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

Yuhan Wang1,2*, Xu Chen1,3*, Junwei Zhu1, Wenqing Chu1, Ying Tai1,†
Chengjie Wang1, Jilin Li1, Yongjian Wu1, Feiyue Huang1 and Rongrong Ji1,4
1Youtu Lab, Tencent
2Zhejiang University
3Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University
4Institute of Artificial Intelligence, Xiamen University

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. Code is available at: https://johann.wang/HifiFace.

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 network 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; Zha 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

*Equal Contribution. Work done when Yuhan and Xu are both interns at Youtu Lab, Tencent.
†Corresponding Author.
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 fitting a 3D morphable face model to both faces. [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 part, 3D shape-aware identity extractor (Sec. 3.1), and SFF
module (Sec. 3.2). First, we set $I_s$ 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:

$$S = S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta, \quad (1)$$

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

As illustrated in Figure 3(a), we regress 3DMM coefficients $c_s$ and $c_r$, 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 $c_{fus}$ with the source’s identity, target’s expression and posture. Note that posture coefficients do not decide face shape but may affect the 3D 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 $c_{fus}$ with the identity feature $v_{sid}$ extracted by $F_{sid}$, 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 $z_{fus}$, formulated as:

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

where $z_{enc}$ means the low-level encoder feature with 1/4 size of the original and $\sigma$ 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 4x 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., $512 \times 512$).

#### 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 slightly 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:

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

---

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 $I_s$, and the decoder fuses the encoder feature and the 3D shape-aware identity feature. Finally, the SFF module helps further improve the image quality.
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 diluted 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 \( \{ q_{low} \} \):

\[
\mathcal{L}_{shape} = \frac{1}{N} \sum_{n=1}^{N} \| q_{fuse}^n - q^n_r \|_1 + \| q_{fuse}^n - q_{low}^n \|_1.
\]

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

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

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:

\[
\mathcal{L}_{sid} = \lambda_{shape}\mathcal{L}_{shape} + \lambda_{id}\mathcal{L}_{id},
\]

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 diluted the masks of the target image to eliminate the limitation in face shape change and get \( M_{tar} \). The segmentation loss is formulated as:

\[
\mathcal{L}_{seg} = \| R(M_{tar}) - M_{low} \|_1 + \| M_{tar} - M_r \|_1,
\]

where \( R(,) \) means the resize operation.

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:

\[
\mathcal{L}_{cyc} = \| I_r - G(I_r, I_s) \|_1,
\]

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:

\[
\mathcal{L}_{real} = \mathcal{L}_{adv} + \lambda_0\mathcal{L}_{seg} + \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{cyc} + \lambda_3\mathcal{L}_{lips},
\]

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

#### Overall Loss.
Our full loss is summarized as follows:

\[
\mathcal{L} = \mathcal{L}_{sid} + \mathcal{L}_{real}.
\]
Table 1: Quantitative Experiments on FaceForensics++. FPS is tested under GPU V100.

<table>
<thead>
<tr>
<th>Method</th>
<th>ID↑</th>
<th>Pose↓</th>
<th>Shape↓</th>
<th>MAC ↓</th>
<th>FPS↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaceSwap</td>
<td>54.19</td>
<td>2.51</td>
<td>0.610</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FaceShifter</td>
<td>97.38</td>
<td>2.96</td>
<td>0.511</td>
<td>121.79</td>
<td>22.34</td>
</tr>
<tr>
<td>SimSwap</td>
<td>92.83</td>
<td>1.53</td>
<td>0.540</td>
<td>55.69</td>
<td>31.17</td>
</tr>
<tr>
<td>Ours-256</td>
<td>98.48</td>
<td>2.63</td>
<td>0.437</td>
<td>102.39</td>
<td>25.29</td>
</tr>
<tr>
<td>Ours-512</td>
<td>98.48</td>
<td>2.63</td>
<td>0.437</td>
<td>102.39</td>
<td>25.29</td>
</tr>
</tbody>
</table>

Table 2: Results in terms of AUC and AP on FF++ and DFDC.

<table>
<thead>
<tr>
<th>Method</th>
<th>FF++</th>
<th>DFDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>FaceSwap</td>
<td>AUC↓</td>
<td>AP↓</td>
</tr>
<tr>
<td>FaceShifter</td>
<td>66.83</td>
<td>65.99</td>
</tr>
<tr>
<td>SimSwap</td>
<td>41.62</td>
<td>42.92</td>
</tr>
<tr>
<td>Ours-256</td>
<td>76.44</td>
<td>72.63</td>
</tr>
<tr>
<td>Ours-512</td>
<td>38.97</td>
<td>41.54</td>
</tr>
</tbody>
</table>

Figure 5: (a) Comparison with AOT. (b) Comparison with DF.

Figure 6: Face shape error between $I_s$ and $I_t$ 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 $\beta 1 = 0; \beta 2 = 0.99$ 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 DeepForensics (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 preservation 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, FaceShifter, 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 $L^2$ 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.
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 $L_{shape}$ on face shape, we train another model Ours-n3d, which replaces the shape-aware identity vector with the normal identity vector from $F_{id}$. 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


