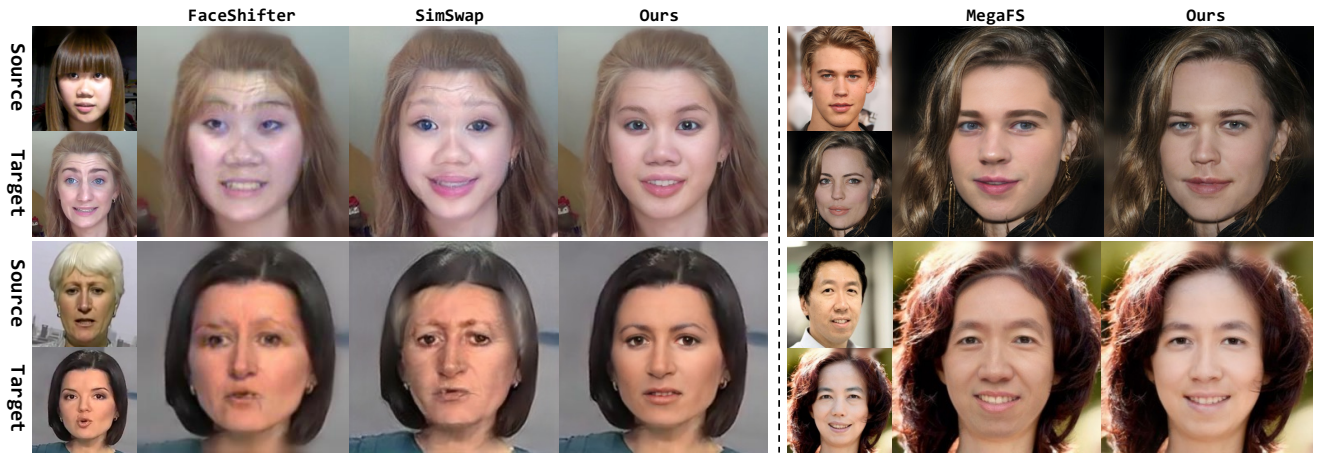


Figure 1. **Comparison with SOTA methods on challenging situations. Left part** shows the results of attribute-distinct cases, *e.g.*, bangs and the white hair from the source identity, where our method is significantly better than FaceShifter [19] and SimSwap [7] with higher quality, better identity-consistency, and fewer artifacts. **Right part** shows high-resolution results of light-changing and in-the-wild situations with SOTA MegaFS [46], which exists artifacts around face contour and light while our method could better preserve attributes of the target face. Images are from official attached results or released codes for fair comparisons. Please zoom in for more details.

## Abstract

*This paper presents a novel Region-Aware Face Swapping (RAFSwap) network to achieve identity-consistent harmonious high-resolution face generation in a local-global manner:* ***1)*** *Local Facial Region-Aware (FRA) branch augments local identity-relevant features by introducing the Transformer to effectively model misaligned cross-scale semantic interaction.* ***2)*** *Global Source Feature-Adaptive (SFA) branch further complements global identity-relevant cues for generating identity-consistent swapped faces. Besides, we propose a Face Mask Predictor (FMP) module incorporated with StyleGAN2 to predict identity-relevant soft facial masks in an unsupervised manner that is more practical for generating harmonious high-resolution faces. Abundant experiments qualitatively and quantitatively demonstrate the superiority of our method for generating more identity-consistent high-resolution swapped faces over SOTA methods,* e.g.*, obtaining 96.70 ID retrieval that outperforms SOTA MegaFS by 5.87↑.*

## 1. Introduction

Face swapping aims at transferring the identity of the source identity to the target identity while keeping the identity-irrelevant attributes of the target face unchanged, which has attracted widespread attention in the film industry and computer games. Recently, many researchers have achieved significant progress in face swapping, especially designing inversion-based methods to generate high-resolution face images. However, **there are two continuously critical issues**: *1) How to maintain identity consistency with the source identity, including local and global facial details.* Almost all current methods [7, 19] perform feature interaction only on global feature representation without modeling identity-relevant local regions, *e.g.*, lips, nose, brows, and eyes, which will limit the model's ability to express identity consistency. *2) How to generate high-resolution swapped faces while keeping the identity-irrelevant details consistent with the target face under the GAN inversion framework,* e.g.*, background and occlusions.* Recent works [38, 46] exploit the StyleGAN2 [17] as the powerful decoder but fail to maintain the consistency of the identity-irrelevant attributes of the target face. In this paper, we are dedicated to solving both the above problems.

* Work done during an internship at Bytedance.
† Equal contribution.
‡ Corresponding Author.

Recent works [7, 18, 19, 34, 37] regard face swapping as a style transfer task that employs global AdaIN [14] to transfer the identity information of the source face into the target face. However, the identity vector produced by the face recognition network is naturally not well-disentangled, which inevitably includes some identity-irrelevant information of the source face, *e.g*., background, light distribution, and hairstyle. This wrong information will be further injected into the target feature in a global manner via AdaIN, resulting in low-quality generation results. As shown in the left part of Fig. 1, recent AdaIN-based methods cannot preserve the source identity well, in which generated faces contain excessive information of the source face in some challenging situations, *e.g*., bangs and white hair. To better preserve the identity consistency of the generated face, we explicitly model the local facial features besides global representation to perform feature interaction more finely, which also excludes the influence of the identity-irrelevant area of the source face at the same time. In this way, our method is well competent for the above challenges, as depicted in the fourth column of Fig. 1. Specifically, we design two parallel branches to process different fine-grained information: **1)** local *Facial Region-Aware* (FRA) branch to model identity-relevant feature interaction between source and target faces, which employs a *Region-Aware Identity Tokenizer* (RAT), transformer layers [31], and a *Region-Aware Identity Projector* (RAP) to realize misaligned cross-scale semantic interaction, *i.e*., lips, nose, brows, and eyes. **2)** global *Source Feature-Adaptive* (SFA) branch to complement global identity-relevant cues, *e.g*., skin wrinkle, for more identity-consistent results. Details can be found in following Sec. 3.1 and 3.2.

To achieve high-resolution face generation for more practical application, we adopt GAN inversion framework [26,30] similar to recent face swapping works [38,46]. But these methods introduce a fatal problem of failing to preserve the background and occlusions, because vector-conditioned progressive generation will inevitably change identity-irrelevant regions. Recent MegaFS [46] blends the high-resolution result to the target face by the pre-existing face mask in a post-processing way, while HifiFace [34] learns to predict face masks in a supervised manner that restricts applications. These methods must rely on ground truth face masks and usually produce artifacts around facial contours, as shown in the right part of Fig 1. Differently, considering that pre-trained StyleGAN2 [17] encapsulate rich facial semantic prior, we design a *Face Mask Predictor* (FMP) to predict identity-relevant soft facial mask in an unsupervised manner, *i.e*., without using specific mask supervision. In this way, our model achieves harmonious high-resolution face generation that keeps identity-irrelevant attributes consistent with the target face. In summary, we make the following three contributions:

- We propose a novel Region-Aware Face Swapping (RAFSwap) network, which consists of a novel *FRA* branch to augment local identity-relevant features by introducing the Transformer to effectively model misaligned cross-scale semantic interaction, and a novel *SFA* branch to further complement global identity-relevant cues for generating identity-consistent swapped faces.
- We propose a *FMP* module incorporated with StyleGAN2 to predict identity-relevant soft facial masks in an unsupervised manner that is more practical.
- Abundant experiments qualitatively and quantitatively demonstrate the superiority of our method for generating more identity-consistent high-resolution swapped faces over SOTA methods.

## 2. Related Work

### 2.1. GAN Inversion

GAN inversion is a task that the latent code from which the well-trained GAN could most accurately reconstruct the original input images. Generally, works [1, 2] directly optimize the latent vector to minimize the errors for the given images. These methods could achieve high reconstruction quality, but they are time-consuming. Subsequently, recent methods employ an encoder to map the given image to the latent space end-to-end. Specifically, pSp and GHF [26, 36] embed real images into a series of style vectors that are fed into a pre-trained StyleGAN2 generator. e4e [30] designs the encoder that generates a single base style code and a series of offset vectors to yield the final style codes.

### 2.2. Face Swapping

Face swapping aims to change the facial identity but to keep other facial attributes constant. Early efforts [5, 6, 8] focus on 3D-based methods, but it requires manual interaction and cannot preserve the target expression. To address this limitation, Face2Face [29] fits a 3D morphable model (3DMM) to both the source and target faces. Nirkin *et al.* [23] combine 3DMM and face segmentation model to achieve robust face swapping under unprecedented conditions. Besides, with the popularity of GANs [12], learning-based methods have enabled significant progress in face swapping. DeepFakes [25] trains an Encoder-Decoder architecture for two specific identities but lacks generalization ability. Some works follow the disentanglement paradigm. IPGAN [4] disentangles the identity from the source face and attributes from the target face separately and recombine them for identity preserving face synthesis. FaceShifter [19] adaptively integrates identity and attribute embeddings with the attentional way. FaceInpainter [18] adopts 3D priors,

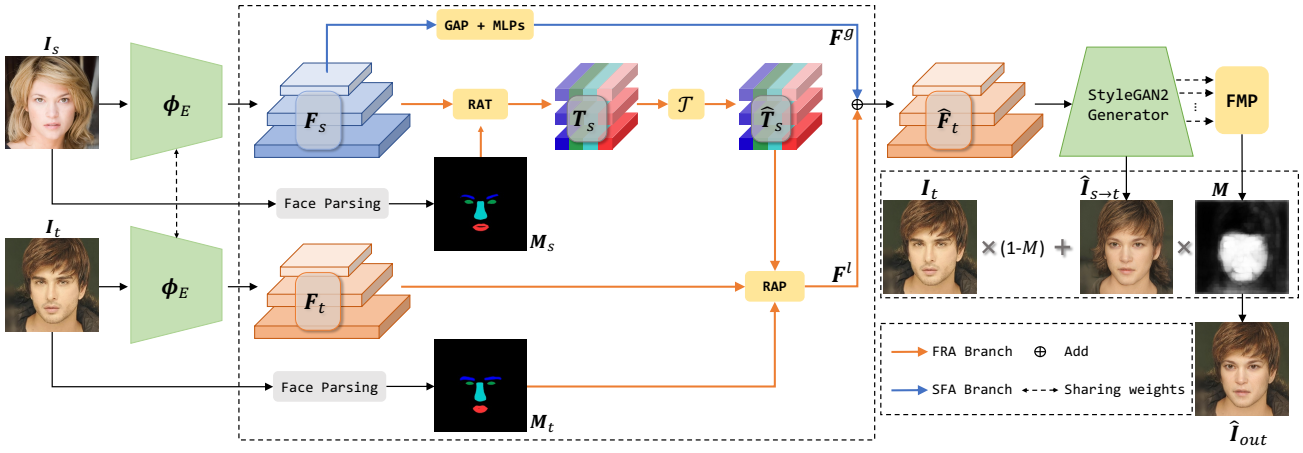


Figure 2. **Overview of the proposed RAFSwap.** The source face $\boldsymbol{I}_s$ and target face $\boldsymbol{I}_t$ firstly go through a weight-sharing hierarchical face encoder $\boldsymbol{\phi}_E$ and a pre-trained face parsing model to obtain hierarchical features, *i.e.*, $\boldsymbol{F}_s$ and $\boldsymbol{F}_t$, and corresponding semantic labels, *i.e.*, $\boldsymbol{M}_s$ and $\boldsymbol{M}_t$, respectively. Then local *Facial Region-Aware* branch in orange and global *Source Feature-Adaptive* branch in blue are employed to integrate identity information of $\boldsymbol{I}_s$ with target attributes of $\boldsymbol{I}_t$ in a local-global manner, obtaining $\boldsymbol{F}^l$ and $\boldsymbol{F}^g$. The fused hierarchical feature $\hat{\boldsymbol{F}}_t$ is mapped into different fine-grained vectors to control the target face generation process by a StyleGAN2 generator. The *Face Mask Predictor* utilize StyleGAN2 feature maps to produce the soft face mask $\boldsymbol{M}$ simultaneously. $\hat{\boldsymbol{I}}_{s\to t}$ is blended to the target face $\boldsymbol{I}_t$ by $\boldsymbol{M}$ to obtain final swapped face $\hat{\boldsymbol{I}}_{out}$.$\hat{}$denotes the generated face instead of the real face.

texture code, and identity code for explicitly disentanglement. Recently, MegaFS [46] first exploits StyleGAN2 as the decoder for high-resolution face swapping. However, the above referential methods struggle to generate highly identity-consistent faces due to the global feature fusion.

### 2.3. Feature Fusion

Feature fusion is an important process in face swapping. Most previous works [7, 18, 19, 34, 37, 40] are inspired by style transfer methods. They employ AdaIN [14] to inject the identity vector into the target face to generate swapped face. Besides, MegaFS [46] proposes FTM to control multiple attributes of identity information, while other methods [22, 24, 38, 42] barely concatenate identity and attribute vectors. However, such global operations do not model the crucial local features interaction. More recently, the attention structure is playing a pivotal role in feature enhancement and interaction in NLP [11, 31, 41] and CV [20, 35]. Based on the attention mechanism, we design our RAT and RAP for feature fusion, which fully fuses the local and global identity-relevant features of the source face while preserving the attributes of the target face.

## 3. Method

In this paper, a novel RAFSwap is proposed to generate high-resolution and identity-consistent swapped face images. Our method is built on a GAN inversion framework, pSp [26]. As depicted in Fig. 2, we first send source face $\boldsymbol{I}_s$ and target face $\boldsymbol{I}_t$ to a Hierarchical Face Encoder $\boldsymbol{\phi}_E$ to extract hierarchical features $\boldsymbol{F}_s = \{\boldsymbol{F}_s^0, \boldsymbol{F}_s^1, \boldsymbol{F}_s^2\}$ and $\boldsymbol{F}_t = \{\boldsymbol{F}_t^0, \boldsymbol{F}_t^1, \boldsymbol{F}_t^2\}$. Superscripts $0, 1, 2$ represent small, medium, and large scale, respectively. All feature maps are mapped to 512 channels. In the meantime, the semantic labels $\boldsymbol{M}_s$ and $\boldsymbol{M}_t$ of the source and target face on lips, nose, brows, and eyes areas are extracted by BiseNet [39]. Second, FRA and SFA are employed to extract local and global discriminative identity features of the source face $\boldsymbol{I}_s$, obtaining $\boldsymbol{F}^l$ and $\boldsymbol{F}^g$, which are then performed element-wise addition to produce fused hierarchical features $\hat{\boldsymbol{F}}_t$. Third, following the pSp, 18 mapping networks are trained to extract the learned styles from the hierarchical feature maps. All style vectors are fed into the StyleGAN2 generator to synthesize raw swapped face $\hat{\boldsymbol{I}}_{s\to t}$. The feature maps of StyleGAN2 are extracted to generate soft mask $\boldsymbol{M}$ simultaneously by FMP. Finally, $\hat{\boldsymbol{I}}_{s\to t}$ and $\boldsymbol{I}_t$ are blended by $\boldsymbol{M}$ to produce swapped face $\hat{\boldsymbol{I}}_{out}$.

### 3.1. Facial Region-Aware Branch

**Region-Aware Identity Tokenizer.** In order to explicitly model facial features by local identity-relevant regions, *i.e.*, lips, nose, brows, and eyes, we propose a *Region-Aware Identity Tokenizer*. As illustrated in Fig. 3, the purpose of RAT is to convert source face features $\boldsymbol{F}_s$ into compact sets of crucial local identity-relevant tokens $\boldsymbol{T}_s \in \mathbb{R}^{N\times L\times 512}$, where $N$ is the number of feature map scales, $L$ is the number of regions. We define three scales and four facial areas, so the $N$ and $L$ are set to 3 and 4. Following the SEAN [45], we adopt a region-wise average pooling layer $\Phi$ to obtain the local semantic representations. Specifically, we resize the semantic labels by using bilinear interpolation to match

the size of each source feature map. Then, each region's pixel-level features are aggregated and averaged into a corresponding token. A linear layer is followed to embed all hierarchical identity-relevant tokens further. The tokenizer operation could be represented in the following formula:

$$\boldsymbol{T}_s^n = \text{Linear}(\Phi(\boldsymbol{F}_s, \boldsymbol{M}_s^n)),$$
$$\text{where } \boldsymbol{M}_s^n \in \{\boldsymbol{M}_s^{lips}, \boldsymbol{M}_s^{nose}, \boldsymbol{M}_s^{brows}, \boldsymbol{M}_s^{eyes}\}. \quad (1)$$

**Transformer Layers.** The AdaIN-based methods lack feature interaction among crucial local features that cause swapped faces in poor identity consistency. Benefiting from our region-aware mechanism, we introduce the Transformer layer $\mathcal{T}$ to model the interaction between tokens across different scales and semantics, which is built upon the Multi-head Self-Attention (MSA) layer, along with Feed-Forward Network (FFN), Layer Normalization (LN), and Residual Connection (RC) operations. In practice, a reshape operation apply on $\boldsymbol{T}_s$ to combine N and L dimensions: $\boldsymbol{T}_s \in \mathbb{R}^{NL\times 512}$. We denote $\boldsymbol{T}_s$ as Query, Key, and Value, respectively. Each attention head is formulated as:

$$\begin{aligned} Attention(\boldsymbol{T}_s) &= \text{Softmax}\left[\frac{\boldsymbol{T}_s\boldsymbol{W}^Q\left(\boldsymbol{T}_s\boldsymbol{W}^K\right)^T}{\sqrt{d_k}}\right]\boldsymbol{T}_s\boldsymbol{W}^V \\ &= \boldsymbol{A}\boldsymbol{T}_s\boldsymbol{W}^V, \end{aligned} \quad (2)$$

where $\boldsymbol{W}^Q \in \mathbb{R}^{d_m\times d_k}, \boldsymbol{W}^K \in \mathbb{R}^{d_m\times d_k}, \boldsymbol{W}^V \in \mathbb{R}^{d_m\times d_v}$ are parameter matrices for feature projections. $d_m$ is the input dimension, while $d_k$ and $d_v$ are hidden dimensions of each projection subspace, $\boldsymbol{A} \in \mathbb{R}^{NL\times NL}$ is the attention matrix, which indicates the relation between all tokens. For FFN, which consists of two cascaded linear transformations with a ReLU activation in between:

$$\text{FFN}(\boldsymbol{x}) = \max(0, \boldsymbol{x}\boldsymbol{W_1} + \boldsymbol{b}_1)\,\boldsymbol{W_2} + \boldsymbol{b}_2, \quad (3)$$

where $\boldsymbol{x}$ is the input tokens, $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are weights of two linear layers, and $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are corresponding bias. The transformed tokens $\hat{\boldsymbol{T}}_s$ is formulated as:

$$\hat{\boldsymbol{T}}_s = \boldsymbol{T}_s + [\text{MSA}|\text{FFN}](\text{LN}(\boldsymbol{T}_s)). \quad (4)$$

Subsequently, each token contains sufficient multi-scale and multi-semantic representation via the Transformer layers.

**Region-Aware Identity Projector.** Corresponding to Tokenizer, we need to project the identity-relevant tokens to the target features spatially while considering the misaligned attributes between source and target faces, *e.g.*, gaze and expression. Unlike SEAN [45] that replaces the style of edited source region by reference style faces, we devise a *Region-Aware Identity Projector* to adaptively transfer identity information to the target face and keep its attributes unchanged. As shown in Fig. 3, the masked target feature map $\boldsymbol{F}_t^m$ is updated by combining weighted $\hat{\boldsymbol{T}}_s$ to refine the $\boldsymbol{F}_t$

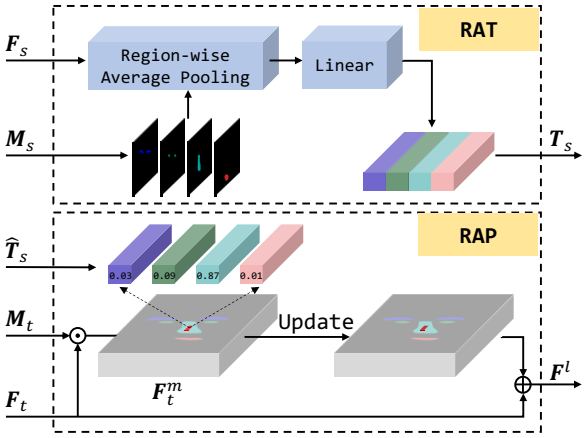


Figure 3. Structure of Region-Aware Identity Tokenizer and Region-Aware Identity Projector on each scale.

for forming local identity-augmented features $\boldsymbol{F}^l$. Specifically, $\boldsymbol{F}_t^m$ is flattened along the height and width dimension: $\boldsymbol{F}_t^m \in \mathbb{R}^{HW\times 512}$ . Given the flattened feature $\boldsymbol{F}_t^m$ of each scale as Query, the identity-relevant tokens $\hat{\boldsymbol{T}}_s$ of each scale as Key and Value, the attention matrix $\boldsymbol{A}^P$ is computed as Eq. 2. Each element of $\boldsymbol{A}_{ij}^P$ indicates the relation between each pixel and token. We show the scores between a pixel located on the nose and all four tokens in Fig. 3. As expected, it has the highest value with the token extracted from the source nose region. The identity-relevant tokens are thus linearly transferred to $\boldsymbol{F}_t^m$, which are then reshaped to the same size as $\boldsymbol{F}_t$ and further added to $\boldsymbol{F}_t$:

$$\boldsymbol{F}^l = \boldsymbol{F}_t + \text{RS}(\boldsymbol{A}^P\hat{\boldsymbol{T}}_s\boldsymbol{W}^P), \quad (5)$$

where $\boldsymbol{W}^P$ is learnable weight, RS is reshape operation

### 3.2. Source Feature-Adaptive Branch

After the FRA, the crucial local identity-relevant features from the source face have been combined into the target face. However, some global facial representations also affect the identity consistency of swapped faces, *e.g.*, skin wrinkle, the relative distance of facial components. Thus, we design a global *Source Feature-Adaptive* branch that captures global information as a complementary cue to distinguish different identities. As shown in Fig. 2, to avoid spatial misalignment between source and target faces, the source feature map with the smallest size first goes through a global averaging pooling (GAP). Then MLPs are followed to further adaptively recombine the global features. Finally, we broadcast the global features as large as three scales and add them to the $\boldsymbol{F}^l$ with the same resolution to obtain integrated target features $\hat{\boldsymbol{F}}_t$:

$$\begin{aligned} \boldsymbol{F}^g &= \text{MLPs}(\text{GAP}(\boldsymbol{F}_s^0)), \\ \hat{\boldsymbol{F}}_t &= \boldsymbol{F}^g + \boldsymbol{F}^l. \end{aligned} \quad (6)$$

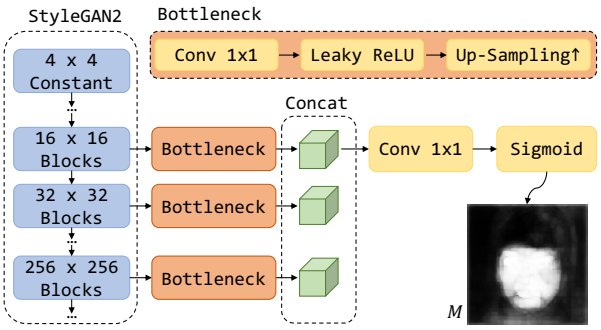


Figure 4. Structure of Face Mask Predictor.

### 3.3. Face Mask Predictor

In order to solve the occlusion and distorted background problem introduced by the GAN inversion framework, MegaFS [46] directly utilize the hard face mask produced by the pre-trained segmentation model for blending, which tends to produce artifacts around the edges and is not computationally-friendly. Instead, inspired by Labels4Free [3], we make full use of the existing structure. First, the layers of a pre-trained StyleGAN2 already contain rich semantic prior. Second, the identity-consistent constraint could force the mask module to focus on the identity-relevant areas. Thus, we exploit the feature maps of StyleGAN2 to produce soft face masks without specific mask supervision. Specifically, as shown in Fig. 4, we first sample feature maps with a resolution ranging from 16 to 256, then apply a bottleneck on each feature map, which reduces the channel to 32 and upsample the resolution to 256. Finally, the concatenated feature maps are fed to a $1 \times 1$ convolution layer and a sigmoid layer sequentially to produce a single channel soft mask $\boldsymbol{M}$. To generate the swapped face, we blend $\hat{\boldsymbol{I}}_{s \to t}$ to the target face $\boldsymbol{I}_t$ by $\boldsymbol{M}$, formulated as:

$$\hat{\boldsymbol{I}}_{out} = \boldsymbol{M} \odot \hat{\boldsymbol{I}}_{s \to t} + (1 - \boldsymbol{M}) \odot \boldsymbol{I}_t. \tag{7}$$

### 3.4. Objective Functions

During the training stage of RAFSwap, we adopt *identity loss*, *reconstruction loss*, and *perceptual loss*.

**Identity Loss.** A well-trained face recognition model can provide representative identity embeddings. We use cosine similarity to estimate the similarity between the identity embedding of the generated face and the source face, which can be written as:

$$\mathcal{L}_{id} = 1 - \cos(R(\boldsymbol{I}_s), R(\hat{\boldsymbol{I}}_{out})), \tag{8}$$

where $R(\cdot)$ is a pre-trained ArcFace [9] network.

**Reconstruction Loss.** If the source and the target faces are from the same identity, the generated face should look the same as the target face. We define a reconstruction loss as pixel-level $\mathcal{L}_2$ distances between the target face and the generated face, which can be written as:

$$\mathcal{L}_{rec} = \begin{cases} \left\| \hat{\boldsymbol{I}}_{out} - \boldsymbol{I}_t \right\|_2 & \textit{if } \boldsymbol{I}_t = \boldsymbol{I}_s \\ 0 & \textit{otherwise} \end{cases}. \tag{9}$$

**Perceptual Loss.** Besides measuring the difference between two faces at the pixel level, we utilize LPIPS [44] loss to calculate semantic errors between the target and generated faces. It can be written as:

$$\mathcal{L}_p = \left\| \phi_p(\hat{\boldsymbol{I}}_{out}) - \phi_p(\boldsymbol{I}_t) \right\|_2, \tag{10}$$

where $\phi_p(\cdot)$ represents the pre-trained VGG16 network.

The total loss is the weighted sum of all the above losses:

$$\mathcal{L}_{total} = \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_p\mathcal{L}_p. \tag{11}$$

## 4. Experiments

### 4.1. Dataset and Implementation Details

**Dataset.** For face swapping, CelebA-HQ [16] is a high-quality version of the CelebA [21], which has 30000 images with 1024 resolution. FaceForensics++ [27] is a forensic dataset consisting of 1000 video sequences from YouTube.

**Implementation Details.** We use CelebA-HQ dataset as the training set, and the values of the loss weights are set to $\lambda_{id} = 0.15$, $\lambda_{rec} = 1$, $\lambda_p = 0.8$, respectively. The ratio of the training data with $\boldsymbol{I}_t = \boldsymbol{I}_s$ and $\boldsymbol{I}_t \neq \boldsymbol{I}_s$ is set to $1 : 4$. The input images are resized to $256 \times 256$. During the training, the StyleGAN2 is fixed and the weights of the rest are updated by using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and learning rate=$1e^{-4}$. RAFSwap is trained with 50K steps, using 1 Tesla V100 GPU and 8 batch size.

### 4.2. Comparison with Previous Methods

**Qualitative Comparison.** We compare our method with FaceShifter [19], SimSwap [7], and MegaFS [46] on FaceForensics++. As shown in Fig. 5, we show some conditions that are prone to produce artifacts, including face shape, hairstyle, and brows with a large difference between source and target faces. We can see that MegaFS and our method can handle these challenges, but MegaFS could not preserve the attributes of the target face, such as skin color. Besides, our results share eye color with the source face much better than other methods in rows 4 and 5. Furthermore, we compare our method with FaceShifter and FaceInpainter [18] on wild face images. As shown in Fig. 6, benefiting from the well-designed identity integration and flexible soft mask generation modules, our results can well preserve the source identity information, *e.g.*, small mouth, target attributes, *e.g.*, hair color, and handle occlusion cases, *e.g.*, eyeglasses.

Since our method can generate high-resolution swapped faces, we compare RAFSwap with MegaFS on CelebA-HQ.


Figure 5. Comparison with FaceShifter [19], SimSwap [7], and MegaFS [46] on FaceForensics++ [27].

| Method | ID Ret.↑ | Pose↓ | Exp.↓ |
|---|---|---|---|
| DeepFakes [25] | 88.39 | 4.46 | 3.33 |
| FaceShifter [19] | 90.68 | 2.55 | 2.82 |
| SimSwap [7] | 89.73 | **1.94** | **2.39** |
| MegaFS [46] | 90.83 | 2.64 | 2.96 |
| Ours | **96.70** | 2.53 | 2.92 |

Table 1. Quantitative comparison results on FaceForensics++ [27]. **Bold** represent optimal result. The up arrow indicates that the larger the value, the better the model performance, and vice versa.

| Method | ID Sim.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|
| MegaFS [46] | 0.4837 | 3.85 | **3.13** | 18.81 |
| Ours | **0.5232** | **3.77** | 3.15 | **13.25** |

Table 2. Quantitative comparison results on CelebA-HQ [16].

As shown in Fig. 7, we sample four pairs of significant gaps between gender, age, skin color, and pose. Obviously, our method achieves higher identity-consistent results that share the same local and global representations with source faces, *e.g.*, eye color and skin wrinkle, and faithfully respect the attributes of the target face. Note that our method produces more harmonious fusion results around edges.

**Quantitative Comparison.** We follow the experiment settings in MegaFS, which is slightly different from FaceShifter in data preprocessing. Firstly, we sample 10 frames from each video and process them by MTCNN [43], resulting in 10K aligned faces. Because some videos display repeated identities and contain multiple faces in one

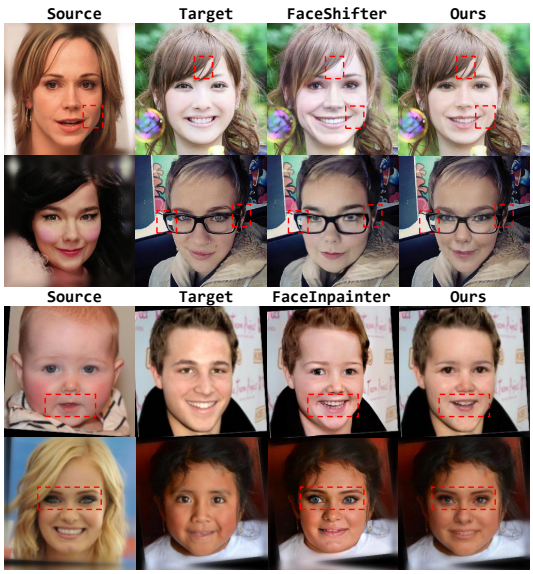


Figure 6. Comparison with FaceShifter [19] and FaceInpainter [18]. Images are from official attached results. Please zoom in the red dotted rectangles for a more clear comparison.

frame, we carefully check the aligned faces and manually categorize all videos into 885 identities. Then, we evaluate the accuracy of identity retrieval (abbrev. ID Ret.), pose, and expression errors (abbrev. Exp.). We apply CosFace [32] to extract identity embedding and retrieve the closest face by using cosine similarity. A pose estimator [28] and 3D facial model [10] are used to extract pose and expression vectors for pose and expression evaluation. We measure the $\mathcal{L}_2$ distances between swapped faces and the corresponding target faces. The comparison results are shown in Tab. 1, SimSwap preserves better attributes of the target face but a poor identity consistency. Our method achieves the highest ID retrieval, outperforming MegaFS with a large margin, and the comparable pose and expression errors with FaceShifter. Note that we have omitted comparisons with FaceInpainter quantitatively since the source codes are not publicly available.

For high-resolution swapped results comparison with MegaFS, we randomly sample 100K pairs of the face images in the CelebA-HQ test set. We report ID similarity (abbrev. ID Sim.), pose errors, expression errors, and FID. ID similarity is measured by calculating the cosine similarity of swapped faces and the corresponding source faces. As shown in Tab. 2, RAFSwap achieves a better performance in ID similarity and pose error than MegaFS but has a higher expression error. Because MegaFS adopts landmark loss, which produces the swapped face that faithfully respects the mouth shape of the target but leads to low identity-consistent with the source. Besides, the lower FID indicates that our method could generate more realistic images.

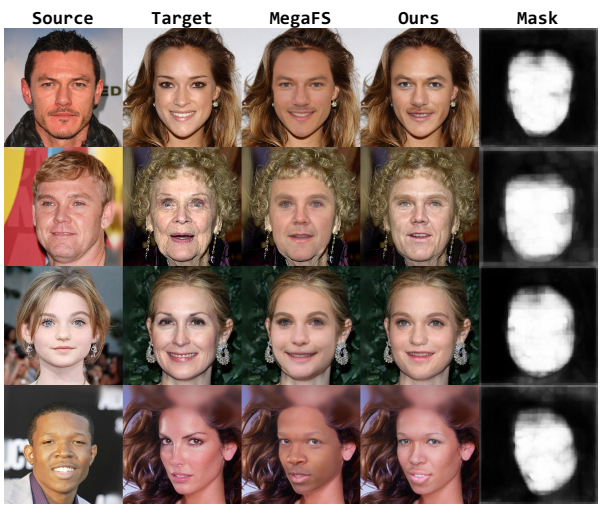


Figure 7. Comparison with MegaFS [46] on CelebA-HQ [16]. We sample some challenging conditions across significant gaps in gender, age, skin color, and pose.

| Method | Identity-Perception↑ | Quality↑ |
|---|---|---|
| DeepFakses [25] | 0.07 | 0.05 |
| FaceShifter [19] | 0.16 | 0.13 |
| SimSwap [7] | 0.13 | 0.09 |
| MegaFS [46] | 0.15 | 0.16 |
| Ours | **0.49** | **0.57** |

Table 3. Human Study results on FaceForensics++ [27].

**Human Study.** We conduct a human study to evaluate the performance of each method. Corresponding to two challenges, we let the users select: i) the one that has the most similar identity with the source face and shares the most similar attributes with the target face; ii) the most high-quality one. For each user, we randomly sample 20 pairs from the 1000 FaceForensics++ videos without duplication. The results reported in Tab. 3 are based on the answers from 50 users, showing that our method significantly surpasses the other four methods.

### 4.3. Ablation Study and Applications

**Feature Fusion module.** To verify that the combination of FRA and SFA is superior to AdaIN in the GAN inversion framework, we conduct qualitative and quantitative experiments. Specifically, we modify an AdaIN-based baseline that injects the identity vector into hierarchical feature maps. As shown in rows 2 and 4 of Fig. 8, the third column images produced by AdaIN could not preserve the facial identity details of the source face. Due to the global AdaIN operation and Fixed StyleGAN2 generator, this baseline could not adaptively maintain the detailed identity features and tend to express the general representation. For comparison, our method could generate more identity-consistent faces thanks to the well-designed feature integration module. Besides, comparing rows 1 and 5 in Tab. 4, our method markedly improves by 4.22 on ID Ret. over the AdaIN-based baseline. As a cost, an extra BiseNet of 13.3M requires 4 hours of training in CelebA-HQ. Furthermore, we analyze the necessity of FRA and SFA branches. As shown in Fig. 8, the isolated SFA exhibits a shortage of identity performance while the isolated FRA has the better identity performance but suffers from source facial texture mismatch. When both FRA and SFA are used, the generated faces preserve the local identity-relevant features and the global facial details of the source face. The quantitative experiments in Tab. 4 consistently demonstrate the effectiveness of each component and the superiority of our module.


Figure 8. Qualitative results of the ablation study. Our full model obtains better results than other variants.

**Attention Structure.** To verify the strong ability of the Transformer to capture token interactions, we conduct a quantitative experiment. Specially, we modify a comparative version that replaces a Transformer layer with a Non-Local layer [33]. As shown in the first three rows of Tab. 5, one Transformer layer improves the performance, while one Non-Local layer could not sufficiently model the token interaction and cause slight performance degradation. Besides, to evaluate the effect of layer numbers, we conduct a controlled experiment. As shown in the last three rows of Tab. 5, as the number of layers increases, the performance does not improve significantly. To balance the performance and calculation, we employ the Transformer with one layer and eight heads experimentally. Besides, we visualize one attention head for a source face. As shown in Fig. 9, the attention map indicates that the Transformer concentrate on different semantic regions across different scales, *i.e.*, the tokens from the large scale focus on eyes, the medium focus on lips, while the small focus on nose. Notably, the brows do not receive much attention due to the small areas and overlapped receptive fields with eyes.

| AdaIN | FRA | SFA | FMP | ID Ret.↑ | Pose↓ | Exp.↓ |
|---|---|---|---|---|---|---|
| ✓ | ✗ | ✗ | ✗ | 92.48 | 2.60 | 2.98 |
| ✗ | ✓ | ✗ | ✗ | 96.63 | 2.58 | 2.94 |
| ✗ | ✗ | ✓ | ✗ | 93.68 | 2.61 | 3.06 |
| ✗ | ✓ | ✓ | ✗ | 96.69 | 2.54 | 2.94 |
| ✗ | ✓ | ✓ | ✓ | **96.70** | **2.53** | **2.92** |

Table 4. Quantitative ablation study of RAFSwap with different proposed components on FaceForensics++ [27].

| Method | ID Ret.↑ | Pose↓ | Exp.↓ |
|---|---|---|---|
| + Non-Local [33] | 96.50 | 2.63 | 3.03 |
| + Tr-0 | 96.62 | 2.60 | 3.01 |
| + Tr-1 | 96.70 | 2.53 | 2.92 |
| + Tr-2 | 96.71 | **2.51** | 2.94 |
| + Tr-3 | **96.73** | 2.54 | **2.90** |

Table 5. Quantitative ablation study of RAFSwap with different attention components on FaceForensics++ [27].

| Method | CPU (s)↓ | GPU (ms)↓ | Params (M)↓ | Flops (G)↓ |
|---|---|---|---|---|
| LADN [13] | 8.70 | 26.8 | 26.99 | 175.77 |
| PSGAN [15] | 8.45 | 128.9 | **12.61** | 91.02 |
| Ours | **0.283** | **9.3** | 13.60 | **71.39** |

Table 6. Efficiency evaluation on makeup transfer. FPS is evaluated on a single Tesla V100.

**Face Mask Predictor.** To demonstrate the effectiveness of FMP, we provide two qualitative comparisons. As shown in Fig. 10, without the guidance of the face mask, our method could not keep some attributes unchanged, *e.g.*, background. Applying the hard ground truth mask on raw swapped faces produces excessive information and unnatural edges, especially on the bangs area. In comparison, our full model with the soft mask module achieves more harmonious fusion faces. FMP also brings improvement quantitatively, as shown in the last two rows of Tab. 4.

**Expanded Application of FRA.** We further apply our FRA branch in makeup transfer. Specifically, we adopt PSGAN [15] as the baseline. For a fair comparison, we only replace the AMM module of PSGAN with the FRA branch. As shown in Fig. 11, compared with LADN [13] and PSGAN, our method precisely transfers makeup colors with realistic results, where the identity and light on the source face are well preserved. Besides, we compare their running efficiency. The results are shown in Tab. 6. Our method is more than ten times faster than PSGAN on GPU. The expanded experiment demonstrates that FRA could also handle texture and color feature transfer due to the flexible token mechanism and sufficient feature interaction.

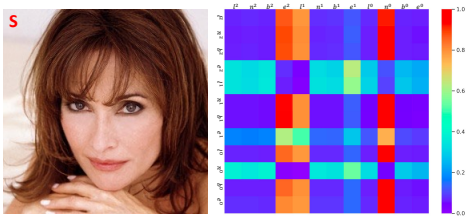


Figure 9. Attention visualization for a source face. Symbols $l, n, b, e$ means lips, nose, brows, eyes, respectively. Superscripts $0, 1, 2$ represent small, medium, and large scales, respectively.

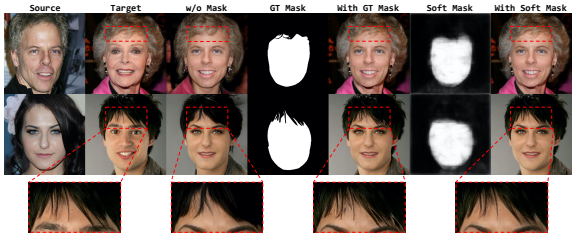


Figure 10. Qualitative results for FMP. We zoom in the red dotted rectangles of the second sample for more clear comparison.

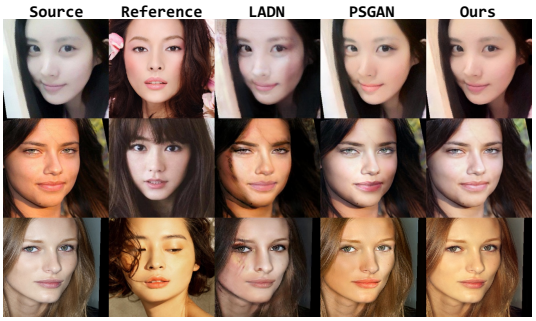


Figure 11. Comparison with SOTA makeup transfer methods on Makeup-Wild dataset [15].

## 5. Conclusion and Future Work

In this paper, we propose a novel RAFSwap built on the GAN inversion to generate high-resolution and identity-consistent swapped faces. Specifically, FRA integrates the identity-relevant local features into the target face, and SFA provides complementary identity-relevant details globally. Furthermore, FMP incorporated with StyleGAN2 is proposed to preserve the background and occlusions of the target unsupervisedly. Extensive experiments demonstrate the superiority of our approach over other SOTA methods.

Due to the limitation of the training dataset, inversion-based methods fail to handle out-range cases, *i.e.*, faces with various perspectives. We will further incorporate prior knowledge to improve the practicability of our method.

## 6. Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61836015.

## References


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.

[3] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. *arXiv preprint arXiv:2103.14968*, 2021.

[4] 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.

[5] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K Nayar. Face swapping: automatically replacing faces in photographs. In *ACM SIGGRAPH 2008 papers*, pages 1–8. 2008.

[6] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. In *Computer Graphics Forum*, volume 23, pages 669–676. Wiley Online Library, 2004.

[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.

[8] Yi-Ting Cheng, Virginia Tzeng, Yu Liang, Chuan-Chang Wang, Bing-Yu Chen, Yung-Yu Chuang, and Ming Ouhyoung. 3d-model-based face replacement in video. In *SIGGRAPH'09: Posters*, pages 1–1. 2009.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[10] 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.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[13] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10481–10490, 2019.

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[15] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5202, 2020.

[16] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[17] 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.

[18] 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.

[19] 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.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[21] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[22] Le Minh Ngo, Sezer Karaoglu, Theo Gevers, et al. Unified application of style transfer for face swapping and reenactment. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[23] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018.

[24] Yotam Nitzan, Amit Bermano, Yangyan Li, and Daniel Cohen-Or. Face identity disentanglement via latent space mapping. *arXiv preprint arXiv:2005.07728*, 2020.

[25] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[26] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[32] 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.

[33] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[34] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021.

[35] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

[36] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4432–4442, 2021.

[37] Zhiliang Xu, Xiyu Yu, Zhibin Hong, Zhen Zhu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Facecontroller: Controllable attribute editing for face in the wild. *arXiv preprint arXiv:2102.11464*, 2021.

[38] Shuai Yang and Kai Qiao. Shapeediter: a stylegan encoder for face swapping. *arXiv preprint arXiv:2106.13984*, 2021.

[39] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[40] Xianfang Zeng, Yusu Pan, Mengmeng Wang, Jiangning Zhang, and Yong Liu. Realistic face reenactment via self-supervised disentangling of identity and pose. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12757–12764, 2020.

[41] Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, and Yong Liu. Analogous to evolutionary algorithm: Designing a unified sequence model. *Advances in Neural Information Processing Systems*, 34, 2021.

[42] Jiangning Zhang, Xianfang Zeng, Mengmeng Wang, Yusu Pan, Liang Liu, Yong Liu, Yu Ding, and Changjie Fan. Freenet: Multi-identity face reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5326–5335, 2020.

[43] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[45] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020.

[46] 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.