Reinforced Disentanglement for Face Swapping without Skip Connection

–Supplementary Material

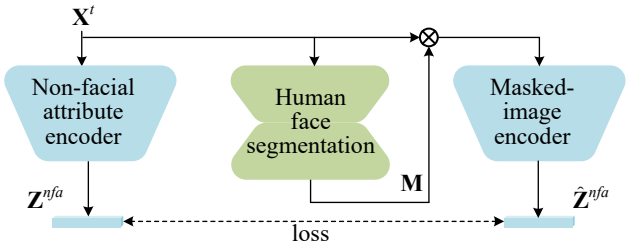

Figure 1. Details of NFA structure in training stage.

A. Training the NFA encoder

To encourage the NFA encoder to extract the non-facial-region features, a masked-image encoder $\mathcal{M}$ is trained simultaneously with the NFA encoder during the training stage. As shown in Fig.1, a pre-trained human face segmentation model is used to predict face mask $\mathbf{M}$, and then $\hat{\mathbf{Z}}^{nfa}$ is generated with $\mathbf{M}$ and $\mathbf{X}^t$:

$$\hat{\mathbf{Z}}^{nfa} = \mathcal{M}((1-\mathbf{M})\cdot\mathbf{X}^t), \tag{1}$$

where $(1-\mathbf{M})$ is the non-face mask.

In this way, facial-region features are excluded in $\hat{\mathbf{Z}}^{nfa}$. Thus we can remove facial region features from $\mathbf{Z}^{nfa}$ by using a regularization loss between $\mathbf{Z}^{nfa}$ and $\hat{\mathbf{Z}}^{nfa}$. The overall loss also encourages $\mathbf{Z}^{nfa}$ to contain non-facial region features as much as possible because $\mathbf{Z}^{nfa}$ is spatially larger than $\mathbf{Z}^{fnid}$ and has a stronger capacity of detail preservation.

B. Training And Evaluation Details

Architecture Details. The detailed network structures of FNID and NFA modules are shown in Tab.1 The FNID encoder contains 7 down-sampling convolutional layers, while the NFA encoder contains 4 down-sampling convolutional layers and 3 ResBlocks. AdvHead is a 3-layer MLP that outputs a 512-dim vector with hidden layer size of 1024. The AdvHead is designed to be stronger to erase ID information from $\mathbf{Z}^{fnid}$. Besides, both ID encoder and AdvHead are the pre-trained ArcFace [4] face recognition model , and the pre-trained 3DMM predictor from [5] is used to form regularization loss $\mathcal{L}_r^{fnid}$. RegHead is an FC layer to output a 67-dim vector with 3-dim $\mathbf{v}^{pose}$ and 64-dim $\mathbf{v}^{exp}$. In the last several layers of our Fusion network, the AAD have three inputs: $\mathbf{F}_k^{fnid}$, $\mathbf{F}_k^{nfa}$ and $\mathbf{Z}^{id}$. $\mathbf{F}_k^{fnid}$ and $\mathbf{F}_k^{nfa}$ are firstly concatenated to predict β and γ, *i.e.*, $\beta,\gamma = \text{Conv}([\mathbf{F}_k^{fnid}, \mathbf{F}_k^{nfa}])$. The usage of $\mathbf{Z}^{id}$ is the same as that in FaceShifter.

	FNID	NFA
Encoder	Conv(c= 32, s=2)	Conv(c= 32, s=2)
	Conv(c= 64, s=2)	Conv(c= 64, s=2)
	Conv(c= 128, s=2)	Conv(c=128, s=2)
	Conv(c= 256, s=2)	Conv(c=256, s=2)
	Conv(c= 512, s=2)	ResBlk(c=512, s=1)
	Conv(c=1024, s=2)	ResBlk(c=512, s=1)
	Conv(c=1024, s=2)	ResBlk(c=512, s=1)
Decoder	TConv(c=1024, s=2)	TConv(c=256, s=2)
	TConv(c= 512, s=2)	TConv(c=128, s=2)
	TConv(c= 256, s=2)	TConv(c= 64, s=2)
	TConv(c= 128, s=2)	TConv(c= 32, s=2)
	TConv(c= 64, s=2)	
	TConv(c= 32, s=2)	

Table 1. The FNID and NFA module details. Conv is the standard convlutional layer. TConv is the transposed convlutional layer. ResBlk is the residual convlutional block [7]. "c" is the number of output channels, and "s" denotes the up/down-sampling scales.

Learnable Parameters. The parameter size for each component is as follows: FNID encoder is 25.2M, FNID decoder is 6.3M, NFA encoder is 6.3M, NFA decoder is 1.6M, NID encoder is 25.2M, and NID decoder is 12.6M. Therefore, the total number of learnable parameters for FNID + NFA is only 4.2% more than that of NID. The NID decoder is larger than the FNID decoder because skip-connections double the input channels in the decoder.

More details of the training losses . The adversary loss $\mathcal{L}_{adv}$ is a Hinge GAN loss from a multi-scale (256, 128, 64) discriminator. The term $\mathcal{L}_{ah}^{fnid}$ is to train the AdvHead. When it is used, the whole FNID encoder is fixed except the AdvHead. In contrast, $\mathcal{L}^{fnid}$ is to train FNID encoder with RegHead, thus $\mathcal{L}_{ah}^{fnid}$ does not contribute to the $\mathcal{L}^{fnid}$.

Hyper-parameters. For balance and stable training, we set $\beta_{adv}^{fnid} = 0.1$ in $\mathcal{L}^{fnid}$; $\beta^{rec} = 0.2$ and $\beta^{attr} = 0.5$ in L^{glb}; $\beta^{glb} = 5$, $\beta^{fnid} = 2$, $\beta^{nfa} = 100$ in the overall loss.

Details of Comparison with StyleGAN-based Methods .

Figure 2. Comparison of ID consistency. The top-left corner is the source image, while the other images in the first row are target images.

Recent SOTA StyleGAN-based face swap methods [13, 14] have not released their inference models or face swap results on FF++, thus we evaluate our method using the same pose and expression metrics of [14] for fair comparison. In terms of FID score, we only compare with [13] because the work [14] did not report the FID evaluation details.

User Study Conduction Details. We conduct a user study to evaluate the face swap performance from three perspectives: source ID similarity, target non-ID preservation, and image quality. To this end, we randomly select 100 source-target pairs from the FF++ [10] testset. The swap results are from faceshifter [8], Simswap [3], InfoSwap [6], and our method. Then, participants are asked to select: (i) the one with the best source-image ID similarity; (ii) the one with the best target-image similarity of the pose and facial expression; and (iii) the one that looks the most like the real photo.

C. More Qualitative Results

Comparison of ID Consistency. We believe ID consistency is important in many applications (*e.g.*, virtual human creation, film-making), and swap identity should be consistent across various contents. Therefore we show additional results in Fig.2, where our method is superior in ID consistency.

Comparison with Prior Arts in FF++. To further visually compare our method with prior methods, we randomly collect source-target pairs in FF++. Referring to Fig. 3, we can see that our swap results are better than that of other methods in terms of source-ID similarity and target-non-ID preservation, indicating our method has advantages in the disentanglement representation.

Comparing with Simswap. Although Simswap [3] achieves slightly lower expression error than our method in quantitative comparison, its performance on source-ID similarity lags considerably behind our method. In Figs. 2 and 3, Simswap has minor visual advantages in expression preservation, but its swap results are not similar to the source ID. For example, referring to the 2nd-8th columns in Fig.2 and the all results in Fig.3, the swapped ID of Simswap are quite close to the target images. In contrast, our results are overall superior when compared to Simswap.

Figure 3. Comparison of face-swapping results on FF++.

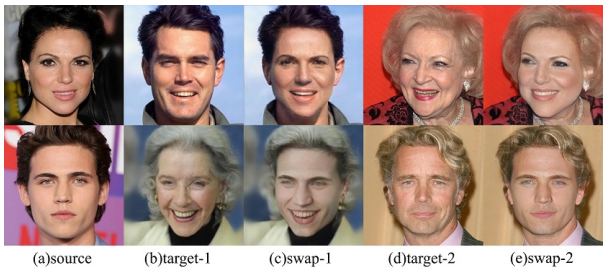

Figure 4. Cross gender and age results by our method.

Cross-Age/Gender/Hairstyle Face Swap Results. As shown in Fig.4, our method can produce impressive face swap results for difficult cross-age/gender cases.

As for cross-hairstyle face swap, there are two situations as shown in Fig.5: (1) Target has bangs while source has no bang. Our method can handle this situation because our NFA encoder can detect the bangs as non-facial attributes

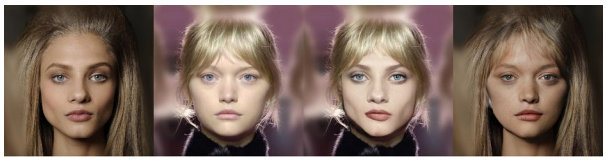

Figure 5. Cross hairstyle results by our method. Result-AB is produced by using image A as the source image and image B as the target image, result-BA is produced by using image B as the source image and image A as the target image.

and our decoder will preserve them. (2) Source has bangs while target has no bang. Our method cannot handle this situation because the pretrained ID encoder regards bangs as a part of facial ID.

D. Discussion on Face shape swap

Face shape is a essential part of face ID. In our method design, the learned masks in AAD ResBLK affect signif-

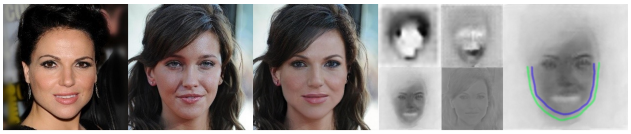

Figure 6. Visualization of the masks when face shape changes.

icantly to face shape swap. Fig. 6(d) shows the learned masks in last four AAD ResBLK when face shape changes. From the masks, we find out the 7th AAD ResBLK plays important role in face shape swap. Fig. 6(e) shows the details of mask of the 7th AAD ResBLK. The GREEN and BLUE lines roughly represent the swap and target face shape. The inner region of the BLUE line are very dark and the outer region of GREEN line is very light, which means those regions are generated mostly according to one input feature (either ID or non-ID). The region between the BLUE and GREEN line is lighter than the facial region yet darker than the background, which means this region should be generated according to both ID and non-ID features. From the qualitative results, our method is with better swapped face shapes than other methods.

E. Ethical Consideration

The goal of this paper is to study high-quality face swaps. It does not intend to manipulate existing images or to create misleading or deceptive content. However, the method, like all other related AI image generation techniques, could still potentially be misused for impersonating humans. We condemn any behavior to create such harmful content. Currently, the synthesized portraits by our method contain certain visual artifacts that can be identified by humans and some deepfake detection algorithms. We encourage to apply this method for learning more advanced forgery detection approaches to avoid potential misusage.

References

[1] Deepfake. *https://github.com/deepfakes/faceswap*, 2020. 3, 5

[2] Xu Chao, Zhang Jiangning, Han Yue, Tian Guanzhong, Zeng Xianfang, Tai Ying, Wang Yabiao, Wang Chengjie, and Liu Yong. Designing one unified framework for high-fidelity face reenactment and swapping. In *European conference on computer vision*, 2022. 2, 3, 5

[3] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 2, 3, 5

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1

[6] Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3404–3413, 2021. 2, 3, 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[8] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5074–5083, 2020. 2, 3, 5

[9] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 3, 5

[10] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2

[11] Truong Vu, Kien Do, Khang Nguyen, and Khoat Than. Face swapping as a simple arithmetic operation. *arXiv preprint arXiv:2211.10812*, 2022. 2, 3, 5

[12] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142, 2021. 3, 5

[13] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022. 2

[14] Zhiliang Xu, Hang Zhou, Zhibin Hong, Ziwei Liu, Jiaming Liu, Zhizhi Guo, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Styleswap: Style-based generator empowers robust face swapping. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[15] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4834–4844, 2021. 3, 5

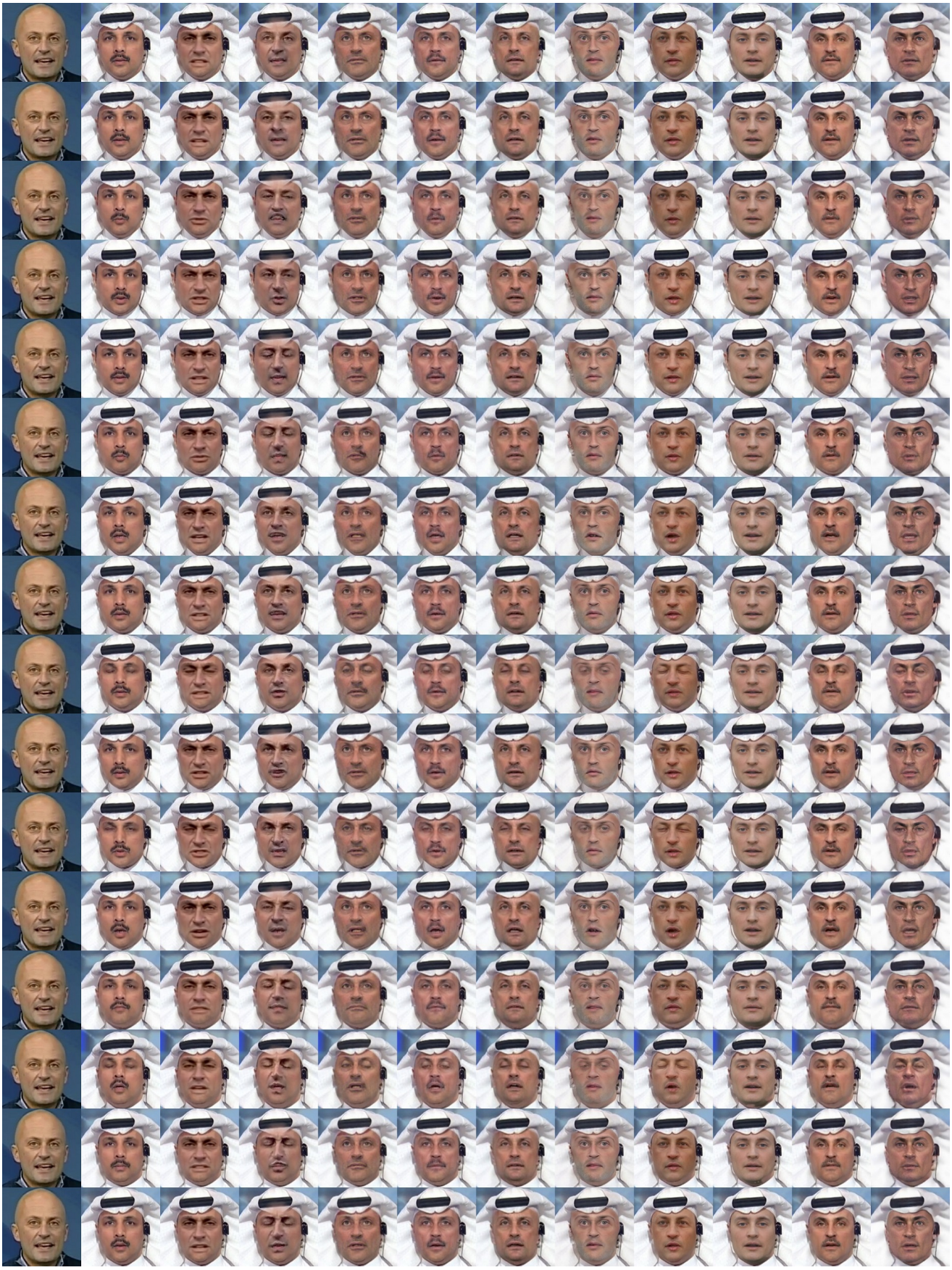

Figure 7. Comparison of face-swapping results on a video from FF++.