

HeadGAP: Few-Shot 3D Head Avatar via Generalizable Gaussian Priors

Supplementary Material

A. Implementation Details

Dataset. We partition the data into training and testing sets, comprising 119 and 45 subjects, respectively. Of these subjects, the data of 11 training subjects and 3 testing subjects are provided by [13] and others are processed by our FLAME tracking algorithm. For more information about the dataset, we highly encourage the reader to refer to the paper of NeRSemble [8] for further details.

Model Detail. We divide the Gaussian primitives into $p = 11$ parts, including 1) “forehead”, 2) “nose”, 3) “eye”, 4) “teeth”, 5) “lip”, 6) “ear”, 7) “hair”, 8) “boundary”, 9) “neck”, 10) “other face region”, and 11) “other”. The part for the primitives is determined by the face masks provided by FLAME [9]. The illustration of Gaussian primitives with different parts is shown in Fig. A. The primitive number is set to $n = 83,651$ by initializing from a UV map with a resolution of 300×300 . The feature dimensions of identity-shared point encoding \mathbf{f} , identity code \mathbf{z} , and point appearance feature \mathbf{h} are set to $c_1 = 48$, $c_2 = 128$, and $c_3 = 34$ respectively. All the MLPs f^M consist of 4 layers. Meanwhile, the CNN f^C contains 6 layers. The identity codebook \mathbf{z} is initialized with zero.

Training Detail. We adopt Adam [7] optimizer for the model training. For prior learning, we utilize $k = 119$ identities and set the batch size to 32. For all the parameters, the learning rate begins at $1e^{-3}$ and decreases with the cosine scheduler. The prior model is trained on 8 A100 GPUs for 100K steps, which takes around 2 days. The loss weights λ_m , λ_{l1} , λ_{ssim} , λ_{lips} , λ_α , λ_s , λ_μ , and λ_{arap} are set to 10, 0.8, 0.2, 0.4, 1, 1, 0.01, and 1 respectively. For few-shot personalization, we set the batch size to 1. We set the learning rate of the identity-shared point encoding \mathbf{f} to $1e^{-3}$ and other parameters’ to $1e^{-5}$. Unless otherwise stated, we take 500 steps for inversion and 500 steps for fine-tuning, which uses about 5 minutes in total with an A100 GPU. For view regularization, we generate $m = 16$ reference views similar to the camera setups of the NeRSemble dataset. The loss weights λ_{ref} is set to 0.01. For 3-shot novel identities’ personalization on NeRSemble, we utilize cameras with id “0”, “8”, and “15”. For our captured data, we select viewpoints similar to those of NeRSemble.

About adaptive density control. To allow full control of primitive numbers, we do not utilize adaptive density control, opacity reset, and point pruning as GaussianAvatars [13].

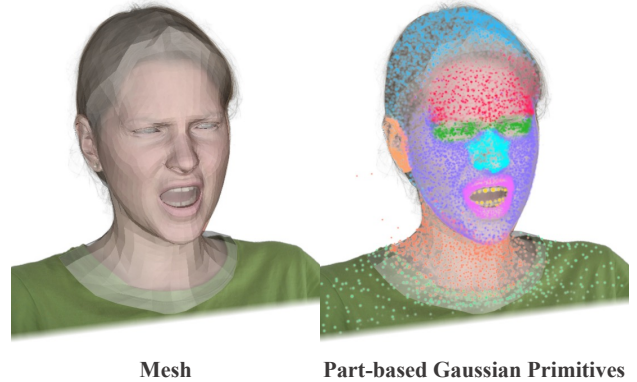


Figure A. Illustration of the FLAME mesh (left) and semantic part of our Gaussian primitives bound on the mesh (right). Different point colors represent different parts.

B. Experiment Results

B.1. Network Comparison

GAPNet is capable of adapting to different numbers of IDs for training. To demonstrate the network capability, we compare its performance for a single person against GaussianAvatars [13]. For a fair comparison of the network, we utilize the same adaptive density control approaches as [13]. We also use the full training data of each single subject, similar to [13]. The mean quantitative results over subject “074”, “175”, and “210” are shown in Tab. A. GAPNet obtains better results in all metrics. We further illustrate the qualitative comparisons in Fig. B. Our model is capable of fitting the dynamic details of the training subject well, as shown in self-reenactment results. Moreover, our cross-reenactment performance is significantly more robust than GaussianAvatars. The robust animations further prove our model design is quite suitable for learning generalizable priors across different subjects.

Method	LPIPS↓	PSNR↑	SSIM↑
GaussianAvatars	0.120	25.21	0.911
Ours	0.091	25.48	0.912

Table A. Comparisons for single avatar creations.

B.2. Prior Learning Results

We show our prior learning results of the 119 identities in Fig. H. The visualized results show that our GAPNet can learn the appearance characteristics of different identities.

B.3. More Qualitative Results

We present additional qualitative experimental results in Fig. C, Fig. D, and Fig. E. All subjects are novel IDs and were not seen during the training process.

Fig. C shows self-reenactment results with novel-view renderings for different identities. Fig. D and Fig. E present cross-reenactment results from frontal and side views, respectively, demonstrating stable animations.

The results indicate that our model effectively generalizes to data that differs from the NeRSemble dataset. Furthermore, it achieves consistent few-shot performance across diverse ethnicities and genders, thereby further reinforcing its capacity for effective generalization.

B.4. Head Avatar Editing

Since our representation models textures using 3D Gaussian Splatting upon the base FLAME mesh, we can perform 1) texture interpolation between different identities using the same FLAME mesh, 2) texture swapping using the same FLAME mesh, and 3) geometry editing by swapping the FLAME mesh. The results are shown in Fig. F.

B.5. More In-the-wild Results

In this section, we present additional results on in-the-wild images. All result IDs are out-of-distribution samples beyond the NeRSemble [8] dataset. Specifically, we capture monocular video data of each identity performing various expressions and select 12 images for avatar personalization. As shown in Fig. G, we present the cross-reenactment driving results when providing the same facial expression motion sequence. The results demonstrate that our method exhibits strong few-shot generalization capability even in in-the-wild settings.

C. Further Discussions on Baselines

We compare multiple baseline approaches for one-shot and few-shot personalization based on the number of input images in the main text. In this section, we further elucidate the details of the experimental comparisons.

C.1. Baseline taxonomy

We categorize the baselines into two types based on whether they involve a process of learning priors.

Type-I includes: ROME [6], GOHA [11], VODOO 3D [14], HiDe-NeRF [10], Portrait4Dv1 [2], Portrait4Dv2 [3], GPAvatar [1] and DiffusionRig [4].

Type-II includes: FlashAvatar [15], GaussianAvatars [13] and NHA [5]

C.2. Comparison with single-view baselines

In one-shot personalization experiments, when driving novel expressions, tri-plane representation-based volume

rendering methods [1–3, 10, 11, 14] require the driving image as input. This might result in appearance leakage (e.g., dynamic details of new expressions). In contrast, our method uses only the tracking mesh of the driving image and models dynamic details through prior learning.

C.3. Comparison with GS-based methods

We show the comparison with GaussianAvatars [13] and FlashAvatar [15] in Fig. J. Although they do not focus on few-shot input like ours, we include comparisons because we all use 3D Gaussian Splatting as a representation. We observe that they require a substantial overlap of input views or monocular videos with human heads rotated to different orientations.

In few-shot personalization experiments, as shown in the Fig. J, all per-subject optimization Gaussian Splatting-based baselines lack prior information and require individual training for each person. It can be observed that all baseline methods tend to overfit the training views and fail to extrapolate to unseen views. This qualitative comparison demonstrates the effectiveness and necessity of constructing priors for Gaussian Splatting. Due to the noticeable artifacts, FlashAvatar[†], GaussianAvatars[‡], and GaussianAvatars[◆] are infeasible for calculating meaningful ID similarity metrics. Therefore, we did not report their corresponding metrics in Tab.1 of the main text.

D. Limitations and Future Works

While our method can quickly construct personalized, high-fidelity, and realistic human head avatars, it still has the following issues: (1) In cases where the subject wears glasses or has noticeable facial accessories, the avatar construction may exhibit artifacts (as depicted in Fig. K). A reason for this incapability is that our prior learning phase does not incorporate such samples for training. Including the corresponding data for training can potentially resolve this problem. (2) The adoption of CNNs for refinement in screen space may result in view-dependent overfitting, which can induce flickering among different viewpoints and lead to quality degradation for certain unseen views during training. Therefore, exploring more consistent refinement techniques in 3D space presents a promising avenue for further investigation. (3) Our method does not focus on modeling the subject’s clothing and hair. We believe that combining methods such as [12] to model hair or clothing separately is a promising research direction. (4) Additionally, lighting variation is important for the realism of head avatars. Currently, we only consider uniform lighting. We believe that integrating relighting into the Gaussian Splatting is also a promising research direction for head avatars.



Figure B. Self- and cross-reenactment comparisons between our method and GaussianAvatars for single-subject modeling.



Figure C. Self-reenactment results. The images inside the red box are the driving expressions. We showcase the renderings from different viewpoints.



Figure D. Cross-reenactment results. The images inside the red box are the driving expressions.



Figure E. Cross-reenactment results. The images inside the red box are the driving expressions.

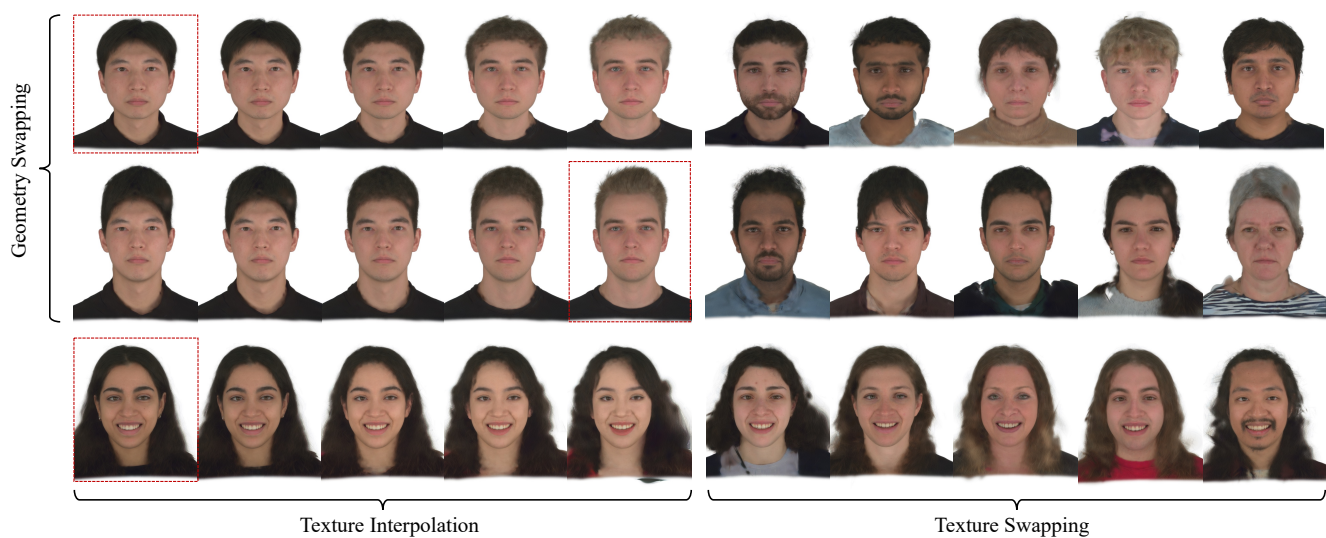


Figure F. Illustration of the GAPNet's 1) texture interpolation, 2) texture swapping, and 3) geometry swapping. The results on the same row are using the same head geometry. The identities inside red boxes use the paired texture and FLAME mesh.



Figure G. Qualitative results of 3D animatable head avatars generated from few-shot in-the-wild images and driven by the same facial expression sequence.



Figure H. The rendered results of the 119 identities used for prior learning.



Figure I. More qualitative experiments on other subjects using 3-shot inputs compared to state-of-the-art methods.



Figure J. Qualitative comparison results. We compare the rendering results from different views using our 3-shot input avatars with the Gaussian Splatting-based baseline methods.

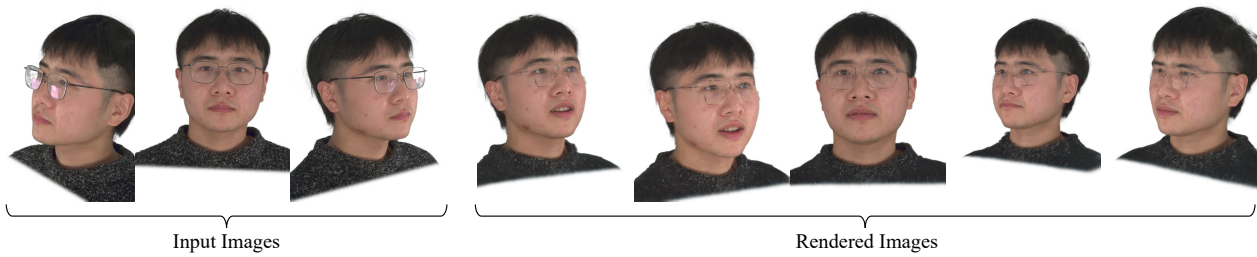


Figure K. Failure cases. Our approach can not resolve subjects with noticeable facial accessories (*e.g.*, glasses).

References

- [1] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *The Twelfth International Conference on Learning Representations*, 2024. 2
- [2] Yu Deng, Duomin Wang, Xiaohang Ren, Xingyu Chen, and Baoyuan Wang. Portrait4d: Learning one-shot 4d head avatar synthesis using synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7119–7130, 2024. 2
- [3] Yu Deng, Duomin Wang, and Baoyuan Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. *arXiv preprint arXiv:2403.13570*, 2024. 2
- [4] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12736–12746, 2023. 2
- [5] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2
- [6] Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision*, pages 345–362. Springer, 2022. 2
- [7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [8] 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. 1, 2
- [9] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 1
- [10] Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, and Xuelong Li. One-shot high-fidelity talking-head synthesis with deformable neural radiance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17969–17978, 2023. 2
- [11] Xueting Li, Shalini De Mello, Sifei Liu, Koki Nagano, Umar Iqbal, and Jan Kautz. Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [12] Haimin Luo, Min Ouyang, Zijun Zhao, Suyi Jiang, Longwen Zhang, Qixuan Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Gaussianhair: Hair modeling and rendering with light-aware gaussians. *arXiv preprint arXiv:2402.10483*, 2024. 2
- [13] 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, 2
- [14] Phong Tran, Egor Zakharov, Long-Nhat Ho, Anh Tuan Tran, Liwen Hu, and Hao Li. Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10336–10348, 2024. 2
- [15] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity head avatar with efficient gaussian embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1802–1812, 2024. 2