HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping

Yuhan Wang[1,2*] , **Xu Chen**[1,3*] , **Junwei Zhu**[1] , **Wenqing Chu**[1] , **Ying Tai**[1†]
Chengjie Wang[1] , **Jilin Li**[1] , **Yongjian Wu**[1] , **Feiyue Huang**[1] and **Rongrong Ji**[3,4]

[1]Youtu Lab, Tencent
[2]Zhejiang University
[3]Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics,
Xiamen University
[4]Institute of Artificial Intelligence, Xiamen University
https://johann.wang/HifiFace

Figure 1: Face swapping results generated by our HifiFace. The face in the target image is replaced by the face in the source image.

Abstract

In this work, we propose a high fidelity face swapping method, called HifiFace, which can well preserve the face shape of the source face and generate photo-realistic results. Unlike other existing face swapping works that only use face recognition model to keep the identity similarity, we propose 3D shape-aware identity to control the face shape with the geometric supervision from 3DMM and 3D face reconstruction method. Meanwhile, we introduce the Semantic Facial Fusion module to optimize the combination of encoder and decoder features and make adaptive blending, which makes the results more photo-realistic. Extensive experiments on faces in the wild demonstrate that our method can preserve better identity, especially on the face shape, and can generate more photo-realistic results than previous state-of-the-art methods.

1 Introduction

Face swapping is a task of generating images with the identity from a source face and the attributes (*e.g.*, pose, expression, lighting, background, *etc.*) from a target image (as shown in Figure 1), which has attracted much interest with great potential usage in film industry [Alexander *et al.*, 2009] and computer games.

In order to generate high-fidelity face swapping results, there are several critical issues: (1) The identity of the result face including the *face shape* should be close to the source face. (2) The results should be *photo-realistic* which are faithful to the expression and posture of the target face and consistent with the details of the target image like lighting, background, and occlusion.

To preserve the identity of the generated face, previous works [Nirkin *et al.*, 2018; Nirkin *et al.*, 2019; Jiang *et al.*, 2020] generated inner face region via 3DMM fitting or face landmark guided reenactment and blend it into the target image, as shown in Figure 2(a). These methods are weak in identity similarity because 3DMM can not imitate the identity details and the target landmarks contain the identity of target image. Also, the blending stage restricts the change of face shape. As shown in Figure 2(b), [Liu *et al.*, 2019; Chen *et al.*, 2020] draw support from a face recognition net-

---

work to improve the identity similarity. However, face recognition network focuses more on texture and is insensitive to the geometric structure. Thus, these methods can not preserve the exact face shape robustly.

As for generating photo-realistic results, [Nirkin *et al.*, 2018; Nirkin *et al.*, 2019] used Poisson blending to fix the lighting, but it tended to cause ghosting and could not deal with complex appearance conditions. [Jiang *et al.*, 2020; Zhu *et al.*, 2020; Li *et al.*, 2019] designed an extra learning-based stage to optimize the lighting or occlusion problem, but they are fussy and can not solve all problems in one model.

To overcome the above defects, we propose a novel and elegant end-to-end learning framework, named HifiFace, to generate high fidelity swapped faces via 3D shape and semantic prior. Specifically, we first regress the coefficients of source and target face by a 3D face reconstruction model and recombine them as shape information. Then we concatenate it with the identity vector from a face recognition network. We explicitly use the 3D geometric structure information and use the recombined 3D face model with source's identity, target's expression, and target's posture as auxiliary supervision to enforce precise face shape transfer. With this dedicated design, our framework can achieve more similar identity performance, especially on face shape.

Furthermore, we introduce a Semantic Facial Fusion (SFF) module to make our results more photo-realistic. The attributes like lighting and background require spatial information and the high image quality results need detailed texture information. The low-level feature in the encoder contains spatial and texture information, but also contains rich identity from the target image. Hence, to better preserve the attributes without the harm of identity, our SFF module integrates the low-level encoder features and the decoder features by the learned adaptive face masks. Finally, in order to overcome the occlusion problem and achieve perfect background, we blend the output to the target by the learned face mask as well. Unlike [Nirkin *et al.*, 2019] that used the face masks of the target image for directly blending, HifiFace learns face masks at the same time under the guidance of dilated face semantic segmentation, which helps the model focus more on the facial area and make adaptive fusion around the edge. HifiFace handles image quality, occlusion, and lighting problems in *one* model, making the results more photo-realistic. Extensive experiments demonstrate that our results surpass other State-of-the-Art (SOTA) methods on wild face images with large facial variations.

Our contributions can be summarized as follows:

1. We propose a novel and elegant end-to-end learning framework, named HifiFace, which can well preserve the face shape of the source face and generate high fidelity face swapping results.

2. We propose a 3D shape-aware identity extractor, which can generate identity vector with exact shape information to help preserve the face shape of the source face.

3. We propose a semantic facial fusion module, which can solve occlusion and lighting problems and generate results with high image quality.

2   Related Work

3D-based Methods. 3D Morphable Models (3DMM) transformed the shape and texture of the examples into a vector space representation [Blanz and Vetter, 1999]. [Thies *et al.*, 2016] transferred expressions from source to target face by fitting a 3D morphable face model to both faces. [Nirkin *et al.*, 2018] transferred the expression and posture by 3DMM and trained a face segmentation network to preserve the target facial occlusions. These 3D-based methods follow a source-oriented pipeline like Figure 2(a) which only generates the face region by 3D fitting and blends it into the target image by the mask of the target face. They suffer from unrealistic texture and lighting because the 3DMM and the renderer can not simulate complex lighting conditions. Also, the blending stage limits the face shape. In contrast, our HifiFace accurately preserves the face shape via geometric information of 3DMM and achieves realistic texture and attributes via semantic prior guided recombination of both encoder and decoder feature.

GAN-based Methods. GAN has shown great ability in generating fake images since it was proposed by [Goodfellow *et al.*, 2014]. [Isola *et al.*, 2017] proposed a general image-to-image translation method, which proves the potential of conditional GAN architecture in swapping face, although it requires paired data.

The GAN-based face swapping methods mainly follow source-oriented pipeline or target-oriented pipeline. [Nirkin *et al.*, 2019; Jiang *et al.*, 2020] followed source-oriented pipeline in Figure 2(a) which used face landmarks to compose face reenactment. But it may bring weak identity similarity, and the blending stage limited the change of face shape. [Liu *et al.*, 2019; Chen *et al.*, 2020; Li *et al.*, 2019] followed target-oriented pipeline in Figure 2(b) which used a face recognition network to extract the identity and use decoder to fuse the encoder feature with identity, but they could not robustly preserve exact face shape and is weak in image quality. Instead, HifiFace in Figure 2(c) replaces the face recognition network with a 3D shape-aware identity extractor to better preserve identity including the face shape and introduces an SFF module after the decoder to further improve the realism.

Among these, FaceShifter [Li *et al.*, 2019] and Sim-Swap [Chen *et al.*, 2020] follow target-oriented pipeline and can generate high fidelity results. FaceShifter [Li *et al.*, 2019] leveraged a two-stage framework and achieved state-of-the-art identity performance. But it could not perfectly preserve the lighting despite using an extra fixing stage. However, HifiFace can well preserve lighting and identity in *one* stage. Meanwhile, HifiFace can generate photo-realistic results with higher quality than FaceShifter. [Chen *et al.*, 2020] proposed weak feature matching loss to better preserve the attributes, but it harms the identity similarity. While HifiFace can better preserve attributes and do not harm the identity.

3   Approach

Let I_s be the source images and I_t the target images, respectively. We aim to generate result image I_r with the identity of the I_s and the attributes of I_t. As illustrated in Figure 2(c), our pipeline consists of four parts: the Encoder part, Decoder
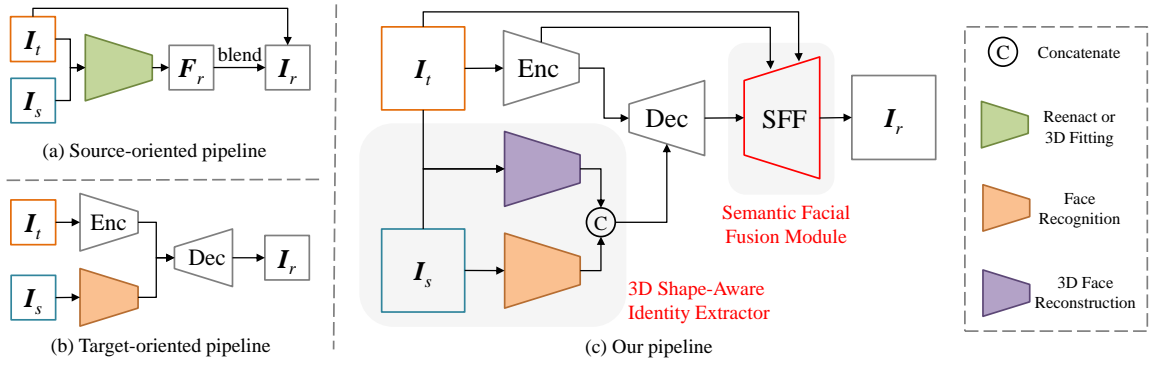
Figure 2: The pipelines of previous works and our HifiFace. (a) Source-oriented pipeline uses 3D fitting or reenactment to generate inner face region and blend it into the target image, in which F_r means the face region of the result. (b) Target-oriented pipeline uses a face recognition network to exact identity and combines encoder feature with identity in the decoder. (c) Our pipeline consists of four parts: the Encoder part, Decoder part, 3D shape-aware identity extractor, and SFF module. The encoder extracts features from I_t, and the decoder fuses the encoder feature and the 3D shape-aware identity feature. Finally, the SFF module helps further improve the image quality.

part, 3D shape-aware identity extractor (Sec. 3.1), and SFF module (Sec. 3.2). First, we set I_t as the input of the encoder and use several res-blocks [He *et al.*, 2016] to get the attribute feature. Then, we use the 3D shape-aware identity extractor to get 3D shape-aware identity. After that, we use res-block with adaptive instance normalization [Karras *et al.*, 2019] in decoder to fuse the 3D shape-aware identity and attribute feature. Finally, we use the SFF module to get higher resolution and make the results more photo-realistic.

3.1 3D Shape-Aware Identity Extractor

Most GAN-based methods only use a face recognition model to obtain identity information in the face swapping task. However, the face recognition network focuses more on texture and is insensitive to the geometric structure. To get more exact face shape features, we introduce 3DMM and use a pre-trained state-of-the-art 3D face reconstruction model [Deng *et al.*, 2019] as a shape feature encoder, which represents the face shape **S** by an affine model:

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

where $\bar{\mathbf{S}}$ is the average face shape; $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$ are the PCA bases of identity and expression; α and β are the corresponding coefficient vectors for generating a 3D face.

As illustrated in Figure 3(a), we regress 3DMM coefficients c_s and c_t, containing identity, expression, and posture of the source and target face by the 3D face reconstruction model F_{3d}. Then, we generate a new 3D face model by c_{fuse} with the source's identity, target's expression and posture. Note that posture coefficients do not decide face shape but may affect the 2D landmarks locations when computing the loss. We do not use the texture and lighting coefficients because the texture reconstruction still remains unsatisfactory. Finally, we concatenate the c_{fuse} with the identity feature v_{id} extracted by F_{id}, a pre-trained state-of-the-art face recognition model [Huang *et al.*, 2020], and get the final vector v_{sid}, called 3D shape-aware identity. Thus, HifiFace achieves well identity information including geometric structure, which helps preserve the face shape of the source image.

3.2 Semantic Facial Fusion Module

Feature-Level. The low-level feature contains rich spatial information and texture details, which may significantly help generate more photo-realistic results. Here, we propose the SFF module to not only make full use of the low-level encoder and decoder features, but also overcome the contradiction in avoiding harming the identity because of the target's identity information in low-level encoder feature.

As shown in Figure 3(b), we first predict a face mask M_{low} when the decoder features z_{dec} are of size $1/4$ of the target. Then, we blend z_{dec} by M_{low} and get z_{fuse}, formulated as:

$$z_{fuse} = M_{low} \odot z_{dec} + (1 - M_{low}) \odot \sigma(z_{enc}), \tag{2}$$

where z_{enc} means the low-level encoder feature with $1/4$ size of the original and σ means a res-block [He *et al.*, 2016].

The key design of SFF is to adjust the attention of the encoder and decoder, which helps disentangle identity and attributes. Specifically, the decoder feature in non-facial area can be damaged by the inserted source's identity information, thus we replace it with the clean low-level encoder feature to avoid potential harm. While the facial area decoder feature, which contains rich identity information of the source face, should not be disturbed by the target, therefore we preserve the decoder feature in the facial area.

After the feature-level fusion, we generate I_{low} to compute auxiliary loss for better disentangling the identity and attributes. Then we use a $4\times$ Upsample Module F_{up} which contains several res-blocks to better fuse the feature maps. Based on F_{up}, it's convenient for our HifiFace to generate even higher resolution results (*e.g.*, 512×512).

Image-Level. In order to solve the occlusion problem and better preserve the background, previous works [Nirkin *et al.*, 2019; Natsume *et al.*, 2018] directly used the mask of the target face. However, it brings artifacts because the face shape may change. Instead, we use SFF to learn a sightly dilated mask and embrace the change of the face shape. Specifically, we predict a 3-channel I_{out} and 1-channel M_r, and blend I_{out} to the target image by M_r, formulated as:

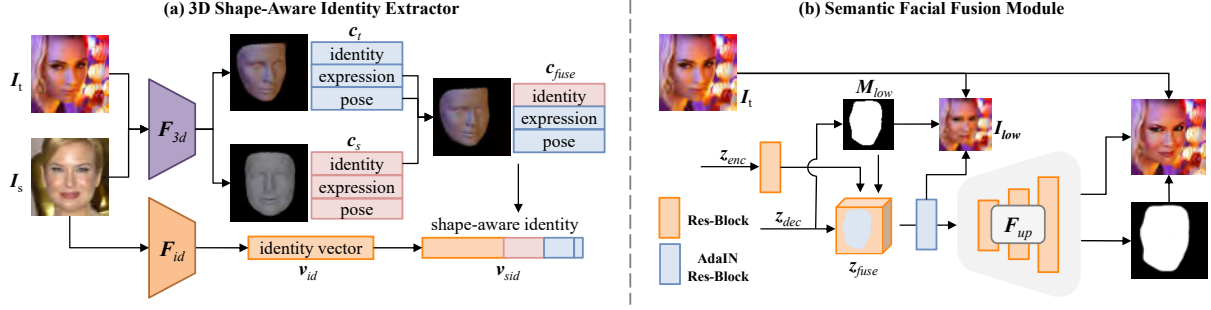$$I_r = M_r \odot I_{out} + (1 - M_r) \odot I_t. \tag{3}$$

Figure 3: Details of 3D shape-aware identity extractor and SFF module. (a) 3D shape-aware identity extractor uses F_{3d} (3D face reconstruction network) and F_{id} (face recognition network) to generate shape-aware identity. (b) SFF module recombines the encoder and decoder feature by M_{low} and makes the final blending by M_r. The F_{up} means the upsample Module.

In summary, HifiFace can generate photo-realistic results with high image quality and well preserve lighting and occlusion with the help of the SFF module. Note that these abilities still work despite the change of face shape, because the masks have been dilated and our SFF benefits from inpainting around the contour of predicted face.

3.3 Loss Function

3D Shape-Aware Identity (SID) Loss. SID loss contains shape loss and ID loss. We use 2D landmark keypoints as geometric supervision to constrain the face shape, which is widely used in 3D face reconstruction [Deng *et al.*, 2019]. First we use a mesh renderer to generate 3D face model by coefficients of source image identity and target image expression and posture. Then, we generate 3D face model of I_r and I_{low} by regressing 3DMM coefficients. Finally we project the 3D facial landmark vertices of reconstructed face shapes onto the image obtaining landmarks $\{q^{fuse}\}$, $\{q^r\}$ and $\{q^{low}\}$:

$$\mathcal{L}_{shape} = \frac{1}{N}\sum_{n=1}^{N}||q_n^{fuse} - q_n^r||_1 + ||q_n^{fuse} - q_n^{low}||_1. \quad (4)$$

Also, we use identity loss to preserve source image's identity:

$$\mathcal{L}_{id} = (1 - cos(v_{id}(I_s), v_{id}(I_r))) + \quad (5)$$
$$(1 - cos(v_{id}(I_s), v_{id}(I_{low}))),$$

where v_{id} means the identity vectors generated by F_{id}, and $cos(,)$ means the cosine similarity of two vectors. Finally, our SID loss is formulated as:

$$\mathcal{L}_{sid} = \lambda_{shape}\mathcal{L}_{shape} + \lambda_{id}\mathcal{L}_{id}, \quad (6)$$

where $\lambda_{id} = 5$ and $\lambda_{shape} = 0.5$.

Realism Loss. Realism loss contains segmentation loss, reconstruction loss, cycle loss, perceptual loss, and adversarial loss. Specifically, M_{low} and M_r in the SFF module are both under the guidance of a SOTA face segmentation network HRNet [Sun *et al.*, 2019]. We dilated the masks of the target image to eliminate the limitation in face shape change and get M_{tar}. The segmentation loss is formulated as:

$$\mathcal{L}_{seg} = ||R(M_{tar}) - M_{low}||_1 + ||M_{tar} - M_r||_1, \quad (7)$$
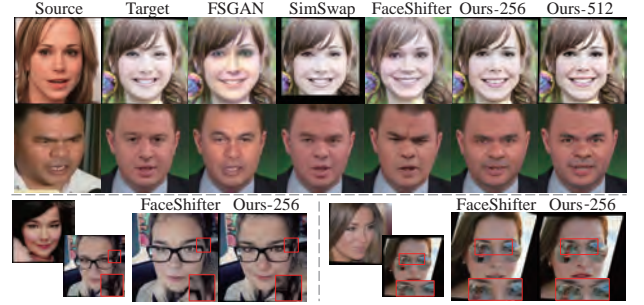
where $R(.)$ means the resize operation.

Figure 4: Comparison with FSGAN, SimSwap and FaceShifter. Our results can well preserve the source face shape, target attributes and have higher image quality, even when handling occlusion cases.

If I_s and I_t share the same identity, the predicted image should be the same as I_t. So we use reconstruction loss to give pixel-wise supervision:

$$\mathcal{L}_{rec} = ||I_r - I_t||_1 + ||I_{low} - R(I_t)||_1. \quad (8)$$

The cycle process can be conducted in the face swapping task too. Let I_r as the re-target image and the original target image as the re-source image. In the cycle process, we hope to generate results with re-source image's identity and re-target image's attributes, which means it should be the same as the original target image. The cycle loss is a supplement of pixel supervision and can help generate high-fidelity results:

$$\mathcal{L}_{cyc} = ||I_t - G(I_r, I_t)||_1, \quad (9)$$

where G means the whole generator of HifiFace.

To capture fine details and further improve the realism, we follow the Learned Perceptual Image Patch Similarity (LPIPS) loss in [Zhang *et al.*, 2018] and adversarial objective in [Choi *et al.*, 2020]. Thus, our realism loss is formulated as:

$$\mathcal{L}_{real} = \mathcal{L}_{adv} + \lambda_0\mathcal{L}_{seg} + \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{cyc} + \lambda_3\mathcal{L}_{lpips}, \quad (10)$$

where $\lambda_0 = 100$, $\lambda_1 = 20$, $\lambda_2 = 1$ and $\lambda_3 = 5$.

Overall Loss. Our full loss is summarized as follows:

$$\mathcal{L} = \mathcal{L}_{sid} + \mathcal{L}_{real}. \quad (11)$$

Method	ID↑	Pose↓	Shape↓	MAC↓	FPS↑
FaceSwap	54.19	2.51	0.610	-	-
FaceShifter	97.38	2.96	0.511	121.79	22.34
SimSwap	92.83	**1.53**	0.540	**55.69**	**31.17**
Ours-256	**98.48**	2.63	**0.437**	102.39	25.29

Table 1: Quantitative Experiments on FaceForensics++. FPS is tested under GPU V100.

Method	FF++		DFDC	
	AUC↓	AP↓	AUC↓	AP↓
FaceSwap	66.83	65.99	92.30	92.54
FaceShifter	41.62	42.92	77.18	76.50
SimSwap	76.44	72.63	78.80	78.44
Ours-256	**38.97**	**41.54**	**62.29**	**59.99**

Table 2: Results in terms of AUC and AP on FF++ and DFDC.
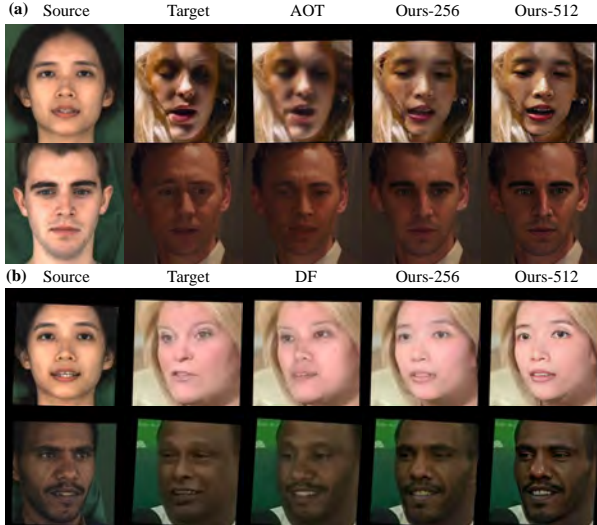
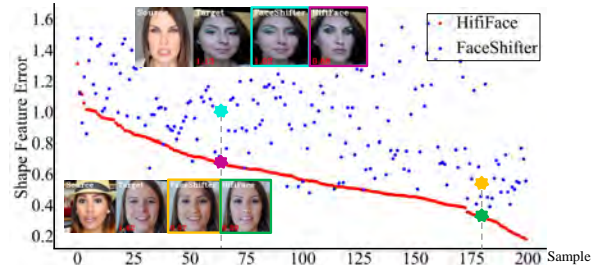Figure 5: (a) Comparison with AOT. (b) Comparison with DF.

Figure 6: Face shape error between I_r and I_s in 200 FF++ pairs with large shape difference. Samples are sorted by shape error of HifiFace. Same column index indicates the same source/target pair.

4 Experiments

Implementation Details. We choose VGGFace2 [Cao *et al.*, 2018] and Asian-Celeb [DeepGlint, 2020] as the training set. For our model with resolution 256 (*i.e.*, Ours-256), we remove images with either size smaller than 256 for better image quality. For each image, we align the face using 5 landmarks and crop to 256×256 [Li *et al.*, 2019], which contains the whole face and some background regions. For our more precise model (*i.e.*, Ours-512), we adopt a portrait enhancement network [Li *et al.*, 2020] to improve the resolution of the training images to 512×512 as supervision, and also correspondingly add another res-block in F_{up} of SFF compared to Ours-256. The ratio of training pairs with the same identity is 50%. ADAM [Kingma and Ba, 2014] is used with $\beta1 = 0$; $\beta2 = 0.99$ and learning rate = 0.0001. The model is trained with 200K steps, using 4 V100 GPUs and 32 batch size.

4.1 Qualitative Comparisons

First, we compare our method with FSGAN [Nirkin *et al.*, 2019], SimSwap [Chen *et al.*, 2020] and FaceShifter [Li *et al.*, 2019] in Figure 4, AOT [Zhu *et al.*, 2020] and Deeper-Forensics (DF) [Jiang *et al.*, 2020] in Figure 5.

As shown in Figure 4, FSGAN shares the same face shape with target faces and it can not well transfer the lighting of the target image either. SimSwap can not well preserve the identity of the source image especially for the face shape because it uses a feature matching loss and focuses more on the attributes. FaceShifter exhibits a strong identity preserva-

tion ability, but it has two limitations: (1) Attribute recovery, while our HifiFace can well preserve all the attributes like face color, expression, and occlusion. (2) Complex framework with two stages, while HifiFace presents a more elegant end-to-end framework with even better recovered images. As shown in Figure 5(a), AOT is specially designed to overcome the lighting problem but is weak in identity similarity and fidelity. As shown in Figure 5(b), DF has reduced the bad cases of style mismatch, but is weak in identity similarity too. In contrast, our HifiFace not only perfectly preserves the lighting and face style, but also well captures the face shape of the source image and generates high quality swapped faces. More results can be found in our supplementary material.

4.2 Quantitative Comparisons

Next, we conduct quantitative comparison on FaceForensics (FF)++ [Rossler *et al.*, 2019] dataset with respect to the following metrics: ID retrieval, pose error, face shape error, and performance on face forgery detection algorithms to again demonstrate the effectiveness of our HifiFace. For FaceSwap [FaceSwap,] and FaceShifter, we evenly sample 10 frames from each video and compose a 10K test set. For SimSwap and our HifiFace, we generate face swapping results with the same source and target pairs above.

For ID retrieval and pose error, we follow the same setting in [Li *et al.*, 2019; Chen *et al.*, 2020]. As shown in Table 1, HifiFace achieves the best ID retrieval score and is comparable with others in pose preservation. For face shape error, we use another 3D face reconstruction model [Sanyal *et al.*, 2019] to regress coefficients of each test face. The error is computed by L2 distances of identity coefficients between the swapped face and its source face, and our HifiFace achieves the lowest face shape error. The parameter and speed comparisons are also shown in Table 1, and our HifiFace is faster than FaceShifter, along with higher generation quality.
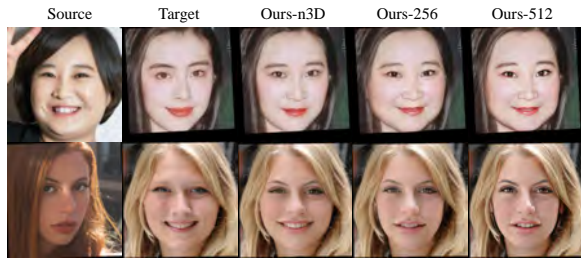
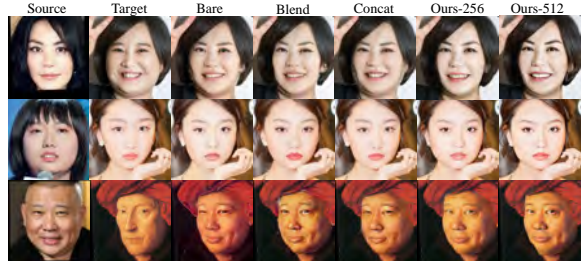Figure 7: Ablation study for 3D shape-aware identity extractor.

Figure 8: Ablation study for SFF module.

Figure 9: Comparison with results using directly mask blending. 'Blend-T', 'Blend-DT', and 'Blend-R' mean blending bare results to the target image by the mask of target, the dilated mask of target and the mask of bare results, respectively.
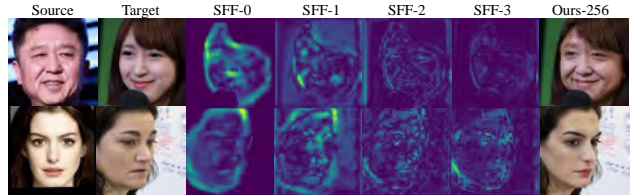

Figure 10: Difference feature maps of SFF.

To further illustrate the ability of HifiFace in controlling the face shape, we visualize the samplewise shape differences between HifiFace and FaceShifter [Li *et al.*, 2019] in Figure 6. The results show that, when the source and target differs much in face shape, HifiFace significantly outperforms Faceshifter with **95%** samples having smaller shape errors.

Besides, we apply the models from FF++ [Rossler *et al.*, 2019] and DeepFake Detection Challenge (DFDC) [Dolhansky *et al.*, 2019; selimsef, 2020] to examine the realism performance of HifiFace. The test set contains 10K swapped faces and 10K real faces from FF++ for each method. As shown in Table 2, HifiFace achieves the best score, indicating higher fidelity to further help improve face forgery detection.

4.3 Analysis of HifiFace

3D Shape-Aware Identity. To verify the effectiveness of shape supervision $\mathcal{L}_{shape}$ on face shape, we train another model Ours-n3d, which replaces the shape-aware identity vector with the normal identity vector from $\boldsymbol{F}_{id}$. As shown in Figure 7, the results of Ours-n3d can hardly change the face shape or have obvious artifacts, while the results of Ours-256 can generate results with much more similar face shape.

Semantic Facial Fusion. To verify the necessity of the SFF module, we compare with 3 baseline models: (1) 'Bare' that removes both the feature-level and image-level fusion. (2) 'Blend' that removes the feature-level fusion. (3) 'Concat' that replaces the feature-level fusion with a concatenate. As shown in Figure 8, 'Bare' can not well preserve the background and occlusion, 'Blend' lacks legibility, and 'Concat' is weak in identity similarity, which proves that the SFF module can help preserve the attribute and improve the image quality without harming the identity.

Face Shape Preservation in Face Swapping. Face shape preservation is quite difficult for face swapping, which is not just because of the difficulty in getting shape information, but also the challenge of inpainting when the face shape has changed. Blending is a valid way to preserve occlusion and background, but it is hard to be applied when the face shape changes. As shown in Figure 9, when the source face is fatter than the target face (row 1), it may limit the change of face shape in Blend-T. If we use Blend-DT or Blend-R, it can not well handle the occlusion. When the source face is thinner than the target (row 2), it is easy to bring artifacts around the face in Blend-T and Blend-DT and may cause a double face in Blend-R. In contrast, our HifiFace can apply the blending without the above issue, because our SFF module has the ability to inpaint the edge of the predicted mask.

To further illustrate how SFF addresses the problem, we show difference feature maps of every stage in the SFF module, named SFF-0∼3, between the input of (I_s, I_t) and (I_t, I_t), where (I_s, I_t) obtains Ours-256 and (I_t, I_t) achieves target itself. In Figure 10, the bright area means where the face shape changes or contains artifacts. SFF module recombines the feature between the face region and non-face area and focuses more on the contour of the predicted mask, which brings great benefits for inpainting areas where the shape changes.

5 Conclusions

In this work, we propose a high fidelity face swapping method, named HifiFace, which can well preserve the face shape of the source face and generate photo-realistic results. A 3D shape-aware identity extractor is proposed to help preserve identity including face shape. An SFF module is proposed to achieve a better combination in feature-level and image-level for realistic image generation. Extensive experiments demonstrate that our method can generate higher fidelity results than previous SOTA face swapping methods both quantitatively and qualitatively. Last but not least, HifiFace can also be served as a sharp spear, which contributes to the development of the face forgery detection community.

References

[Alexander *et al.*, 2009] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. Creating a photoreal digital actor: The digital emily project. In *2009 Conference for Visual Media Production*, pages 176–187. IEEE, 2009.

[Blanz and Vetter, 1999] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[Cao *et al.*, 2018] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *FG*, pages 67–74. IEEE, 2018.

[Chen *et al.*, 2020] 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.

[Choi *et al.*, 2020] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[DeepGlint, 2020] DeepGlint. http://trillionpairs.deepglint.com. *Accessed: 2020-12-20*, 2020.

[Deng *et al.*, 2019] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[Dolhansky *et al.*, 2019] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[FaceSwap,] FaceSwap. https://github.com/ondyari/faceforensics/tree/master/dataset/faceswapkowalski. *Accessed: 2020-12-20*.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Huang *et al.*, 2020] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[Jiang *et al.*, 2020] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2886–2895. IEEE, 2020.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2019] 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.

[Li *et al.*, 2020] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *European Conference on Computer Vision*, pages 399–415. Springer, 2020.

[Liu *et al.*, 2019] Jialun Liu, Wenhui Li, Hongbin Pei, Ying Wang, Feng Qu, You Qu, and Yuhao Chen. Identity preserving generative adversarial network for cross-domain person re-identification. *IEEE Access*, 7:114021–114032, 2019.

[Natsume *et al.*, 2018] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Fsnet: An identity-aware generative model for image-based face swapping. In *Asian Conference on Computer Vision*, pages 117–132. Springer, 2018.

[Nirkin *et al.*, 2018] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *FG*, pages 98–105. IEEE, 2018.

[Nirkin *et al.*, 2019] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *ICCV*, 2019.

[Rossler *et al.*, 2019] 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 International Conference on Computer Vision*, pages 1–11, 2019.

[Sanyal *et al.*, 2019] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019.

[selimsef, 2020] selimsef. https://github.com/selimsef/dfdc_deepfake_challenge. *Accessed: 2021-01-10*, 2020.

[Sun *et al.*, 2019] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019.

[Thies *et al.*, 2016] 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.

[Zhang *et al.*, 2018] 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.

[Zhu *et al.*, 2020] Hao Zhu, Chaoyou Fu, Qianyi Wu, Wayne Wu, Chen Qian, and Ran He. Aot: Appearance optimal transport based identity swapping for forgery detection. *Advances in Neural Information Processing Systems*, 33, 2020.
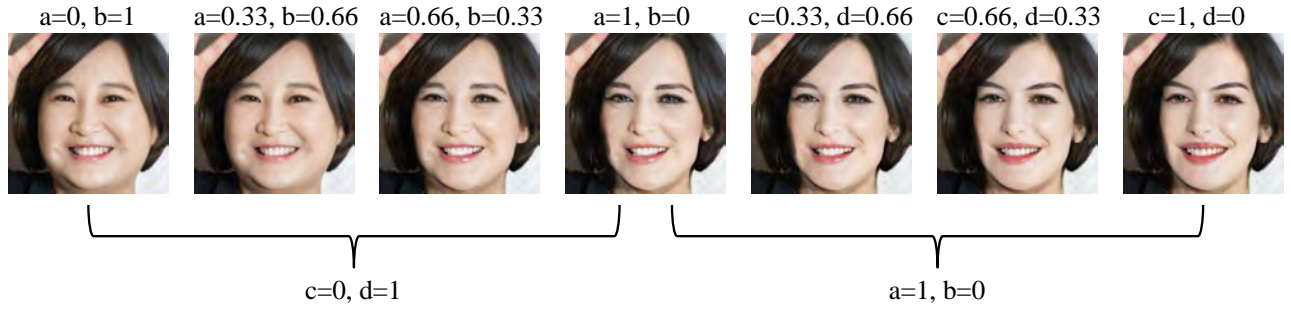
a=0, b=1   a=0.33, b=0.66   a=0.66, b=0.33   a=1, b=0   c=0.33, d=0.66   c=0.66, d=0.33   c=1, d=0

c=0, d=1   a=1, b=0

Figure 11: Interpolated results with different compositions of SID.

Network Structures

Detailed structures of our HifiFace are given in Figure 12. For all residual units, we use the Leaky ReLU (LReLU) as the activation function. Resample means the Average Pooling or the Upsampling, which is used to change the size of feature maps. Res-Blocks with the Instance Normalization (IN) are used in encoder, while Res-Blocks with the Adaptive Instance Normalization (AdaIN) are used in decoder.

More Results

To analyse the specific impacts of the shape information from 3D face reconstruction model and the identity information from face recognition model, we adjust the composition of SID to generate interpolated results. It is formulated as:

$$\psi_{in} = a\psi_s + b\psi_t, \tag{12}$$

$$v_{in} = cv_s + dv_t, \tag{13}$$

where ψ_s, ψ_t and ψ_{in} means the 3D identity coefficients of source, target and interpolated image, v_s, v_t and v_{in} means source, target and interpolated image's identity vector from recognition model.

As we can see in Figure 11 rows $1 - 4$, we first fix $c = 0$ and $d = 1$, the face shape can still change but lake of identity detail. Then in rows $4 - 7$, we fix $a = 1$ and $b = 0$, and the identity becomes more similar. The results prove that the shape information control the basic of shape and identity, while the identity vector is helpful to the identity texture.

In the end, we download lots of wild face images from Internet and generate more face swapping results in Figure 13 and Figure 14 to demonstrate the strong capability of our methods. And more results can be found at https://johann.wang/HifiFace.
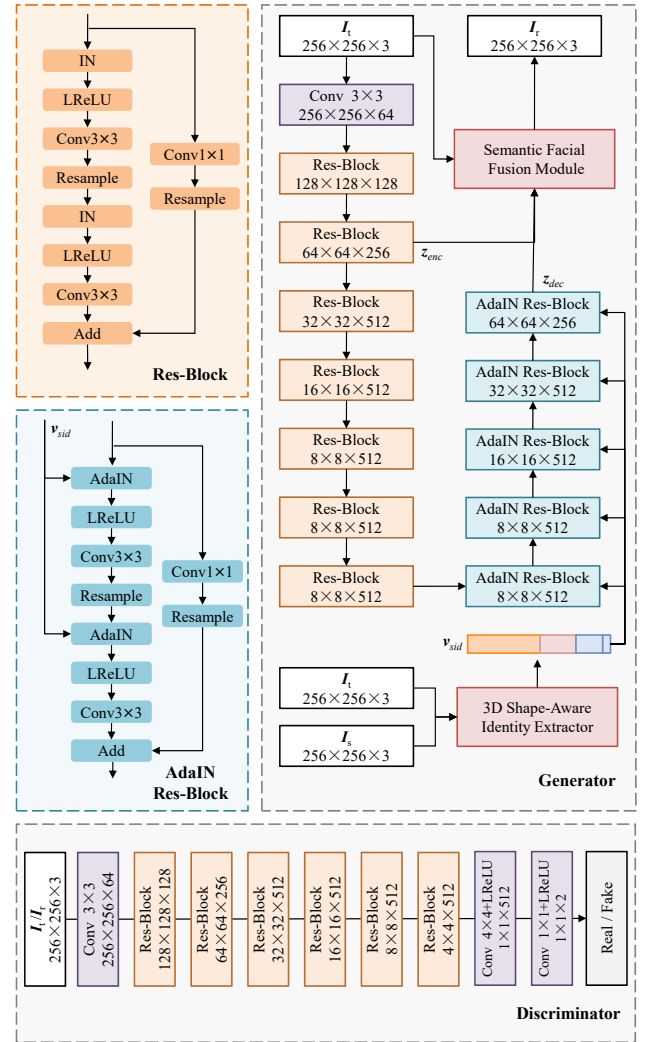

Figure 12: Architectural details of HifiFace.

Figure 13: More results on high resolution wild faces. The face in the target image is replaced by the face in the source image.

Figure 14: Some results on high quality real world photos. The face in the target image is replaced by the face in the source image. This is to show that the faces generated by our method can be very naturally integrated into high-resolution shot real world scenes.