Synthetic Prior for Few-Shot Drivable Head Avatar Inversion – Supplemental Document –

Figure 1. Linearly interpolating $z_{\mathbf{q}_{\text{id}}}$ and $z_{\mathbf{q}_{\text{expr}}}$ between the leftmost and rightmost avatars demonstrates that our latent manifold exhibits smooth transitions in both expression and identity.

A. Appendix

This supplementary material includes additional comparisons with monocular methods such as INSTA [15], Flash Avatar (FA) [13], and Splatting Avatar [11], as well as comparisons with single-image-based reconstruction methods like PanoHead [1], MoFaNeRF [14], and HeadNeRF [4]. Additionally, we present inversions on a more diverse set of subjects, along with failure cases.

All our inversion results used only three input images (Figure 12) unless stated otherwise. Figure 7 compares monocular baseline methods trained on the entire dataset with our inversion approach. Furthermore, we provide additional examples of cross-reenactment comparisons, demonstrating the advantages of our method compared to baselines trained on only 13 frames. Next, we present results with progressively varying numbers of training frames, illustrating how this influences the quality of reconstruction. Figures 8 and 9 highlight the importance of our synthetic prior.

We include comparisons to single-image inversion methods in Figure 10, and the losses diagrams for each stage in Figure 2. We also present additional samples from our synthetic dataset in Figure 11, as well as more interpolation steps for our identity $z_{\mathbf{q}_{\text{id}}}$ and expression $z_{\mathbf{q}_{\text{expr}}}$ latent spaces, shown in Figure 1. Finally, we complement the reconstruction error evaluation with additional metrics Figure 3.

Inversion Objectives We depict the inversion optimization loss for one subject using three images as input. We show two stages of our pivotal fine-tuning: Figure 2a presents identity encoder optimization in the first stage, and Figure 2b presents the second stage, where the decoding part of our pipeline is optimized. In this particular case, the optimization took around 5 minutes on a single Nvidia H100.

Additional Results Figure 4 illustrates a challenging inversion for identities with long hair and beards, where *SynShot* successfully models these features using subjects from Preface [2] dataset. Additionally, we present the failure cases of our method, categorized into the primary scenarios where *SynShot* may fail. As shown in Figure 5:

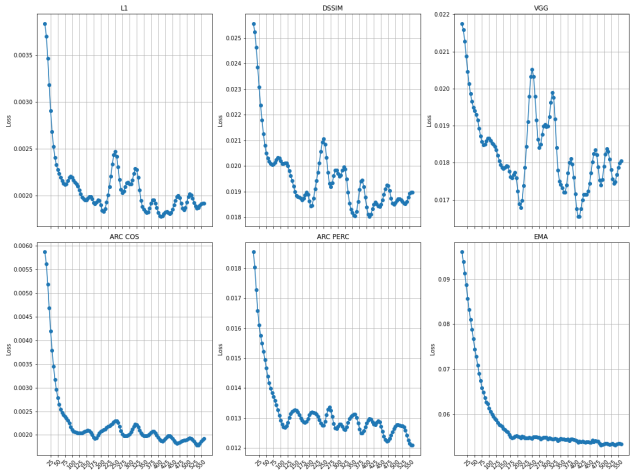
- **A)** Input images with facial accessories like glasses are not supported currently as they were not used in our synthetic dataset.
- **B)** Challenging input images, such as those with squinting eyes or closed eyes, can introduce artifacts in the final avatar due to difficulties in faithfully reproducing these details.
- **C)** Missing hairstyles in the synthetic dataset often result in errors during inversion, particularly for uncommon or complex hairstyles, further exacerbated by artifacts in hair segmentation.

B. 3D Gaussian Splatting Preliminaries

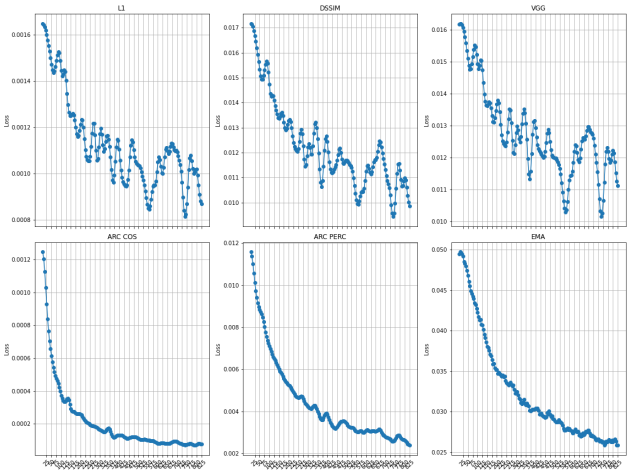
3D Gaussian Splatting (3DGS) [5] provides an alternative to Neural Radiance Field (NeRF) [7] for reconstructing and rendering static multi-view scenes from novel perspectives. Kerbl *et al.* [5] represent the 3D space using scaled 3D Gaussians [6, 12], defined by a 3D covariance matrix $\mathbf{\Sigma}$ and a mean μ:

$$G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)}. \tag{1}$$

To render this representation, Zwicker *et al.* [16] project 3D Gaussians onto the image plane using the formula $\mathbf{\Sigma}' = \mathbf{AW\Sigma W}^T\mathbf{A}^T$, where $\mathbf{\Sigma}'$ denotes the 2D covariance matrix. Here, $\mathbf{W}$ is the view transformation, and $\mathbf{A}$ is the projective transformation. Rather than directly optimizing the covariance matrix $\mathbf{\Sigma}$, which must remain positive semidefinite, Kerbl *et al.* [5] parameterize it in terms of scale $\mathbf{S}$ and rotation $\mathbf{R}$. This reformulation expresses the 3D Gaussian as a 3D ellipsoid: $\mathbf{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$. Finally, 3DGS



(a) Our pivotal fine-tuning first stage: In this part, we optimize only the identity encoder to find the optimal projection of the input image onto our synthetic latent space.



(b) Our pivotal second stage of fine-tuning involves fixing the optimization latent code and focusing on optimizing the decoder to bridge the domain gap between the synthetic avatar and real subjects. During this phase, we typically address global illumination, identity texture, teeth color, and hair appearance by refining the decoders.

Figure 2. An overview of the two pivotal fine-tuning stages. (a) The first stage optimizes the identity encoder. (b) The second stage optimizes the decoder to bridge the domain gap between synthetic avatars and real subjects.

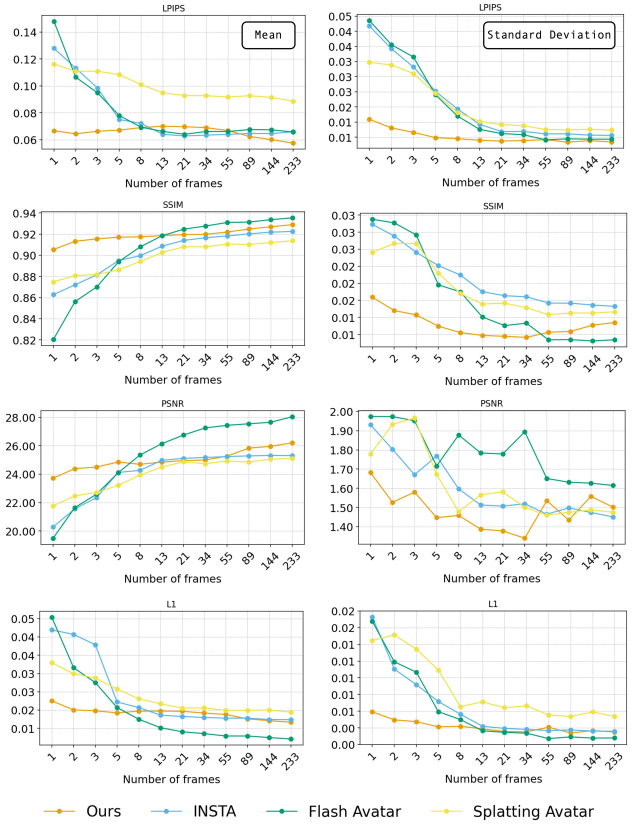


Figure 3. We evaluated the reconstruction error with respect to the number of frames using LPIPS, SSIM, L1, and PSNR metrics. For each frame count, we report the average error (left) and standard deviation (right) over 600 frames across 11 subjects.

leverages the approach of Ramamoorthi *et al*. [8] to approximate the diffuse component of the BRDF [3] using spherical harmonics (SH) for modeling global illumination and view-dependent color. Four SH bands are utilized, resulting in a 48-element vector.

C. Broader Impact

Our project centers on reconstructing highly detailed human face avatars from multiview videos, allowing for the extrapolation of expressions beyond those originally captured. While our technology is intended for constructive applications, such as enhancing telepresence and mixed reality experiences, we recognize the risks associated with misuse. To mitigate these risks, we advocate for progress in digital media forensics [9, 10] to support the detection of synthetic media. We also stress the importance of conducting research in this field with transparency and integrity.

Figure 4. Novel view evaluation of long hair and beard inversion using only three input images demonstrates the strong generalization capability of *SynShot*, which accurately models both long hair and beards.

Figure 5. Additionally, we present failure cases of our method, categorized into primary scenarios where *SynShot* may fail (from the top): (1) input images with facial accessories, such as glasses, which are absent from our synthetic dataset; (2) challenging inputs, such as squinting or closed eyes, which introduce artifacts in the final avatar; and (3) missing hairstyles in the dataset, leading to inversion errors for uncommon styles, further exacerbated by artifacts in hair segmentation.

Figure 6. Cross-Reenactment on a Limited Number of Frames: We compare *SynShot* inversion using only 3 views to SOTA methods that utilize 13 frames. While the baseline methods produce good qualitative results on the test sequence with 13 frames, they all fail severely in novel view and expression evaluation.

Figure 7. Test View Evaluation: When comparing the test views, which are very close to the training distribution, all baselines perform comparably well. Our method also achieves good results, despite the prior model being insufficiently refined in some cases (e.g., teeth).

Figure 8. We trained each method on a different number of frames to demonstrate the importance of our prior model using test sequences. In this experiment, we progressively increased the number of training frames up to 377. The frames were sampled from the training set using Farthest Point Sampling defined on the 3DMM expression space. The comparison includes INSTA [15], Flash Avatar (FA) [13], and Splatting Avatar (SA) [11].

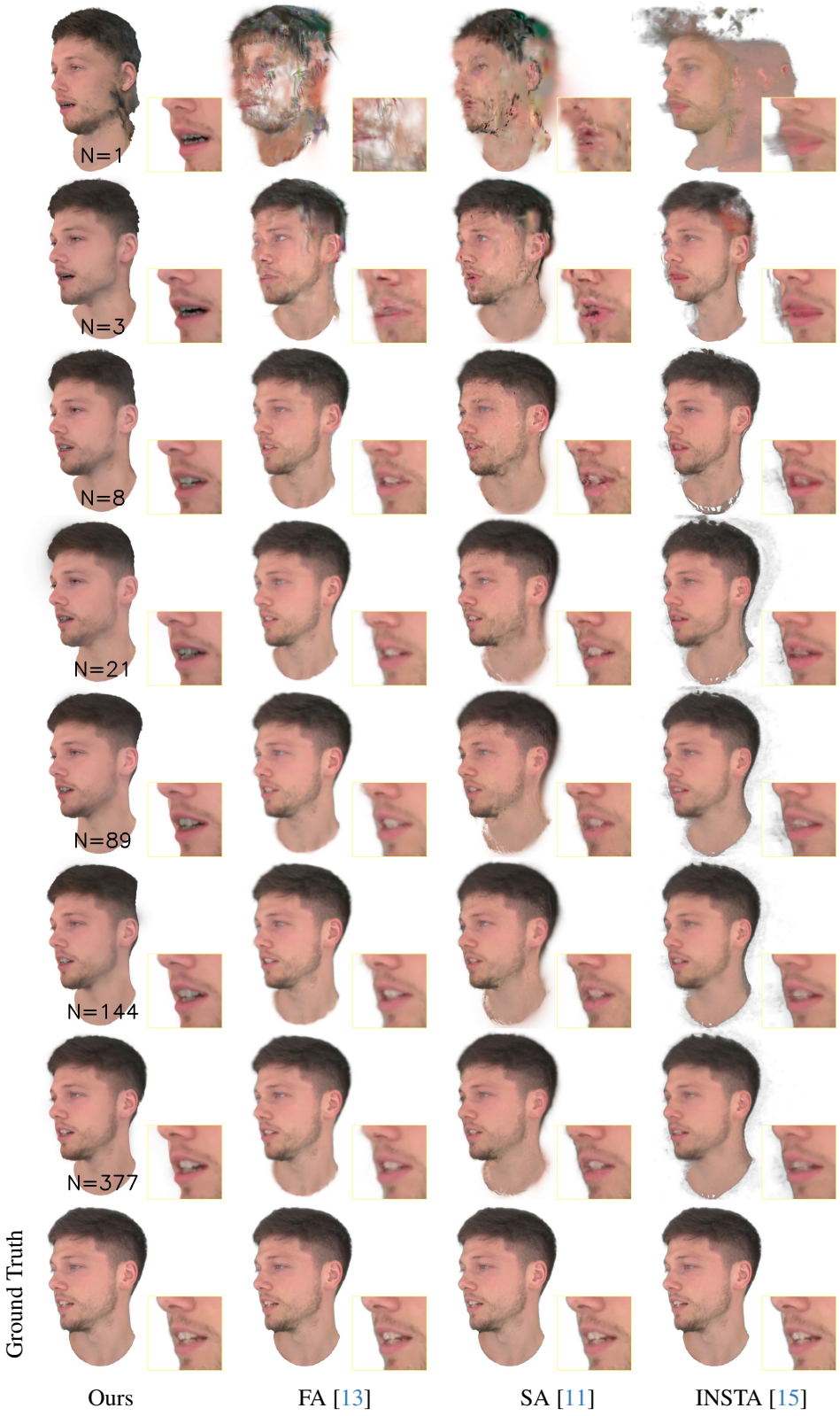


Figure 9. We trained each method on a different number of frames to demonstrate the importance of our prior model using test sequences. In this experiment, we progressively increased the number of training frames up to 377. The frames were sampled from the training set using Farthest Point Sampling defined on the 3DMM expression space. The comparison includes INSTA [15], Flash Avatar (FA) [13], and Splatting Avatar (SA) [11].

Figure 10. Additional results of single image inversion.

Figure 11. Random samples from our synthetic dataset, showcasing a diverse range of identities, expressions, and hairstyles that would be challenging to capture in an in-house studio with real subjects.

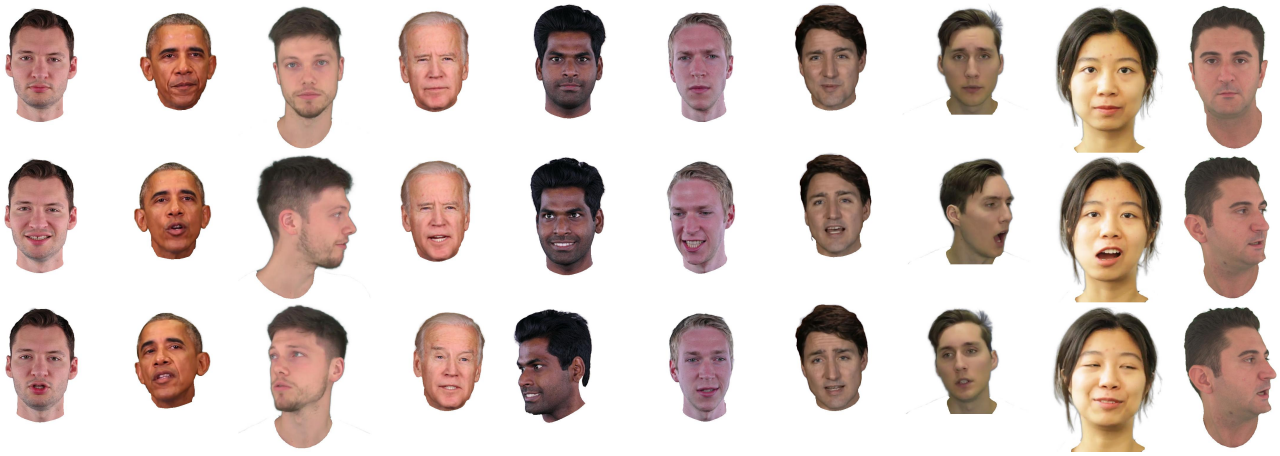


Figure 12. Unless otherwise stated, all experiments in this paper used three input images. Here, we present these images for each actor.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20950–20959, 2023. 1, 9

[2] Marcel C. Buehler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, and Abhimitra Meka. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 1

[3] Cindy M. Goral, Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile. Modeling the interaction of light between diffuse surfaces. *SIGGRAPH*, 1984. 2

[4] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A real-time nerf-based parametric head model. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 9

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *Transactions on Graphics (TOG)*, 42(4), 2023. 1

[6] Georgios Kopanas, Julien Philip, Thomas Leimkühler, and George Drettakis. Point-based neural rendering with per-view optimization. *Computer Graphics Forum (CGF)*, 40, 2021. 1

[7] 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. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1

[8] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. *SIGGRAPH*, 2001. 2

[9] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *ArXiv*, 2018. 2

[10] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *International Conference on Computer Vision (ICCV)*, pages 1–11, 2019. 2

[11] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 5, 6, 7, 8

[12] Yifan Wang, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *Transactions on Graphics (TOG)*, 38:1 – 14, 2019. 1

[13] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. FlashAvatar: High-fidelity head avatar with efficient gaussian embedding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 5, 6, 7, 8

[14] Yiyu Zhuang, Hao Zhu, Xusen Sun, and Xun Cao. MoFaNeRF: Morphable facial neural radiance field. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 9

[15] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4574–4584, 2022. 1, 5, 6, 7, 8

[16] Matthias Zwicker, Hans Rüdiger Pfister, Jeroen van Baar, and Markus H. Gross. Surface splatting. *SIGGRAPH*, pages 371–378, 2001. 1