Reinforced Disentanglement for Face Swapping without Skip Connection

Xiaohang Ren, Xingyu Chen, Pengfei Yao, Heung-Yeung Shum, Baoyuan Wang

Xiaobing.AI

Figure 1. Face swap results by our method. Note that, the swapped images consistently preserve the identity of the source and the non-identity attributes (i.e., hair, expression, gaze, pose, etc.) from a variety of target images. Best viewed with zoom-in and a colored version.

Abstract

The SOTA face swap models still suffer the problem of either target identity (i.e., shape) being leaked or the target non-identity attributes (i.e., background, hair) failing to be fully preserved in the final results. We show that this insufficient disentanglement is caused by two flawed designs that were commonly adopted in prior models: (1) counting on only one compressed encoder to represent both the semantic-level non-identity facial attributes (i.e., pose) and the pixel-level non-facial region details, which is contradictory to satisfy at the same time; (2) highly relying on long skip-connections [50] between the encoder and the final generator, leaking a certain amount of target face identity into the result. To fix them, we introduce a new face swap framework called “WSC-swap” that gets rid of skip connections and uses two target encoders to respectively capture the pixel-level non-facial region attributes and the semantic non-identity attributes in the face region. To further reinforce the disentanglement learning for the target encoder, we employ both identity removal loss via adversarial training (i.e., GAN [18]) and the non-identity preservation loss via prior 3DMM models like [11]. Extensive experiments on both FaceForensics++ and CelebA-HQ show that our results significantly outperform previous works on a rich set of metrics, including one novel metric for measuring identity consistency that was completely neglected before.

1. Introduction

Deepfake [27] is a double-edged sword, despite the fact that various negative use cases are currently dominating the conversations to steer people’s attention to their social impact [21], it’s arguably true that AI-synthesized faces, bodies, and voices have huge potential for a variety of positive applications, such as virtual instructor or health counseling as discussed in [48], or other content generations in movie/game industry (i.e., Paul Walker in “Fast and Furious 7”) and etc. Keep improving the generation quality also inspires new ways to detect [7, 15, 20, 57] deepfakes for negative use cases. Face swapping has been studied intensively in both academia and industry, where the quality improvement is remarkable over the years thanks to the advances in deep generative learning. As defined in [32], face swapping must achieve three goals simultaneously (1) fully preserve the face identity from the source image (2) fully preserve everything else except the identity (identity-irrelevant) from the target image, and (3) ensure the final result is both artifacts-free and photo-realistic.
Besides, in real-world scenarios, a source image could be swapped to multiple targets, where cross-target consistency becomes fairly important. Therefore, we additionally argue that the disentanglement of ID and non-ID properties is essential for cross-target consistent face swapping. Prior works [8, 17, 32, 35, 58] investigated various methodologies for face identity and non-identity attributes disentanglement as well as their fusions to synthesize the final swapped image. Despite the impressive progress, the above three goals are hardly achieved at the same time. One of the key challenges is that those prior works are essentially playing the seesaw-style game, where improving on the identity preservation from the source is usually at the cost of sacrificing the non-identity preservation from the target, and vice versa. Fundamentally, we argue that this is still caused by the insufficient disentanglement between face identity and non-identity representations. Specifically, on one hand, works that try to improve the source identity preservation through 3D prior [58] or information compression [17] or smoothing regularization [32] on top of a pre-trained face recognition model [35] usually don’t provide sufficient forces to remove the target identity while preserving other non-identity attributes. On the other hand, work like Faceshifter [35] that tries to preserve the low-level non-identity details through attribute loss would also likely leak target identity information into the results. Such phenomena are evident from Fig. 3, where the target face silhouette is clearly leaked to the final swapped faces of works [8, 35].

By delving deep into the model structures, we noticed that most of the prior works [8, 17, 32, 35, 58] leverage a bottleneck encoder and decoder structure on the target image, where the encoder is expected to (1) fully remove the face identity (2) fully preserve everything else (i.e., pixel-level background appearance, hair, facial expression, head pose, eye gaze, and other non-identity facial attributes) from the target image. While the decoder (also called the generator) is in charge of generating the final swapped result by fusing the source identity representation from a face recognition model as well as all the compressed representations from the target decoder. To restore more details, symmetrical long skip connections are often employed by copying fine-grained low-level features from the encoder [8, 17, 35]. However, we argue that it is very difficult, if not impossible, to simultaneously achieve the above two goals using only one single compressed bottleneck encoder (see $Z_{\text{assist}}$ in Fig. 2 (a)). In addition, the skip connections used in the decoder would inevitably bring the target identity information into the results together with other non-identity attributes, therefore, further hurting the disentanglement learning.

To encourage thorough disentanglement, in this paper, we designed a new framework that gets rid of the skip connections in our entire model structure. Specifically, we proposed two separate encoders that respectively capture the pixel-level attributes outside the face region and the semantic-level non-identity facial attributes inside the face region. Each of which is tailored to dedicated capturing its own desired representation without compromising with each other. Meanwhile, a target identity removal loss and a few non-identity attribute preservation losses are explicitly employed to compensate for the missing details due to the lack of skip connections. To be specific, we leverage adversarial training techniques to penalize the target representation if it contains any target identity, meanwhile, we use prior models like 3DMM predictor [14] to explicitly preserve facial expression and head pose, in addition to the attribute loss proposed in Faceshifter [35].

To summarize, our contributions are (1) we analyzed that prior works are still suffering poor disentanglement where improving one goal may hurt other goals; (2) we unveiled that the skip connection is one root cause for such insufficient disentanglement in prior model architecture; (3) we proposed new network structures, as well as new training strategies to reinforce the identity disentanglement while preserving more identity-irrelevant attributes to compensate the lack of skip connections; lastly (4) we introduced a new evaluation metric and conducted extensive experiments, the results validated the effectiveness of our method.

2. Related works

**Face Swap.** Early face swap works [3, 10, 37, 44] mainly leverage the 3D facial model to transfer facial identity from the source into target images/videos. Recent works heavily rely on Generative Adversarial Networks (GAN) to improve the face swap visual quality. Korshunova et al. [33] design a person-specific network for face swapping, in which each source identity model is trained separately. Deepfakes [1] shares a similar idea of training a face-swapping GAN with paired source video and the target video. Naruniec et al. [41] introduce the first approach that can swap facial identity at the megapixel resolution, again with the help of GAN. RSGAN [42] makes an early attempt of extracting facial identity and non-identity attributes separately for face swapping by the face and hair latent extractors. FSGAN [43] improves its previous work [44] by replacing the 3DMM fitting and re-rendering with a GAN-based face reenactment network. Although rapid progress has been made, there is still much room to improve in order to get realistic and high-fidelity results. StyleGAN [29] based works [59, 60, 62, 64] can generate high-fidelity swapped images, but they suffer from the limited representation space of StyleGAN which results in unsatisfied identity and non-identity preserving. FaceInpainter [34] explores face swap in heterogeneous domains based on StyleGAN [29] and 3DMM [4]. On the other hand, FaceShifter [35] introduces an adaptive framework that can integrate more faithful non-identity attributes from the target with the identity embedding of the source...
from a pre-trained face recognition model. This framework shows strong generalization and therefore many subsequent works [8, 17, 32, 58] follow a similar design. Simswap [8] designs Weak Feature Matching (weak FM) loss to preserve more detailed target attributes. However, it hurts the identity similarity between the source and the final swapped face. MobileFaceSwap [61] proposes a lightweight face swap model based on both Faceshifter [35] and Simswap [8]. Hi-fiFace [58] improves identity similarity by adding 3DMM shape parameters into the facial identity feature, but the target facial identity still leaks into the resulting image. InfoS- swap [17] tries to remove the target facial identity by masking out the identity-relevant features in all the convolution stages of the pre-trained face recognition model, but other target attributes such as background and expression can’t be fully preserved, because, by definition, any face recognition model is trained to be invariant to all the non-identity attributes (i.e., pose, expression). In Smooth-swap [32], the face recognition model is further trained for face swapping by contrastive learning to smooth the identity representation space. Although the source identity preservation is improved, the preservation of non-identity attributes from the target also gets hurt. Xu et al. [6] propose a unified framework for face reenactment and face swapping, which shows that pixel-level and semantic attribute disentanglement can both help remove target facial identity and preserve non-identity attributes. In contrast, we delve into the U-net of encoder-decoder structure which is commonly used in previous works [17,35], and unveil that skip-connection is one root cause for entanglement of the target identity and its non-identity attributes. In this regard, we propose novel regularization to encourage further disentanglement learning after removing the skip connections.

**Skip Connection.** Skip connection is widely adopted in modern neural-net architectures, such as Highway [53], ResNet [22], DenseNet [24] and U-net [50] etc. It was initially designed to address the vanishing gradient issue when training very deep neural nets for classification tasks (i.e., image recognition [22]), but later extended to more broad applications including segmentation [19], optical flow prediction [46] and image synthesis [25, 35, 40]. Compared with short skip connections used in ResNet [22] (via summation) and DenseNet [24] (via concatenation), the symmetrical long skip connections used in U-net [50] directly copy fine-grained low-level details from the encoder to the decoder, results in more accurate and sharper dense predictions. However, we discover that even though skip connection benefits faster convergence, stable training, and fine-grained details preservation, it is one root cause for identity leaking and face attributes entanglement in the state-of-the-art face swap models [35].

**Face Disentanglement.** One mainstream face disentanglement idea is to leverage a geometry prior model, such as 3DMM [4], to disentangle identity, facial expression, head pose as well as lighting. Given a 2D face image, the goal is to estimate the 3DMM parameters either through direct optimization or deep learning models [11, 13, 14, 16, 63], then based upon the estimated parametric model, subsequent facial attributes editing may become easier. However, those prior models [36, 49] usually generalize poorly for wild images due to their limited capacity. Moreover, the fitting process also contains inevitable errors and ambiguities (i.e., between face identity and expression), limiting many applications that require precisely controllable editing. Therefore, a new line of work starts to learn disentangled facial attributes directly from 2D image-sets [9, 23, 31]. These generative approaches typically build latent representations for each specific facial attribute (i.e., pose, glass, hairstyle, elevation, etc.). Recent works [2, 5, 26, 45] focus more on identity and non-identity disentanglement. Specifically, Nitzen et al. [45] propose a latent space mapping network to map both the identity and attribute representations into StyleGAN latent code. Likewise, because of the limited information contained in StyleGAN latent code, this approach cannot preserve all the non-identity attributes, especially in hair and background. To better preserve the non-identity attributes, FICGAN [26] uses a much larger latent code to better restore expression and pose. However, the facial identity is not fully disentangled, meanwhile, the hair and background details are not totally preserved either.

### 3. Method

Our method aims to swap the facial identity of a source image $X^s$ to a target image $X^t$. The face-swapped result $Y^{s,t}$ is a seamlessly blended version of the facial identity of $X^s$ and all other identity-irrelevant (also called "non-
identity") properties (e.g., background, hair, pose, expression, and illumination) of $X$. As discussed in Sec. 1, previous works commonly employ skip connections (see Fig. 2(a)), where shallow convolution embedding is fed to the swap image decoder. As shown in Fig. 2(b), instead of using skip connections, we propose a novel framework consisting of a Facial Non-ID (FNID) network and a Non-Facial attributes (NFA) network to perform face swapping. The skip-connection-free framework exhaustively preserves target non-identity information, while at the same time preventing the target facial identity from leaking into the swap image decoder. In the following, we first analyze the ID vs. Non-ID disentanglement of prior works through a performance calibration based on consolidated metrics. Then we dissect the effect of skip connections to motivate our new designs (Sec. 3.1). Finally, we introduce our proposed network structure and training strategies (Sec. 3.2) in detail.

**ID vs. Non-ID disentanglement.** Although face swap has been vigorously studied for years, the evaluation methods for non-ID performance are distinct in previous reports [8, 17, 35, 58]. To consolidate the evaluations and perform fair comparisons, we employ off-the-shelf pose and expression estimations from the state-of-the-art 3D face reconstruction model [39] to serve as the non-ID metrics. And run their publicly released face swap results/models to observe all results by treating Faceshifter [35] as a baseline. As shown in Tab. 1, after calibration, we discover that existing works still suffer from insufficient disentanglement between ID-relevant and ID-irrelevant representations, causing the improvement of a single metric always comes with the performance drop of others.

3.1. Dissection of Skip Connections

We noticed that in Faceshifter [35] and most follow-up works [17, 32, 58], skip connections play an important role in preserving non-identity attributes. Hence, we delve deep into this design. After diagnostic evaluations, we realized that because shallow convolutional features contain both ID and non-ID information, the skip connection that introduces shallow features to image decoding could be the “reason” behind such entanglement. By directly feeding the shallow convolutional features into the swap image decoder through skip connections, the detailed non-ID information (e.g., background, hair, and expression) can be easily preserved. However, as shown in Fig. 3, we observe that the swapped facial identities of Faceshifter are noticeably affected by the target face shape, resulting in very low identity similarity to the source image. Alternatively, If we remove the skip connections in Faceshifter, surprisingly, the swapped facial identities are much closer to the source image, indicating less identity information from the target being leaked into the result. Nevertheless, it’s also clear from Fig. 3 that such improvement comes with the cost of worse non-ID preservation (e.g., hair). That is, image details out of the face region are inconsistent with target images. This phenomenon is also statistically confirmed in Tab. 1. As shown, removing the skip connection brings a better identity score but worse non-identity performance. Therefore, we conclude that the skip connection is at least one cause that leads to the entangled representation.

3.2. Network Designs and Training Strategies

As we argued in Sec. 1, $Z^{\text{nid}}$ alone (see Fig. 2(a)) is not sufficient to preserve target non-ID properties without skip connections. In our method, we define two types of non-ID properties based on the relation with face ID, i.e., face-relevant or face-irrelevant properties. Then, we design two encoders to extract the two-fold non-ID features, respectively. One is the FNID encoder that derives $Z^{\text{fnid}}$ and the other is NFA encoder for generating $Z^{\text{nfa}}$.

**Facial Non-ID (FNID) Encoder.** Facial Non-ID attributes (e.g., expression, pose) are spatially coupled with the facial identity, thus we need deep convolution layers to remove the identity and preserve Non-ID attributes. Following this motivation, our FNID encoder only outputs the

<table>
<thead>
<tr>
<th>Methods</th>
<th>Use skip connections</th>
<th>ID-ret</th>
<th>pose</th>
<th>expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faceshifter [35]</td>
<td>✓</td>
<td>/</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>Hitface [58]</td>
<td>✓</td>
<td>5%</td>
<td>-7%</td>
<td>+0.3%</td>
</tr>
<tr>
<td>InfoSwap [17]</td>
<td>✓</td>
<td>+10%</td>
<td>-77%</td>
<td>-11%</td>
</tr>
<tr>
<td>Simswap [8]</td>
<td>✓</td>
<td>-19%</td>
<td>+9%</td>
<td>+23%</td>
</tr>
<tr>
<td>Faceshifter*</td>
<td>×</td>
<td>+8%</td>
<td>-24%</td>
<td>-48%</td>
</tr>
</tbody>
</table>

Table 1. Calibrated results among different prior works under the same evaluation method. * represents a version of Faceshifter after removing skip connections.
deepest representation $Z^{fnid}$ without shallow features. We design training strategies tailored for FNID encoder. As shown in Fig. 4(a), to encourage the FNID encoder to extract facial non-ID features and remove facial identities, we design two additional heads to regularize $Z^{fnid}$, namely regularization head (RegHead) and adversarial head (AdvHead). RegHead maps $Z^{fnid}$ into a vector that represents pose and expression information. In addition, a pre-trained 3DMM parameter predictor [14] is also used to predict pose and expression vectors $v^{pose}$, $v^{exp}$. Then, we use $L_2$ loss to regularize the mapped vector with the predicted pose and expression vectors:

$$L_{reg}^{fnid} = ||\text{RegHead}(Z^{fnid}) - \text{concat}(v^{pose}, v^{exp})||_2^2.$$  

AdvHead is to map $Z^{fnid}$ into facial identity code, which is used to remove facial identities from $Z^{fnid}$. In detail, AdvHead attempts to extract ID feature $Z^{id}_X$, of $X$ from $Z^{fnid}$ and we use an adversary loss to prevent $Z^{fnid}$ from containing the ID information:

$$L_{adv}^{fnid} = \text{CSim}(\text{AdvHead}(Z^{fnid}), Z^{id}_X),$$

$$L_{ah}^{fnid} = 1 - \text{CSim}(\text{AdvHead}(Z^{fnid}), Z^{id}_X),$$

where $L_{ah}^{fnid}$ is only applied on the adversary head while $L_{adv}^{fnid}$ is applied on the FNID encoder. CSim denotes cosine similarity. The overall loss to train FNID can be expressed as:

$$L^{fnid} = L_{t}^{fnid} + \beta_{adv}^{fnid} L_{adv}^{fnid}.$$  

**Non-Facial Attribute (NFA) Encoder.** Non-facial attributes are spatially separated from the facial regions, so they are weakly associated with the identity. Thus, a straightforward way to extract non-facial features without identity information is to mask out the facial regions of the target image $X$ and use a masked-image encoder for feature extraction. However, this method would require a face mask predictor during the inference stage. To reduce inference complexity, instead of masking out face region from $X$, we encourage our NFA encoder to process full $X$ and extract similar representations to the masked-image encoder. In terms of network structure, NFA encoder has the same depth as FNID encoder but only the first few convolution layers have a down-sampling operation. Thus, the produced $Z^{nfa}$ is a feature map with large spatial resolution. NFA encoder is trained in parallel with the masked-image encoder in the training stage (see suppl. material for details). We use a regularization loss on the $Z^{nfa}$ and the $Z^{nfa}$ from masked-image encoder:

$$L_{reg}^{nfa} = ||Z^{nfa} - \hat{Z}^{nfa}||_2.$$  

**Swap Image Decoder.** The swap image decoder takes $Z^{fnid}$, $Z^{nfa}$, $Z^{id}$ as the input, and generate face-swapped image $Y^{s,t}$. Three components are included, i.e., FNID decoder, NFA decoder, and fusion network (shown in Fig. 4(b)). FNID decoder decodes $Z^{fnid}$ into $N^{f}$ feature maps $\{F^{fnid}_i \}_{i=1}^{N^{f}}$ for fusion network. NFA decoder has similar structure that decodes $Z^{nfa}$ into $N^{a}$ feature maps $\{F^{nfa}_i \}_{i=1}^{N^{a}}$. Because Non-Facial attribute features are usually represented by shallow features, we set $N^{a} < N^{f}$. The fusion network integrates $\{F^{fnid}_i\}$, $\{F^{nfa}_i\}$ and $Z^{id}$ for generating $Y^{s,t}$. AAD layers [35] is employed as the fusion module. The number of AAD layers is $N^{s}$ to fully preserve non-ID attributes. In the first $N^{f} - N^{a}$ AAD layers, $F^{k}^{fnid}(N^{f} > k > N^{a})$ and $Z^{id}$ are integrated. While in the last $N^{a}$ AAD layers, we concatenate $F^{k}^{fnid}$ and $F^{k}^{nfa}(k \leq N^{a})$ and then integrate them with $Z^{id}$. Following [35], we use the three global losses: ID loss, reconstruction loss, and attribute loss to train the whole network:

$$L^{id} = \text{CSim}(Z^{id}_X, Z^{id}_Y),$$

$$L^{rec} = ||X - Y||_2^2,$$

$$L^{attr} = \sum_{i=1}^{N^{f}} ||F^{fnid}_i - F^{fnid}_{Y,i}||_2^2 + \sum_{j=1}^{N^{a}} ||F^{nfa}_j - F^{nfa}_{Y,j}||_2^2,$$

$$L^{gb} = L^{id} + \beta_{rec} L^{rec} + \beta_{attr} L^{attr}.$$  

**ID Encoder.** Like Faceshifter [35], a pretrained face recognition model [12] is used as the ID encoder in our method. The normalized feature vector $z^{id}$ extracted by the ID encoder is then fed to swap image decoder.

**Total Loss.** The overall training loss function is defined as:

$$L = L^{adv} + \beta_{gb} L^{gb} + \beta_{fnid} L^{fnid} + \beta_{nfa} L^{nfa}.$$  

4. Results

4.1. Datasets and Metrics

**Training Data.** We use three face image datasets to train our model, i.e., CelebA-HQ [28], FFHQ [29] and VGGFace
The best and second best results are marked in bold. The bold and second best results are marked in Table 2. Quantitative Comparison on FaceForensics++. The best and second best results are marked in bold and underline. As VGGFace contains many small and blur images, we only use 127.7K images from VGGFace with high facial resolution for training. Additionally, we select the first 100 images in CelebA-HQ for evaluation, so they are not involved in training. Following [35], all training images are cropped and aligned as 256×256 resolution to cover the entire face and a part of background regions. See suppl. material for more training details.

**Evaluation Data and Metrics.** FaceForensics++ (FF++) [51] is a standard face swap test dataset. On FF++, the overall performance of face swap methods is evaluated on three metrics: ID retrieval, pose error, and expression error. We follow the ID retrieval setting of most recent methods [1, 6, 8, 17, 35, 43, 58, 60, 62, 64], which uses a different face recognition model [56] as the ID vector extractor. However, the above-mentioned recent methods use different pose and expression predictors to evaluate pose and expression errors. For a fair comparison, we use the state-of-the-art 3D face reconstruction model [39] to predict pose and expression parameters and then employ Euclidean distance to define pose and expression errors. Also, we use pose [52] and expression [54] metrics reported by [62] to compare with recent methods [60, 62].

In addition, we develop the CelebA-HQ test split with

\[
\text{evaluation w/ 3D face reconstruction model [39]}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>ID retrieval↑</th>
<th>pose↓</th>
<th>expression↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deepfakes [1]</td>
<td>81.96</td>
<td>0.375</td>
<td>7.270</td>
</tr>
<tr>
<td>FSGAN [43]</td>
<td>76.57</td>
<td>0.169</td>
<td>5.402</td>
</tr>
<tr>
<td>Faceshifter [35]</td>
<td>97.38</td>
<td>0.195</td>
<td>5.467</td>
</tr>
<tr>
<td>Simswap [8]</td>
<td>92.83</td>
<td>0.167</td>
<td>5.943</td>
</tr>
<tr>
<td>Hiiface [58]</td>
<td>98.48</td>
<td>0.216</td>
<td>5.474</td>
</tr>
<tr>
<td>InfoSwap [17]</td>
<td>99.67</td>
<td>0.441</td>
<td>5.733</td>
</tr>
<tr>
<td>MegaFS [64]</td>
<td>90.83</td>
<td>0.465</td>
<td>5.615</td>
</tr>
<tr>
<td>AFS [55]</td>
<td>88.81</td>
<td>0.171</td>
<td>5.493</td>
</tr>
<tr>
<td>UniFace [6]</td>
<td>99.45</td>
<td>0.174</td>
<td>5.352</td>
</tr>
<tr>
<td>Ours</td>
<td>99.88</td>
<td>0.146</td>
<td>5.290</td>
</tr>
</tbody>
</table>

Table 2. Quantitative Comparison on FaceForensics++. The best and second best results are marked in bold and underline.

\[
\text{evaluation w/ the metrics of [62]}
\]

<table>
<thead>
<tr>
<th>Method</th>
<th>ID-CSim ↑</th>
<th>ID-Consis ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSLSD [60]</td>
<td>0.205</td>
<td>0.196</td>
</tr>
<tr>
<td>StyleSwap [62]</td>
<td>0.556</td>
<td>0.588</td>
</tr>
<tr>
<td>InfoSwap [17]</td>
<td>0.611</td>
<td>0.599</td>
</tr>
<tr>
<td>MegaFS [64]</td>
<td>0.466</td>
<td>0.543</td>
</tr>
<tr>
<td>AFS [55]</td>
<td>0.445</td>
<td>0.547</td>
</tr>
<tr>
<td>UniFace [6]</td>
<td>0.663</td>
<td>0.669</td>
</tr>
<tr>
<td>Ours</td>
<td>0.743</td>
<td>0.761</td>
</tr>
</tbody>
</table>

Table 3. Quantitative Comparison on CelebA-HQ test split. The best and second best results are marked in bold and underline.

We use their original resolution in comparison because resolution has tiny effect on our ID and non-ID metrics. The works [60, 62] that release neither FF++ swapped images/videos nor pre-trained models are only compared with reported results. As shown in Tab. 2, our method achieves superior performances in terms of ID retrieval, pose error, and expression error. When compared with InfoSwap [17] which has strong ID swap ability, our results are still significantly better on all three metrics.

Although Simswap [8] has the lowest expression error,
Figure 6. Comparison of the face-swapping results of various models. The top left corner is the source image, while the other images in the first row are target images. As shown, our results are noticeably better in terms of ID consistency across various target images.

Table 4. Quantitative Comparison of image quality.

<table>
<thead>
<tr>
<th>Method</th>
<th>ID (%)</th>
<th>Non-ID (%)</th>
<th>Fidelity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimSwap [8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFS [55]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>InfoSwap [17]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniFace [6]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>52.2</td>
<td>32.4</td>
<td>27.4</td>
</tr>
</tbody>
</table>

Table 5. User Study. Percentage of each method being selected.

<table>
<thead>
<tr>
<th>Method</th>
<th>ID (%)</th>
<th>Non-ID (%)</th>
<th>Fidelity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Styleshifter [35]</td>
<td>17.6</td>
<td>22.0</td>
<td>25.4</td>
</tr>
<tr>
<td>SimSwap [8]</td>
<td>8.8</td>
<td>30.8</td>
<td>26.0</td>
</tr>
<tr>
<td>Ours</td>
<td>52.2</td>
<td>32.4</td>
<td>27.4</td>
</tr>
</tbody>
</table>

Compare with StyleGAN-based methods. In terms of image quality, though it is not our primary focus, the FID of our method is on par with the prior GAN-based methods, as shown in Tab. 4. Note that StyleGAN-based methods can produce significantly higher image quality than GAN-based methods due to the advantage of pre-trained StyleGAN [30] model. However, StyleGAN-based methods are poor in ID and non-ID preserving, which are more essential metrics for face swapping. Particularly, the SOTA StyleGAN-based method [62] reports much worse ID retrieval (2.01 drop) and expression error (0.26 increase) than ours (Tab. 2). This gap is mainly brought by the fact that the latent is not designed for ID and non-ID disentanglement and it is very difficult to preserve both ID and non-ID by latent swapping.

Qualitative Evaluations. In Fig. 1 and Fig. 6, we show examples of our results and the visual comparison with prior methods. Benefiting from the disentangled representation of our framework, our results have a noticeable advantage in terms of ID consistency across different target images when swapped with the same source.
Table 6. Ablation results. The best results are marked in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Skip connections</th>
<th>Facial Non-ID encoder</th>
<th>Regularization of $Z^{\text{fnid}}$</th>
<th>Non-facial attribute encoder</th>
<th>FaceForensics++</th>
<th>CelebA-HQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ID-retrieval↑ pose↓ expression↓</td>
<td>PSNR↑</td>
</tr>
<tr>
<td>Setting1</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>97.59</td>
<td>0.161</td>
</tr>
<tr>
<td>Setting2</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td></td>
<td>99.41</td>
<td>0.175</td>
</tr>
<tr>
<td>Setting3</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>99.83</td>
<td><strong>0.143</strong></td>
</tr>
<tr>
<td>Ours(full)</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>99.88</strong></td>
<td>0.146</td>
</tr>
</tbody>
</table>

Figure 7. Visualization results of ablation models

4.3. Ablation Studies

We perform ablation studies on FF++ and CelebA-HQ test split with all their evaluation metrics. To this end, three additional models are designed with different network components and training strategies, whose configurations and quantitative results are shown in Tab. 6. In Ablation, Setting1 is treated as the baseline model. Besides, a few randomly selected rich expression results are shown in Fig 7.

Effects of Skip Connections. With skip connections, target facial ID could leak into the decoder through shallow features, leading to poor source-ID similarity. As shown in Tab. 6, compared to Setting1, skip-connection free of Setting2 achieves significant improvements in all source-ID-related metrics. In contrast, skip connections also have a noticeable effect on non-ID attribute preservation from target images, especially for background details. As expected, Setting2 is worse than Setting1 in terms of PSNR, pose, and expression errors. Therefore, skip connections could lead to a seesaw-like trade-off by leading to the aforementioned compromises, which is not conducive to overall face-swapping performance.

Effects of $Z^{\text{fnid}}$ Regularization. $Z^{\text{fnid}}$ regularization strategy can simultaneously enhance the non-ID preservation and ID removal from target images. From Setting2 to Setting3, as shown in Tab. 6, the pose and expression errors are both reduced thanks to this regularization, which is also confirmed by visual examples in Fig 7 (i.e., eye closure & gaze is preserved in Setting3). Meanwhile, source-ID-related metrics are also elevated because target-ID are successfully removed by our adversarial head. Nevertheless, because of the limited capacity of $Z^{\text{fnid}}$, the quality of facial non-ID and non-facial attributes can hardly be improved simultaneously, resulting in a poor PSNR value.

Effects of NFA Modules. The NFA module can model non-facial regions so as to improve the performances on both non-facial and non-ID information preservation. Referring to our full method in Tab. 6, PSNR is significantly elevated by NFA module and the value is on par with that of Setting1. In the meantime, The expression error is also noticeably reduced thanks to the NFA module. The reason behind this is that the non-facial attributes are represented by $Z^{nfa}$, so $Z^{\text{fnid}}$ can now dedicate modeling facial non-ID features. As the expression is a highly abstract non-ID property, a high-capacity $Z^{\text{fnid}}$ is beneficial in modeling the extensive-expression variation. In contrast, the pose error remains nearly unchanged with NFA modules because the pose is a weak semantic property that is not sensitive to the capacity of $Z^{\text{fnid}}$. As shown in Fig 7, a significant improvement in expression and background detail preservation is introduced by our full method.

5. Conclusions

In this work, we unveil that the skip connection that was widely used in prior works is one root cause for poor disentanglement between ID and non-ID representation. We proposed a new framework to address this issue from both network structure and regularization loss perspectives. The experimental results confirm both our hypothesis and the effectiveness of our method.

Limitation and future works. StyleGAN-based face swap methods are superior in terms of rendering quality (i.e., FID), although still suffer the disentanglement challenges. Combining our framework with StyleGAN would be interesting future work.
References


[34] Jia Li, Zhaoyang Li, Jie Cao, Xingguang Song, and Ran He. Faceinpainter: High fidelity face adaptation to heterogeneous domains. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5089–5098, 2021. 2


[40] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. Pagan: Real-time avatars using dynamic textures. ACM Trans. Graph., 37(6), dec 2018. 3


[54] Raviteja Vemulapalli and Aseem Agarwala. A compact embedding for facial expression similarity. In Proceedings of


