

# Feed-Forward One-Shot Animatable Textured Mesh Avatar Reconstruction

Yisheng He  
Tongyi Lab, Alibaba Group

## A. More Results

**Visualization of Reconstructed Geometry and Texture Map.** As shown in Fig. 7, our method produces high-quality geometry and texture outputs from a single input image. The reconstructed mesh faithfully captures the subject’s facial structure, hair volume, and accessory shapes, demonstrating significant deformation from the FLAME template while maintaining topological integrity. The corresponding texture map displays sharp, high-resolution details with consistent color and lighting, effectively transferring the input appearance to the UV atlas. This visualization underscores the advantage of our explicit mesh and texture representation, which jointly enables detailed geometry reconstruction and photorealistic texture synthesis in a compact and efficient form.

**Animation Results Visualization.** We provide visualization comparisons of the in-the-wild animation results from the proposed framework and previous state-of-the-art methods, LAM [3], in the supplementary videos. We can see from the videos that our mesh-based framework can better preserve identity and texture details, such as texts and tattoos on the head.

**Experiments on low-quality, in-the-wild inputs.** Fig. 8 shows our method with low-quality out-of-domain (OOD) input. Our method can generalize to both OOD AI-generated and real-world in-the-wild low-quality images.

**More baselines.** We present the qualitative comparisons with more baselines in Fig. 9, including GAGAvatar [1] and Portrait4Dv2 [2]. The quantitative results are shown in Tab. 3. Compared with previous work, our methods better reconstruct the texture details of the person.

**Reconstruction Speed.** Our method demonstrates significant computational efficiency, reconstructing a complete 3D avatar in only *0.7 seconds* from a single input image.

In comparison, the Gaussian-based LAM [3] requires 1.4 seconds for 20K Gaussians and 5.8 seconds for 80K Gaussians. This substantial speedup stems from our efficient mesh-based representation: our model operates on only 5K vertices and a compact texture token grid (4K tokens) for cross-attention, whereas Gaussian-based methods must process orders of magnitude more primitives (20K or 80K) to reconstruct details but still fail on high-frequency textures like texts and tattoos. The lightweight nature of our representation reduces both memory footprint and computational overhead, enabling fast avatar generation without sacrificing reconstruction quality.

## B. More Implementation Details

Our framework utilizes a learnable token grid  $T_0$  to extract texture features from the input image to reconstruct the texture map. We initialize  $T_0$  to a shape of  $64 \times 64$ , with  $H_t = W_t = 64$ , and decode them into a  $1024 \times 1024 \times 3$  texture map, by setting the hyperparameters  $H_a = W_a = 1024$ . The convolution operation  $\varphi$  in the iterative texture synthesis blocks,  $GRU_{\text{tex}}$  denoted in Formula 4, is implemented with 2-layer convolutions with kernel size 3. The MLP operation  $\vartheta$ , for iterative geometric deformation denoted in Formula 5, is implemented with two-layer MLPs.

## C. Ethical Impact

Generating realistic 3D avatars from single images raises ethical concerns regarding privacy, consent, and potential misuse for deceptive content. While enabling positive applications in virtual communication, our method also carries risks of unauthorized identity replication and manipulation. We stress the importance of ethical frameworks with consent mechanisms and safeguards to ensure responsible use. Continued dialogue and proactive measures are crucial to mitigate harms while preserving societal benefits.

Table 3. Quantitative comparison with more baselines.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$
Portrait4Dv2	24.650	0.835	0.076	30.137
GAGAvatar	25.020	0.864	0.066	25.014
Ours	25.233	0.879	0.061	22.699

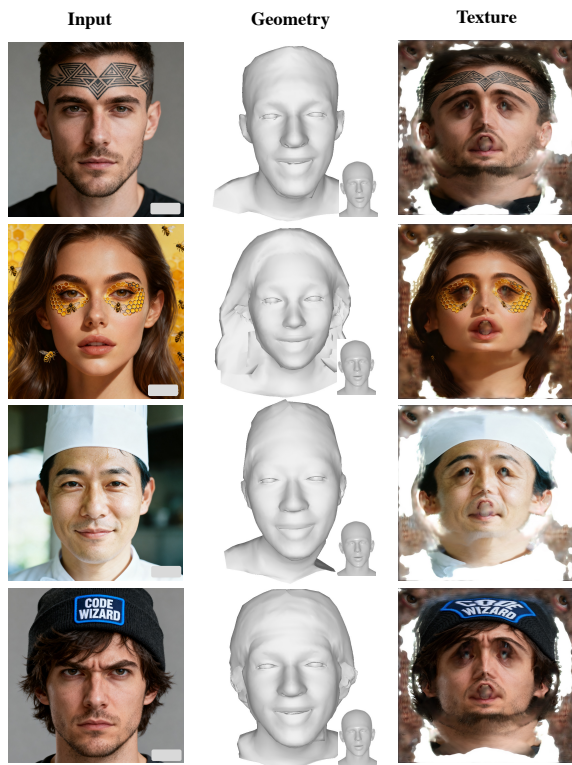


Figure 7. Reconstructed geometry and texture map visualization. Our method produces a mesh with high-fidelity texture, even when animated to an expression different from the input image. The sharp, consistently unwrapped texture details demonstrate the effectiveness of our dual-branch design.



Figure 8. In the wild challenging lighting and occlusion.

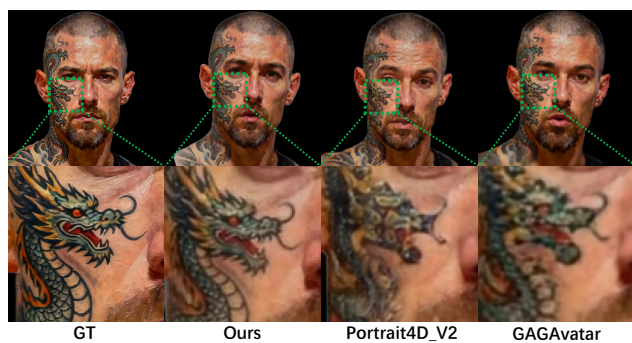


Figure 9. Qualitative comparison with more baselines.

## References

- [1] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. [1](#)
- [2] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, pages 316–333. Springer, 2024. [1](#)
- [3] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. Lam: Large avatar model for one-shot animatable gaussian head. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–13, 2025. [1](#)