VOODOO 3D: Volumetric Portrait Disentanglement for One-Shot 3D Head Reenactment

Supplementary Material

6. Training Details

Training Data. We fine-tune Lp3D using CelebV-HQ dataset [115]. For the expression modules, we also use the CelebV-HQ dataset but adopt an expression re-sampling process to make the expressions of the sources and drivers during training more different. Specifically, for a given video, we use EMOCA [23] to reconstruct the mesh of every frame without the head pose. Let these obtained meshes be $\{M_1, M_2, ..., M_n\}$, we first pick two frames x^* and y^* such that the distance between their meshes are maximized:

$$x^*, y^* = \arg\max_{x,y} \|M_x - M_y\|_2.$$

Then we pick the third frame z^* such that:

$$z^* = \arg\max_{z} min\left(\|M_{x^*} - M_z\|, \|M_{y^*} - M_z\|\right).$$

We use this frame selection process for all the videos in the CelebV-HQ dataset [115] and use the re-sampled frames to train the expression modules. A few examples from this selection process are shown in Fig. 9.

Figure 9. Some examples of our training data extracted from the CelebV-HQ dataset [115]

Driver Augmentation. To prevent identity leaking from the driver to the output, we apply several augmentations to the frontalized driver images, including: (1) Kornia color jiggle[1] with parameters for brightness, contrast, saturation, hue set to 0.3, 0.4, 0.3, and 0.4, respectively; (2) random channel shuffle; (3) random warping[2]; and (4) random border masking with the mask ratio uniformly sampled from 0.1 to 0.3. During testing, we removed all the augmentations except the random masking and fixed the mask ratio to 0.25. This random masking greatly improves the consistency in the output, especially for border regions. In addition, since we mask the border with a fixed rate, we can modify the renderer to only generate the center of the frontalized driver and further improve the performance.

Conv2d(96, 96, kernel_size=3, stride=2, padding=1)
ReLU()
Conv2d(96, 96, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(96, 128, kernel_size=3, stride=2, padding=1)
ReLU()
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)

Table 4. Architecture of E_T

Architecture Details. Our architecture design is inspired by Lp3D [84]. Specifically, for $\mathbf{E}_s$ and $\mathbf{E}_d$, we use two separate DeepLabV3 [22] with all normalization layers removed. Since the triplane already captures deep 3D features of the source, we adopt a simple convolutional network for $\mathbf{E}_t$, which is given in Tab. 4. Recall that:

$$F = F_s \oplus F_d \oplus F_t$$

For the final transformer that is applied on the concatenations of the feature maps F, we use a slight modification of $\mathbf{E}_{\text{low}}$ (light-weight version) in Lp3D [84]. The architecture of this module is given in Tab. 5 where block used is the transformer block in SegFormer [93]. As mentioned in our paper, we use a pretrained GFPGAN as the super-resolution module. This module is loaded from a public pretrained weight GFPGAN v1.4 [88] and fine-tuned end-to-end with the network.

[1] https://kornia.readthedocs.io/en/latest/augmentation.module.html#kornia.augmentation.ColorJiggle

[2] https://github.com/deepfakes/faceswap/blob/a62a85c0215c1d791dd5ca705ba5a3fef08f0ffd/lib/training/augmentation.py#L318

```
PatchEmbed(64, patch=3, stride=2, in=640, embed=1024)
Block(dim=1024, num_heads=4, mlp_ratio=2, sr_ratio=1)
Block(dim=1024, num_heads=4, mlp_ratio=2, sr_ratio=1)
PixelShuffle(upscale_factor=2)
upsample(scale_factor=2, mode=bilinear)
Conv2d(256, 128, kernel_size=3, stride=1, padding=1)
ReLU()
upsample(scale_factor=2, mode=bilinear)
Conv2d(128, 128, kernel_size=3, stride=1, padding=1)
ReLU()
Conv2d(128, 96, kernel_size=3, stride=1, padding=1)
```

Table 5. Architecture of the transformer network used in the expression module.

Training Losses. To train the model used in our experiments, we set $\lambda_{\text{syn}} = 0.1$, $\lambda_{\text{tri}} = 0.01$, and $\lambda_{\text{CIR}} = 0.01$. For GAN-based losses, we use hinge loss [56] with projected discriminator [71].

7. Implementation Details for Holographic Display System

We implement our model on a Looking Glass monitor 32"[3]. To visualize results on a holographic display, we must render multiple views for each frame using camera poses with a yaw angle that spans the range from $-17.5°$ to $17.5°$. In our case, we find that using 24 views is sufficient for the user experience. While our model can run at 32FPS using a single NVIDIA RTX 4090 on a regular monitor, which only requires a single view at a time, it cannot run in real-time when rendering 24 views simultaneously. Thus, to achieve real-time performance for the Looking Glass display, we ran the holographic telepresence demo on seven NVIDIA RTX 6000 ADA GPUs.

We parallelize the rendering process to four GPUs, so each one needs to render six views in a batch. We dedicate one GPU for driving image pre-processing and another one for disentangled tri-plane estimation. We use the last GPU to run the looking-glass display itself. This setup results in 25 FPS for the whole application. We showcase the results rendered on the holographic display in the supplementary videos.

8. Additional Comparisons with LPR [55]

In this section, we compare our method with the current state-of-the-art in 3D aware one-shot head reenactment, LPR [55] using their test data from HDTF [109] and CelebA-HQ datasets [41]. In particular, for CelebA-HQ, they use even-index frames as sources and odd-index frames as drivers, while in contrast, in our experiment section, we use the first half as sources and the rest as drivers. For the HDTF dataset, they use a single driver (WRA_EricCantor_000) and the first frame of each video as source image. Compared to our split, this reduces the diversity in the driver images. We provide the comparison results in Tab. 6 and Tab. 7. The ECMD scores on both datasets show that our method is more accurate in transferring expression from the driver to the source images. On the HDTF dataset, our results have much higher CSIM. Our FID score is better than LPR [55] on CelebA-HQ but worse on the HDTF dataset. We found that the HDTF's ground-truth images have poor quality while our outputs are higher in quality; this mismatch causes our FID to be unimpressive on this dataset. Hence, this FID arguably does not correctly reflect the performance of our model. According to the qualitative examples in Fig. 14, our method captures the driver's expression more accurately than LPR. However, we note that our quality is even higher than the input, as can be observed in Fig. 14.

We also provide extensive qualitative comparisons in Fig. 16 and Fig. 14. The expression of our output images is more realistic and faithful to the driver, which is particularly more visible in the mouth/teeth/jaw region, as well as for driver or source side views. Notably, in Fig. 15, it can be observed that LPR fails to remove the smiling from the source, resulted in inaccurate expression in the reenacted output while our method can still successfully transfer the expression from the driver to the source image.

Method	Cross-reenactment		
	CSIM	ECMD	FID
LPR [55]	0.531	0.912	**25.26**
Ours	**0.774**	**0.860**	54.15

Table 6. Quantitative comparisons with LPR [55] on HDTF dataset using the test split proposed in [55].

Method	Cross-reenactment		
	CSIM	ECMD	FID
LPR [55]	**0.643**	0.483	47.39
Ours	0.628	**0.473**	**34.27**

Table 7. Quantitative comparisons with LPR [55] on CelebA-HQ dataset using the test split proposed in [55].

9. Additional Qualitative Comparisons

We provide additional qualitative comparisons with other methods in Fig. 18, Fig. 19, Fig. 20, Fig. 21, Fig. 22, Fig. 23, Fig. 24, Fig. 25, Fig. 26, Fig. 27, Fig. 28, Fig. 29, Fig. 30, and Fig. 31.

In Fig. 17, we evaluate the ability to synthesize novel views of our method. In addition, we also reconstruct the 3D mesh of the reenacted results.

[3]https://lookingglassfactory.com/looking-glass-32

In Fig. 10, we evaluate our model on self-reeactment task using HDTF and our collected datasets.

In Fig. 11, we compares our method with the others on source images that have jewelries. As can be seen, other methods struggle to reconstruct the jewelries while our results still have the jewelries from the source input.

10. Addtional Experiments with PTI [70]

Our method can achieve high-quality results without noticeable identity change without additional fine-tuning, which is known to be computaionally expensive. In this section, we try to fine-tune [70] the super-resolution module using PTI [70] for 100 iterations, which takes around 1 minute per subject. Without PTI, our pipeline runs instantly similarly to [55]. For most cases, the difference between results with and without fine-tuning is negligible. However, for out-of-domain images such as Mona Lisa, PTI fine-tuning helps retain the oil-painting style and fine-scale details from the input source. For the fine-tuning results, please refer to the supplementary video.

11. Additional Limitations

Besides the limitations that we discussed in the paper, we also notice that the model cannot transfer tongue-related expressions or certain asymmetric expressions due to limited training data for our 3D lifting and expressions module. Since our method is not designed to handle the shoulder pose, the model uses the head pose as a single rigid transformation for the whole portrait. This issue would be an interesting research direction for future work. Also, our model sometimes fails to produce correct accessories when the input has out-of-distribution sunglasses. These failure cases are illustrated in Fig. 12.

References

[1] ItSeez3D AvatarSDK, https://avatarsdk.com. 1

[2] in3D, https://in3d.io. 1

[3] Leia, https://www.leiainc.com. 2

[4] Looking Glass Factory, https://lookingglassfactory.com. 2

[5] Pinscreen Avatar Neo, https://www.avatarneo.com. 1

[6] ReadyPlayerMe, https://readyplayer.me. 1

[7] Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, and Linjie Luo. Panohead: Geometry-aware 3d full-head synthesis in 360deg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 8

[8] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20364–20373, 2022. 2

[9] Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, and Linchao Bao. Ffhq-uv: Normalized facial uv-texture dataset for 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 362–371, 2023. 1

[10] Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, Rohit Pandey, Ping Tan, Thabo Beeler, Sean Fanello, and Yinda Zhang. Learning personalized high quality volumetric head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[11] Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. FLARE: Fast learning of animatable and relightable mesh avatars. *ACM Transactions on Graphics*, 42:15, 2023. 2

[12] Ananta R Bhattarai, Matthias Nießner, and Artem Sevastopolsky. Triplanenet: An encoder for eg3d inversion. *arXiv preprint arXiv:2303.13497*, 2023. 3

[13] Sai Bi, Stephen Lombardi, Shunsuke Saito, Tomas Simon, Shih-En Wei, Kevyn McPhail, Ravi Ramamoorthi, Yaser Sheikh, and Jason M. Saragih. Deep relightable appearance models for animatable faces. *ACM Transactions on Graphics (TOG)*, 40, 2021. 2

[14] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, page 187–194, USA, 1999. ACM Press/Addison-Wesley Publishing Co. 1

[15] Volker Blanz and Thomas Vetter. *A Morphable Model For The Synthesis Of 3D Faces*. Association for Computing Machinery, New York, NY, USA, 1 edition, 2023. 2

[16] 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 *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017. 7

[17] Egor Burkov, Igor Pasechnik, Artur Grigorev, and Victor Lempitsky. Neural head reenactment with latent pose descriptors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13786–13795, 2020. 2

[18] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41, 2022. 2

[19] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, 2021. 3

[20] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 3, 4, 5

[21] Chuhan Chen, Matthew O'Toole, Gaurav Bharaj, and Pablo Garrido. Implicit neural head synthesis via controllable local deformation fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 416–426, 2023. 2

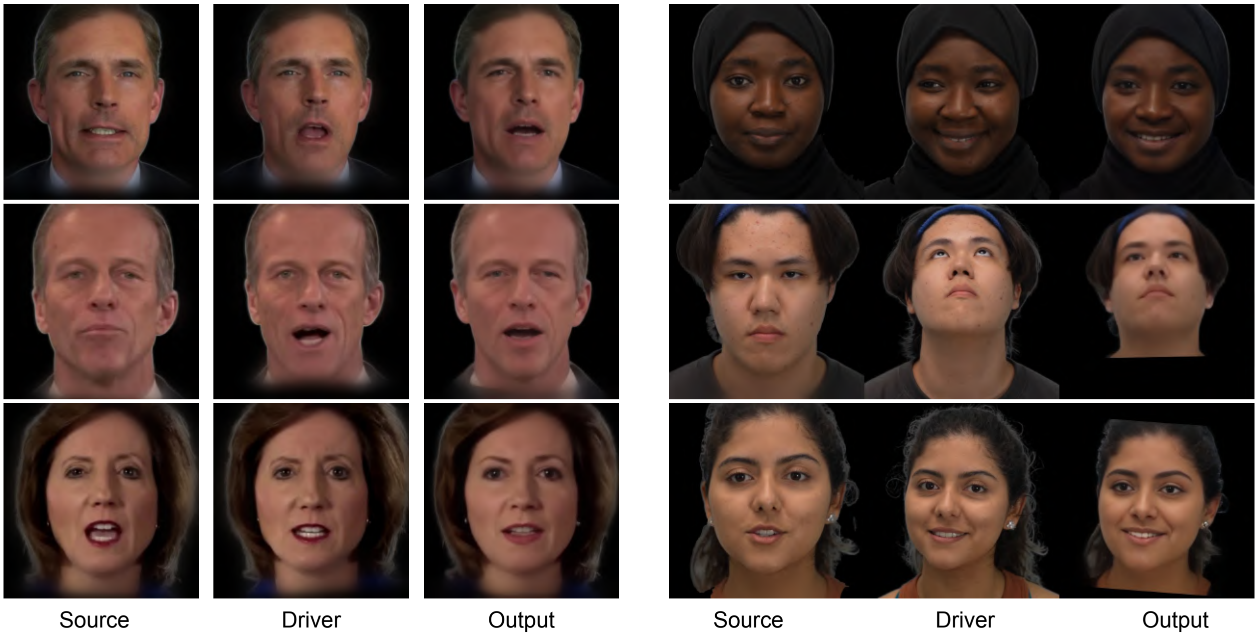

Figure 10. Qualitative results of our method on self-reenactment

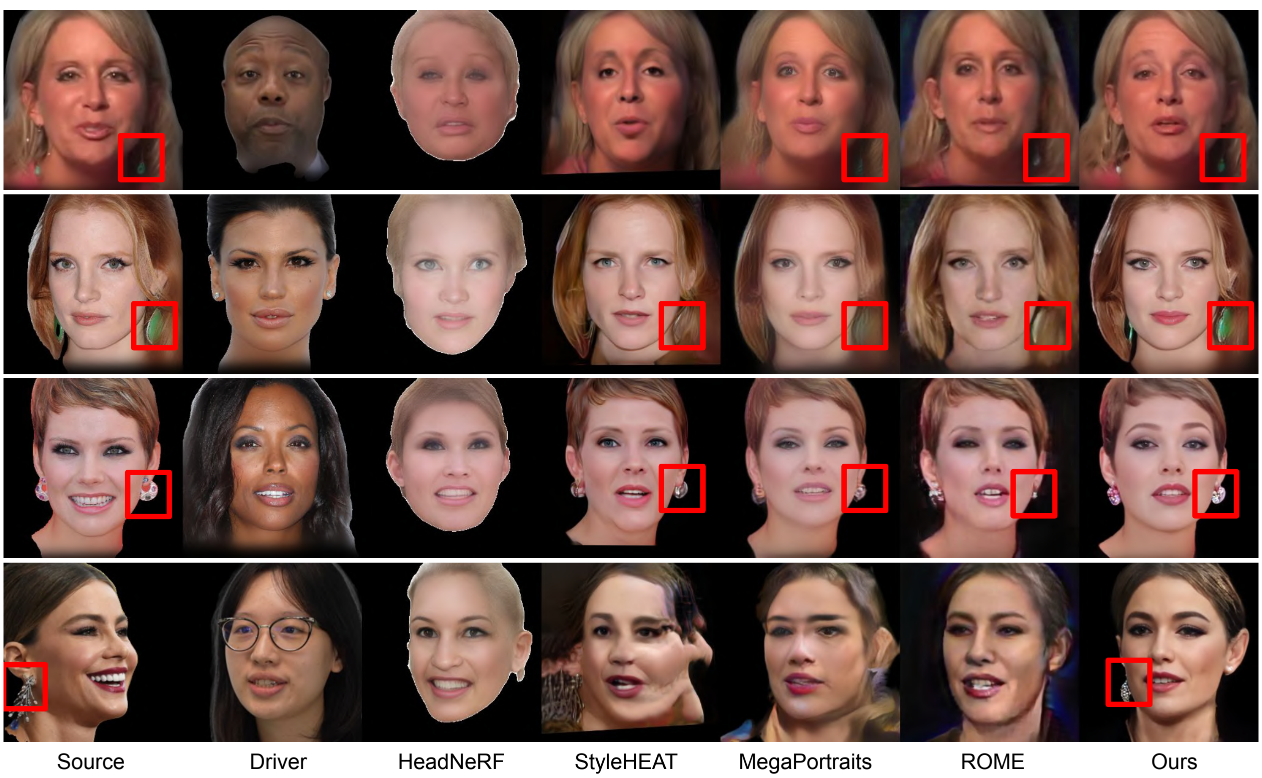

Figure 11. Our method faithfully retains the jewelries from the source image

[22] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 1

[23] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–

Figure 12. Additional Limitations: our method cannot handle the driver's tongue and sometimes produces wrong accessories that are out-of-domain, such as exotic sunglasses. Also, our head pose uses a single rigid transformation instead of a multi-joint body rig, which leads to the shoulders always moving together with the head pose.

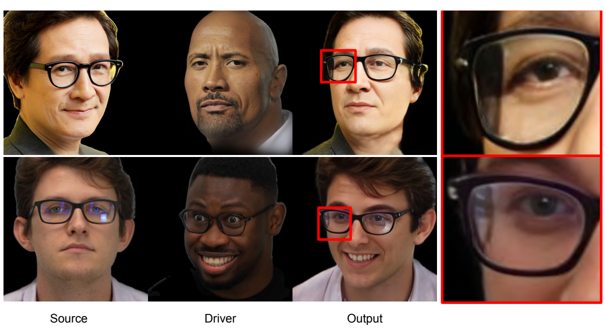

Figure 13. Our method can handle glass's refraction

20322, 2022. 7, 1

[24] 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. 4

[25] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10673–10683, 2022. 3

[26] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 2

Figure 14. Qualitative comparisons with LPR [55] on HDTF dataset.

[27] Michail Christos Doukas, Stefanos Zafeiriou, and Viktoriia Sharmanska. Headgan: One-shot neural head synthesis and editing. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022. 2, 3, 5, 7

[29] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 3

[30] Yonggan Fu, Yuecheng Li, Chenghui Li, Jason Saragih, Peizhao Zhang, Xiaoliang Dai, and Yingyan (Celine) Lin. Auto-card: Efficient and robust codec avatar driving for real-time mobile telepresence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21036–21045, 2023. 2

[31] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[32] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d

Figure 15. Novel view synthesis comparison with LPR. In this example, LPR fails to remove the smiling expression from the source while our method successfully transfer the expression from the driver to the source due to better disentanglement.

facial avatar reconstruction. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2021. 2

[33] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5609–5619, 2023. 2

[34] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos P Zafeiriou. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[35] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18653–18664, 2022. 2

[36] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. 2022. 2

[37] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 2, 3, 7

[38] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Trans. Graph.*, 36(6), 2017. 1

[39] Yiyu huang, Hao Zhu, Xusen Sun, and Xun Cao. Mofanerf: Morphable facial neural radiance field. In *ECCV*, 2022. 3

[40] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, 2022. 2

[41] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 7, 2

[42] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[43] 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 (CVPR)*, 2020. 3

[44] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference of Computer vision (ECCV)*, 2022. 2, 3, 7

[45] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Péerez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics 2018 (TOG)*, 2018. 2

[46] Jaehoon Ko, Kyusun Cho, Daewon Choi, Kwangrok Ryoo, and Seungryong Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023. 3

[47] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Abhijeet Ghosh, and Stefanos P Zafeiriou. Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[48] Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Fitme: Deep photorealistic 3d morphable model avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8629–8640, 2023. 1

[49] Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project starline: A high-fidelity telepresence system. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6), 2021. 2

[50] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 394–403, 2023. 1

[51] Hao Li, Laura Trutoiu, Kyle Olszewski, Lingyu Wei, Tristan Trutna, Pei-Lun Hsieh, Aaron Nicholls, and Chongyang Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2015)*, 34(4), 2015. 2

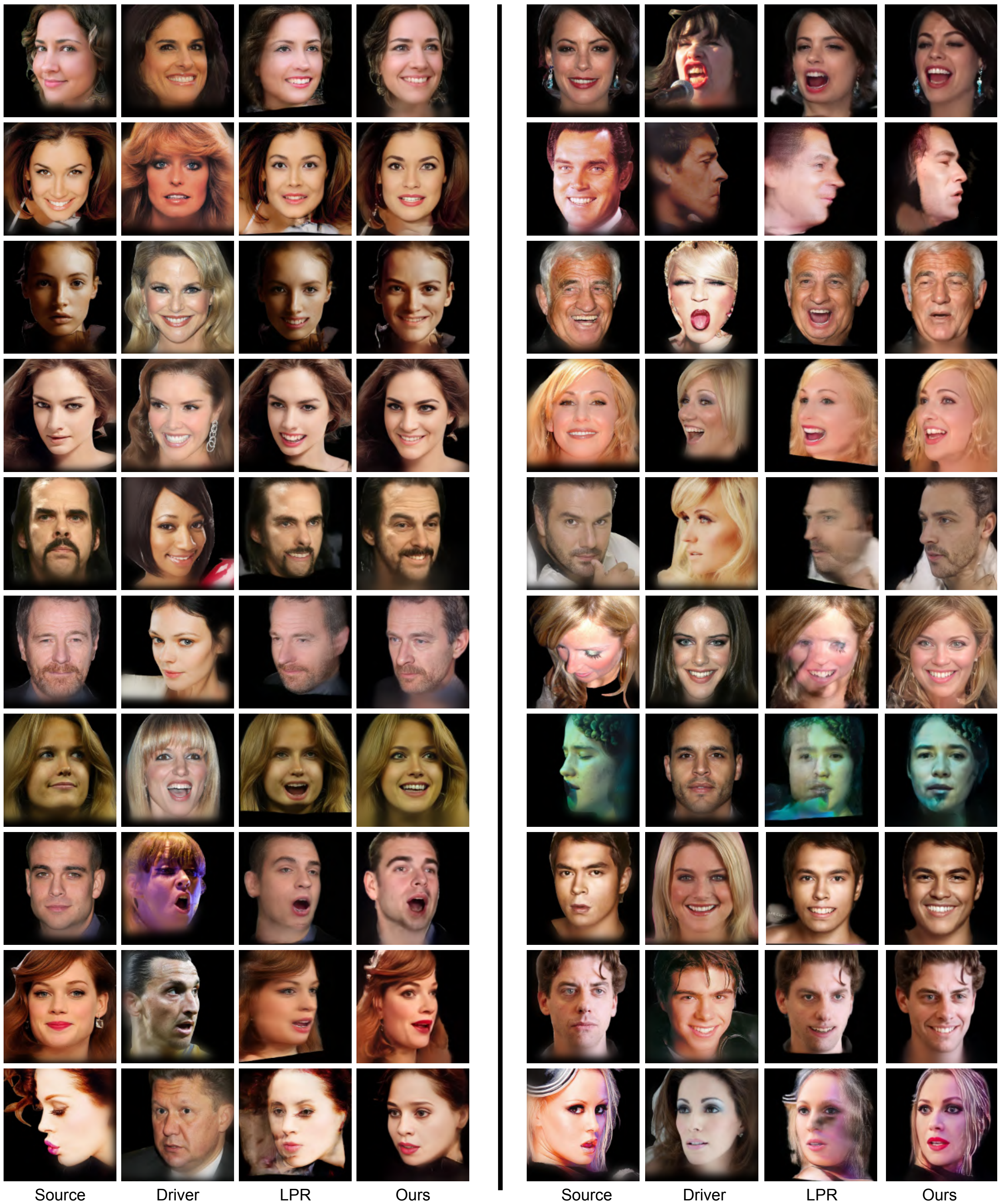

Figure 16. Qualitative comparisons with LPR [55] on CelebA-HQ dataset.

[52] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3

[53] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and ex-

Figure 17. Synthesizing novel views using our method.

pression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 7

[54] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17969–17978, 2023. 2, 3

[55] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot neural head avatar. *arXiv preprint arXiv:2306.08768*, 2023. 2, 3, 4, 5, 7, 8

[56] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv*

Figure 18. Qualitative results on various datasets.

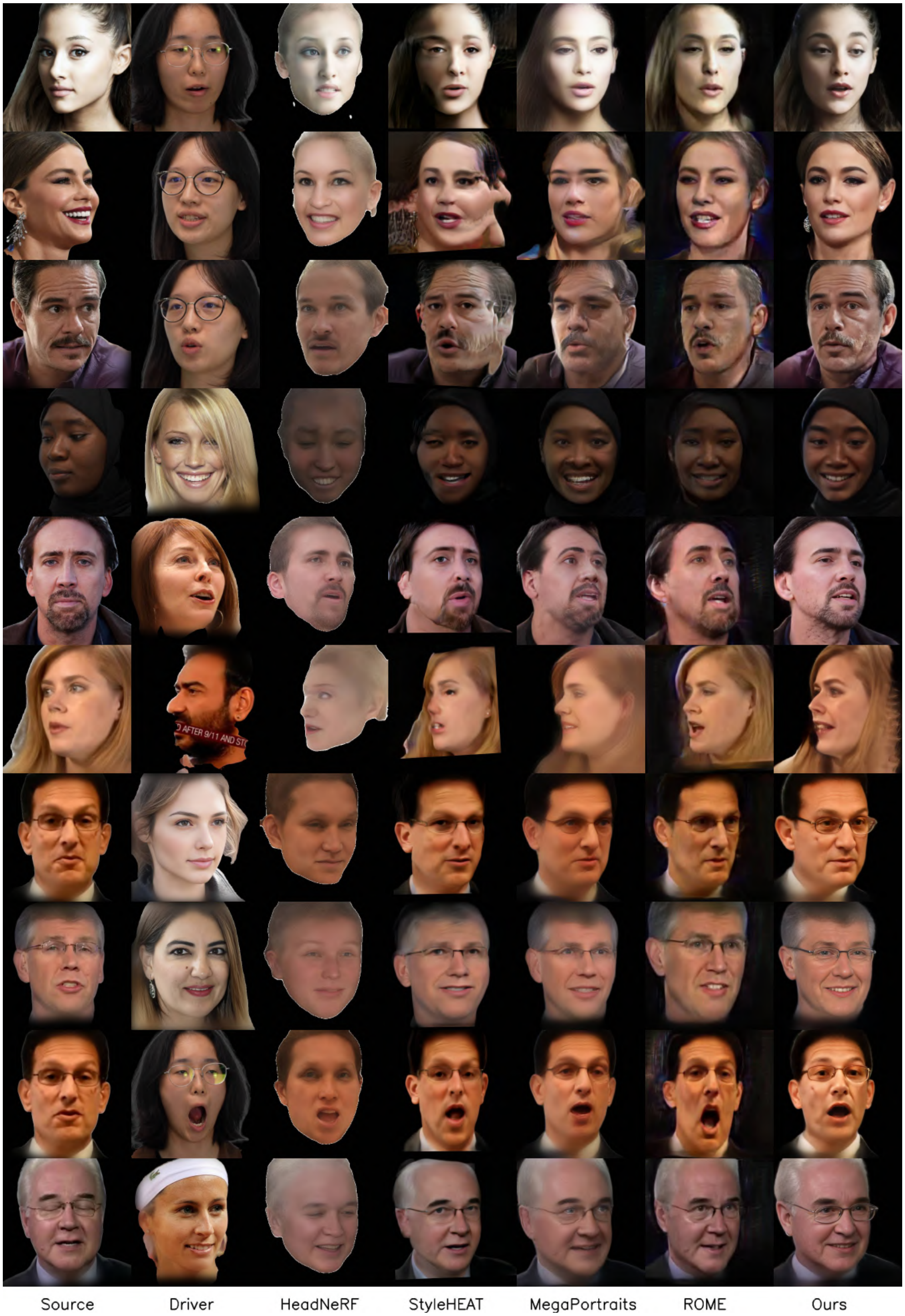

Figure 19. Qualitative results on various datasets.

Figure 20. Qualitative results on various datasets.

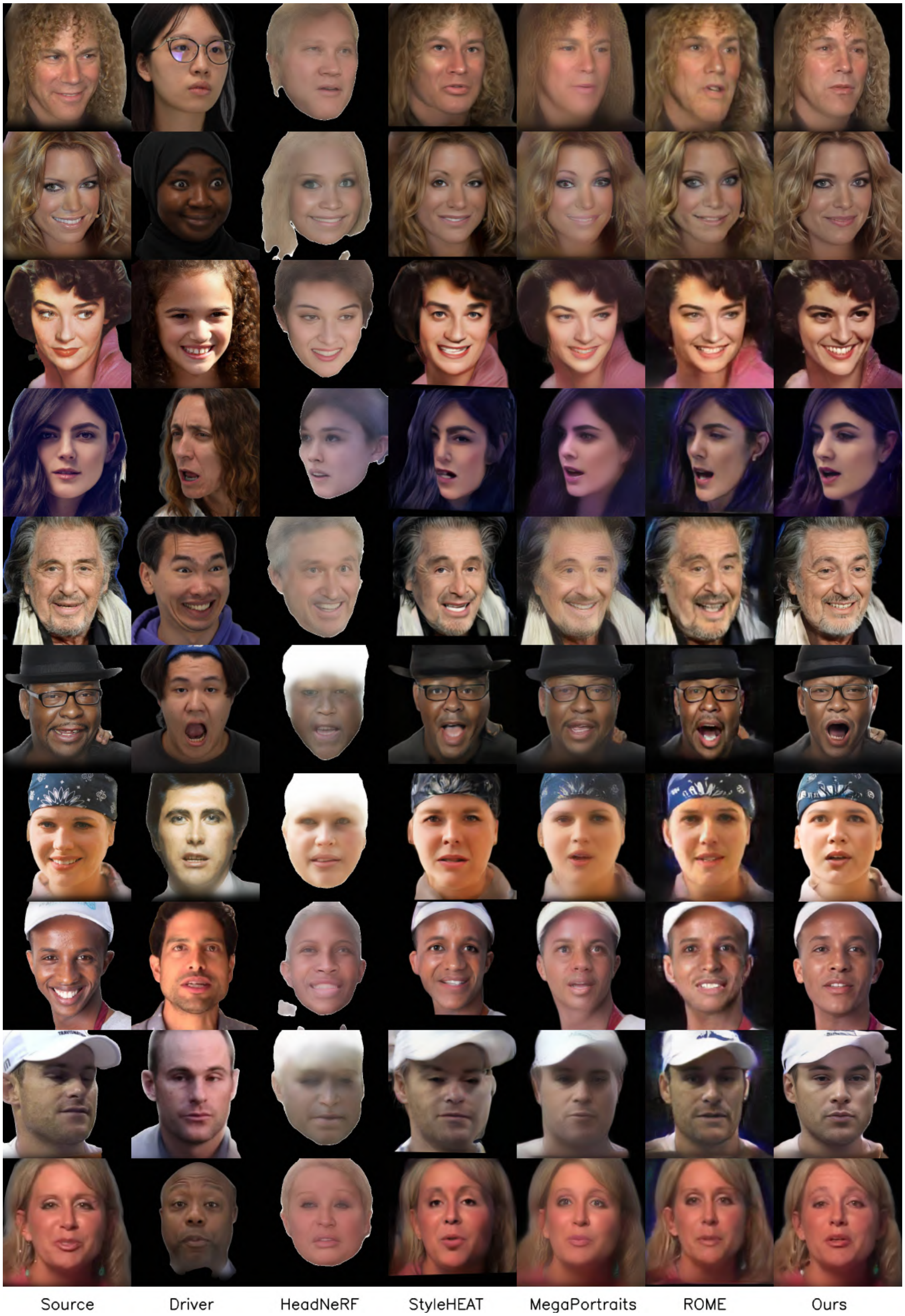

Figure 21. Qualitative results on various datasets.

Figure 22. Qualitative results on various datasets.

Figure 23. Qualitative results on various datasets.

Figure 24. Qualitative results on various datasets.

Figure 25. Qualitative results on various datasets.

Figure 26. Qualitative results on various datasets.

Figure 27. Qualitative results on various datasets.

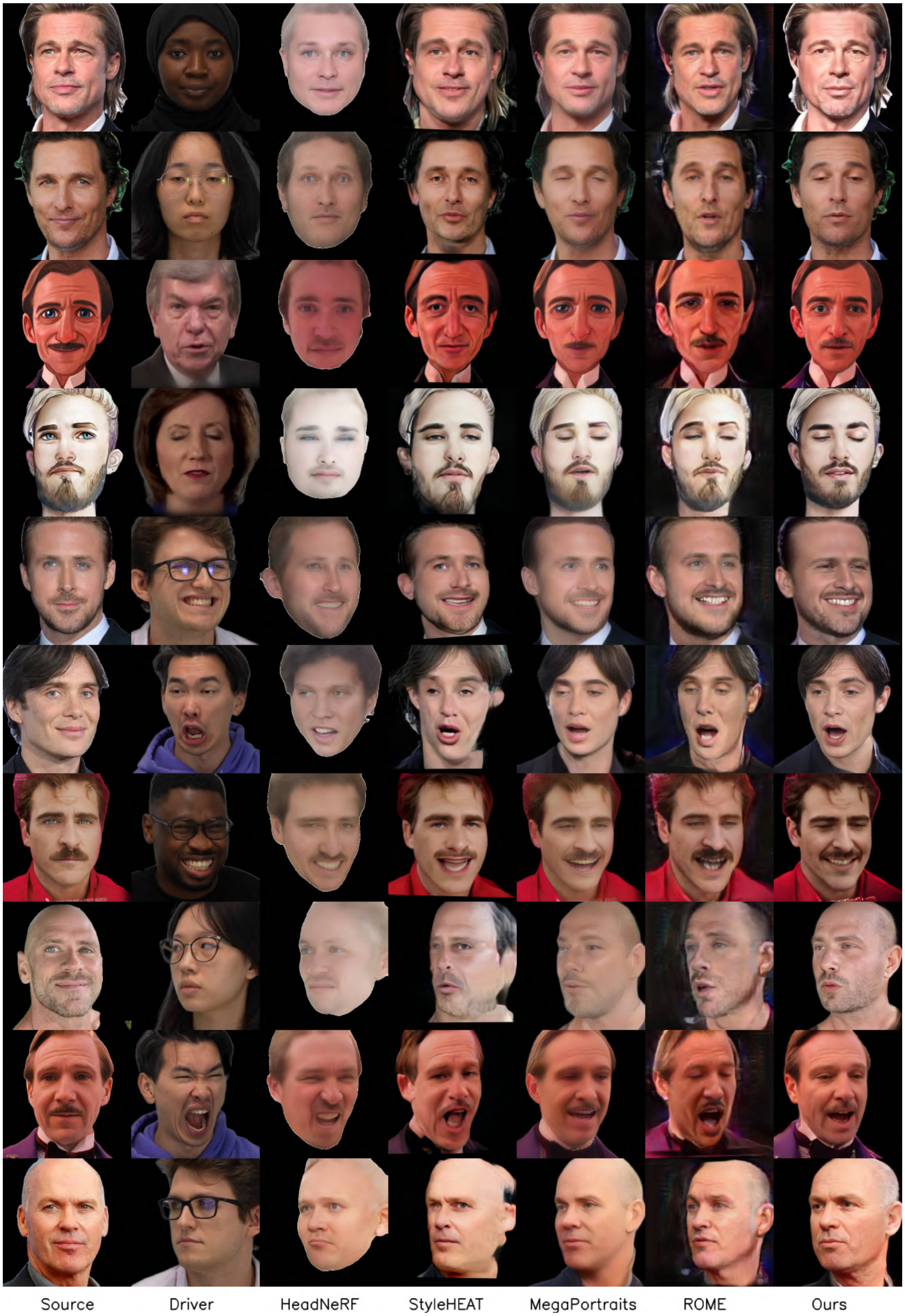

Figure 28. Qualitative results on various datasets.

Figure 29. Qualitative results on various datasets.

Figure 30. Qualitative results on various datasets.

Figure 31. Qualitative results on various datasets.

preprint arXiv:1705.02894, 2017. 2

[57] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4), 2018. 2

[58] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[59] Huiwen Luo, Koki Nagano, Han-Wei Kung, Mclean Goldwhite, Qingguo Xu, Zejian Wang, Lingyu Wei, Liwen Hu, and Hao Li. Normalized avatar synthesis using stylegan and perceptual refinement. *CoRR*, abs/2106.11423, 2021. 1

[60] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73, 2021. 2

[61] Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, and Lei Zhang. Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16910, 2023. 2, 3, 7

[62] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 4

[63] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[64] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature

fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11453–11464, 2021. 3

[65] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 2

[66] Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG)*, 35:1 – 14, 2016. 2

[67] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, 2022. 3

[68] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13759–13768, 2021. 2

[69] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[70] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 2, 3

[71] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 2

[72] Gabriel Schwartz, Shih-En Wei, Te-Li Wang, Stephen Lombardi, Tomas Simon, Jason Saragih, and Yaser Sheikh. The eyes have it: An integrated eye and face model for photorealistic facial animation. *ACM Trans. Graph.*, 39(4), 2020. 2

[73] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020. 3

[74] Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. Voxgraf: Fast 3d-aware image synthesis with sparse voxel grids. *Advances in Neural Information Processing Systems*, 35:33999–34011, 2022. 3

[75] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *CVPR*, 2019. 2

[76] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[77] Aliaksandr Siarohin, Oliver Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 2

[78] Ivan Skorokhodov, Sergey Tulyakov, Yiqun Wang, and Peter Wonka. EpiGRAF: Rethinking training of 3d GANs. In *Advances in Neural Information Processing Systems*, 2022. 3

[79] Linsen Song, Wayne Wu, Chaoyou Fu, Chen Qian, Chen Change Loy, and Ran He. Pareidolia face reenactment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[80] Jiale Tao, Biao Wang, Borun Xu, Tiezheng Ge, Yuning Jiang, Wen Li, and Lixin Duan. Structure-aware motion transfer with deformable anchor model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3637–3646, 2022. 2

[81] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2016. 2

[82] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Headon: Real-time reenactment of human portrait videos. *ACM Transactions on Graphics 2018 (TOG)*, 2018. 2

[83] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40 (4):1–14, 2021. 3

[84] Alex Trevithick, Matthew Chan, Michael Stengel, Eric Chan, Chao Liu, Zhiding Yu, Sameh Khamis, Manmohan Chandraker, Ravi Ramamoorthi, and Koki Nagano. Real-time radiance fields for single-image portrait view synthesis. *ACM Transactions on Graphics (TOG)*, 42(4):1–15, 2023. 2, 3, 4, 7, 1

[85] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition, 2018. 5

[86] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 2

[87] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 5

[88] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021. 4, 1

[89] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022. 2

[90] O. Wiles, A.S. Koepke, and A. Zisserman. X2face: A network for controlling face generation by using images, audio, and pose codes. In *European Conference on Computer Vision*, 2018. 2

[91] Jianfeng Xiang, Jiaolong Yang, Yu Deng, and Xin Tong. Gram-hd: 3d-consistent image generation at high resolution with generative radiance manifolds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2195–2205, 2023. 3

[92] Sitao Xiang, Yuming Gu, Pengda Xiang, Mingming He, Koki Nagano, Haiwei Chen, and Hao Li. One-shot identity-preserving portrait reenactment. *arXiv preprint arXiv:2004.12452*, 2020. 2

[93] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and

efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 1

[94] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023. 3

[95] Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, and Linjie Luo. Omniavatar: Geometry-guided controllable 3d head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12814–12824, 2023. 3

[96] Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Han Huang, Guojun Qi, and Yebin Liu. Latentavatar: Learning latent expression code for expressive neural head avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*. Association for Computing Machinery, 2023. 2

[97] Zhongcong Xu, Jianfeng Zhang, Junhao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *The Tenth International Conference on Learning Representations*, 2023. 3

[98] Yang Xue, Yuheng Li, Krishna Kumar Singh, and Yong Jae Lee. Giraffe hd: A high-resolution 3d-aware generative model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18440–18449, 2022. 3

[99] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. 2022. 2

[100] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *ECCV*, 2022. 2, 7

[101] Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, and Yun Fu. Nerfinvertor: High fidelity nerf-gan inversion for single-shot real image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8539–8548, 2023. 3

[102] Wangbo Yu, Yanbo Fan, Yong Zhang, Xuan Wang, Fei Yin, Yunpeng Bai, Yan-Pei Cao, Ying Shan, Yang Wu, Zhongqian Sun, and Baoyuan Wu. Nofa: Nerf-based one-shot facial avatar reconstruction. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 2, 3

[103] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. *arXiv preprint arXiv:2303.12326*, 2023. 3

[104] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 2, 7

[105] Egor Zakharov, Aleksei Ivakhnenko, Aliaksandra Shysheya, and Victor Lempitsky. Fast bi-layer neural synthesis of one-shot realistic head avatars. In *European Conference on Computer Vision*, pages 524–540. Springer, 2020. 2

[106] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22096–22105, 2023. 2

[107] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Fdnerf: Few-shot dynamic neural radiance fields for face reconstruction and expression editing. *arXiv preprint arXiv:2208.05751*, 2022. 2

[108] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[109] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 7, 2

[110] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3657–3666, 2022. 2

[111] Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. Havatar: High-fidelity head avatar via facial model conditioned neural radiance field. *ACM Trans. Graph.*, 2023. 2

[112] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I m avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13545–13555, 2022.

[113] Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. Pointavatar: Deformable point-based head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[114] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pages 650–667. Springer, 2022. 5

[115] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022. 2, 5, 7, 1

[116] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2