Designing One Unified Framework for High-Fidelity Face Reenactment and Swapping

Chao Xu[1⋆], Jiangning Zhang[1,3⋆], Yue Han[1], Guanzhong Tian[2], Xianfang Zeng[1], Ying Tai[3], Yabiao Wang[3], Chengjie Wang[3], and Yong Liu[1†]

[1] APRIL Lab, Zhejiang University
[2] Ningbo Research Institute, Zhejiang University
[3] YouTu Lab, Tencent
{21832066, 186368, 22132041, gztian, zzlongjuanfeng}@zju.edu.cn
{yingtai, caseywang, jasoncjwang}@tencent.com yongliu@iipc.zju.edu.cn

In our supplemental materials, we first attach the training and inference codes of our proposed method in Pytorch. Because the maximum file size is 100 Mb, we could not upload any checkpoint. The detailed network structure could refer to our codes. Then we provide the following contents, which were not presented in the paper due to the space limitations:

- Additional swapped faces on FaceForensics++ [6] and CelebA-HQ [3] test set. Details can be found in following Sec. 1
- Additional reenacted faces of the reconstruction and reenactment tasks on VoxCeleb2 [2] test set. Details can be found in following Sec. 2
- Additional video face swapping results sampled from FaceForensics++ and video face reenactment results sampled from VoxCeleb2 test set. Details can be found in following Sec. 3

1 More Swapped Results

We show more swapped faces on FaceForensics++ [3] and CelebA-HQ [3] in Fig. 1 and Fig. 2. MegaFS [10] fails to produce realistic facial texture due to the directly blending post-processing operation. Compared to the FaceShifter [4] and SimSwap [1], our results share the local (*i.e.*, mouth and eyes areas) and global (*i.e.*, facial textures) identity information with the source faces much better while keeping the attributes of the target unchanged. Besides, thanks to the sufficient identity-related feature interaction and powerful face generator, our method does not introduce inappropriate cues from the source and generates more realistic results, *i.e.*, fewer artifacts and higher sharpness.

2 More Reenacted Results

We show more reenacted faces of the reconstruction task and the reenactment task on VoxCeleb2 [2] test set in Fig. 3 and Fig. 4. X2Face [8] can only handle the

⋆ indicates equal contributions.
† indicates corresponding author.

case when two faces are in similar poses. Bi-layer [9] fails to produce authentic textures and background, and the face shapes are not faithful to the source. FOMM [7] and PIRenderer [5] successfully animate the source into the attributes of the target, but they struggle to keep the identity consistency when two faces are quite different of the poses and face shape, thus the synthesized faces are prone to low realism. For comparison, our method generates more realistic results with accurate pose and expression while still preserving the source identity in various conditions.

3 Video Face Swapping and Reenactment

For face swapping, we sample two pairs from FaceForensics++ [6] for video face swapping: $540 - 536$, $025 - 067$. The former index provides the target face, while the latter provides the source face. Note that our model is not trained on FaceForensics++, and no temporal constraints are used during training and inference. The results of our method are highly consistent with the source identity and target attributes. For face reenactment, we sample eight pairs from VoxCeleb2 [2] test set, *i.e.*, four pairs for the reconstruction and four pairs for the reenactment, which vividly show that our method successfully learns accurate motions from the target and maintain the identity of the source face under some challenging conditions, leading to the better video coherence. Please refer to the supplementary video for more details. Notably, reenacted videos are presented in 10 fps for better comparison.

Fig. 1. More results compared to FaceShifter [4], SimSwap [1], and MegaFS [10] on FaceForensics++ [6]. Please zoom in for more details.

Fig. 2. More results compared to SimSwap [1], and MegaFS [10] on CelebA-HQ [3] test set. Please zoom in for more details.

Fig. 3. More reconstruction results compared to X2Face [8], Bi-layer [9], FOMM [7], and PIRenderer [5] on VoxCeleb2 [2] test set. Please zoom in for more details.

Fig. 4. More reenactment results compared to X2Face [8], Bi-layer [9], FOMM [7], and PIRenderer [5] on VoxCeleb2 [2] test set. Please zoom in for more details.

References

1. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2003–2011 (2020)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018)
3. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
4. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
5. Ren, Y., Li, G., Chen, Y., Li, T.H., Liu, S.: Pirenderer: Controllable portrait image generation via semantic neural rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13759–13768 (2021)
6. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
7. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in Neural Information Processing Systems **32**, 7137–7147 (2019)
8. Wiles, O., Koepke, A., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–686 (2018)
9. Zakharov, E., Ivakhnenko, A., Shysheya, A., Lempitsky, V.: Fast bi-layer neural synthesis of one-shot realistic head avatars. In: European Conference on Computer Vision. pp. 524–540. Springer (2020)
10. Zhu, Y., Li, Q., Wang, J., Xu, C.Z., Sun, Z.: One shot face swapping on megapixels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4834–4844 (2021)