From Blurry to Believable: Enhancing Low-quality Talking Heads with 3D Generative Priors

Supplementary Material

In this supplementary material, we provide additional details and results omitted in the main text.

A. Contribution and Limitations

Main contribution. While many previous works have explored super-resolution (SR) in 2D content, e.g., images, or 3D representation static representation, e.g., 3D Gaussian, super-resolution in dynamic 3D representation remains an unexplored topic. The main challenge lies in the fact that 2D SR not only struggles with multi-view but also temporal inconsistencies, when up-sampling a dynamic 3D representation. Our method addresses this challenge by performing multi-view and multi-expression 3D GAN inversion, ensuring that the synthesized 3D head preserves high-frequency details even when the up-sampled anchor images are inconsistent. To the best of our knowledge, this is the first attempt on super-resolution of dynamic 3D avatar representation.

Limitation. The major limitation is that 3D GAN cannot synthesize complete 3D head, i.e., it struggles to generate back of a human head. The main reason is that 3D GAN is trained on FFHQ [2], which consists only of frontal human faces. Building a large-scale human face dataset that includes views of back head is a possible way to extend 3D GAN's ability of synthesizing back views of human heads. As shown in Figure S1, while GSGAN [1] can synthesize frontal views of high-fidelity details, it struggles to synthesize the back of the human head.

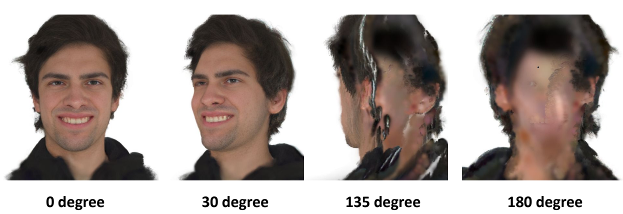

Figure S1. 3D GAN struggles to synthesize the back of a human head. We rotate a synthesized head before camera to show quality gap between views of frontal and back of a head.

B. Additional Implementation Details

We adopt GSGAN [1] as our 3D GAN backbone. To make 3D GAN robust towards side views of a 3D head and hairstyles, we processed FFHQ [2] by cropping the image source to include full head in the image. Then, we fine-tuned the GSGAN checkpoint on the re-cropped FFHQ dataset. Figure S3 shows that the fine-tuned 3D GAN can not only synthesize finer details on facial parts but also accurate hairstyles. All of our experiments, including the 3D GAN fine-tuning, were conducted on a RTX A6000 GPU.

C. Additional Results and Analyses

Additional qualitative results. We show additional visual comparison of various baselines introduced in the main paper in Figure S3. Our method demonstrates superior capability in recovering detailed facial expressions, e.g., corner of the mouth, but also accurate geometry of the hair.

Table S1. SuperHead achieves identical performance when applying to SplattingAvatar [5] on INSTA dataset [6], proving SuperHead's generalizability to enhance diverse 3D avatar models.

Setting	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SplattingAvatar (LR) [5]	19.24	0.825	0.251
SuperHead + SplattingAvatar [5]	23.04	0.834	0.167
SuperHead + GaussianAvatars [4]	23.76	0.864	0.135

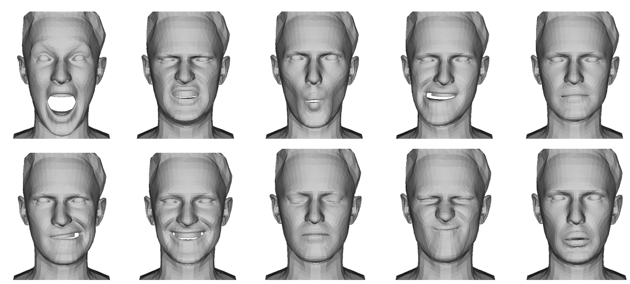
Figure S2. Expressions we used to sample anchor images.

Comparability to other 3D avatar model. We further evaluate our method on an alternative 3D avatar model to demonstrate its generalizability. Specifically, we adopt SplattingAvatar [5], which, similar to GaussianAvatars [4], rigs 3D Gaussians onto the FLAME mesh with a learnable normal offset to the surface. The upsampling procedure follows the same pipeline as with GaussianAvatars: we first sample and enhance anchor images from a SplattingAvatar trained on low-resolution captures, and then perform multi-view inversion along with dynamics-aware 3D refinement to optimize a 3D Gaussian head rigged on the underlying FLAME mesh. The results are reported in Table S1. We compare SplattingAvatar (LR), a 3D head model trained on low-resolution captures, with SuperHead + SplattingAvatar, the corresponding upsampled 3D head model. We show that our method successfully enhances low-quality 3D head

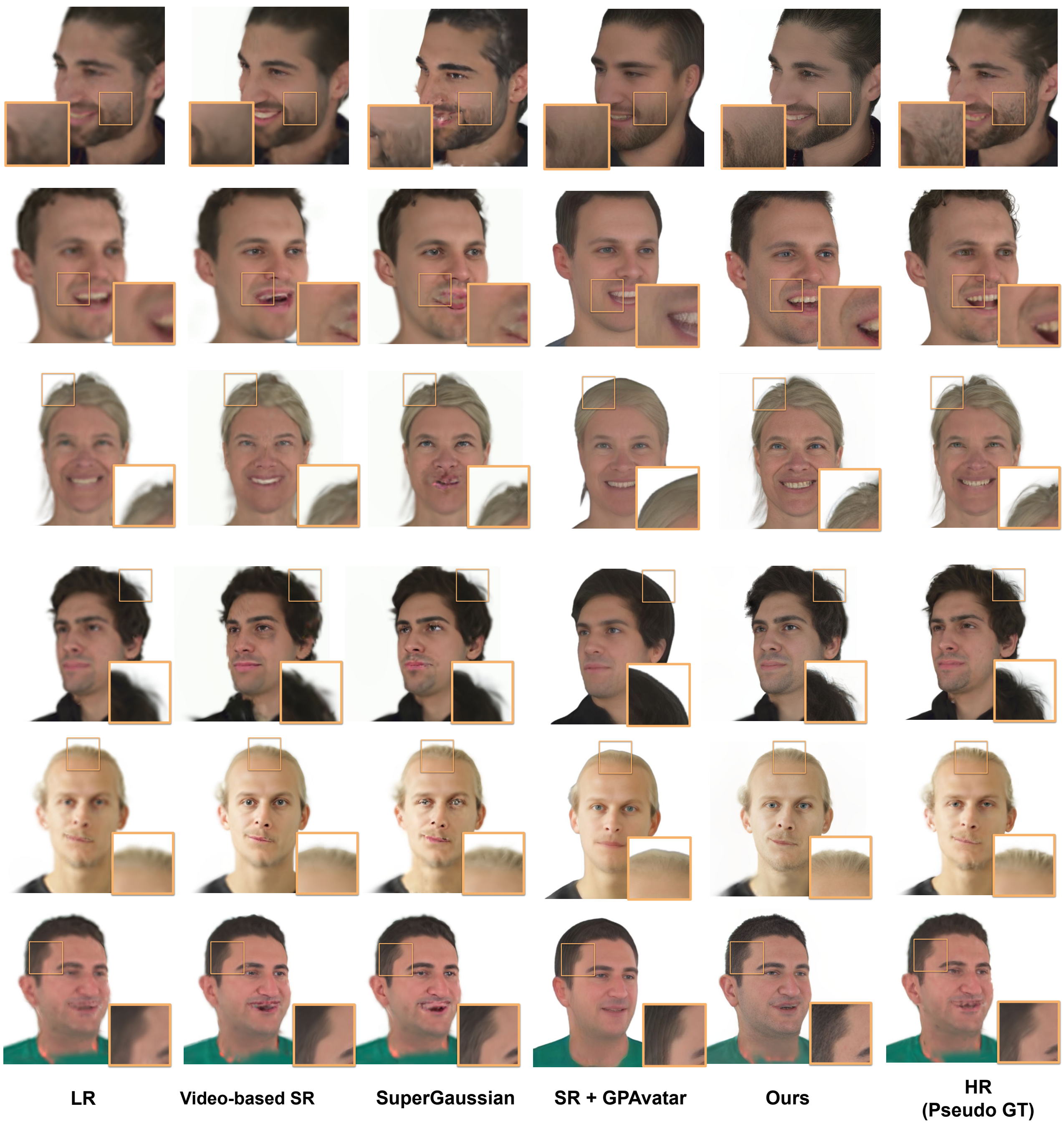

Figure S3. Additional qualitative results on the NeRSemble dataset [3] and INSTA [6]. In addition to zooming in facial parts of results, we also show the holistic view of upsampled 3D avatar, indicating that our method can not only enhance facial expressions but also details such as hair strands. Please zoom in to check details.

models across different design choices, thereby demonstrating its strong generalizability.

Anchor image sampling As mentioned in Section 4.3 of the main paper, we perform dynamics-aware 3D GAN refinement to improve the synthesized 3D head under different expressions and motions. For this purpose, we carefully select a set of expressions to form an ”expression pool”, from which we sample anchor images with different camera poses. We found that a set of 10 expressions is sufficient to achieve good performance. We show the expressions we use throughout our experiments in Figure S2. The selected expressions cover a range of facial motions, from screaming to smiling and eye-closing.

D. Ethical and Societal Impacts

Our work improves the quality of 3D head avatar reconstruction, which has potential benefits in areas such as telecommunication and digital content creation. At the same time, we are aware of possible risks, including issues of privacy, misuse for non-consensual content, and bias in representation. We emphasize the importance of developing and applying such techniques responsibly and with appropriate safeguards. Furthermore, like many generative methods, reconstruction results may contain certain biases if not carefully addressed. We believe it is important for future research and deployment of these techniques to be guided by principles of responsible AI, including fairness, transparency, and safeguards against malicious use.

References

[1] Sangeek Hyun and Jae-Pil Heo. Gsgan: Adversarial learning for hierarchical generation of 3d gaussian splats. *Advances in Neural Information Processing Systems*, 37:67987–68012, 2024. 1

[2] 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*, pages 4401–4410, 2019. 1

[3] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2

[4] 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

[5] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[6] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4574–4584, 2023. 1, 2