3D-FM GAN: Towards 3D-Controllable Face Manipulation Supplementary Material

Yuchen Liu[1], Zhixin Shu[2], Yijun Li[2], Zhe Lin[2], Richard Zhang[2], S.Y. Kung[1]

[1]Princeton University [2]Adobe Research
[1]{yl16, kung}@princeton.edu [2]{zshu, yijli, zlin, rizhang}@adobe.com

We organize our supplementary as follows. We include a video presentation of 3D-FM GAN in video link. In Sec. 1, we include more implementation details. In Sec. 2, we show more comparison results among different designs of co-modulation architectures. In Sec. 3, we include more results for single factor disentangled editing as well as reference based generation. More comparisons with CONFIG [9] and VariTex [2] are shown in Sec. 4. We discuss the limitation and potential societal impacts of our work in Sec. 5.

1 Detailed Implementations

1.1 Modules and Training Strategies

We adopt implementations of the face reconstruction network **FR**, the BFM 3DMM, and the renderer **Rd** all from DiscoFaceGAN [5] released repository[1]. We use a public implementation[2] of StyleGAN2 [7] generator and discriminator. The ResNet-18 [6] encoder for $\mathbf{E_T}$ and $\mathbf{E_W}$ is provided by official release[3] in PyTorch [10]. We use the official implementation[4] of PSP encoder [11].

Our StyleGAN generator $\mathbf{G_s}$ and discriminator $\mathbf{D_s}$ are initialized with the pre-trained weights from the unconditional noise-to-image unpaired training regime. All encoders, $\mathbf{E_T}$, $\mathbf{E_W}$, and $\mathbf{E_{W^+}}$ are initialized randomly. We adopt two Adam optimizers [8] to update the parameters in $\mathbf{G}$ ($\mathbf{G_s}$ and $\mathbf{E}$) and $\mathbf{D_s}$ separately. In phase-1 training, we set our learning rate to be 0.0001 while in phase-2 training, the learning rate is set to be 0.001.

The face recognition network [4], the landmark detection model [3], and the LPIPS module [12] are from [5], [6], and [7].

[1] https://github.com/microsoft/DiscoFaceGAN
[2] https://github.com/rosinality/stylegan2-pytorch
[3] https://pytorch.org/vision/stable/models.html
[4] https://github.com/eladrich/pixel2style2pixel
[5] https://github.com/ronghuaiyang/arcface-pytorch
[6] https://github.com/1adrianb/face-alignment
[7] https://github.com/richzhang/PerceptualSimilarity

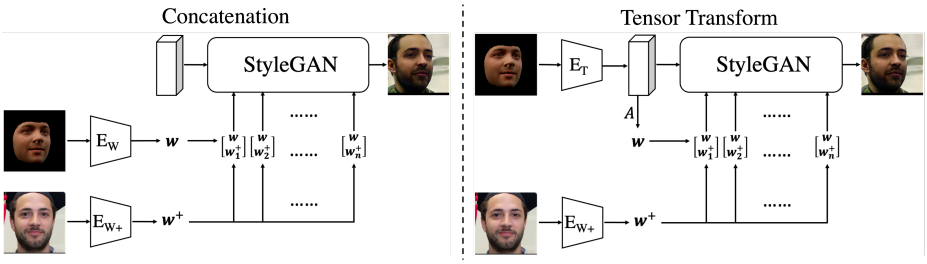

Fig. 1. Two types of co-modulation architectures. **Left:** Concatenation co-modulation architecture. **Left:** Tensor transform co-modulation architecture.

1.2 Co-Modulation Architectures

In addition to the multiplicative co-modulation architecture, we also investigate the concatenation and tensor transform co-modulation, as shown in Fig. 1. For concatenation scheme, we encode R by $\mathbf{E_W}$ into the $\mathcal{W}$ space. Then, for layer l, the modulation signal is provided by concatenating $\mathcal{W}$ and W_l^+ as $[\mathcal{W}, \mathcal{W}_l^+] \in \mathbb{R}^{1024}$. The tensor transform scheme originally proposed in [13] is similar to the concatenation scheme in terms of generating the co-modulation signals, while its R is encoded into $\mathcal{T}$ and an additional linear transformation layer A transforms the flattened $\mathcal{T}$ into $\mathcal{W}$.

1.3 Evaluation of DiscoFaceGAN

FID. We follow a similar procedure as Sec. 4.2 of the main paper to generate manipulated images with the generator $\mathbf{G_d}$ from DiscoFaceGAN (DFG) [5]. The process is similar except that we need to sample an extra noise n for generating each edited image $\hat{P}_d = \mathbf{G_d}(\hat{p}, n)$. We then measure the FID between $\hat{\mathbb{P}}_d$ and $\mathbb{P}$.
Image Manipulation. Since DFG does not provide codes for its image editing, we implement it on our own, strictly following Eqn. 11 of its paper: (1) Given a photo P, obtain its 3DMM parameter by face reconstruction network $p = \mathbf{FR}(P)$, and its latent code w^+ by StyleGAN inversion [1][8]. (2) with the desired manipulation $\hat{p}$, offset w^+ by $\Delta w(p, \hat{p})$ to generate the manipulated face.
Analysis of λ and $\mathcal{W}+$ Space. We conduct an analysis of DFG's λ and $\mathcal{W}^+$ space in Fig. 13 of the main paper. Specifically, given a photo P, we extract its 3DMM parameter by face reconstruction and $p = \mathbf{FR}(P)$ and do the following for the two spaces: (1) $\mathcal{W}+$ space: follow the above image manipulation step with a series of manipulation $\hat{p}$. (2) λ space: sample a noise n and conduct forward-only inference (no back-propagated optimization) with its generator to synthesize images of the original parameter $\mathbf{G_d}(p, n)$ and from the manipulated parameters $\mathbf{G_d}(\hat{p}, n)$.
Run-Time Efficiency. To manipulate an image, DFG would take around 120s to retrieve the latent code on a P100 GPU, followed by a 0.5s synthesis process.

[8] We run 3000 optimization steps to fully retrieve the latent code. The implementation is at: `https://github.com/Puzer/stylegan-encoder`

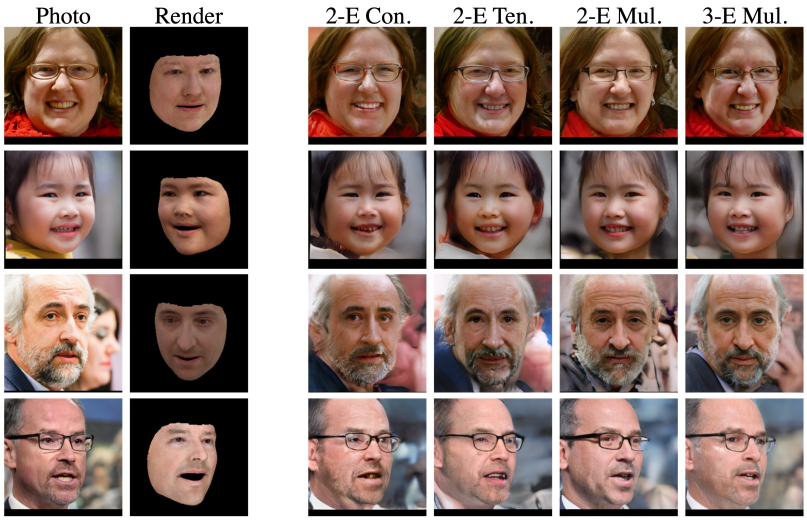

Fig. 2. Visual comparison among different co-modulation architectures. While concatenation (**Col. 3**) and tensor transform (**Col. 4**) schemes have obvious photo-realism issues (**Row 2 & 4**), the multiplication scheme (**Col. 5 & 6**) generally synthesizes images with higher quality, where the 3-encoder architecture (**Col. 6**) further enhances the editability (**Row 1 & 2**).

On the contrary, our method only takes less than 1s for face reconstruction and around 0.7s for image generation on the same hardware. Hence, our method enjoys a speedup of 70× (1.7s vs. 120.5s) for single image editing.

2 Co-Modulation Comparison

In addition to the quantitative results in Tab. 1 of the main paper, we further show visual comparisons of the 4 proposed co-modulation schemes in Fig. 2, where we compare 2-encoder concatenation (**Col. 3**), 2-encoder tensor transform (**Col. 4**), 2-encoder multiplication (**Col. 5**), and 3-encoder multiplication (**Col. 6**) co-modulation architectures. As shown in the figure, the concatenation and tensor transform schemes would have photo-realism issues with significant amount of artifacts (**Row 4**) and unrealistic poses (**Row 2**). On the contrary, the multiplicative scheme performs much better, and the 3-encoder multiplicative co-modulation further demonstrates better editability in lighting (**Row 1**) and pose (**Row 2**).

3 Additional Image Manipulations

We show more results of disentangled editing in Fig. 3. Our model again provides highly disentangled manipulation with high photo-realism and strong identity

preservation. We also show additional reference based face generation results with more identities in Fig. 4. Noticeably, even the identity images with extreme poses (**2nd and 8th identites**) can be well re-posed with the expression and illumination transferred. Our model also well preserves the eyeglasses in manipulating the 7th identity image.

4 More Comparisons to SOTA Methods

We show additional comparisons with CONFIG [9] in Fig. 5 on both yaw and pitch rotations for real image. Clearly seen from the plot again, our method enjoys larger range of editability, stronger identity preservation, and higher photo-realism.

We further include comparisons with VariTex [2] in Fig. 6 on real image manipulation. Again, we find a clear advantage in identity preservation for our model over VariTex for all inputs. As shown in the **Top** example, while VariTex could not synthesize background and the generated hair has a unrealistic texture, our method demonstrates a much higher photo-realism with better background and hair. Moreover, we compare the editability between our model and VariTex by manipulating images with extra rigid bodies, the eyeglasses, which represents a harder task in the **Bottom** example. While VariTex could not properly synthesize the faces with glasses, our model provide a decent control to generate high-quality images.

5 Limitations & Societal Impacts

Although 3D-FM GAN shows a strong ability for 3D-controllable, identity-preserved face editing, there remains certain limitations and potential negative impacts.

Limitations. Our 3DMMs can not model fine details like wrinkles and hair styles, and thus our model can not explicitly control those attributes. We also see a gap between the reconstructed images and the inputs in face shapes, which might be caused by the imprecise 3D estimation in face reconstruction. Moreover, our model would inherit bias from the training data, and due to lack of public availability, we can only use synthetic data for disentangled learning.

Potential Negative Impacts. Face manipulation techniques has in the past helped creating deep-fakes and spread disinformation. Our work is intended for intelligent content creation for portrait photography and we believe it does not improve the accessibility of deep-fakes and disinformation. Moreover, our discoveries of identity-editability trade-off might also offer new viewpoints on future development for deep-fake detection techniques.

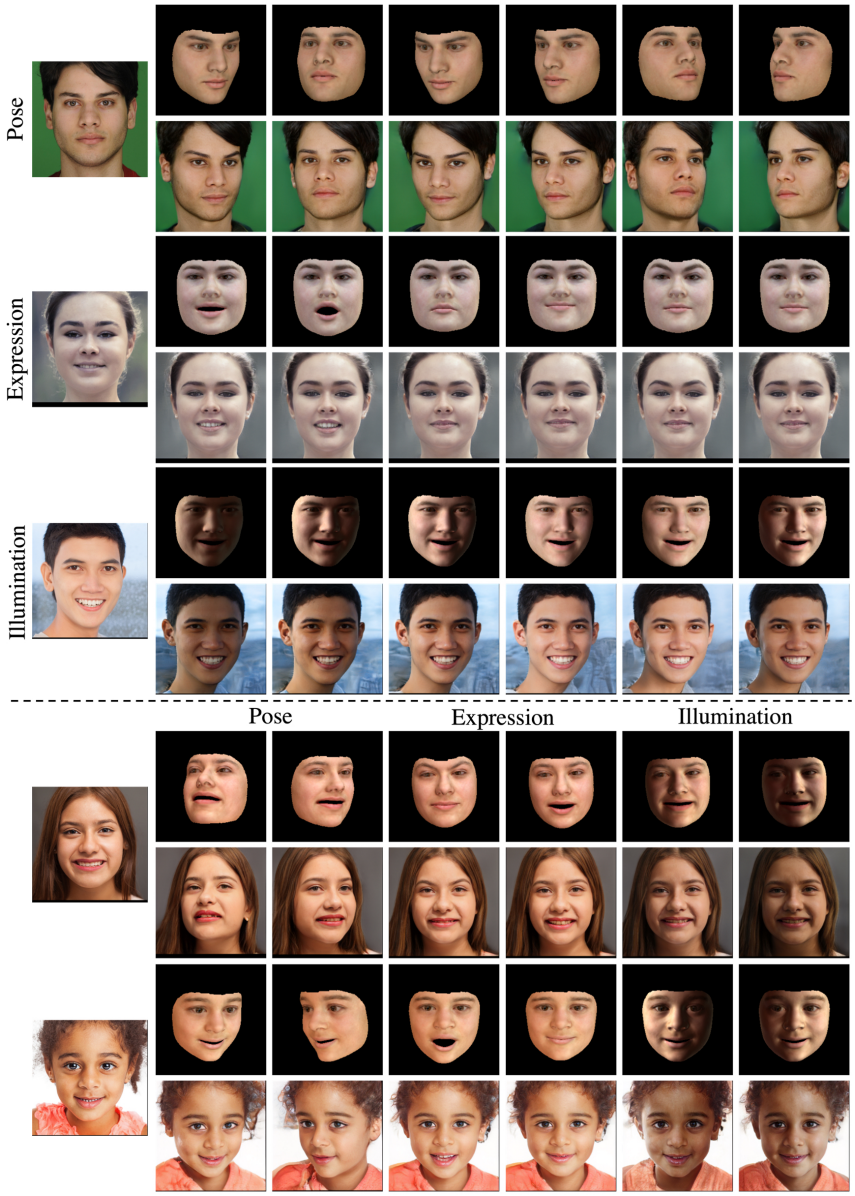

Fig. 3. More results for disentangled editing. Our model again achieves good disentangled editability, high photo-realism, and strong identity preservation.

Fig. 4. Additional examples for reference based face generation. Our model demonstrates a good editability with identities with extreme poses (**2nd and 8th**) and with eyeglasses (**7th**).

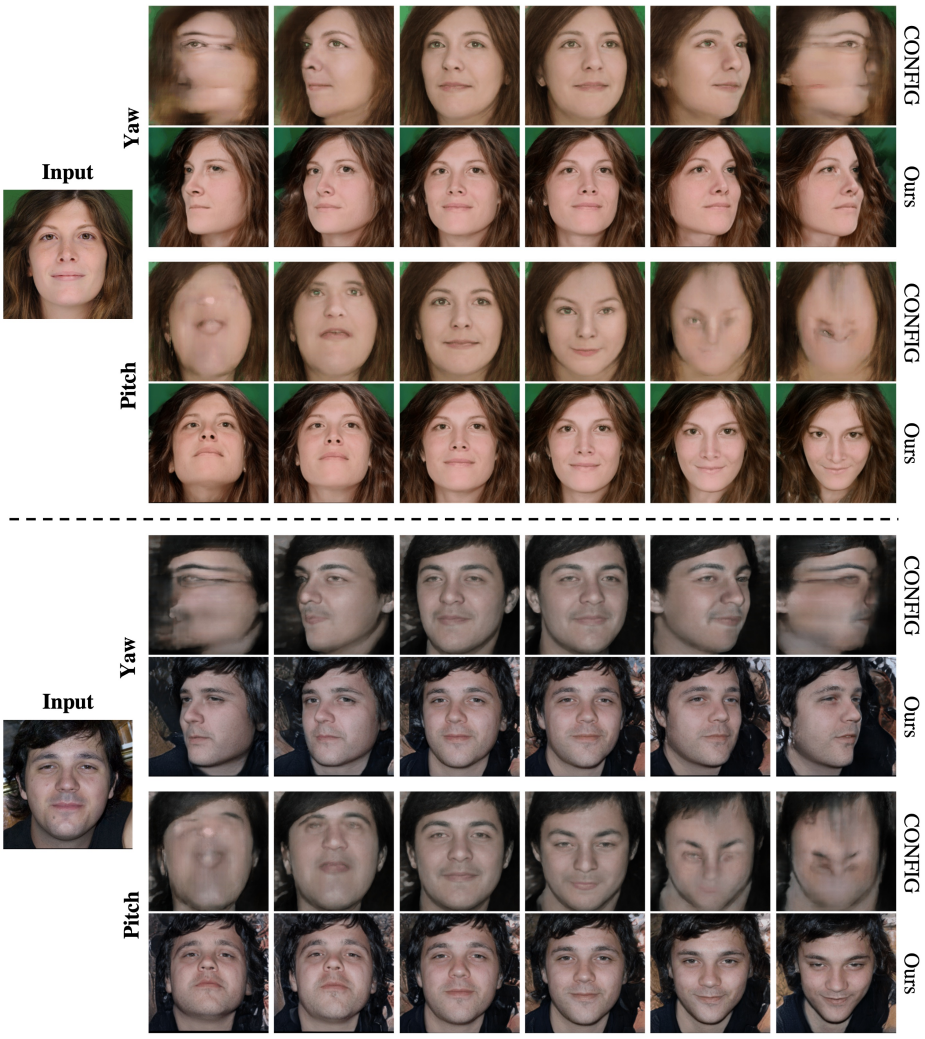

Fig. 5. More comparison with CONFIG [9] on real image editing. Our method again outperforms CONFIG with larger range of editability, stronger identity preservation, and higher photo-realism.

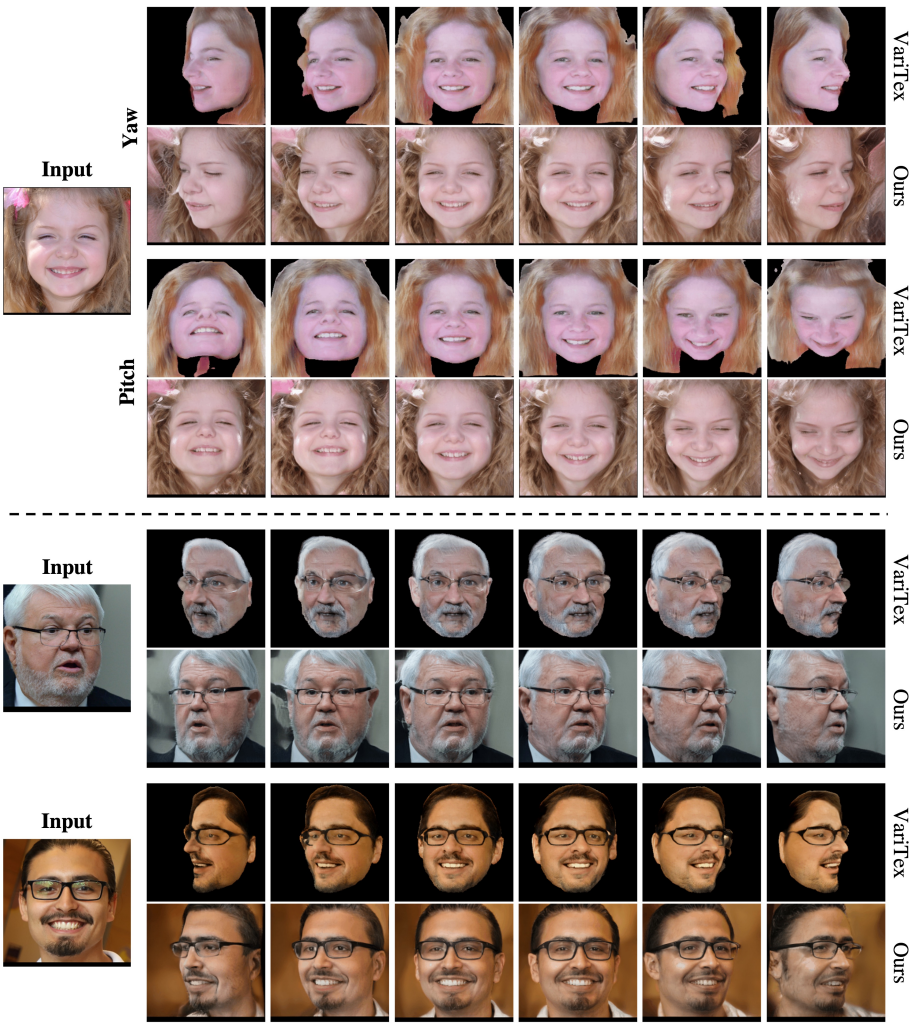

Fig. 6. Manipulating the same real images with our model and VariTex [2], where our model shows better identity preservation in all examples. **Top:** While VariTex could not synthesize realistic hair and background due to the absence of 3DMM modelling, our model provides a much better synthesis result on these regions, demonstrating a higher photo-realism and better editability. **Bottom:** We manipulate faces with additional rigid bodies like glasses. Our method again generates images with much higher photo-realism, even at extreme poses.

References

1. Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020.
2. Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *International Conference on Computer Vision (ICCV)*, 2021.
3. Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.
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.
5. Yu Deng, Jiaolong Yang, Dong Chen, Fang Wen, and Xin Tong. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5154–5163, 2020.
6. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
7. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
8. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
9. Marek Kowalski, Stephan J Garbin, Virginia Estellers, Tadas Baltrušaitis, Matthew Johnson, and Jamie Shotton. Config: Controllable neural face image generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 299–315. Springer, 2020.
10. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.
11. Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021.
12. Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017.
13. Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021.