StyleFace: Towards Identity-Disentangled Face Generation on Megapixels

Yuchen Luo[1], Junwei Zhu[2], Keke He[2], Wenqing Chu[2], Ying Tai[2]
Chengjie Wang[2], and Junchi Yan[1*]

[1] Department of CSE and MoE Key Lab of Artificial Intelligence
Shanghai Jiao Tong University
[2] Youtu Lab, Tencent
{592mcavoy,yanjunchi}@sjtu.edu.cn wqchu16@gmail.com
{junweizhu,katehe,yingtai,jasoncjwang}@tencent.com

Abstract. Identity swapping and de-identification are two essential applications of identity-disentangled face image generation. Although sharing a similar problem definition, the two tasks have been long studied separately, and identity-disentangled face generation on megapixels is still under exploration. In this work, we propose StyleFace, a unified framework for 1024^2 resolution high-fidelity identity swapping and de-identification. To encode real identity while supporting virtual identity generation, we represent identity as a latent variable and further utilize contrastive learning for latent space regularization. Besides, we utilize StyleGAN2 to improve the generation quality on megapixels and devise an Adaptive Attribute Extractor, which adaptively preserves the identity-irrelevant attributes in a simple yet effective way. Extensive experiments demonstrate the state-of-the-art performance of StyleFace in high-resolution identity swapping and de-identification.

1 Introduction

A face image can be semantically separated into two parts, including the *identity* that contains the identifiable characteristics, and the identity-irrelevant *attributes*, such as pose, expression, background, *etc.*. Although many works are devoted to disentangling and editing the facial attributes, identity-disentangled face generation is still not well investigated.

Identity-disentangled face generation constrains the generation randomness on the identity property with conditioned attributes. It has two important applications, including *Identity Swapping* and *De-identification* (De-ID). Identity swapping changes the identity in the original image to that of a specific person, while De-ID changes it to a nonexistent one. Most identity swapping methods [22, 5, 38, 42, 9] learn by maximizing the identity similarity to a specific identity and maintaining the attributes in the original image, which formulates an effective semi-supervised learning scheme. On the contrary, due to the ambiguity

* Correspondence author.

Fig. 1. This paper presents ***StyleFace***, the first unified framework for high-fidelity identity swapping (col 3) and de-identification (col 4) on megapixels. Both the embedded and sampled identities are visually realistic, and the attributes (*e.g.*, lighting, occlusion, expression, *etc.*) are faithfully preserved.

of anonymization, De-ID methods modify the original identity by feature-level repulsion [8], representation manipulation [3], or indirect identity guidance [24]. These methods often suffer from poor anonymization diversity [8] and lack supervision on attribute retention, leading to unrealistic visual effect [8, 24, 3].

Identity swapping and De-ID have been treated as two independent tasks for a long time. Nevertheless, both tasks require the generated faces to faithfully preserve the identity-irrelevant attributes in the original image, differing only on the generated identity. Therefore, we wonder if it is possible to *unify the two tasks in one framework* and promote the De-ID performance with the supervision signals in the identity swapping scheme.

In addition, *high-fidelity identity-disentangled face generation on megapixels* is still an unresolved problem, albeit the rapid improvement of generative techniques [17, 18]. Existing De-ID methods [8, 27, 10, 39, 24] are mostly cursed with limited resolution and poor fidelity. Early identity swapping methods utilize feature matching [5] or self-supervised refinement [22, 21] to improve the fidelity, but mainly focus on 256^2 resolution generation or need extra super-resolution [38, 21]. Currently, MegaFS [42] and InfoSwap [9] make early attempts on megapixel-level identity swapping, but they produce visible artifacts and have difficulty in detailed attributes recovery.

In this paper, we propose a novel framework *StyleFace*, which unifies identity swapping and de-identification in one model and renders identity-disentangled face images on mega-pixels. To bridge the gap between identity swapping and De-ID, we first design a Variational Auto-Encoder (VAE)-based projector, which encodes the identity priors from the face recognition model as a latent variable. On this basis, we can *embed the identity of a real person for identity swapping* and *sample virtual identities for de-identification*. We apply a hierarchical augmentation on the identity latent space to improve the effectiveness on different scales. Moreover, we introduce contrastive learning [11] to promote the uniformity in the intermediate latent space and improve the quality of de-identification.

With the unified framework design, we can train the model by the identity swapping objectives but directly apply it to de-identification at test time.

Thanks to the effective supervision signals on attribute retention and identity distinctiveness from the identity swapping scheme, the fidelity and realism of de-identified faces are largely promoted. Also, the diversity of de-identified faces is improved with the infinite sampling power of the latent identity space.

Next, to achieve megapixel-level generation, we utilize StyleGAN2 [18] as the generator. We formulate the projected identity as style and the identity-irrelevant attributes extracted by a carefully devised Adaptive Attribute Extractor (AAE) as noise input. Unlike [9] that erases redundant information with mutual information compression, which is time-consuming and often fails to maintain the details, AAE adaptively preserves desired attributes under simple constrain. We show that AAE is explainable and can effectively preserve even the fine-level facial details (*e.g.*, hair and wrinkles).

With the disentangled latent representation of identity, *StyleFace* unifies identity swapping and de-identification in one framework and conducts megapixel-level generation (see Fig. 1). Experiments on high-resolution identity swapping show the superiority of our model in synthesizing high-fidelity face images with precise identity control. Besides, *StyleFace* achieves state-of-the-art performance for de-identification and the de-identified faces are realistic and diverse.

In summary, our main contributions include:

- We represent identity as a latent variable and introduce contrastive learning for latent regularization. In this way, we propose *StyleFace*, to our best knowledge, the first unified high-fidelity face generation framework for both identity swapping and de-identification.
- We devise an attribute extractor to cooperate with a powerful generator (StyleGAN2) and achieve high-fidelity generation on megapixels.
- Extensive experiments show that the proposed model can generate visually appealing results with both real and virtual identities, achieving state-of-the-art performance in identity swapping and de-identification, respectively.

2 Related Works

Identity Swapping. Identity swapping is a long researched task. Early methods [22, 5] mainly focus on 256-res face swapping or need extra super resolution [38, 21], therefore cannot meet the requirement in a real-world application. Recently, InfoSwap [9] leverages the information-bottleneck principles and proposes an identity contrastive loss to promote the disentanglement. MegaFS [42] proposes a Face Transfer Module to modify identity by latent code manipulation. These methods achieve precise identity control with the carefully designed supervision objectives but are not satisfactory in maintaining the attributes.

Following this line, we train the model with the identity swapping objective but have mainly two differences. Firstly, we do not directly use the deterministic identity representations of the FR model but embed the identity priors into a latent distribution. Secondly, we preserve the attributes and details with a carefully designed feature extractor. Therefore, our model can generate new identities and have better fidelity on megapixels.

Face De-identification. Conventional de-identification methods use pixelation, blurring, or masking to conceal identifiable characteristics, which harm the original facial attributes. Current methods attach more importance to the quality and realism of the anonymized face but are still far from satisfactory. The faces generated by [8] and [27] are not natural enough and lack diversity. CIA-GAN [24] shows better diversity but cannot handle complicated facial attributes and has poor fidelity. Currently, [10, 39, 3] focus on recoverable de-identification, but still produce visible artifacts. In this work, we attempt to increase the fidelity of anonymized faces with the help of identity swapping supervisions.

Face Identity Embedding with GANs. GAN has been widely used in face image manipulation. [1, 2, 33] propose to enlarge the latent space for better editing or transferring. For manipulating facial semantics, most works focus on changing the attributes [36, 32, 14]. [23] provides inspirational findings in learning identity-distilled and identity-dispelled features, but it focuses more on attribute editing and is usually applied at regular resolution. For editing the identity, SD-GAN [7] trains with a pair of images of the same identity and disentangles identity and attributes with the specialized discriminator. Recently, DiscoFaceGAN [6] embeds the 3D prior into adversarial learning with several VAE-based encoders. These methods conduct identity-specific generation but do not precisely disentangle identity from the other image content (*e.g.*, background, haircut, *etc.*). In this work, we follow [6] to embed identity to the latent space but focus on the fine-grained control of the identifiable characteristics.

High-resolution Face Generation. The image quality of generative methods, particularly Generative Adversarial Networks (GAN), have improved rapidly. StyleGAN [17] adopts a novel intermediate latent space and a *style ingcontrol* mechanism, which allows more disentangled and scale-specific control. Besides, it facilitates stochastic variation by providing additional random noise maps and further improves the generation fidelity. Recently, StyleGAN2 [18] and StyleGAN3 [16] fix some characteristics artifacts in [17] and yield the state-of-the-art generation quality. In this work, we utilize the powerful StyleGAN2 network and devise a new attribute extractor to improve the image fidelity.

3 Approach

We train StyleFace with the identity swapping scheme: Given the source image $\mathbf{X}_s$ and the target image $\mathbf{X}_t$, we change the identity of $\mathbf{X}_t$ to that of $\mathbf{X}_s$ while preserving all the identity-irrelevant attributes, thus producing image $\mathbf{Y}_{s,t}$. Once trained, we can anonymize the identity in image $\mathbf{X}_t$ by directly sampling a virtual identity. As presented in Fig. 2 (a), we utilize the StyleGAN2 model for high-resolution generation, with the identity as style and attribute as noise input. To construct the latent identity space and improve the generation fidelity, we devise the *Identity Projector* and the *Adaptive Attribute Extractor* (AAE). Next, we will explain the proposed framework in detail.

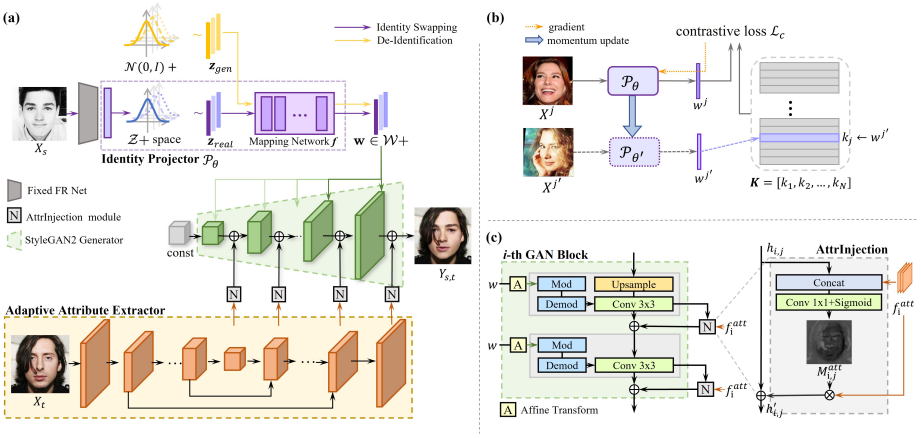

Fig. 2. The architecture of ***StyleFace***. (a) The generation pipeline for identity swapping (with embedded z_{real}) and de-identification (with sampled z_{gen}). (b) Illustration of the contrastive loss. (c) Detailed structures of the GAN Block and the *AttrInjection* module. Please refer to [18] for details of the GAN block.

3.1 Identity Projector

To unify identity swapping and de-identification in one framework, we desire to represent identity as a latent variable to embed real identity and generate virtual identities. Inspired by DiscoFaceGAN [6], we devise a Variational Auto-Encoder (VAE)-based projector to project the identity prior from a pretrained Face Recognition (FR) network to the latent space of the StyleGAN2 model.

$\mathcal{Z}+$ Space. The original latent space $\mathcal{Z}$ of StyleGAN is a standard Gaussian distribution. Inspired by [1, 2, 33] that enlarge the latent space to increase the model's expressiveness, we expand $\mathcal{Z}$ by adopting three different latent codes from $\mathcal{Z}$ in the low (4^2-16^2), middle (32^2-128^2), and high-level (256^2-1024^2) layers. This is equivalent to sampling from a $\mathcal{Z}+$ space that consists of three versions of $\mathcal{Z}$. We empirically find that $\mathcal{Z}+$ provides a hierarchical identity control and increases the distinctiveness of identity change (Sec. 4.4).

To embed identity to the $\mathcal{Z}+$ space, we regard deep features from the pre-trained FR model [13] as identity priors and devise a simple VAE-based projector. The projector is implemented as a one-layer MLP, which maps the features to the means and covariances of the $\mathcal{Z}+$ space. We regularize the latent space by the Kullback-Leibler divergence loss $\mathcal{L}_{kl}$:

$$\mathcal{L}_{kl} = \frac{1}{2}\sum_i (\mu_i^2 + \sigma_i^2 - \log \sigma_i^2 - 1), \tag{1}$$

where $\mu_i, \sigma_i \in \mathbb{R}^{1\times 512}$ and $i \in \{l, m, h\}$. For simplicity, we use l, m, h to denote the **l**ow, **m**iddle, **h**igh-level layers, respectively. We do not add an extra reconstruction task as the typical VAE does but use the identity preserving objective

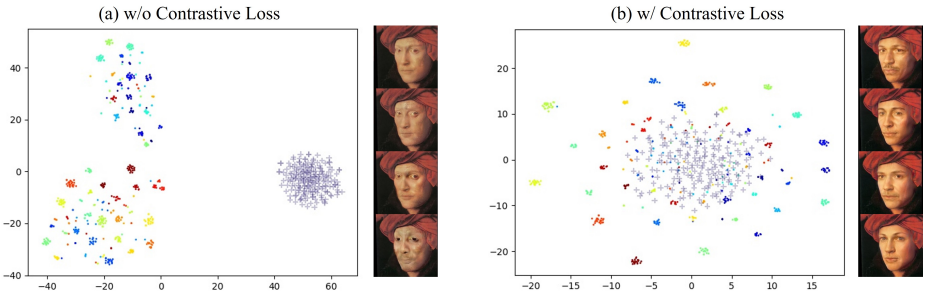

Fig. 3. Distributions of $\mathbf{w}_{real}$ and $\mathbf{w}_{gen}$ without/with the contrastive loss. '.' and '+' denote $\mathbf{w}_{real}$ and $\mathbf{w}_{gen}$, respectively. Different identities are marked by different colors. Examples of generated identity are shown on the right.

to guarantee that the valid identity information is not lost. In the following sections, $\mathbf{z} = \{\mathbf{z}_i\}, i \in \{l, m, h\}$ is used to represent the latent code in $\mathcal{Z}+$ space.

Non-uniformity in the $\mathcal{W}+$ Space. There are two latent spaces in StyleGAN: the original latent space $\mathcal{Z}$, and the less entangled intermediate latent space $\mathcal{W}$. $\mathcal{W}$ is produced from $\mathcal{Z}$ by a non-linear mapping f. With identity embedded to the $\mathcal{Z}+$ space, we subsequently map $\mathbf{z}$ to $\mathbf{w}$ and compose the intermediate latent space $\mathcal{W}+$. The $\mathbf{w}$ vectors modulate the weights of corresponding convolution layers in the generator to control the generated identity.

Denoting the latent identity code embedded from a real image as $\mathbf{z}_{real}$ and that randomly sampled from the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$ as $\mathbf{z}_{gen}$, we find that $\mathbf{z}_{real}$ can recover the corresponding identity for identity swapping, but $\mathbf{z}_{gen}$ fails to produce a feasible identity for de-identification. To figure out the reason, we use t-SNE to visualize the distribution of $\mathbf{w}_{real}$ and $\mathbf{w}_{gen}$ vectors. We randomly pick 100 identities from the training dataset and get 626 $\mathbf{w}_{real}$ vectors. Then we randomly sample 200 $\mathbf{z}_{gen}$ vectors from $\mathcal{N}(0, \mathbf{I})$ and map them to $\mathbf{w}_{gen}$.

As shown in Fig. 3 (a), the $\mathbf{w}_{real}$ codes are clustered to different centers and there is no overlap between $\mathbf{w}_{gen}$ and $\mathbf{w}_{real}$. Accordingly, the generated virtual identities are not feasible, indicating the $\mathbf{w}_{gen}$ codes are lying out of the reasonable intermediate space. We think the reason for the non-uniformity of $\mathcal{W}+$ is mainly two-fold: i) $\mathcal{W}+$ space is a complex non-Gaussian distribution that has no constrain on the uniformity. ii) The amount of different IDs in the training dataset is limited, thus $\mathbf{w}_{real}$ may not span the whole $\mathcal{W}+$ space but only a small subspace. Hence, we try to resolve this issue via *contrastive learning*.

Contrastive Constrain. Currently, [37] points out that the contrastive loss optimizes the alignment of features from positive pairs and the uniformity of the induced distribution. To generate reasonable identities from randomly sampled latent code, we desire the intermediate identity representation $\mathbf{w}$ to meet the

following requirements: i) The $\mathbf{w}$ codes for samples of the same identity gather together. ii) All the $\mathbf{w}$ codes distribute uniformly in the $\mathcal{W}+$ space. To this end, we introduce a contrastive constrain on $\mathbf{w}$.

We parameterize the process of embedding an image $\mathbf{X}$ to the intermediate identity representation $\mathbf{w}$ as an Identity Projector $\mathcal{P}_\theta$:

$$\mathbf{w} = \mathcal{P}_\theta(\mathbf{X}) = f(\phi(\omega(FR(\mathbf{X})))), \tag{2}$$

where FR is the fixed FR net, ω is the VAE-based projector, ϕ is the reparamterization process, f is the non-linear mapping from $\mathcal{Z}+$ to $\mathcal{W}+$, and θ denotes the learnable parameters in ω and f. Inspired by MoCo [11] that facilitates unsupervised contrastive learning with a large and consistent queue and a moving-averaged encoder, we build a dynamic list $\mathbf{K} = [\mathbf{k}_i]_{i=1}^{N}, \mathbf{k}_i \in \mathcal{R}^{1\times 512}$, where N is the amount of distinct identity in the training set. Note that we build a dynamic list $\mathbf{K}_i$ for each $\mathbf{w}_i \in \mathbf{w}, i \in \{l, m, h\}$ and omit the subscript i for simplicity.

Fig. 2 (b) illustrates the contrastive constrain. We create another encoder $\mathcal{P}_{\theta'}$, which has the same structure with $\mathcal{P}_\theta$. Given image $\mathbf{X}^j$ with the identity label j, we randomly pick another image $\mathbf{X}^{j'}$ of the same person to compose a positive pair. Then, we update $\mathcal{P}_{\theta'}$ by a momentum-based moving average of $\mathcal{P}_\theta$ so that $\theta' \leftarrow m\theta' + (1-m)\theta$, where $m \in [0, 1)$ is the momentum coefficient. We encode $\mathbf{X}^j$ and $\mathbf{X}^{j'}$ with the projector $\mathcal{P}_\theta$ and the moving-averaged projector $\mathcal{P}_{\theta'}$, respectively. In this way, we get $\mathbf{w}^j = \mathcal{P}_\theta(\mathbf{X}^j)$ and $\mathbf{w}^{j'} = \mathcal{P}_{\theta'}(\mathbf{X}^{j'})$.

Unlike MoCo, which updates the dynamic queue by replacing the oldest sample in an unsupervised manner, we update the j-th item in $\mathbf{K}$ by $\mathbf{K}[j] \leftarrow \mathbf{w}^{j'}$. With $\mathbf{w}^j$ as the query, we regard $\mathbf{K}[j]$ as the positive key and the other items in $\mathbf{K}$ as the negative keys. Then, we normalize all the vectors to the unit space and measure the similarity between the query $\mathbf{w}^j$ and the dynamic list $\mathbf{K}$ by the InfoNCE [25] loss. In this way, the contrastive constrain $\mathcal{L}_c$ is formulated as:

$$\mathcal{L}_c = -\log \frac{\exp(\mathbf{w}^j \cdot \mathbf{K}[j]/\tau)}{\sum_{k=1}^{N} \exp(\mathbf{w}^j \cdot \mathbf{K}[k]/\tau)}, \tag{3}$$

where τ is the temperature. $\mathcal{L}_c$ encourages the $\mathbf{w}$ codes of the same identity to be similar to each other and dissimilar to those of other identities. Note that we empirically set $m = 0.999$ and $\tau = 0.07$.

As shown in Fig. 3 (b), with the intermediate representations constrained by the contrastive loss $\mathcal{L}_c$, the $\mathbf{w}_{real}$ codes uniformly distribute throughout the whole space and overlap with the $\mathbf{w}_{gen}$ codes. In this way, we can generate realistic identities with the randomly sampled latent codes.

3.2 Adaptive Attribute Extractor

A critical issue in megapixel-level face generation is to faithfully preserve the identity-irrelevant attributes and facial details, which is crucial for face realism and image quality. In this subsection, we introduce the Adaptive Attribute Extractor (AAE), which adaptively preserves the necessary information.

Multi-level Attribute Encoding. The attributes of a face image often span a large range of spatial resolution, such as the global-level position, the middle-level expression, and the fine-level details. Early work [22] demonstrates that multi-level features better preserve the image details than compressed single vectors. Therefore, we devise a lightweight U-shape DNN to extract features in various resolutions. Inspired by [40], we carefully design the DNN so that the feature map $\mathbf{f}_i^{att}$ in the i-th layer has the same shape as that in the i-th GAN block. Unlike [22] that injects attributes in a SPADE [26]-like design, we treat $\mathbf{f}_i^{att}$ in the i-th layer as the noise input of the corresponding i-th GAN block. In this way, we name the module as *AttrInjection*.

Adaptive Attribute Disentangle. The extracted multi-level features contain redundant information, such as the identity information of the target image. We desire to preserve just the least sufficient information of the target attributes. Therefore, we predict a control mask $\mathbf{M}_{i,j} \in [0, 1]$ for the corresponding *AttrInjection* module, where $\mathbf{M}_{i,j}$ has the same shape as $\mathbf{f}_i^{att}$:

$$\mathbf{M}_{i,j} = \sigma(Conv(\mathbf{h}_{i,j} \circ \mathbf{f}_i^{att})). \tag{4}$$

$\mathbf{h}_{i,j}$ is the output of the j-th modulated convolution in i-th GAN block, σ is the $Sigmoid(\cdot)$ function and $\circ$ is the channel-wise concatenation. We compress the extracted attribute features $\mathbf{f}_i^{att}$ with the control mask $\mathbf{M}_{i,j}$ by

$$\mathbf{h}'_{i,j} = \mathbf{h}_{i,j} + \mathbf{M}_{i,j} \times \mathbf{f}_i^{att}. \tag{5}$$

As shown in Fig. 2 (c), we incorporate the distilled attribute information into the generation process without modifying the GAN structure.

Recent work [9] supervises the information compression by mutual information, but we observe that the imperfect information compression harms both the identity and attributes. Differently, we simply constrain the control mask by minimizing the activation in $\mathbf{M}_{i,j}$:

$$\mathcal{L}_{mask} = \sum_{i,j} ||\mathbf{M}_{i,j}||_1. \tag{6}$$

In this way, the multi-level information and spatial correspondence in the target image are maintained, and the redundant information is filtered.

3.3 Loss Function

Attribute Preserving Loss. When the source image $\mathbf{X}_s$ and the target image $\mathbf{X}_t$ have the same identity, we expect the output $\mathbf{Y}_{s,t}$ to be identical with $\mathbf{X}_t$, thus define the pixel-wise reconstruction loss as,

$$\mathcal{L}_{rec} = ||\mathbf{Y}_{s,t} - \mathbf{X}_t||_1 \quad if\, ID(\mathbf{X}_t) = ID(\mathbf{X}_s). \tag{7}$$

Following [5], we define the feature matching loss by minimizing the $L2$ distance between the multi-level features from the discriminator D for $\mathbf{X}_t$ and $\mathbf{Y}_{s,t}$. To

eliminate the ghosting artifacts, we use a background mask $\mathbf{M}_{bg}$ from the segmentation model [34] in the shadow layers:

$$\mathcal{L}_{FM}^{low} = \sum_{i=1}^{m} \mathbf{M}_{bg} \cdot ||D^{(i)}(\mathbf{X}_t) - D^{(i)}(\mathbf{Y}_{s,t})||_2. \tag{8}$$

In deep layers, we match the features in the whole image:

$$\mathcal{L}_{FM}^{high} = \sum_{i=m}^{M} ||D^{(i)}(\mathbf{X}_t) - D^{(i)}(\mathbf{Y}_{s,t})||_2. \tag{9}$$

We define the total feature matching objective as the equally weighted sum:

$$\mathcal{L}_{FM} = \mathcal{L}_{FM}^{low} + \mathcal{L}_{FM}^{high}. \tag{10}$$

Identity Preserving Loss. To encourage the swapped identity to be more distinctive, we adopt the Identity Contrastive Loss (ICL) in [9]:

$$\begin{aligned} \mathcal{L}_{ICL} =& 1 - \cos < z_{id}(\mathbf{Y}_{s,t}), z_{id}(\mathbf{X}_s) > \\ &+ (\cos < z_{id}(\mathbf{Y}_{s,t}), z_{id}(\mathbf{X}_t) > - \cos < z_{id}(\mathbf{X}_s), z_{id}(\mathbf{X}_t) >)^2, \end{aligned} \tag{11}$$

where z_{id} is the 512-dim vector extracted by the FR net.

Overall Loss. We adopt the same GAN loss $\mathcal{L}_{GAN}$ for adversarial training as StyleGAN2, and the total objective is formulated as:

$$\begin{aligned} \mathcal{L}_{total} =& L_{GAN} + \mathcal{L}_c + \mathcal{L}_{mask} + \mathcal{L}_{FM} \\ &+ \lambda_{rec}\mathcal{L}_{rec} + \lambda_{ICL}\mathcal{L}_{ICL} + \lambda_{KL}\mathcal{L}_{KL}, \end{aligned} \tag{12}$$

where the contrastive loss $\mathcal{L}_c$ is defined in Eq. (3), the mask loss $\mathcal{L}_{mask}$ is defined in Eq. (6) , and the KL-divergence loss is defined in Eq. (1), We train the whole model end-to-end with $\mathcal{L}_{total}$. Once the training is finished, the model can be directly used for de-identification (see Fig. 2 (a)).

4 Experiments

4.1 Implementation Details and Protocols

We train the model on a combination of FFHQ [17], VGGFace2 [4], and CelebAHQ [15], with all images aligned and cropped to 1024×1024. We devise the AAE (Sec. 3.2) to have the same spatial resolution as the StyleGAN2 model, but only one layer in each resolution and 1/8 the channel dimension. We use a 1×1 Conv to adjust the channel dimension of the DNN feature map before sending it to the $AttrInjection$ module. The VAE-liked projector is a one-layer MLP. More details of the model's architecture are provided in the supplementary.

Fig. 4. Qualitative comparison with MegaFS [42] and InfoSwap [9] on FF++ [28]. Our model maintains the (a) eye color and (b) face shape of source identity, and better preserves target attributes, such as (c) expression and (d) skin color.

Fig. 5. Qualitative comparison about identity swapping on the CelebAMaskHQ dataset. The first row shows source-target image pairs, and the last three rows show the results of MegaFS, InfoSwap, and *StyleFace* (ours) from top to bottom.

At the start of training, we set the ratio of source-target pairs with the same identity to 100% for a warm-up and linearly decrease it to 50%. Adam [19] is used with $\beta_1 = 0$, $\beta_2 = 0.99$. We first pretrain the generator on FFHQ for 20K steps and then train the whole model end-to-end. The learning rates of AAE and the generator are $1e-4$, while that of the identity projector is $1e-6$. For Eq. (12), we set $\lambda_{rec} = 10$, $\lambda_{ICL} = 5$, and $\lambda_{KL} = 1e-4$. The 1024^2-res model is trained using 4 A100 GPUs for 2 days with a batch size of 4.

4.2 High-resolution Identity Swapping

In this section, we compare *StyleFace* with state-of-the-art high-resolution identity swapping methods, including MegaFS [42] and InfoSwap [9]. The public model and processing scripts are used in the following experiments.

Table 1. Quantitative test w.r.t.identity preserving, attribute preserving, and image quality. Values underlined are from [42] due to lack of ground-truth segmentation of FF++ [28]. Inference speed is tested on one V100 GPU over 1,000 independent runs. For MegaFS, the time for segmentation is excluded.

Model	ID↑	Shape↓	Pose↓	Exp.↓	FID ↓	Inference Speed (ms/1024^2 Image)	User Study
MegaFS [42]	<u>90.83</u>	-	<u>2.64</u>	-	16.64	91 ± 1.7	6.7%
InfoSwap [9]	**98.70**	0.57	3.11	0.31	3.39	329 ± 0.7	18.7%
Ours	96.34	**0.50**	**2.52**	**0.28**	**2.04**	**86 ± 1.8**	**74.6%**

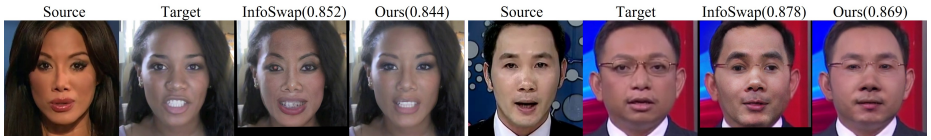

Fig. 6. Numbers in the bracket denotes the cosine similarity between the generated identity and the source identity. Our model produces more natural and perceptually similar faces than InfoSwap, though it has numerically lower scores.

Qualitative Comparison. We first compare on the FaceForensics++ (FF++) [28] dataset. As shown n Fig. 4, MegaFS produces distinct face contour and ignores the source face shape. Besides, InfoSwap produces visible skin artifacts and inconsistent eye color. Both two methods cannot preserve the target skin color. In contrast, our model maintains identity-level characteristics such as face shape and eye color and faithfully preserves the target attributes like pose and expression. In addition, the face images rendered by our model have distinctly better quality and are more visually appealing.

For megapixel-level identity swapping, we randomly compose 30K pairs of the source-target images from CelebAMaskHQ [20] and generate 1024^2 resolution results in Fig. 5. We observe that MegaFS [42] occasionally fails (col 5) due to unstable GAN-inversion. InfoSwap produces twisted hair (col 4), while our model preserves the detailed hair strands. Our model can better retain the lighting (col 1), expression (col 2), the source face shape (col 3/4), and better handle the occlusion (col 6/7). Besides, it produces fewer artifacts and maintains the image details, showing superior fidelity on megapixels.

Quantitative Comparison. Following [22, 38, 5], we conduct quantitative comparison on the FF++ [28] dataset on the following metrics: *ID retrieval*, *pose error*, *face shape error*, and *expression error*. For the ID retrieval rate, we use [35] to extract the identity embedding and report the Top-1 matching rate of the swapped image and the source image. We estimate the 3D pose by [29], and the expression and shape by [30]. We report the $L2$ distance between the regressed coefficients of swapped image and the ground truth for these three metrics. To further evaluate image fidelity and model efficiency on high-resolution

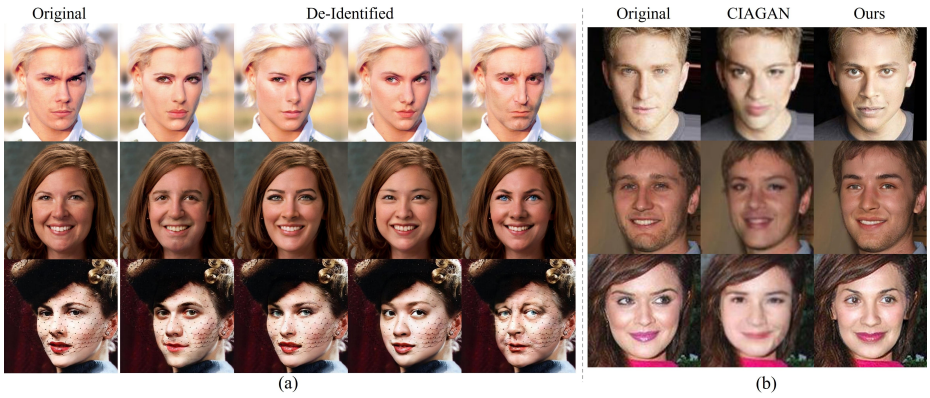

Fig. 7. (a) Examples of de-identified faces. (b) Comparison with CIAGAN [24]. Our method better preserves the original attributes (*e.g.*, lighting, expression, and occlusion), and generates identities that are more realistic and diverse.

Table 2. Comparison with recent de-identification methods on LFW [12].

Method	VGGFace2 [4]↓	CASIA [41]↓
Original	0.986±0.010	0.965±0.016
Gafni *et al.* [8]	0.038±0.015	0.035±0.011
CIAGAN [24]	0.034±0.016	0.019±0.008
Ours	**0.013±0.006**	**0.012±0.008**

generation, we compute the Fréchet Inception Distance (FID) score on the CelebMaskHQ dataset and the inference speed of generating one 1024^2-res image.

As shown in Table 1, our model has the lowest pose and expression error, indicating that the target attributes are well maintained. Besides, the FID scores imply that images generated by our model have high quality and fewer artifacts. Fig. 6 shows that our model produces more natural and perceptually similar faces, although it has a slightly lower ID retrieval rate than InfoSwap, Besides, our model has the fastest inference speed, showing good efficiency. Moreover, we conduct a user study among 20 users on 50 source-target pairs from the CelebAHQ dataset, and each user selects the best one from three methods. As reported in Table 1, our method significantly outperforms the other methods.

4.3 Face De-identification

Qualitative Comparison. In Sec. 3.1, we design an Identity Projector to construct a latent identity space. Thus we can sample infinite virtual identities from $\mathcal{Z}+$ space for face de-identification. Here we present some examples of the de-identified faces in Fig. 7 (a) and the qualitative comparison with the current state-of-the-art method CIAGAN [24] in Fig. 7 (b). It can be observed that

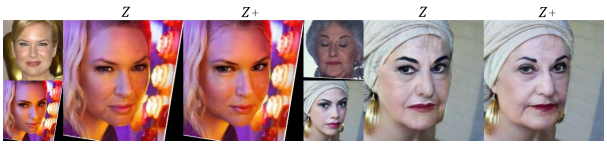

Fig. 8. Identity swapping results with $\mathcal{Z}$ and $\mathcal{Z}+$ spaces.

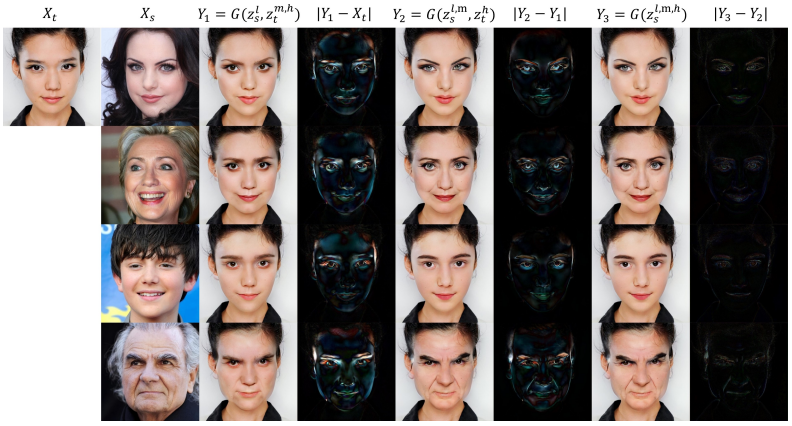

Fig. 9. Visualization of the hierarchical identity control of the $\mathcal{Z}+$ space. The brighter pixel indicates a larger difference.

faces de-identified by our model are more diverse and realistic, showing better preservation of the original attributes and better image quality.

Quantitative Comparison. Following [24], we anonymize the second image of each positive pair in the LFW [12] dataset. We utilize two FaceNet [31] models, which are pretrained on VGGFace2 [4] and CASIA-WebFace [41], respectively. The true acceptance rate is reported in Table 2 that lower value indicates better anonymization. We compare with the state-of-the-art De-ID methods, including Gafni *et al.* [8] and CIAGAN [24]. As presented in Table 2, when the face is anonymized by our model, the identification rate is lower than the other two methods, showing better de-identification ability.

4.4 Analysis

$\mathcal{Z}+$ Space. To verify the effectiveness of the $\mathcal{Z}+$ space (Sec. 3.1), we train another model with the original $\mathcal{Z}$ space. As shown in Fig. 8, the model with $\mathcal{Z}$ space has lower identity similarity and fails to recover the gaze direction. Furthermore, we analyze the hierarchical identity control of the $\mathcal{Z}+$ space. With X_t as the attribute reference, we change the identity code from $\mathbf{z}_t$ to $\mathbf{z}_s$ by gradually replacing the **l**ow, **m**iddle, and **h**igh-level component of $\mathbf{z}_t$ with those

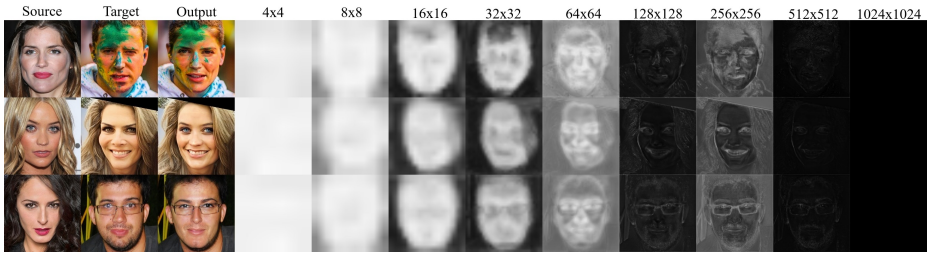

Fig. 10. Visualizing the control masks in AAE at different resolutions. The lighter pixel indicates a higher weight for attribute preserving.

of $\mathbf{z}_s$, generating $\mathbf{Y}_1$, $\mathbf{Y}_2$ and $\mathbf{Y}_3$. Then, we compute the differences to show the impact of each component in Fig. 9. The low-level code $\mathbf{z}^l$ affects the coarse-level attributes, such as the shape of face and eyebrow. The middle-level code $\mathbf{z}^m$ primarily affects the perceptual similarity, with the most facial attributes (*e.g.*, eye color, lips, *etc.*) changed. Finally, the high-level codes $\mathbf{z}^h$ further strengthen some facial details, and the identity completely changes to $\mathbf{X}_s$.

Control Mask Visualization. In Sec. 3.2, we predict a control mask $M_{i,j}$ (Eq. (4)) to select the identity-irrelevant information from feature maps. Here we visualize the mean value of control masks at each resolution in Fig. 10. The mask highlights the whole face region in the low-level layers, indicating that the model learns to recover the global pose and facial layout. As the resolution increases, it focuses more on the background and facial decorations (*e.g.*, makeup and glasses). In the highest layers, the activation becomes sparser that only some edges and details are highlighted. The visualization shows that the AAE adaptively preserves the desired attributes at different resolutions.

5 Conclusion

In this paper, we have proposed a novel framework *StyleFace*, which unifies identity swapping and de-identification in one model and achieves high-fidelity face rendering on megapixels. To bridge the gap between identity swapping and de-identification, we embed identity prior into the latent space and introduce a contrastive constrain for further regularization. We utilize the StyleGAN2 for megapixel-level generation and devise an adaptive attribute extractor to preserve the identity-irrelevant information. We show that the proposed model can generate high-fidelity results with both embedded real identities and sampled virtual identities. Extensive experiments demonstrate the state-of-the-art performance of StyleFace in identity swapping and de-identification.

Acknowledgements This work was partly supported by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102) and the National Science of Foundation China (61972250, 72061127003).

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? ICCV (October 2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? CVPR (June 2020)
3. Cao, J., Liu, B., Wen, Y., Xie, R., Song, L.: Personalized and invertible face de-identification by disentangled identity information manipulation. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3334–3342 (October 2021)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. FG 2018 pp. 67–74 (2018)
5. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. ACM MM pp. 2003–2011 (2020)
6. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. CVPR (2020)
7. Donahue, C., Lipton, Z.C., Balsubramani, A., McAuley, J.J.: Semantically decomposing the latent spaces of generative adversarial networks. ICLR (2018)
8. Gafni, O., Wolf, L., Taigman, Y.: Live face de-identification in video. ICCV pp. 9377–9386 (2019)
9. Gao, G., Huang, H., Fu, C., Li, Z., He, R.: Information bottleneck disentanglement for identity swapping. CVPR pp. 3404–3413 (June 2021)
10. Gu, X., Luo, W., Ryoo, M.S., Lee, Y.J.: Password-conditioned anonymization and deanonymization with face identity transformers. European Conference on Computer Vision (2020)
11. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
12. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
13. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Jilin Li, F.H.: Curricularface: Adaptive curriculum learning loss for deep face recognition. CVPR pp. 1–8 (2020)
14. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. Proc. NeurIPS (2020)
15. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. ICLR (2018), https://openreview.net/forum?id=Hk99zCeAb
16. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. NIPS (2021)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CVPR pp. 4401–4410 (2019)
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. CVPR (2020)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. CVPR (2020)
21. Li, J., Li, Z., Cao, J., Song, X., He, R.: Faceinpainter: High fidelity face adaptation to heterogeneous domains. CVPR pp. 5089–5098 (June 2021)

22. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. CVPR (June 2020)
23. Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X.: Exploring disentangled feature representation beyond face identification. In: CVPR (2018)
24. Maximov, M., Elezi, I., Leal-Taixé, L.: CIAGAN: conditional identity anonymization generative adversarial networks. CVPR pp. 5446–5455 (2020)
25. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
26. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. CVPR (2019)
27. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. ECCV (September 2018)
28. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. ICCV (2019)
29. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. CVPRW pp. 2074–2083 (2018)
30. Sanyal, S., Bolkart, T., Feng, H., Black, M.: Learning to regress 3d face shape and expression from an image without 3d supervision. CVPR (2019)
31. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. CVPR pp. 815–823 (2015)
32. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. TPAMI (2020)
33. Song, G., Luo, L., Liu, J., Ma, W.C., Lai, C., Zheng, C., Cham, T.J.: Agilegan: Stylizing portraits by inversion-consistent transfer learning. SIGGRAPH (jul 2021)
34. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. In: arXiv (2019)
35. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. CVPR pp. 5265–5274 (2018)
36. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. arxiv:2109.06590 (2021)
37. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. ICML pp. 9929–9939 (2020)
38. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hififace: 3d shape and semantic prior guided high fidelity face swapping. IJCAI-21 pp. 1136–1142 (8 2021)
39. Yamaç, M., Ahishali, M., Passalis, N., Raitoharju, J., Sankur, B., Gabbouj, M.: Reversible privacy preservation using multi-level encryption and compressive sensing. EUSIPCO pp. 1–5 (2019)
40. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. CVPR (2021)
41. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR **abs/1411.7923** (2014)
42. Zhu, Y., Li, Q., Wang, J., Xu, C., Sun, Z.: One shot face swapping on megapixels. CVPR pp. 4834–4844 (June 2021)