Fine-Grained 3D Gaussian Head Avatars Modeling from Static Captures via Joint Reconstruction and Registration

Supplementary Material

Yuan Sun[1,2] Xuan Wang[2] Cong Wang[3] WeiLi Zhang[1] Yanbo Fan[4] Yu Guo[1,*] Fei Wang[1]
[1]Xi'an Jiaotong University [2]Ant Group [3]Tsinghua University [4]Nanjing University

A. More Details

Merging strategy. The FLAME [3] team provides a predefined template that specifies the facial region associated with each mesh triangle. Building on this, we further customized distinct facial regions, such as the teeth and inner lips. Each Gaussian primitive is rigged to a triangle, enabling easy identification of splats within specific regions using the binding information and template indices. To obtain the complete result, the invisible regions of the prior-based Gaussian are merged with the visible regions of the prior-free Gaussian.

Teeth enhancement details. Given a specific expression (with the mouth fully open), we enhance the details of the inner mouth region by distilling the outputs of HeadGAP's[8] 2D enhancement network across multiple virtual viewpoints. These outputs are used as pseudo labels to supervise the model through L1 loss and SSIM loss.

In-the-wild data processing. BiRefNet[7] was utilized for matting, while VHAP[4, 5] was employed for FLAME fitting. Although VHAP is designed for monocular or multi-view video sequences, it can still achieve accurate fitting for static captures by relying solely on landmark matching loss and photometric losses.

Input	Method	SSIM ↑	PSNR ↑	LPIPS ↓
1 image	GAGAvatar	0.833	17.52	0.241
15 images	FlashAvatar	0.837	20.43	0.213
	HeadGAP	0.886	23.25	0.147
	Ours	0.890	25.03	0.113

Table A. Quantitative comparison across 30 subjects.

Figure A. Self-reenactment results.

B. Additional Results and Visualization

We have carried out additional experiments under 802×550 resolution conditions on 30 subjects, comparing with the representative Gaussian-based methods GAGAvatar[2] and FlashAvatar[6]. The average metrics are summarized in Table A, and the visualization is provided in Figure A. As GAGAvatar supports only single-view input, we used the front view as the input; the results are thus only for reference. The results demonstrate that our method achieves superior performance in terms of identity preservation, expression fidelity, and the level of detail representation.

C. Additional Discussion

Effect of driving signal noise. Our method significantly enhances the visual quality of details. However, mismatches between the images and the FLAME mesh in the test data limit the capability of PSNR and SSIM metrics to fully capture this improvement. As shown in Figure B, although the rendered image and the ground truth appear visually similar, noticeable misalignments caused by noise in the FLAME parameters become evident when overlaid. As noted in [1], while PSNR and SSIM are highly sensitive to pixel misalignments, LPIPS demonstrates greater robustness by measuring differences in deep feature representations.

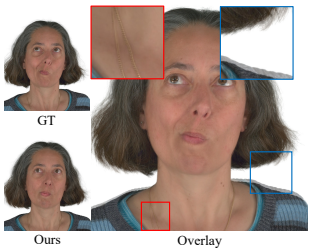

Figure B. Misaligned.

The impact of registration-aware regularization terms. The regularization terms that we designed specifically address irregularities in dynamic details. These anomalies exhibit distinct characteristics: they are exclusively associated with specific expressions, become noticeable only from certain viewpoints, and occupy a very small region of the image. This leads to slight metric variations in Table 2 which record the average metrics. However, Figure 5 in the main paper offers a clearer and visual demonstration of how the regularization terms effectively mitigate these dynamic anomalies, thereby improving overall quality and preserving finer details across different expressions and viewpoints.

Figure C. Results with different numbers of input viewpoints.

Influence of input quantity. As discussed in the limitations section of the main manuscript and further illustrated in Figure C, our method demonstrates satisfactory performance when five or more views are provided as input. Compared to reducing the amount of input, we focus more on how to leverage higher-resolution images to achieve more detailed geometric and texture modeling in few-shot scenarios. A potential solution for further reducing the number of input views is to incorporate more advanced sparse Gaussian splatting techniques. Moreover, enabling the model to reconstruct 3D head avatars from inconsistent inputs, such as imperfect smartphone-captured images with slight unintentional motion, represents a promising direction for future research.

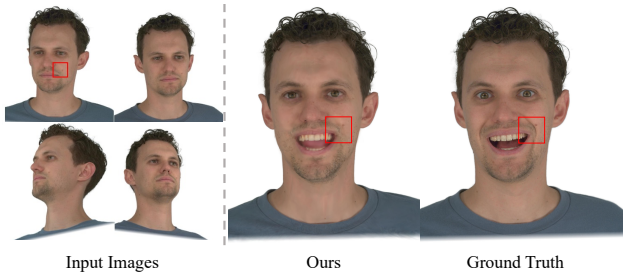

Figure D. Limitation of our method.

Failure Cases. Since the texture modeling is based on static inputs with a near-neutral expression rather than a multi-expression video sequence, our method encounters challenges in effectively capturing wrinkles, as shown in Figure D. Furthermore, due to the limitations of the Gaussian-on-mesh driving paradigm, our method is unable to handle components not modeled by FLAME, such as eyeglasses, as illustrated in Figure E.

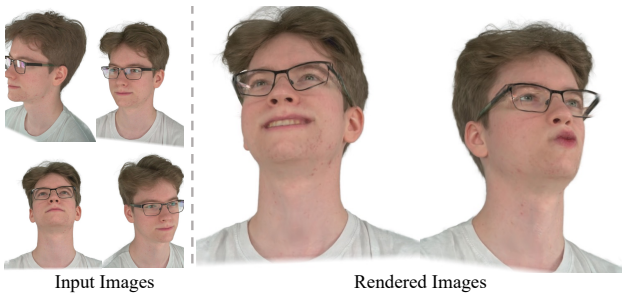

Figure E. Failure Cases.

References

[1] Jianchuan Chen, Jingchuan Hu, Gaige Wang, Zhonghua Jiang, Tiansong Zhou, Zhiwen Chen, and Chengfei Lv. Taoavatar: Real-time lifelike full-body talking avatars for augmented reality via 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10723–10734, 2025. 2

[2] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[3] 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. 1

[4] Shenhan Qian. Vhap: Versatile head alignment with adaptive appearance priors, 2024. 1

[5] 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

[6] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[7] Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. *CAAI Artificial Intelligence Research*, 3:9150038, 2024. 1

[8] Xiaozheng Zheng, Chao Wen, Zhaohu Li, Weiyi Zhang, Zhuo Su, Xu Chang, Yang Zhao, Zheng Lv, Xiaoyuan Zhang, Yongjie Zhang, et al. Headgap: Few-shot 3d head avatar via generalizable gaussian priors. In *International Conference on 3D Vision 2025*, 2025. 1