# Avat3r: Large Animatable Gaussian Reconstruction Model for High-fidelity 3D Head Avatars

## Supplementary Material

| | Inputs | | PSNR↑ | SSIM↑ | LPIPS↓ | AKD↓ | CSIM↑ |
|---|---|---|---|---|---|---|---|
| Ava256 | 1 | GAGAvatar | 18.1 | 0.66 | 0.37 | 7.0 | **0.45** |
| | 1 | GAGAvatar† | **19.6** | 0.68 | **0.31** | **5.6** | 0.33 |
| | 1 | Ours$^{3DGAN}$ | 19.1 | 0.68 | 0.39 | 6.0 | 0.37 |
| | 1 | Ours[1] | **19.6** | **0.70** | 0.46 | 5.7 | 0.40 |
| NeRSemble | Video | FlashAvatar | 15.0 | 0.42 | 0.61 | 8.8 | 0.17 |
| | 4 | Ours | 20.5 | 0.75 | 0.33 | **3.7** | 0.50 |
| | 4 | Ours$^{+\ 984}$ | **21.6** | **0.77** | **0.29** | 3.8 | **0.76** |
| | 1 | HeadNeRF | 9.7 | 0.69 | 0.48 | 5.2 | 0.18 |
| | 1 | Portrait4D-v2 | 17.5 | 0.58 | 0.36 | 5.4 | 0.41 |
| | 1 | GAGAvatar | 18.7 | 0.70 | 0.35 | 5.4 | 0.44 |
| | 1 | GAGAvatar† | 18.9 | 0.70 | **0.32** | 4.8 | 0.24 |
| | 1 | Ours$^{3DGAN}$ | 19.5 | 0.71 | 0.38 | **4.3** | 0.31 |
| | 1 | Ours$^{3DGAN\ +\ 984}$ | 19.4 | 0.71 | 0.38 | 4.6 | **0.46** |
| | 1 | Ours[1] | **19.8** | **0.73** | 0.38 | 4.4 | 0.30 |

†re-trained on Ava256 [1]trained on 1 input view

Table 3. **Quantitative Comparison on 3D head avatar creation from various.** Our approach performs competitively when only a single input image is available, despite not being designed for a single image use-case. It also performs much better compared to monocular methods that receive a full video as input. Ours$^{3DGAN}$ denotes a 4-shot model where the 4 required images are obtained from the single input via 3D lifting with a 3D GAN. Ours$^{+\ 984}$ denotes a model that is additionally fine-tuned on 984 neutral identities for better identity preservation (CSIM metric).

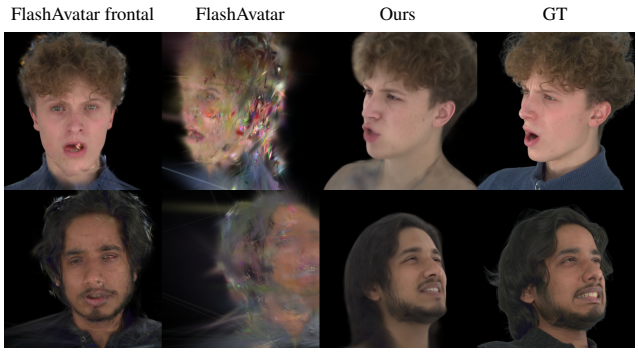



Figure 9. **Comparison with FlashAvatar** on NeRSemble.

# A. Additional Results

## A.1. Avat3rs from Phone Captures and Accessories

Fig. 12 showcases additional phone scans captured by users on their own devices, including challenging cases with glasses, rotated inputs, and a NeRSemble subject wearing a headscarf. Despite not being trained on accessories, the model handles glasses reasonably well and reconstructs the headscarf with high fidelity.

## A.2. Single-shot 3D Head Avatar Creation

To make Avat3r amenable for inference on only a single input image, we make use of a pre-trained 3D GAN [11] to first lift the single image to 3D and then render four views of the head. These renderings then constitute the input for Avat3r. We conduct comparisons with the recent 3D-aware portrait animation method GAGAvatar [1]. Specifically, we compare with two version of GAGAvatar: One provided by the authors which is trained on VFHQ [10], and another version, denoted as GAGAvatar†, that we train on the Ava256 dataset in the same setting as our method. To drive GAGAvatar, we use their monocular FLAME tracker to obtain tracked meshes. We also compare with another 3D portrait animation method, Portrait4Dv2 [2], and HeadNeRF [4]. Fig. 10 and Fig. 11 show qualitative comparisons between our method and the baselines for single input images of hold out persons. Note that our method performs competitively compared to the single-input baselines despite never being trained for a single-shot scenario. We also include a version of our model that was trained on a single input image without DUSt3R. In general, we find that Avat3r produces more realistic facial expressions than GAGAvatar and Portrait4Dv2 which are limited by FLAME's expression space. Furthermore, our method allows much more extreme viewpoint changes without sacrificing rendering quality. Tab. 3 shows quantitative results. Note that, in contrast to portrait animation methods like GAGAvatar and Portrait4Dv2, our method can benefit when more input views are available (see Appendix B and tab. 1 in the main paper).

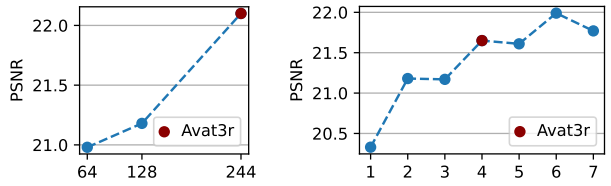
## A.3. Comparison with Monocular Methods

We compare with the recent monocular approach FlashAvatar and provide it with a full video sequence from NeRSemble. As shown in Fig. 9, the monocular method performs well from the training view but fails on novel views due to overfitting. Please also refer to the Gaussian Avatar Fusion or HeadGAP papers, which made similar observations. In contrast, our learned reconstruction prior produces a plausible 3D head avatar. Quantitative comparisons are shown in Tab. 3.


Figure 10. **Single-image comparison on Ava256.** We compare Avat3r with the recent 3D-aware portrait animation method GAGAvatar [1] in a self-reenactment scenario on hold-out persons from the Ava256 dataset. GAGavatar† denotes a version of the baseline that we trained on the Ava256 dataset. Ours[1] is a version of our model that was trained on only 1 input image (see Appendix B). Our method with 3D lifting (Ours[3DGAN]) shows better rendering quality than the baselines, especially for extreme expressions and viewing angles.


Figure 11. **Single-image comparison on NeRSemble.** We compare Avat3r with the recent 3D-aware portrait animation methods GAGAvatar [1] (GAGavatar† denotes a version of the baseline that we trained on the Ava256 dataset) and Portrait4Dv2 [2] as well as the NeRF-based face model HeadNeRF [4] in a self-reenactment scenario on persons from the NeRSemble dataset [6] Note that the NeRSemble dataset has not been used during training and therefore constitutes an evaluation scenario where both source and driver image are out-of-domain. Ours[1] is a version of our model that was trained on only 1 input image (see Appendix B). Our method with 3D lifting (Ours[3DGAN]) shows better rendering quality than the baselines in these challenging scenarios where input and target view are from opposite sides of the face.

Figure 12. **More Avat3r reconstruction results** from phone scans with different phone cameras and persons with challenging clothing.




(a) Training subjects (b) Input views

Figure 13. **Analysis of Data Efficiency.** We study how reconstruction and animation performance behaves when changing the number of training subjects and input views.


Figure 14. **Effect of Number of Train Subjects.** Training on a larger and more diverse set of people enhances the Avat3r's generalization capabilities, as expected. This leads to more accurate reconstructions, with avatars better matching the identities shown in the input images. For instance, when dealing with complex hairstyles, a model trained on a broader range of individuals reproduces the hairstyle more accurately. All ablations are trained without LPIPS loss.


Figure 15. **Effect of Number of Input Views.** Training with just a single input image noticeably impairs quality. On the other hand, using more than 4 input images during training does not lead to significant improvements. Models are trained without DUSt3R position maps and without LPIPS loss in the interest of comparability.


Figure 16. **Improved identity preservation** by including 984 additional identities with neutral expression during training.

## B. Analysis of Data Efficiency

### B.1. Scaling Subjects and Views from Ava256

In Fig. 13 we show how our model scales with the number of training subjects and input views available in the Ava256 dataset. For the analysis on the number of input views, we disable DUSt3R as it produces less reliable position maps for 2 input views, and cannot be executed at all for 1 input view. We see a clear improvement when using more training subjects as well as using more input views. However, further scaling the number of input views also has drawbacks, as it drastically increases runtime due to dense attention inside the transformer and the increased number of Gaussians that have to be rendered. We qualitatively analyze the effect of using more train subjects in Fig. 14 and the effect of the number of input views in Fig. 15.

### B.2. Effect of Adding More Neutral Subjects

Since Avat3r is only trained on 244 subjects from the Ava256 dataset, it is at risk of overfitting to those identities during training time. In our experiments, we observe that while there is a slight identity shift between the final avatar and the person in the input images, the expression transfer works quite well. We therefore hypothesize that for further improvements on the model's generalization capabilities it is not necessary to also add thousands of expressions for each additional person. To test this, we fine-tune our model with 984 additional identities from an internal dataset — just one expression each — adding only 0.08% to the training data. As shown in Fig. 16 and confirmed by improved CSIM metrics in Tab. 3, this small addition noticeably improves identity retention for these challenging cases.

(a) No skip connections (b) No position skip (c) No color skip (d) Ours full GT

Figure 17. **Qualitative Ablation of Skip Connections.** Not employing skip connections (a) causes misalignments, blurry renderings, and a slight color shift. Adding the color skip connection (b) already noticeably improves sharpness and color fidelity. On the other hand, if the position skip connection is added (c), geometric details are improved but the overall color slightly off. Using both skip connections (d) yields the best result.

| | Creation↓ in [s] | Driving↑ in [fps] |
|---|---|---|
| HeadNeRF | 65 | **111** |
| Portrait4D-v2 | 0.2 | 4 |
| GAGAVatar | **0.1** | 63 |
| GPAvatar | 0.2 | 9.5 |
| Ours[1] | 1.1 | 53 |
| Ours$^{3DGAN}$ | 17.9 | 7.9 |
| Ours | 12.3 | 7.9 |

Table 4. **Runtime analysis.** Our method can create a high-quality avatar in a few seconds, and animate it at interactive rates.

| | $\mathcal{C}$ | $\mathcal{P}$ | PSNR↑ | SSIM↑ | LPIPS↓ | JOD↑ | AKD↓ | CSIM↑ |
|---|---|---|---|---|---|---|---|---|
| No skip | ☐ | ☐ | 21.39 | 0.740 | 0.456 | 4.99 | 9.24 | 0.60 |
| No pos. skip | ☒ | ☐ | 21.76 | 0.746 | 0.443 | 5.03 | 9.04 | 0.611 |
| No col. skip | ☐ | ☒ | 21.55 | 0.745 | 0.435 | 5.00 | **7.69** | 0.648 |
| Avat3r | ☒ | ☒ | **22.05** | **0.751** | **0.421** | **5.15** | 7.99 | **0.689** |

Table 5. **Quantitative Ablation of Skip Connections.** We analyze the effect of the color ($\mathcal{C}$) and position ($\mathcal{P}$) skip connections. All ablation models are trained without LPIPS loss. Metrics are computed on $667 \times 667$ renderings.

## C. Inference efficiency & driving speed.

While Avatar creation takes several seconds for our method due to DuSt3R and Sapiens, we can cache all activations up to the final cross-attention layers afterwards, leading to expression driving at 7.9fps for our 4-shot model and 53fps for our single-shot model, see Tab. 4. Runtimes measured on a single RTX3090 GPU.

## D. Effect of Skip Connections

We analyze the effect of the proposed skip connections, i.e., omitting Eq. (11), Eq. (12), or both of the main paper. The results are listed in Tab. 5. We observe a noticeable hit in performance when either skip connection is removed. Furthermore, we qualitatively analyze the effect of skip connections in Fig. 17.

## E. Training Details

**Dataset processing.** We use the 4TB version of the Ava256 dataset [8] which contains 256 persons, 80 cameras, and roughly 5000 frames per person that are sampled at 7.5 fps. We compute foreground segmentation masks with BackgroundMattingV2 [7] and replace the background in all images with black pixels. We use the provided tracked mesh to find a $512 \times 512$ head-centered square crop for input images and $667 \times 667$ head-centered square crop for supervision views. This ensures that the pixels in the input images are used efficiently to show as much as possible of the head, leading to more 3D Gaussians. The reason for also cropping the target images is to remove parts of the torso, as it is not the focus of this work.

**DUSt3R and Sapiens.** Since both DUSt3R [9] and Sapiens [5] are expensive foundation models, we pre-compute the position and feature maps for the input frames. For Dust3r, it is prohibitive to pre-compute all possible combinations of 4 input views out of the available 80 cameras. Instead, we choose 3 "reasonable partner views" for each input and only store the position map for that viewpoint. This assigns each input view exactly one position map, which is conceptually wrong since the position map from DUSt3R should depend on the other 3 selected views. Nevertheless, we did not observe any disadvantages from this simplification strategy.


Figure 18. **Performance on NeRSemble dataset.** We show reconstructed and animated avatars using 4 images from the NeRSemble dataset. Note that this dataset was not used during training and contains images with different lighting conditions, viewpoints, and camera intrinsics than the Ava-256 dataset that was used to train Avat3r.

| | Hyperparameter | Value |
|---|---|---|
| ViT & GRM | ViT patch size | $8 \times 8$ |
| | hidden dimension $D$ | 768 |
| | #self-attention layers | 8 |
| | #cross-attention layers | 8 |
| | #GRM transformer upsampler step | 1 |
| Input & Output | Sapiens version | 2b |
| | Sapiens feature dimension | 1920 |
| | Input image resolution | $512 \times 512$ |
| | Gaussian attribute map resolution | $512 \times 512$ |
| | Train render resolution | $667 \times 667$ |
| Expression MLP | Dimension of expression code | 256 |
| | #expression sequence MLP layers | 2 |
| | Dimension of expression sequence MLP | 256 |
| | Expression sequence MLP activation | ReLU |

Table 6. **Hyperparameters.**

**Head-centric coordinates.** We further simplify the task by factoring the head poses from the provided tracked mesh into the camera poses instead of letting the network predict them. That way, our model can always predict the head in canonical pose, making the task easier. This is possible because modeling the torso, which in head-centric coordinates moves a lot when the person shakes their head, is not the focus of this work.

**Expression codes.** Our architecture is agnostic to the specific choice of animation signal. For experiments on ava256, we used the dataset's expression codes that were originally predicted by a generalized expression encoder providing a driving signal beyond FLAME's topology. For experiments on NeRSemble and in-the-wild driving videos, we fine-tuned our model using FLAME codes obtained by running GAGAvatar's version of Metrical Tracker [1, 12]. This shows that Avat3r learns a general notion of facial expressions that can be adapted to fit a specific driving signal.

**k-farthest viewpoint sampling.** To ensure that the 4 input images always follow a reasonable viewpoint distribution, we employ k-farthest viewpoint sampling. Specifically, we first start from a random camera and collect a set of 10 candidate cameras that are evenly spread out using farthest point sampling. From this candidate set, we then randomly select 4 cameras as input. This two-stage approach ensures that the input cameras are sufficiently random during training but also reasonably spread out to avoid seeing a person only from one side. During sampling input viewpoints, we exclude cameras that only observe the person from the back since those are not realistic inputs during test-time.

**Input timestep sampling.** To improve robustness of our model, we sample different timesteps for each of the 4 input images. This ensures that the model can deal with inconsistencies in the input. To maximize the diversity in the input expressions, we uniformly sample 10 timesteps in the segments: `EXP_eye_wide`, `EXP_tongue001`, and `EXP_jaw003` from the recordings of the Ava256 dataset. This covers the most extreme facial expressions while avoiding having to pre-compute DUSt3R and Sapiens maps for every single image in the dataset.

**Speed.** To speed-up training, we employ the 3D Gaussian Splatting performance improvements of DISTWAR [3].

## F. Hyperparameters

In Tab. 6, we list the most important hyperparameters for training Avat3r.

## References


[1] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *arXiv preprint arXiv:2410.07971*, 2024. 1, 2, 5

[2] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 1, 2

[3] Sankeerth Durvasula, Adrian Zhao, Fan Chen, Ruofan Liang, Pawan Kumar Sanjaya, and Nandita Vijaykumar. Distwar: Fast differentiable rendering on raster-based rendering pipelines. *arXiv preprint arXiv:2401.05345*, 2023. 5

[4] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20374–20384, 2022. 1, 2

[5] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision*, pages 206–228. Springer, 2025. 4

[6] 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

[7] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. *arXiv*, pages arXiv–2012, 2020. 4

[8] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired Human Captures for Complete, Driveable, and Generalizable Avatars. *NeurIPS Track on Datasets and Benchmarks*, 2024. 4

[9] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 4

[10] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 1

[11] Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, and Chun Yuan. Make encoder great again in 3d gan inversion through geometry and occlusion-aware encoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2437–2447, 2023. 1

[12] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022. 5