Supplementary Material for
FaceCraft4D: Animated 3D Facial Avatar Generation from a Single Image

Fei Yin[1,2], Mallikarjun B R[2], Chun-Han Yao[2], Rafał Mantiuk[1†], Varun Jampani[2†]
[1]University of Cambridge, [2]Stability AI

1. Implementation Details

1.1. Data Pre-processing

Following [1], we crop head regions for GAN inversion. Specifically, we use dlib [5] to detect 68 facial landmarks. The landmarks are then aligned to ensure the face is centered in the image. To isolate the face, we apply matting to remove the background, replacing it with a white color. For SV3D [11], which was trained on general objects, its training set typically centers objects in the image. To adapt to this domain, we add extra white padding around the face, ensuring better alignment with the model's original training conditions.

1.2. Multiview Image Generation

We set the DDIM inversion steps to $T = 25$ and the image strength to 0.4, which corresponds to adding noise at timestep 10. Then, we use a diffusion model to denoise the added noise feature and apply cross-view mutual attention through the remaining steps.

In the warping-based control generation module, we employ a copy-pasting operation to blend the warping features with the generation features. A mask is extracted during the warping process to identify regions in the novel view that correspond to reference textures in the original view. This mask is then downsampled to match the feature size. To handle boundary artifacts, we set the boundary values of the mask to 0.1. This adjustment mitigates the effects of disconnections or distortion artifacts commonly observed at the edges of warping features. By using a non-binary mask, we enable a smooth transition between the warping texture and the inpainted regions, effectively blending generative features with the warping values.

1.3. Gaussian Training

We choose GaussianAvatar [7] as our consistent Gaussian representation, which binds Gaussian features to a FLAME [6] template mesh. This approach allows animating the avatar with controllable FLAME expression and pose parameters.

We utilize the 2023 version of FLAME [6], which includes revised eye regions. Additionally, following [7], we manually add 168 triangles to represent the teeth in the FLAME template mesh. The upper and lower teeth triangles are rigidly attached to the neck and jaw joints, respectively, improving the avatar's fidelity.

For FLAME tracking, we use the tracking code from [7]. The optimization process includes per-frame parameters (translation, joint poses, and expression) and shared parameters (shape, vertex offset, and an albedo map). The optimization combines a landmark loss, a color loss, and regularization terms. We optimize all the parameters on the first time step of the video sequence until convergence, then optimize per-frame parameters for 50 iterations for each following time step with the previous one as initialization. Afterward, we conduct global optimization for 30 epochs by randomly sampling time steps to fine-tune all parameters. For more details, please refer to [7].

Table 1. Quantitative comparison on static 3D head generation on MEAD [12].

Method	PSNR ↑	SSIM ↑	LPIPS ↓	ID ↑
Portrait3D [13]	9.51	0.5622	0.5275	0.2646
SV3D [11]	13.54	0.6893	0.3174	0.7622
PanoHead [1]	15.37	0.7342	0.2104	0.7981
Ours	**15.74**	**0.7495**	**0.2010**	**0.8006**

2. Additional Experiments

2.1. Comparison with Controllable 3D GANs

We compare our method with controllable 3D GANs: 3DFaceShop [10] and Next3D [9]. For a fair comparison, since controllable 3D GANs do not develop proper GAN inversion methods, we randomly sampled images from 3D GANs and used them as identity source images, then used other videos as expression signals. The qualitative results are shown in Fig. 1, demonstrating that our method achieves

[†]Equal advising.
*No biometric data was used to train, validate, or evaluate the model described in this work.

Figure 1. Qualitative comparison with controllable 3D GANs.

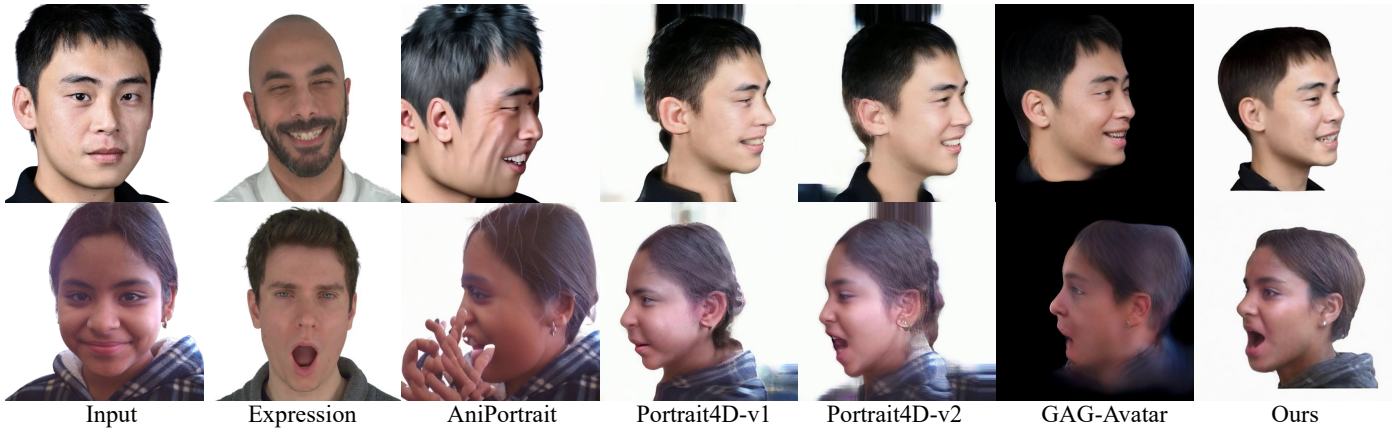

Figure 2. Qualitative comparison on animation of novel views.

more accurate expressions from the driving images.

2.2. Extended Animation Comparisons.

We incorporate additional baselines like GAGAvatar [2] and Portrait4D-v1 [4] as well as metrics such as AED [3], APD [3], IQA [8]. Note that the reported AED/APD are computed on near-frontal frames to avoid 3DMM tracking failures under extreme poses. Fig. 2 and Tab. 2 show that our method achieves the *SotA* performance.

2.3. Evaluation on Multiview Dataset

To get a comprehensive understanding of the performance of our method, we evaluate on MEAD [12], a multi-view dataset. The quantitative comparison between the reconstruction portraits and the ground truth is shown in Tab. 1. The qualitative results are shown in Fig. 4 and Fig. 5. The results demonstrate that our method can generate a consistent geometry and texture in novel views, which aligns with the conclusion of the manuscript.

2.4. Portrait3D Implementation Clarification

Portrait3D is originally designed for *text-to-3D*. For adaptation to *image-to-3D*, we incorporated an image $\mathcal{L}_2$ loss and mask $\mathcal{L}_1$ loss. We conducted ablation studies by replacing head-only inputs with properly cropped upper-body images using FFHQ in-the-wild data and Portrait3D's preprocessing code. The metrics showed marginal improvements, as demonstrated in Fig. 3 and Tab. 3. Despite adhering to the default 20-epoch optimization and testing extended 40-epoch runs, geometric/textural artifacts persist.

We attribute this to a *domain mismatch*: The SDS loss struggles with non-celebrity image inputs due to misalignment between visual data and text-aligned priors in the diffusion model, causing optimization conflicts. This differs from text-to-3D cases, where optimization is only guided by semantically coherent prompts.

2.5. Video Results

To better demonstrate the performance and robustness of our method, we provide a demo webpage in the supplementary.

2.6. Limitations Discussion

PanoHead-driven PTI inversion can occasionally produce Janus-like artifacts (degraded back-head texture or another face shown on the back of the head), which then propagate to the downstream framework. This failure mode likely stems from the optimized latent codes deviating outside the

Figure 3. Ablation study on Portrait3D adaptation.

Table 2. Quantitative comparison on animation of novel views.

Method	CLIP-I ↑	ID ↑	FID ↓	AED ↓	APD ↓	HyperIQA ↑
AniPortrait	0.4653	0.4171	364.99	1.461	0.119	56.69
Portrait4D-v1	0.5060	0.4295	291.21	1.363	0.097	59.69
Portrait4D-v2	0.5236	0.4592	248.36	1.291	0.121	61.96
GAG-Avatar	0.5588	0.4524	399.62	1.076	**0.073**	59.74
Ours	**0.5737**	**0.4602**	**201.76**	**0.930**	0.085	**63.59**

Table 3. Ablation study on Portrait3D adaptation.

Method	CLIP-I ↑	ID ↑	FID ↓
Portrait3D (Head)	0.7066	0.3719	**302.74**
Portrait3D (Upper body)	**0.7152**	**0.3932**	315.82

GAN's original training manifold. Additionally, since our approach involves multiple modules, the entire process requires approximately 2 hours, making computational efficiency a clear limitation. Future work may explore feed-forward models to streamline the entire avatar generation process.

References

[1] 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*, pages 20950–20959, 2023. 1

[2] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[3] 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. 2

[4] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[5] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 1

[6] 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), 2017. 1

[7] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20299–20309, 2024. 1

[8] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *CVPR*, 2023. 1

[10] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3dfaceshop: Explicitly controllable 3d-aware portrait generation. *IEEE transactions on visualization and computer graphics*, 30(9):6020–6037, 2023. 1

[11] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 1

[12] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020. 1, 2

[13] Yiqian Wu, Hao Xu, Xiangjun Tang, Xien Chen, Siyu Tang, Zhebin Zhang, Chen Li, and Xiaogang Jin. Portrait3d: Text-guided high-quality 3d portrait generation using pyramid representation and gans prior. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 1

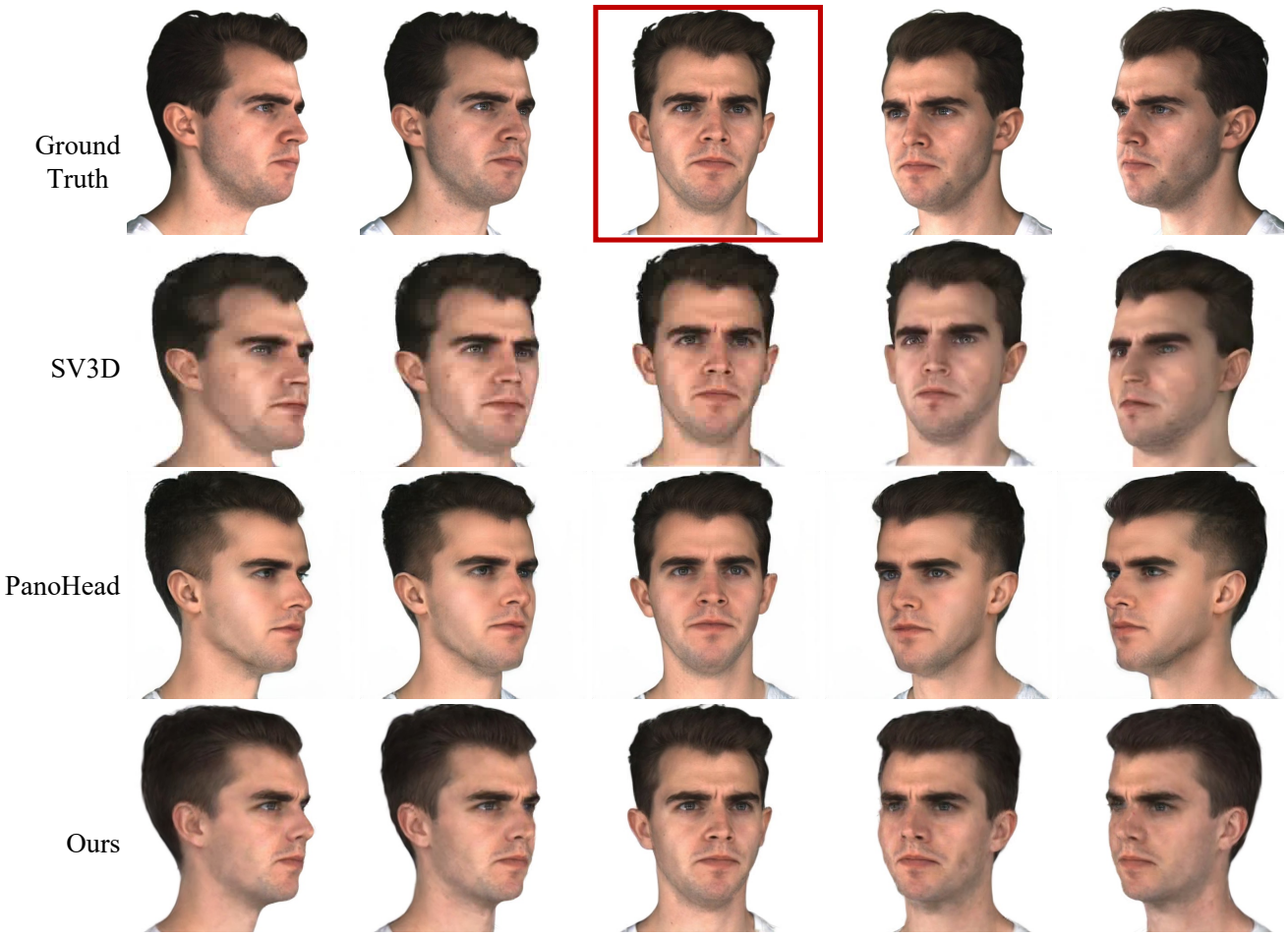

Figure 4. Qualitative comparison on static 3D head generation from a single image on MEAD (1). Red box indicates the input image.

Figure 5. Qualitative comparison on static 3D head generation from a single image on MEAD (2). Red box indicates the input image.