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

Appendix

1 Implementation Details

We propose a novel framework *StyleFace*, which unifies identity swapping and de-identification in one model and achieves high-fidelity face rendering on megapixels. In the main script, we introduce the generation mechanism in detail and present the comparisons with state-of-the-art methods for identity swapping and identity anonymization, respectively. Here we elaborate on the practical details of the proposed model.

1.1 Network Architecture

We integrate the disentangled generation into the StyleGAN2 [5, 6] model and control the identity and attributes in two separated modules, namely the *Identity Projector* and *the Adaptive Attribute Extractor* (AAE). Here we present the detailed architecture in Fig. 1.

Identity Projector. We resize and crop the source image $\mathbf{X}_s$ to 112×112 following the preprocessing procedure in [3]. Then, we extract the identity prior of $\mathbf{X}_s$ from the pretrained face recognition model [3] and map these features to the means and covariances of the $\mathcal{Z}+$ space. To compensate for the potential information loss in the identity embedding procedure, we use the output features of the last convolution layers as the ID prior, not the 512-dim deep vector. After that, three different latent codes are sampled from $\mathcal{Z}+$ for the low (4^2-16^2), middle (32^2-128^2), and high-level (256^2-1024^2) layers in the generator, respectively. Following the style-control mechanism in StyleGAN2, we map the latent $\mathbf{z}$ vectors to the $\mathcal{W}+$ space and control the generated identity by modulating the convolution weights in the GAN block.

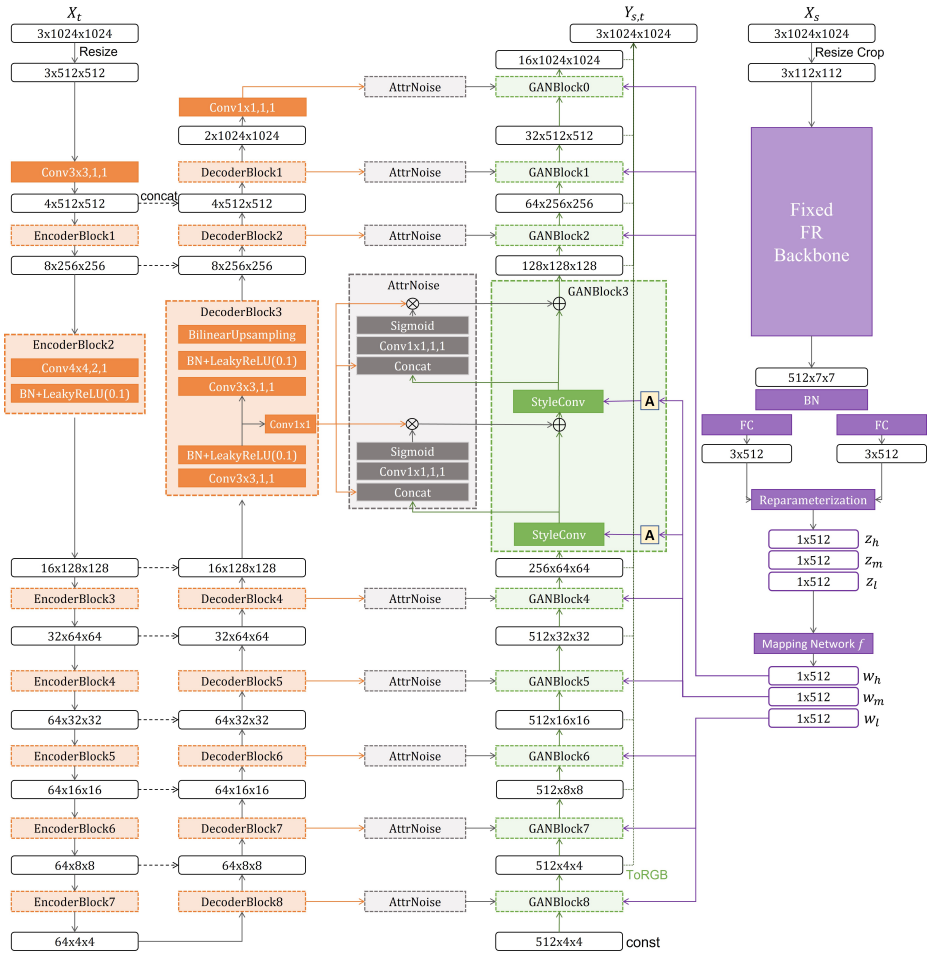


Fig. 1. Detailed architecture of the proposed ***StyleFace***.

Adaptive Attribute Extractor. We devise a U-shape DNN to extract the multi-level attribute features from the reference image $\mathbf{X}_t$. To eliminate the ghosting effect, we set both the input size of the DNN encoder and the shortcut connection to 512 resolution, and then we upsample the feature to 1024×1024. After that, we adaptively filter the redundant information and inject the attributes to the generator through the *AttrInjection* module, which is demonstrated in Sec. 3.2 in the main paper. To reduce the computation burden, we set the channel dimension in each layer to one-eighth of that in the corresponding GAN block and use a 1×1 convolution to adjust the channel dimension before sending the feature maps to the *AttrInjection* module.

Algorithm 1 Pseudocode of the contrastive constrain in Eq. (3) in a PyTorch-like style.

$\#\mathcal{P}_\theta$: the identity projector
$\#\mathcal{P}_{\theta'}$: the moving-averaged identity projector
$\#\mathbf{K}$: the dynamic list
$\#m$: the momentum
$\#\tau$: the temperature
Require: $x^j, x^{j'}$ of the j-th identity
$\mathbf{w}^j = \mathbf{P}_\theta(x^j)$ ▷ the query: 1xC
$\mathbf{w}^{j'} = \mathbf{P}_{\theta'}(x^{j'})$ ▷ the positive key: 1xC
$\mathbf{w}^{j'} = \mathbf{w}^{j'}.detach()$
$\mathbf{K}[j] = \mathbf{w}^{j'}$ ▷ update the j-th item
$logits = mm(\mathbf{w}^j.view(1, C), \mathbf{K}.view(C, N))$
▷ logits: 1xN
$\mathcal{L}_c = CrossEntropyLoss(logits/\tau, j)$
▷ contrastive loss in Eq. (3), positive is the j-th item
$\mathcal{L}_c.backward()$
$update(\theta)$ ▷ update the identity projector
$\theta' = m \cdot \theta' + (1 - m) \cdot \theta$ ▷ momentum update

1.2 Method Details

Contrastive Constrain. To promote the uniformity in the intermediate latent space $\mathcal{W}+$ and thus improve the feasibility of the generated identities, we introduce a MoCo [2]-like constrain on the $\mathbf{w}$ vectors in Sec 3.1 in the main paper. Here we provide the pseudo-code of the contrastive constrain in Algorithm 1. Given the two images $\mathbf{X}^j$ and $\mathbf{X}^{j'}$ of the same identity j, we embed them to the $\mathcal{W}+$ space, forming the positive sample pairs. We update the j-th item in the dynamic list $\mathbf{K}$, and regard the other items as the negative samples. The contrastive constrain is formed as the InfoNCE [7] loss, which is the log loss of a N-way softmax-based classifier that tries to classify $\mathbf{w}^j$ as $\mathbf{K}[j]$, where N is the number of distinct identities in the training set. In training, we get a total of N=12,119 distinct IDs, with 9,130 IDs from the VGGFace2 [1] dataset and another 2,989 IDs from the CelebAHQ [4] dataset, which are not overlapped with those from the VGGFace2.

Feature Matching. As presented in Sec. 3.3 in the main paper, we define the feature matching loss by the L_2 distance between the multi-level features from the discriminator D for the target image $\mathbf{X}_t$ and the output $\mathbf{Y}_{s,t}$. Here we constrain the feature matching to the background region for the first $m = 4$ layers (*i.e.*, 1024^2-128^2 resolution). Besides, we compute the whole-image feature matching only for pairs that have the same identities.

2 Experiments

In this section, we present more qualitative results for identity swapping. Besides, we explore the latent space of the identity by conducting identity mixing in the $\mathcal{W}+$ space and identity interpolation in both the $\mathcal{Z}+$ and $\mathcal{W}+$ space.

Identity Mixing. The *style mixing* operation in the original StyleGAN [5] means to generate an image with different style codes ($\mathbf{w} \in \mathcal{W}$) at different layers. Given that we have adopted a hierarchical $\mathcal{W}+$ space design, we investigate to mix the intermediate latent codes from two identities at the low, middle, and high-level layers, respectively. Fig. 2 presents some examples of identity swapping results with mixed $\mathbf{w}$ codes. It can be seen that the $\mathbf{w}$ codes at different scales influence different characteristics.

Identity Interpolation. We perform a linear interpolation in the $\mathcal{Z}+$ (Fig. 3 (a)) space and the $\mathcal{W}+$ space (Fig. 3 (b)) , respectively. Technically, we interpolate between the embedded latent identity codes of the source image and those of the target image, where λ is the proportion of the source codes. Then, the image is generated with the interpolated identity codes and the conditioned attributes from the target image. It can be observed that the interpolated codes in both latent space generate feasible results, showing that the latent space $\mathcal{Z}+$ and $\mathcal{W}+$ are uniformly distributed.

More Quantitative Results. We provide more results of *StyleFace* w.r.t.1024^2 resolution identity swapping in Fig. 4. It can be observed that the proposed model can well preserve the target expression (Fig. 4(a)), facial occlusion (Fig. 4(b)), texture style (Fig. 4(d)), and the source face shape (Fig. 4(c)), while producing realistic and visually appealing effect.

Quantitative test about De-ID. In Table 1, we compare with SOTA De-ID method CIAGAN on the LFW dataset w.r.t.*expression error* and *pose error*. Our method outperforms CIAGAN in both metrics.

Table 1. Quantitative comparison with CIAGAN.

Method	Expression ↓	Pose ↓
CIAGAN	1.26	6.18
Ours	**0.35**	**2.75**

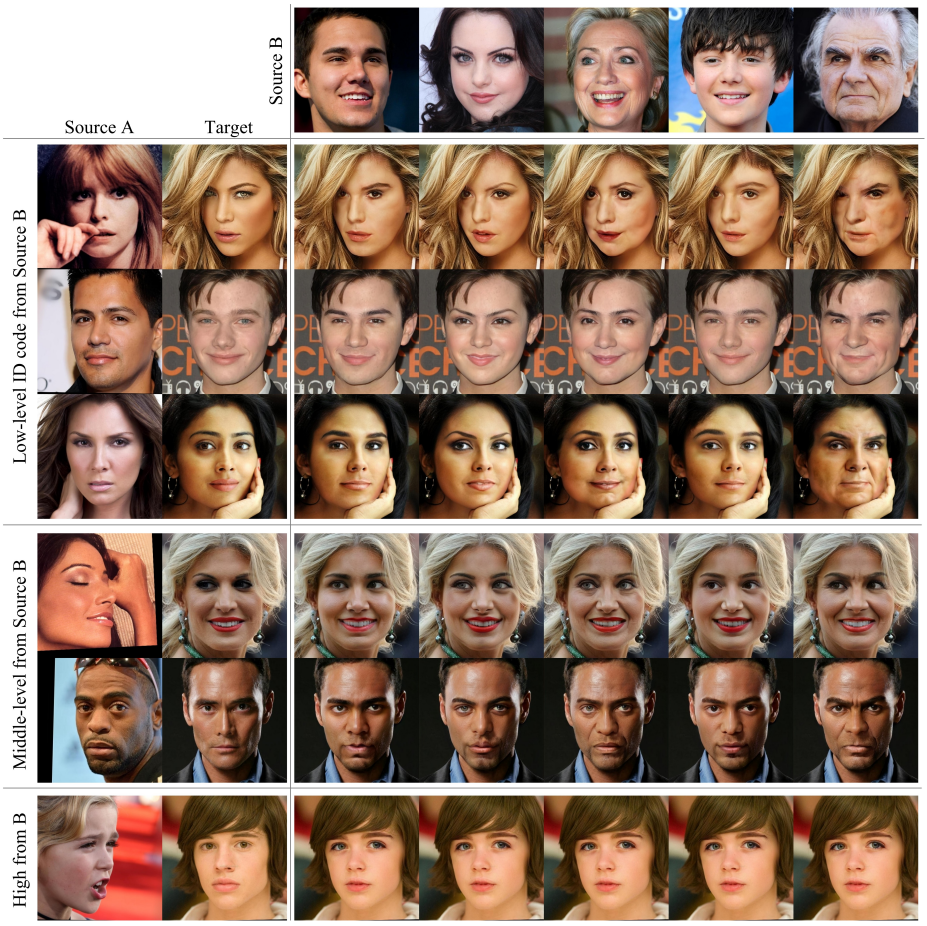
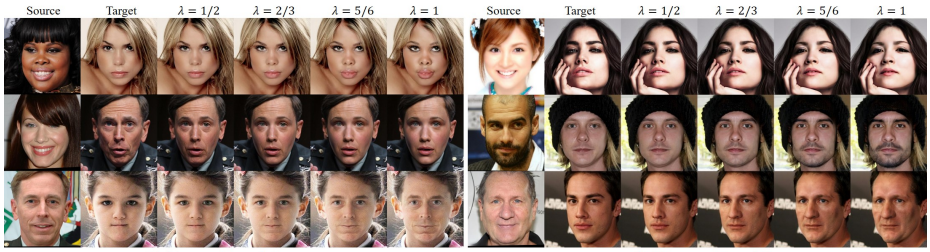


Fig. 2. The identity of the *Target* image is swapped by a combination of latent identity codes from *Source A* and *Source B*. We adopt the **w** codes of Source A by default and replace the code at a specific level with that of Source B. As we can see, the low-level (4^2-16^2) codes bring the most coarse-level characteristics, such as the shape of the face, eye, mouth, *etc.*, from B. The middle-level (32^2-128^2) code changes the smaller scale facial features (*e.g.*, eye color and wrinkles) to that of B. Finally, changing only the high-level codes (256^2-1024^2) does not produce a visible difference, indicating that the high-level code controls some micro-structure.



(a) $Z+$ space interpolation (b) $W+$ space interpolation

Fig. 3. Linear interpolation between the source identity and the target identity in (a) $\mathcal{Z}+$, and (b) $\mathcal{W}+$ space. λ denotes the proportion of the source codes.

(a) Expression

(b) Occlusion

(c) Face shape

(d) Texture style

Fig. 4. More qualitative results for 1024^2-resolution identity swapping under several challenging situations: (a) large facial expression; (b) facial occlusion; (c) different face shape; (d) stylized target image.

References

1. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG 2018. pp. 67–74. IEEE (2018)
2. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
3. 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 pp. 1–8 (2020)
4. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018), https://openreview.net/forum?id=Hk99zCeAb
5. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR. pp. 4401–4410 (2019)
6. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR (2020)
7. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)