AttFace: High-resolution Face Swapping with Attention Network

Shaoting Zhu*, Linzhan Mou*, Junyi Shen, Chao Xu, Yong Liu

***Abstract*—In recent years, face-swapping technology has undergone significant advancement driven by the success of generation models. However, the prevailing focus of face-generation tasks remains limited to low-resolution images of 256×256 primarily due to technological constraints and imperfect training data for high-resolution images. In this paper, we address two challenges in face-swapping: (1) using source images that have non-ideal attributes such as unusual facial expressions and head poses, and (2) obtaining high-quality face-swapping results on high-resolution images. To solve these problems, we propose IDentity transfer based on ATTention (IDATT), a method for high-fidelity face-swapping even with non-ideal source images. In addition, we introduce a new two-stage post-processing strategy to keep the consistency of non-identity information. IDATT produces high-quality results on high-resolution images by considering the relevance between the source and target. Experiment results show that our method produces more robust and visually pleasing results than previous methods.**

***Index Terms*—Face Swapping, self-attention, computer vision, deep learning.**

I. Introduction

FACE swapping is a complex process that aims to transfer the facial identity from a source image to a target image while preserving the target's non-identity attributes (e.g., pose and expression). Recently, this task has garnered significant interest, primarily due to its wide range of applications in fields like entertainment, the film industry, and privacy protection. Moreover, it is worth highlighting that advancements in face-swapping techniques have the potential to bolster the detection of manipulated or forged faces in contemporary data-driven technological advancements. However, despite the immense interest and significance surrounding face swapping, there remain unresolved challenges within the realm of research and development.

In recent years, the focus of research on face generation tasks [1]–[3] has predominantly revolved around low-resolution images, typically measuring 256×256 pixels. This emphasis can be attributed to the myriad limitations and challenges associated with generating high-resolution facial images. Firstly, existing techniques for generating high-resolution facial images often encounter certain restrictions that give rise to problems such as image sharpening or distortion. Furthermore, there is ample room for improvement in terms of preserving facial details, textures, and skin tones in these high-resolution outputs. Secondly, a significant challenge lies in ensuring the authenticity and realism of the generated high-resolution face images. Imperfections in generative models or shortcomings in the training data can lead to unnatural appearances, which is unacceptable in applications that require a high level of fidelity. Moreover, the generation of high-resolution images necessitates considerable computational complexity, making the training and inference processes for high-resolution face image generation tasks both time-consuming and challenging. This becomes particularly problematic for real-time generation applications where efficiency is crucial.

The emergence of StyleGAN [4], [5] has brought about a consensus regarding the utilization of powerful generators to produce high-quality face images. StyleGAN offers a convenient means to generate high-resolution face images measuring 1024×1024 pixels. However, due to the uncontrolled nature of the original StyleGAN2, the prevailing approach involves an inverse mapping scheme. This approach incorporates an encoder that maps images to latent codes and a fixed-parameter StyleGAN2 generator designed for high-resolution outputs, which receives these latent codes. These latent codes, located at various layers, contain crucial information pertaining to facial identity, attributes, and textures, thereby facilitating specific face-swapping tasks.

Building upon these insights, this paper focuses primarily on introducing IDentity transfer based on ATTention (IDATT) as a module to extract and exchange facial identity information. This module addresses the challenges of identity inconsistency in isolated face reenactment and the trade-off between attribute preservation and face swapping. The proposed approach enables the generation of high-resolution face images measuring 1024×1024 pixels. Moreover, a two-stage post-processing strategy is implemented to ensure consistency in non-identity information, thereby enhancing the quality and practicality of the generated high-resolution face images. Specifically, we employ a post-processing strategy on the source features to align their attributes with the target face before transmitting them to the decoder. By aligning the source face, our method achieves excellent preservation of the target face's attributes without compromising the performance of identity modification.

In summary, we make the following three contributions:

- We propose IDentity transfer based on ATTention (IDATT) and design the novel joint-learned IDTR (Identity Transformer) that transfers the attributes and maintains identity consistency for face-swapping.

S. Zhu, L. Mou, J. Shen, C. Xu and Y. Liu are with the APRIL Lab, Zhejiang University, Hangzhou, Zhejiang 310058, China (e-mail: zhust@zju.edu.cn; moulz@zju.edu.cn; j1shen@zju.edu.cn; 21832066@zju.edu.cn; yongliu@iipc.zju.edu.cn).

S. Zhu and L. Mou contribute equally to the article. Y. Liu is the corresponding author.

Manuscript received April 19, 2022; revised August 16, 2022.

- To further transfer the background of the target face, we propose the two-stage post-processing strategy. To be specific, we use an encoder to generate multi-resolution features from the target image and blend them with the corresponding features from the upsampling blocks of the StyleGAN generator.
- Abundant experiments qualitatively and quantitatively demonstrate the superiority of our method to generate high-fidelity faces. Specifically, our approach surpasses state-of-the-art unified and isolated methods in terms of generating reenacted faces that preserve identity with utmost accuracy and swapped faces that maintain consistent attribute preservation. These findings highlight the superiority of our method in capturing and reproducing a broader range of attributes, resulting in highly realistic and faithful face generation.

II. Related Works

A. Identity Swapping

The objective of identity swapping is to transform the identity of a source image into that of a target image. Early works [6]–[9] primarily focused on 3D-based methods. However, these methods often suffer from poor visual quality due to their reliance on the accuracy of non-trainable external 3D models. In recent years, significant advancements have been made with GAN-based [10] methods [3], [11], [12]. Notably, Faceshifter [1] adeptly incorporates identity and attribute embeddings using a carefully designed learning model. SimSwap [13] introduces a feature matching loss aimed at preserving more attribute embeddings, albeit at the expense of sacrificing identity similarity. Hififace [14] and FaceInpainter [15] leverage 3D face descriptors to enhance the geometric structure of the swapped results. These recent advancements in GAN-based methods have significantly pushed the boundaries of identity swapping, addressing issues related to visual quality, identity preservation, and geometric structure enhancement. However, the aforementioned methods overlook the modeling of identity-related feature interaction, which is susceptible to introducing identity-unrelated cues, leading to inadequate identity consistency and attribute preservation. To address these issues, we propose an innovative approach called Identity Transfer based on Self-Attention to mitigate the problems associated with poor identity consistency and attribute preservation, resulting in improved overall performance.

B. Attention-Based Style Migrator

Our IDATT method is closely related to recent self-attention methods [16], [17] used for image generation and style transfer. These methods operate in an embedding space where they determine the response at a specific position within a sequence or image by considering all positions and calculating their average weights. In a related work [18], a style transfer algorithm was proposed that effectively and flexibly modifies local style patterns based on the semantic spatial distribution of the content image. Drawing inspiration from this work, we attempted to employ SAnet (Self-Attention Network) as an identity migrator for the face-swapping task, to finely integrate identity information and further plug with attribute transfer to help maintain attributes.

C. High-resolution Face Generation

In light of the achievements brought about by StyleGAN [4], [19], several approaches have emerged as effective solutions for high-resolution face-swapping. MegaFS [20] leverages StyleGAN2 as the decoder, while subsequent studies [21], [22] employ the pSp [23] framework and style migration strategies to enhance attribute preservation. However, these approaches are limited by the inflexible nature of the fixed StyleGAN generator, prompting efforts to overcome this limitation. StyleFace [24] addresses this issue by introducing trainable parameters in the StyleGAN2 module during training. StyleSwap [25] improves the accuracy and resilience of face-swapping through the incorporation of a mask branch and an ID inversion strategy, resulting in high-fidelity outcomes. Nevertheless, the visual quality of the results on 1024×1024 images remained unsatisfactory. To address the aforementioned limitations and tackle the previously unaddressed problem of multi-source face swapping, we propose IDATT. Our method aims to overcome these challenges and achieve superior performance in high-resolution face-swapping tasks.

III. Method

A. High-efficiency Face-swapping Module

Generator. Previous approaches have employed identity and structure priors to extract disentangled representations that remain fixed during training. However, these fixed representations may be inaccurate and lead to performance degradation in challenging conditions. To overcome this limitation, we propose pSp [26] as the Encoder. Specifically, the pSp encoder consists of an encoder $\boldsymbol{\Phi}_E^{feat}$ that converts input images $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$ into low-resolution semantic feature maps at three levels: coarse $\boldsymbol{F}_c \in \mathbb{R}^{128\times H/4\times W/4}$, medium $\boldsymbol{F}_m \in \mathbb{R}^{256\times H/8\times W/8}$, and fine $\boldsymbol{F}_f \in \mathbb{R}^{512\times H/16\times W/16}$.

$$\boldsymbol{F}_c, \boldsymbol{F}_m, \boldsymbol{F}_f = \boldsymbol{\Phi}_E^{feat}(\boldsymbol{I}), \tag{1}$$

After that, the second part of pSp $\boldsymbol{\Phi}_E^{lat}$ can fuse these three feature maps into $W+$ space $\boldsymbol{Z} \in \mathbb{R}^{18\times 512}$.

$$\boldsymbol{Z} = \boldsymbol{\Phi}_E^{lat}(\boldsymbol{F}_c, \boldsymbol{F}_m, \boldsymbol{F}_f), \tag{2}$$

Finally, the style vector in $\boldsymbol{W}+$ space could directly be used to generate a high-resolution face image $\boldsymbol{O} \in \mathbb{R}^{3\times 1024\times 1024}$ taking advantage of the StyleGAN2 $\boldsymbol{\Phi}_D^{style}$.

$$\boldsymbol{O} = \boldsymbol{\Phi}_D^{style}(\boldsymbol{Z}). \tag{3}$$

IDentity transfer based on ATTention(IDATT). To effectively model the interaction of identity-related features and accurately aggregate identity information between identity and reference faces, we have developed a novel Identity Transfer module that is based on the self-attention mechanism.

The module's design, as shown in Fig. 1, involves extracting the query from the target features $\boldsymbol{F}_t$ using one convolution

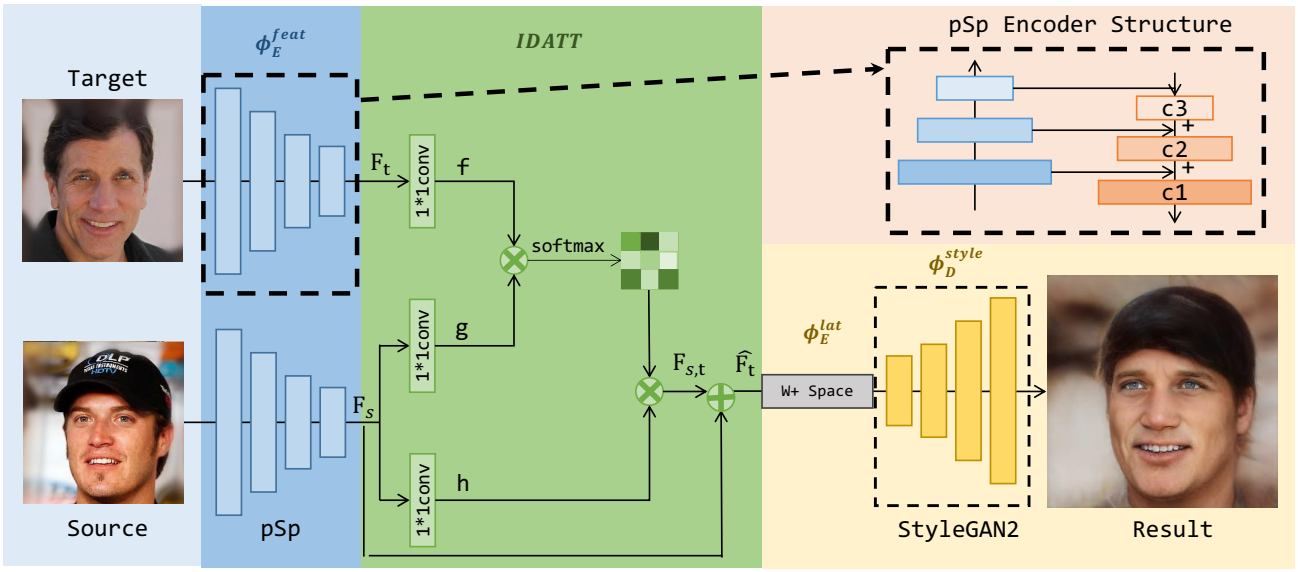

Fig. 1. **Overall architecture of our face-swapping framework.** The overall architecture is divided into three parts: 1) pSp backbone for feature extraction; 2) IDATT performing face-swapping; and 3) StyleGAN2 for image generation.

and obtaining the key and value from the source features $\boldsymbol{F}_s$, resulting in $\boldsymbol{Q}_t, \boldsymbol{K}_s, \boldsymbol{V}_s \in \mathbb{R}^{C/4\times H\times W}$ with reduced channel numbers.

$$\boldsymbol{F}_{s,t} = \text{Softmax}\left(\boldsymbol{Q}_t \left(\boldsymbol{K}_s\right)^T\right) \boldsymbol{V}_s = \boldsymbol{M}\boldsymbol{V}_s, \tag{4}$$

We then compute the correlation matrix $\boldsymbol{M}$ using $\boldsymbol{Q}_t$ and $\boldsymbol{K}_s$, which is used to multiply $\boldsymbol{V}_i$ and obtain $\boldsymbol{F}_{s,t}$. Finally, we apply a learned scale parameter γ, which is initialized to zero, to $\boldsymbol{F}_{s,t}$ to control the flow of identity transfer when it is added to $\boldsymbol{F}_t$.

$$\hat{\boldsymbol{F}}_t = \gamma \boldsymbol{F}_{s,t} + \boldsymbol{F}_t. \tag{5}$$

This innovative approach enables us to transfer identity information between faces with precision, using self-attention to effectively model the interaction between identity-related features. By aggregating identity information in this way, we are able to achieve more accurate and convincing results in our face-swapping application.

Face Swapping. Using the pSp (pixel2Style2pixel) encoder, we can exploit the capability of global context information by different-region-based context aggregation. Specifically, we can extract feature maps at three resolution levels from the source image and the target image through our pyramid pooling module.

$$\boldsymbol{F}_{sc}, \boldsymbol{F}_{sm}, \boldsymbol{F}_{sf} = \boldsymbol{\Phi}_E^{feat}\left(\boldsymbol{I}_s\right), \tag{6}$$

$$\boldsymbol{F}_{tc}, \boldsymbol{F}_{tm}, \boldsymbol{F}_{tf} = \boldsymbol{\Phi}_E^{feat}\left(\boldsymbol{I}_t\right), \tag{7}$$

Benefiting from representing identity with feature maps, IDATT is allowed to learn the explicit correlation between identity and reference faces on identity-related regions and transfer the identity information to the reference face adaptively. We perform IDATT feature fusion on three levels of feature maps.

$$\boldsymbol{F}_{sc,tc} = \text{IDATT}\left(\boldsymbol{F}_{sc}, \boldsymbol{F}_{tc}\right), \tag{8}$$

$$\boldsymbol{F}_{sm,tm} = \text{IDATT}\left(\boldsymbol{F}_{sm}, \boldsymbol{F}_{tm}\right), \tag{9}$$

$$\boldsymbol{F}_{sf,tf} = \text{IDATT}\left(\boldsymbol{F}_{sf}, \boldsymbol{F}_{tf}\right). \tag{10}$$

Then, using the lat layer of pSp $\boldsymbol{\Phi}_E^{lat}$ and StyleGAN2 $\boldsymbol{\Phi}_D^{style}$, we could learn more sufficient facial prior, enabling us to generate more realistic faces. Such a powerful generator successfully embeds the transformed features into the high-fidelity faces $\boldsymbol{O}$:

$$\boldsymbol{O} = \boldsymbol{\Phi}_D^{style}\left(\boldsymbol{\Phi}_E^{lat}\left(\boldsymbol{F}_{sc,tc}, \boldsymbol{F}_{sm,tm}, \boldsymbol{F}_{sf,tf}\right)\right), \tag{11}$$

The IDATT proposed in this paper operates directly on the feature map level and is therefore compatible with StyleGAN. In the training process, $\boldsymbol{\Phi}_E^{feat}$ and $\boldsymbol{\Phi}_D^{style}$ are fixed, while IDATT and $\boldsymbol{\Phi}_E^{lat}$ are trainable. This approach is advantageous due to its simple network structure, which absorbs the benefits of the self-attention mechanism and has a larger receptive field than the traditional convolutional network. Consequently, training time is fast, solving the problem of limited computing resources for high-resolution image generation and facilitating application deployment. In experiments, IDATT converges to a better result within 1 day, with a resolution of 1024×1024. Comparatively, SimSwap, a classic open-source face-changing method, needs 7 days under the same training conditions to achieve a similar effect, with a resolution of only 256×256.

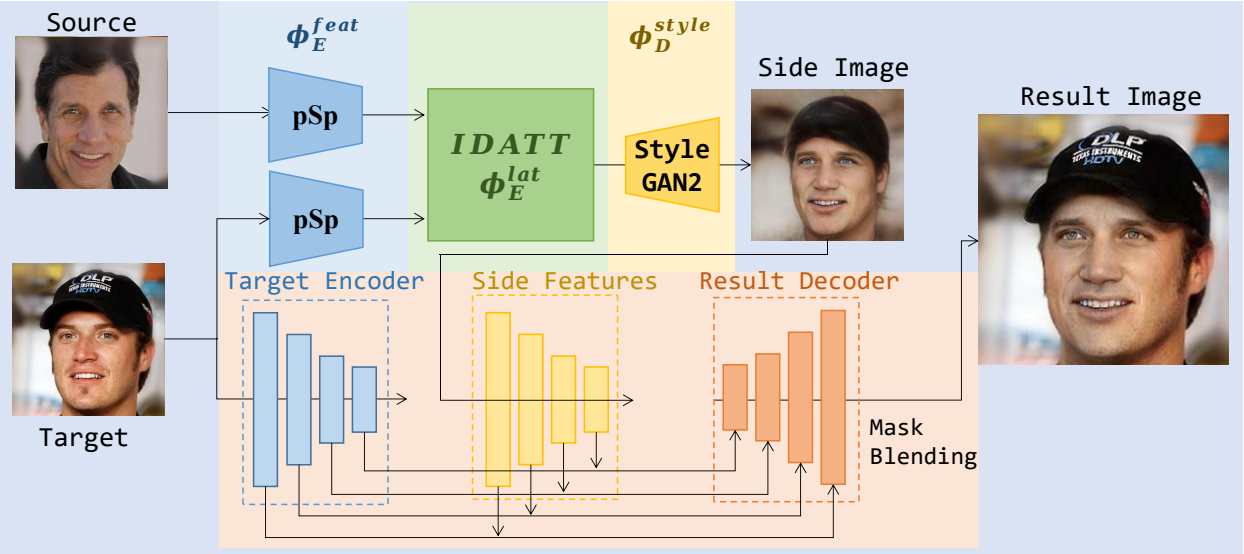

Fig. 2. **Two-stage post-processing.** To eliminate the blending boundary, we employ a multi-scale blending approach that combines target features with side features. Specifically, we discard the side image but retain its corresponding features produced by each upsampling block of the StyleGAN generator, extracted from the latent code $\widehat{w}_s$ with different levels of details. Simultaneously, we apply a target encoder to the target face such that its layers generate corresponding features of the same dimension. We then aggregate each pair of corresponding features by replacing the side image components for the inner-face region with our well-designed Mask Blending Module. All aggregated features are fed into a decoder to produce the result image.

B. Two-stage Post-processing Strategy

During our experiments, we discovered that the style transfer capability of IDATT was exceedingly strong, resulting in the over-editing of facial images. Consequently, aspects such as hairstyles, backgrounds, and skin tones were entirely altered, as illustrated in Fig. 1. To address this issue, we introduced a Two-stage post-processing strategy. This strategy involved incorporating a lateral encoder, decoder, and mask blending structure, which effectively addressed the problem and yielded significant enhancements in the overall face-swapping quality.

Building upon the original network structure, our approach initially generates a side image (pre-edited face) and subsequently utilizes an encoder-decoder to blend masks within each feature map. This process prevents abrupt changes at the stitching edges, ultimately producing the resulting image (final face). The Two-stage post-processing strategy effectively mitigates the over-editing phenomenon observed in IDATT's style transfer, ensuring that the generated facial images maintain a high degree of realism while incorporating the desired modifications.

Furthermore, the rapid training capabilities of IDATT enabled us to enhance the resolution of the images to 1024×1024 pixels, successfully achieving high-resolution face-swapping. The high resolution not only bolsters the visual fidelity of the synthesized images but also expands the potential applications of our method to various domains that require high-resolution facial images. The integration of the two-stage post-processing strategy, combined with the high-resolution capabilities, makes our approach a robust and versatile solution for face-swapping tasks.

C. Objective Functions

We use several loss terms to train our framework: adversarial loss $\mathcal{L}_{adv}$, reconstruction loss $\mathcal{L}_{rec}$ and perceptual loss $\mathcal{L}_p$ for face reenactment. Based on these three losses, identity loss $\mathcal{L}_{id}$ and 3D-aware swapping loss $\mathcal{L}_{3d}$ based on D3DFR [27] are adopted for face-swapping. Thus, the **total loss** is defined as follows:

$$L_{all} = \lambda_{adv}L_{adv} + \lambda_{rec}L_{rec} + \lambda_p L_p + \lambda_{id}L_{id} + \lambda_{3d}L_{3d}. \quad (12)$$

Adversarial Loss. We adopt adversarial training to ensure the overall quality and authenticity of the transformed faces:

$$\mathcal{L}_{adv} = \mathbb{E}_{x\sim p_x}[\log D(x)] + \mathbb{E}_{\tilde{x}\sim p_{\tilde{x}}}[\log(1 - D(\tilde{x}))], \quad (13)$$

Reconstruction Loss. The reconstruction loss can be defined by the pixel-level distance to measure the consistency when the source and target faces are from the same identity.

$$L_{rec} = \left\| I_{Sw} - I_t \right\|_2, \quad (14)$$

Perceptual Loss. Besides measuring the difference between two faces at the pixel level, we adopt LPIPS [28] loss to calculate the semantic errors to comprehensively measure the similarity of two faces:

$$L_p = \left\| \Phi_{vgg}\left(\hat{I_{Sw}}\right) - \Phi_{vgg}\left(I_t\right) \right\|_2, \quad (15)$$

Identity Loss. We calculate the cosine similarity to estimate the identity consistent between swapped and source faces:

$$L_{id} = 1 - \cos\left(R\left(\hat{I_{Sw}}\right), R\left(I_s\right)\right), \quad (16)$$

where $R(\cdot)$ is a pre-trained ArcFace [29] network.

Fig. 3. The result of our method. The first column on the left is the target, and the first row on the top is the source. It is noted that all of the images are 1024 × 1024!

3D-aware Swapping Loss. We use the cosine distance to further measure the similarity of 3D information between the generated face and the source face based on the pre-trained D3DFR [27] model, which can extract decoupled identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, expression $\boldsymbol{\beta} \in \mathbb{R}^{64}$, texture $\boldsymbol{\delta} \in \mathbb{R}^{80}$, illumination $\boldsymbol{\gamma} \in \mathbb{R}^{27}$, and pose $\boldsymbol{p} \in \mathbb{R}^{6}$. We utilize the vector of identity $\boldsymbol{\alpha} \in \mathbb{R}^{80}$ as our 3D-aware Swapping Loss.

$$L_{3d} = 1 - \cos\left(D\left(\hat{I_{Sw}}\right), D\left(I_s\right)\right). \tag{17}$$

where $D(\cdot)$ is a pre-trained D3DFR [27] network.

IV. Expetiment

A. Datasets

This article uses the commonly used datasets CelebA-HQ [30] and CelebAMask-HQ [31] for the face-swapping task, each dataset is briefly introduced as follows:

CelebA-HQ CelebA [32] is a large-scale celebrity feature dataset, which contains more than 10,000 identities and 200,000 face images, and detailed key points and face attributes expressions. 30,000 1024 × 1024 high-resolution face images are selected from CelebA to form the CelebA-HQ dataset, of which 28,000 are used for training and the rest are used for testing. This dataset is commonly used in tasks

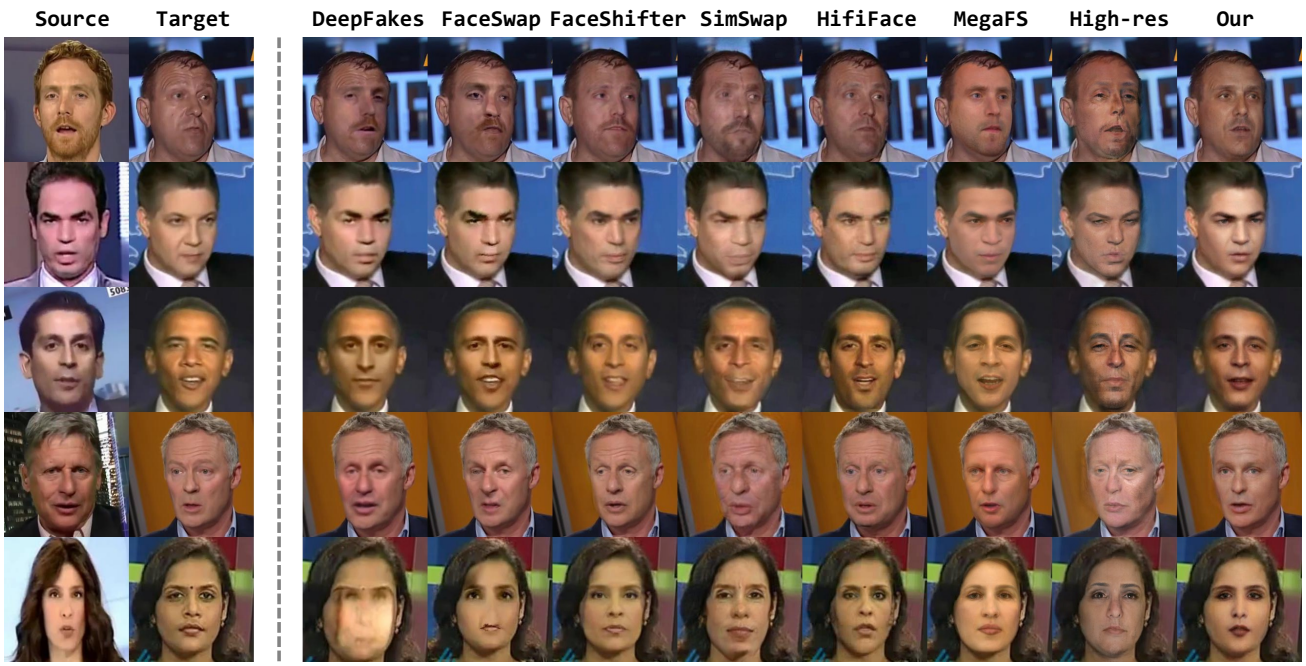

Fig. 4. Qualitative comparison with DeepFakes [33], FaceSwap, FaceShifter [1], SimSwap [2], HifiFace [34], MegaFS [20] and High-res [35] on FaceForensics++ [36]. Overall, our method and HifiFace generate more visually pleasing results.

such as face recognition, face editing, and high-resolution face generation.

CelebAMask-HQ CelebAMask-HQ performs face semantic labeling on the large-scale face dataset CelebA-HQ, including fine-grained mask labels, in which each face image is marked with 19 facial component categories, which contain the eyebrow, eye, nose, and lip regions used in this work. This dataset contains 30,000 512×512 resolution faces and corresponding 512×512 mask images, which are commonly used in face parsing, recognition, generation, and editing tasks. High-quality images and fine-grained mask annotations fit perfectly with the proposed method in this study.

B. Metrics

In order to compare with other advanced methods more fairly, this work adopts mainstream evaluation indicators, using ID Retrieval to evaluate identity transfer and pose and expression errors to evaluate attribute preservation. Considering that the accuracy of identity retrieval is computationally complex when the number of face pairs is large, the face-swapping results on CelebAMask-HQ use ID Similarity to measure identity consistency, and use FID [37] (Fréchet Inception Distance) to evaluate the quality of generated images. Specifically, both identity retrieval accuracy and identity similarity require face identity embedding representations.

Unlike using the ArcFace face recognition network during training, CosFace [38] is used during index evaluation to prevent unfair factors caused by fitting network bias. The former matches the closest face on the basis of the similarity measure, and calculates the ratio of the number of matching successes to the total data pairs to obtain the retrieval accuracy, while the latter directly reports the cosine distance of the data pairs. Pose error and expression error require face pose [39] and expression estimation [40] expert networks to extract pose and expression vector representations, and use L2 distance to measure the corresponding attributes between the generated face and the target face error. The FID indicator uses the Inception model to extract intermediate features to measure the correlation between generated images and real images.

C. Implement Details

In this study, the 30,000 faces in CelebAMask-HQ are divided into 28,000 faces for training and the remaining 2,000 faces for testing. During the training phase, the proportion of source and target faces being the same for all training face pairs is 0.5. The input of the network is uniformly scaled to 256 × 256 resolution, and the output is 1024 × 1024 resolution. All codes are built using Pytorch [41], the weights of StyleGAN2 remain fixed, and the remaining parameters are learned and updated using the Adam [42] optimizer, where $\beta_1 = 0.9, \beta_2 = 0.999$, and the learning rate is set is 1e-4, and the final model is obtained by training 50,000 iterations with a batch size of 6 on an A40 GPU. In the performance evaluation stage, different methods use the same crop alignment and scale to the same size for metric calculation.

D. Result Display

Fig. 3 presents additional visualization results that provide further evidence of the effectiveness of IDATT in generating high-resolution face-swapping outcomes. These results demonstrate the absence of noticeable fusion artifacts, as well as the preservation of the source face's identity and the attributes of the target face (all of the results are 1024 × 1024, so the enlarged view is recommended).

Fig. 5. Ablation study of the two-stage post-processing strategy. After applying the two-stage post-processing, there is a remarkable improvement in the consistency of non-identity elements between the generated result image and target image.

E. Comparision with Baselines

Qualitative comparison. We perform qualitative comparisons with seven SOTA unified methods. Fig. 4 shows a qualitative comparison of FF++ results. DeepFakes and FaceSwap have noticeable blending inconsistencies, distortion, and visual artifacts. FaceShifter and SimSwap show reasonable results. In the case of MegaFS, the results are generally clear, but there are some unnatural cases due to the difference in the color of the target and result faces. As for High-res, it can be observed that the transformation of facial features is not very natural, leading to noticeable distortions in facial expressions, as shown in row 1. Overall, our method and HifiFace generate more visually pleasing results. However, HifiFace fails to effectively transfer identity information into the result, as shown in rows 1, 4, and 5. Additionally, as shown in row 5, HifiFace's excessive consideration of the face shape caused visual discomfort when the shape of the target faces and the source faces greatly differ.

Quantitative comparison. When the face-swapping method was proposed in this paper, there were few open sources works on 1024 × 1024 high-definition resolution face-swapping in this field, mainly only MegaFS [20] and High-resolution [35]. For quantitative comparison, we just compare those two. Table. I shows that our method achieves state-of-the-art performance in ID-retrieval, Angle Error, Expression Error, and FID.

TABLE I
QUANTITATIVE COMPARISON

Method	ID-C↑	Angle↓	Exp↓	FID↓
MegaFS [20]	0.4837	3.8	3.13	18.81
High-res [35]	0.3181	4.2	3.45	25.74
Ours	**0.5879**	**3.5**	**3.12**	**15.54**

F. Ablation Study

In this subsection, we do ablation studies to verify that the network components and the structure in this study are reasonable and effective.

Face-swapping Module Based on Self-attention The pre-trained pSp encoder can decode the original face image and map it to three levels of feature maps: coarse, medium, and fine. This subsection performs ablation on whether to perform feature fusion on each level.

TABLE II
ABLATION STUDY ON CELEBA-HQ DATASETS

Method	ID-C↑	Angle↓	Exp↓	FID↓
Fine	0.4769	**3.4**	3.25	15.65
Medium+Fine	0.5208	3.5	**3.09**	15.56
Coarse+Medium+Fine	**0.5879**	**3.5**	3.12	**15.54**

It can be seen from the results that the result obtained by using all three has the best identity similarity, and far exceeds the other two. At the same time, although the post error and expression error are slightly worse, the gap is not large, and compared with other works it also has obvious advantages. Therefore, we add a face-swapping module to the three feature maps.

Two-stage Post-processing Strategy By employing the IDATT face-swapping module to obtain the implicit vector, followed by inputting it into StyleGAN2 to generate the resulting image, we encounter the issue of face over-editing. To address this concern, we employ a two-stage post-processing strategy for further refinement. It is worth noting that the side image may exhibit higher identity similarity with the source image compared to the resulting image. This phenomenon arises from altering the background of a portion of the target image, which is an undesired outcome. Currently, there is a lack of a suitable measurement index to evaluate the background consistency between the resulting image and the target image. Hence, this aspect can only be evaluated qualitatively.

The image result without post-processing is depicted on the left side of Fig. 5, while the image result after undergoing the two-stage post-processing is shown on the right side. It is evident that after applying the two-stage post-processing, there is a remarkable improvement in the consistency between

the generated result image and non-identity elements, such as the background and hair of the target image. Additionally, the overall quality of the generated image is significantly enhanced.

V. Conclusion And Future Work

This paper is dedicated to addressing high-resolution face-swapping tasks and introduces IDATT, a high-quality face-swapping module based on self-attention. IDATT enables the generation of high-definition faces at the megapixel level (1024×1024) while significantly accelerating the training process.

However, this study still has some limitations, as shown in Fig. 6. The usage of a fixed StyleGAN2 [19] in this paper subjects it to the constraints of prior knowledge during the training of StyleGAN2. Notably, StyleGAN2 struggles in generating side profiles and images with occluded faces. Although a two-stage post-processing strategy has been implemented to mitigate some of these issues, complete elimination of this limitation remains challenging. The specific limitations are depicted in the figure.

In future research, our aim is to enhance face-swapping by redesigning the StyleGAN2 module with trainable parameters and refining the mask blending process. This involves improving the model's capability to generate realistic swaps and exploring more effective blending methods for seamless integration. Thorough evaluations will be conducted to validate our proposed enhancements, ultimately enabling more realistic and seamless face swaps for various applications.

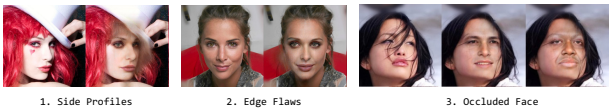

Fig. 6. Some failure cases caused by the limitation of the fixed StyleGAN2. Under the extreme lateral face and occluded face situations, our method may have some flaws.

References

[1] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5074–5083.

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

[3] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang *et al.*, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *arXiv preprint arXiv:2005.05535*, 2020.

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

[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[6] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," in *Computer Graphics Forum*, vol. 23, no. 3. Wiley Online Library, 2004, pp. 669–676.

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

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

[9] Y. Lin, S. Wang, Q. Lin, and F. Tang, "Face swapping under large pose variations: A 3d model based approach," in *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 2012, pp. 333–338.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[11] R. Natsume, T. Yatagawa, and S. Morishima, "Rsgan: face swapping and editing using face and hair representation in latent spaces," *arXiv preprint arXiv:1804.03447*, 2018.

[12] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "Towards open-set identity preserving face synthesis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6713–6722.

[13] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2003–2011.

[14] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *arXiv preprint arXiv:2106.09965*, 2021.

[15] J. Li, Z. Li, J. Cao, X. Song, and R. He, "Faceinpainter: High fidelity face adaptation to heterogeneous domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5089–5098.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.

[18] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5880–5888.

[19] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[20] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4834–4844.

[21] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7642–7651.

[22] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu, "Region-aware face swapping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7632–7641.

[23] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. 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*, 2021, pp. 2287–2296.

[24] Y. Luo, J. Zhu, K. He, W. Chu, Y. Tai, C. Wang, and J. Yan, "Styleface: Towards identity-disentangled face generation on megapixels," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*. Springer, 2022, pp. 297–312.

[25] Z. Xu, H. Zhou, Z. Hong, Z. Liu, J. Liu, Z. Guo, J. Han, J. Liu, E. Ding, and J. Wang, "Styleswap: Style-based generator empowers robust face swapping," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 2022, pp. 661–677.

[26] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. 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*, 2021, pp. 2287–2296.

[27] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. 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*, 2019, pp. 0–0.

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[29] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
[30] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
[31] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.
[33] R. Winter and A. Salter, "Deepfakes: uncovering hardcore open source on github," *Porn Studies*, vol. 7, no. 4, pp. 382–397, 2020.
[34] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "Hififace: 3d shape and semantic prior guided high fidelity face swapping," *arXiv preprint arXiv:2106.09965*, 2021.
[35] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7642–7651.
[36] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
[37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.
[38] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.
[39] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2074–2083.
[40] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. 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*, 2019, pp. 0–0.
[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.