REFERENCES

ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. 2023. Flame-in-nerf: Neural control of radiance fields for free view face animation. In *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. 1–8.

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. 2022. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 20364–20373.

Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, et al. 2023. Learning Personalized High Quality Volumetric Head Avatars from Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16890–16900.

Jonathan T Barron and Jitendra Malik. 2011. High-frequency shape and albedo from shading using natural image statistics. In *CVPR 2011*. IEEE, 2521–2528.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-Quality Single-Shot Capture of Facial Geometry. *ACM Trans. Graph.* 29, 4, Article 40 (jul 2010), 9 pages. https://doi.org/10.1145/1778765.1778777

Alexander Bergman, Petr Kellnhofer, Wang Yifan, Eric Chan, David Lindell, and Gordon Wetzstein. 2022. Generative neural articulated radiance fields. *Advances in Neural Information Processing Systems* 35 (2022), 19900–19916.

Shrisha Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. 2023. FLARE: Fast learning of Animatable and Relightable Mesh Avatars. *ACM Transactions on Graphics* 42 (Dec. 2023), 15. https://doi.org/10.1145/3618401

Volker Blanz and Thomas Vetter. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 157–164.

Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. 2010. High Resolution Passive Facial Performance Capture. 29, 4, Article 41 (jul 2010), 10 pages. https://doi.org/10.1145/1778765.1778778

Zhiqin Chen and Hao Zhang. 2019. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5939–5948.

Hao-Bin Duan, Miao Wang, Jin-Chuan Shi, Xu-Chuan Chen, and Yan-Pei Cao. 2023. BakedAvatar: Baking Neural Fields for Real-Time Head Avatar Synthesis. *arXiv preprint arXiv:2311.05521* (2023).

Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8649–8658.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022a. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–12.

Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022b. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)* 41, 6 (2022). https://doi.org/10.1145/3550454.3555501

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. *ACM Trans. Graph.* 30, 6 (dec 2011), 1–10. https://doi.org/10.1145/2070781.2024163

Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural head avatars from monocular RGB videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18653–18664.

Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2335–2342.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794.

Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. 2023. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint arXiv:2312.02134* (2023).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 694–711.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *TOG* 42, 4 (2023), 1–14.

Taras Khakhulin, Vanessa Sklyarova, Victor Lempitsky, and Egor Zakharov. 2022. Realistic One-shot Mesh-based Head Avatars. In *European Conference of Computer vision (ECCV)*.

Marc Levoy, Kari Pulli, Brian Curless, Szymon Rusinkiewicz, David Koller, Lucas Pereira, Matt Ginzton, Sean Anderson, James Davis, Jeremy Ginsberg, Jonathan Shade, and Duane Fulk. 2000. The Digital Michelangelo Project: 3D Scanning of Large Statues. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 131–144. https://doi.org/10.1145/344779.344849

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. 2023. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. *arXiv preprint arXiv:2311.16096* (2023).

Xian Liu, Yinghao Xu, Qianyi Wu, Hang Zhou, Wayne Wu, and Bolei Zhou. 2022. Semantic-Aware Implicit Neural Audio-Driven Video Portrait Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep Appearance Models for Face Rendering. *ACM Trans. Graph.* 37, 4, Article 68 (July 2018), 13 pages.

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of Volumetric Primitives for Efficient Neural Rendering. *ACM Trans. Graph.* 40, 4, Article 59 (jul 2021), 13 pages.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. 2024. Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis. In *3DV*.

Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 64–73.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4460–4470.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 165–174.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.

Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2023. Relightable gaussian codec avatars. *arXiv preprint arXiv:2312.03704* (2023).

Ruizhi Shao, Jingxiang Sun, Cheng Peng, Zerong Zheng, Boyao Zhou, Hongwen Zhang, and Yebin Liu. 2023. Control4d: Dynamic portrait editing by learning 4d gan from 2d diffusion-based editor. *arXiv preprint arXiv:2305.20082* (2023).

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. *arXiv preprint arXiv:2309.16653* (2023).

Marshall Tappen, William Freeman, and Edward Adelson. 2002. Recovering intrinsic images from a single image. *Advances in neural information processing systems* 15 (2002).

Cong Wang, Di Kang, Yanpei Cao, Linchao Bao, Ying Shan, and Song-Hai Zhang. 2023. Neural Point-based Volumetric Avatar: Surface-guided Neural Points for Efficient and Photorealistic Volumetric Head Avatar. In *ACM SIGGRAPH Asia 2023 Conference Proceedings*.

Chenyang Wang, Zerong Zheng, Tao Yu, Xiaoqian Lv, Bineng Zhong, Shengping Zhang, and Liqiang Nie. 2024. DiffPerformer: Iterative Learning of Consistent Latent Guidance for Diffusion-based Human Video Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. 2022. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20333–20342.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 2023. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. *arXiv preprint arXiv:2310.08528* (2023).

Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. 2023. FlashAvatar: High-Fidelity Digital Avatar Rendering at 300FPS. *arXiv preprint arXiv:2312.02214* (2023).

Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023b. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Yuelang Xu, Hongwen Zhang, Lizhen Wang, Xiaochen Zhao, Huang Han, Qi Guojun, and Yebin Liu. 2023c. LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In *ACM SIGGRAPH 2023 Conference Proceedings*.

Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 2023a. 4K4D: Real-Time 4D View Synthesis at 4K Resolution. (2023).

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2023. Deformable 3D Gaussians for High-Fidelity Monocular Dynamic Scene Reconstruction. *arXiv preprint arXiv:2309.13101* (2023).

Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems* 33 (2020), 2492–2502.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Xiaochen Zhao, Jingxiang Sun, Lizhen Wang, and Yebin Liu. 2023a. InvertAvatar: Incremental GAN Inversion for Generalized Head Avatars. *arXiv preprint arXiv:2312.02222* (2023).

Xiaochen Zhao, Lizhen Wang, Jingxiang Sun, Hongwen Zhang, Jinli Suo, and Yebin Liu. 2023b. HAvatar: High-Fidelity Head Avatar via Facial Model Conditioned Neural Radiance Field. *ACM Trans. Graph.* (oct 2023). https://doi.org/10.1145/3626316 Just Accepted.

Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. 2023b. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. *arXiv preprint arXiv:2312.02155* (2023).

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13545–13555.

Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J Black, and Otmar Hilliges. 2023a. Pointavatar: Deformable point-based head avatars from videos. In *CVPR*. 21057–21067.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. 2023. Instant volumetric head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4574–4584.

Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. 2001. EWA volume splatting. In *Proceedings Visualization, 2001. VIS'01.* IEEE, 29–538.

In this supplementary document, we present additional experiments conducted with our method, specifically exploring different rendering resolutions in Sec. A.1 and examining training and rendering efficiency in Sec. A.6. Furthermore, the implementation details are provided, encompassing the strategies for point insertion and deletion in Sec. B.1, the network architecture in Sec. B.2, training details in Sec. B.3, evaluation details in Sec. B.4, and the data preprocessing in Sec. B.5. For a more in-depth exploration, we recommend referring to our supplemental video.

A ADDITIONAL RESULTS

A.1 Different rendering resolution

We present diverse resolution results of rendering on the same case, demonstrating the capability of our method to achieve higher resolutions with enhanced details. It is noteworthy that all the results in the main paper are rendered at a resolution of 512×512. In this section, we conduct a comparative analysis of the same case using different rendering resolutions, namely, 1024×1024 and 512×512. As depicted in Fig. 9, the case rendered at a resolution of 1024×1024 exhibits superior quality in terms of appearance details.

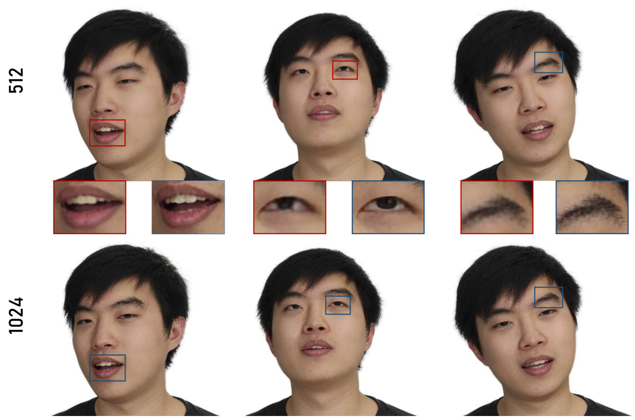

Fig. 9. **Qualitative comparison of different resolutions.** We render our Gaussian point-based avatar representation with different resolutions. Compared to the lower resolution, the higher resolution recovers more details.

A.2 More Quantitative Result

Considering article space constraints, we present only two results in the main paper. We list other three results in Table 2. Case 3 indicates the second row in Fig. 10, case 4 indicates the third row in Fig. 10, and case 5 indicates the first row in Fig. 10.

A.3 Long hair and Beards

Our methods can also handle cases with long hair (without fluttering) or beards. We present the results of rendering on the case with beards (first row) and another case with long hair (second row). As depicted in Fig. 11, both cases exhibit high quality in terms of appearance details.

Error Metric (case 3)	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Nerface[Gafni et al. 2021]	0.016	0.1353	0.9213	26.32
IMavatar[Zheng et al. 2022]	0.016	0.1384	0.9127	25.31
PointAvatar[Zheng et al. 2023a]	0.014	0.0649	0.9313	28.53
Ours	**0.010**	**0.0566**	**0.9514**	**30.83**
Error Metric (case 4)	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Nerface[Gafni et al. 2021]	0.059	0.2655	0.8015	18.78
IMavatar[Zheng et al. 2022]	0.029	0.1509	0.8755	24.48
PointAvatar[Zheng et al. 2023a]	0.025	0.0954	0.8708	25.61
Ours	**0.019**	**0.0896**	**0.8908**	**27.47**
Error Metric (case 5)	L1 ↓	LPIPS ↓	SSIM ↑	PSNR ↑
Nerface[Gafni et al. 2021]	0.012	0.1004	0.8702	28.26
IMavatar[Zheng et al. 2022]	0.015	0.1351	0.8974	27.11
PointAvatar[Zheng et al. 2023a]	0.013	0.0862	0.9033	28.97
Ours	**0.009**	**0.0835**	**0.9293**	**31.76**

Table 2. **Quantitative evaluation.** We report the other 3 quantitative results on test poses and expressions. Our method also achieves better rendering quality compared to SOTA methods.

A.4 Duration and Expression Diversity

Most of the video-based head avatar literature (including ours) emphasizes the diversity of facial expressions and poses in videos. Giving a comprehensive range of expressions and poses is crucial to achieving impressive performance. Similar to other video-based person-specific portrait reenactment methods, such a dataset typically requires at least one minute long. In this section, we select two clips from the training set of a specific case in Sec. 4 to serve as a comparison group. Through this comparison, we aim to underscore the significance of expression diversity. In Fig. 12, the first column depicts the ground truth data, while the second column corresponds to models trained on comprehensive videos lasting approximately 120 seconds. The third column corresponds to models trained using data from the first clip, which spans 30 seconds and predominantly features frontal facial expressions. The fourth column represents models trained using data from the second clip, which extends over 60 seconds and showcases a mixture of both frontal and side facial expressions.

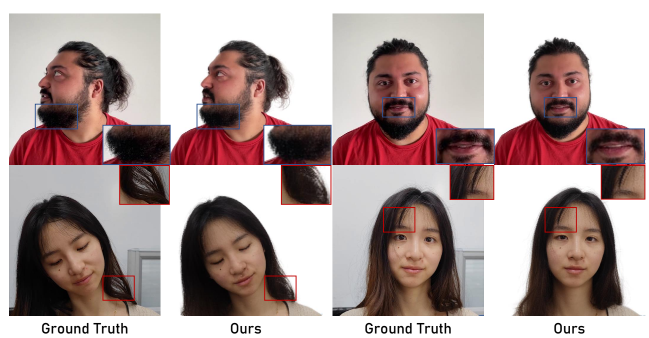

Fig. 11. **Long Hair and Beards.** We train our method on the dataset of a man with a beard and a woman with long hair.

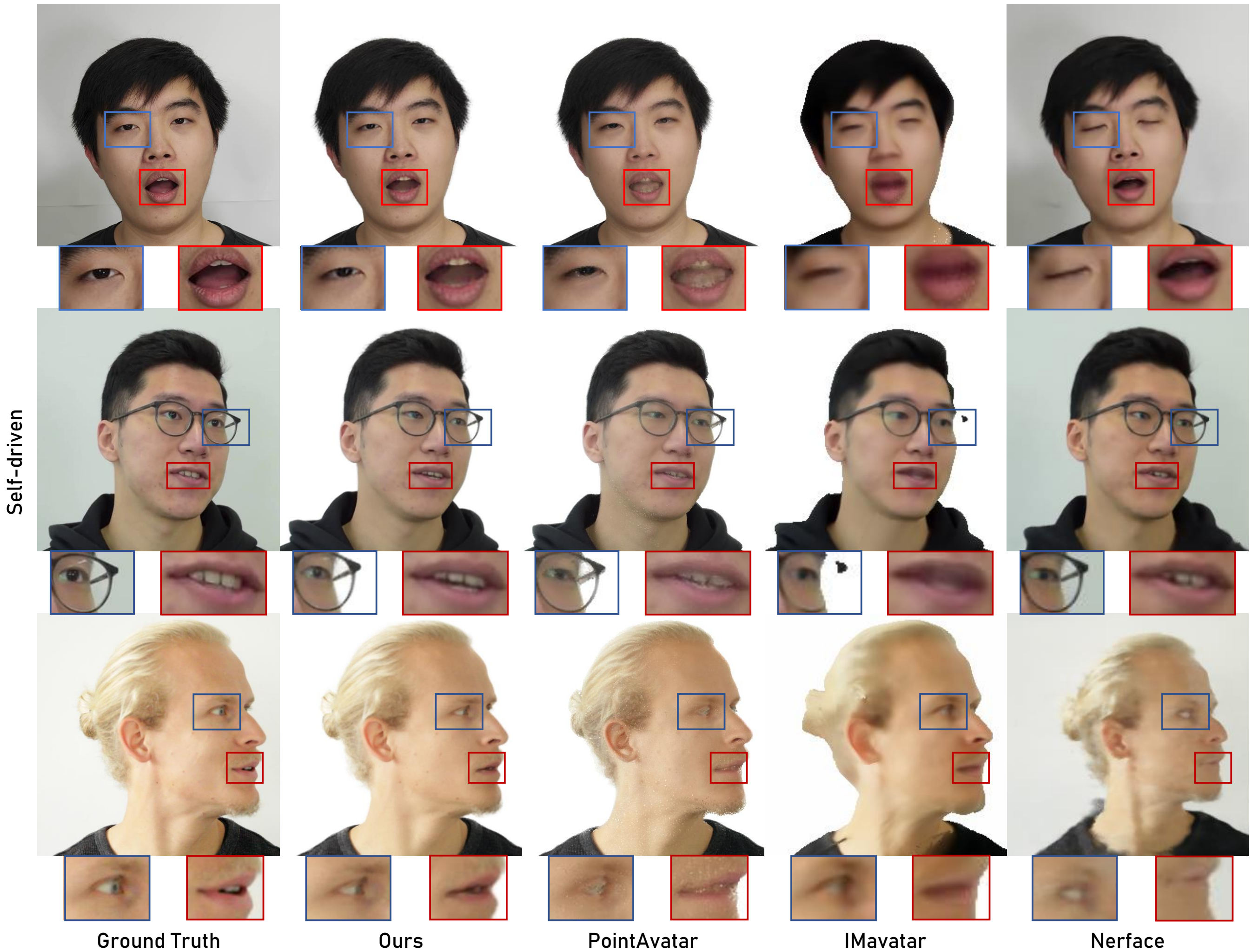

Fig. 10. In qualitative comparisons, MonoGaussianAvatar demonstrates superior performance in producing photo-realistic and detailed appearances compared to state-of-the-art methods.

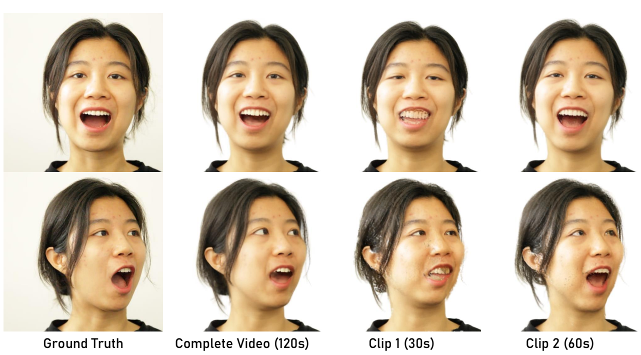

Fig. 12. **Duration and Expression Diversity.** We train our method on the same dataset with different duration.

A.5 User Study

To enhance the quantification of the quality of our method, we conduct a user study concentrating on two metrics: expression accuracy and video rendering quality, comparing with state-of-the-art (SOTA) methods. We randomly sample 10 video clips from 5 cases (2 subjects from IMavatar [Zheng et al. 2022], 1 subject from NeRFace [Gafni et al. 2021], and 2 subjects captured by us) mentioned in Sec. 4. In our user study, 15 participants are enlisted to evaluate each video, focusing on two aspects: "visual quality" and "expression accuracy." Notably, we provide the participants with ground truth data as a reference for their assessments. The results, as depicted in Table 3, underscore our approach's superiority, as it attained the highest ratings for both visual quality and expression accuracy.

Method	Visual Quality ↑	Expression Accuracy ↑
Nerface[Gafni et al. 2021]	12.67	23.33
IMavatar[Zheng et al. 2022]	0	0
PointAvatar[Zheng et al. 2023a]	8.67	8.00
Ours	**78.67**	**68.67**

Table 3. **User study.** The table exhibits the percentage of participant evaluations for each method concerning both visual quality and expression accuracy.

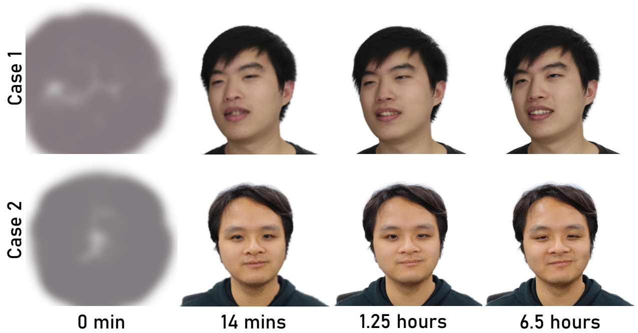

Fig. 13. **Qualitative comparison of training stages.** We document the convergence process of our MonoGaussianAvatar. In comparison with the implicit-based method detailed in Table 4, our approach exhibits significantly faster convergence.

Method	Training time (hour)	Runtime(s)
Nerface[Gafni et al. 2021]	48h	2s
IMavatar[Zheng et al. 2022]	54h	38s
PointAvatar[Zheng et al. 2023a]	11h	0.05s
Ours	9h	0.03s

Table 4. **Training time and Runtime (per image).** We provide comprehensive insights into the training time and animation runtime of both our method and state-of-the-art (SOTA) methods. Notably, our approach attains superior efficiency in both training and animation compared to the existing state-of-the-art methods.

A.6 Training and Animation Efficiency

As illustrated in Table 4, we present a comprehensive comparison of training time and runtime of animation per image for the same case, underscoring the notable efficiency of our method in both training and animation processes. Moreover, we depict the training convergence process of our method in Fig. 13, illustrating its efficient training performance in two distinct cases.

B IMPLEMENTATION DETAILS

In this section, we provide implementation details on the strategy of point insertion and deletion, network architecture, and training details. Furthermore, our code will be made available for research purposes. It is pertinent to note that we implemented our approach in PyTorch utilizing an NVIDIA GTX 3090.

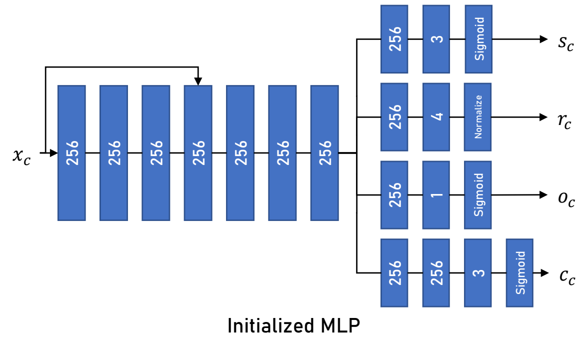

Fig. 14. **The initialized MLP.** In the initialized MLP, each linear layer is succeeded by weight normalization, and the activation function utilized is the Softplus, with the exception of the final layer.

B.1 Point Insertion and Deletion

We elaborate on the detailed process of point insertion and deletion, elucidating the settings for rendering radius and sampling radius.

We randomly initialize 400 points on a sphere. During the initial 40 epochs, we employ a two-fold strategy: pruning points with opacity below 0.1 and upsampling the points to a predetermined quantity (specified as 400, 800, 1600, 3200, 6400, 10000, 20000, 40000; the designated quantity is updated every 5 epochs). Notably, we utilize existing points as centers and sample additional points on the sphere, ensuring that the sampling radius equals the radius of the sphere during the upsampling process. Simultaneously, the radius for both sampling and rendering is systematically reduced by a factor of $\lambda_f = 0.75$ every 5 epochs. Over the subsequent 20 epochs, we configure the designated point quantity to be 80000 and 100000, with an update occurring every 10 epochs. Additionally, the reduction in epochs for both sampling and rendering is set at 10. During the final stage of training, we consistently upsample points and maintain the point number (100000) after pruning points each epoch. In the 61-100 epoch stage, both the sampling radius and rendering radius undergo a reduction by the same factor every 5 epochs. Beyond the 100th epoch, the sampling radius is maintained at a constant value of 0.004. In our rendering process, we integrate the scales of our Gaussian points with the rendering radius.

B.2 Network Architecture

We show the architecture of the initialized MLP in Fig. 14 and the deformation MLPs in Fig. 15. The initialized MLP, discussed in Sec. 3.1, serves as a Gaussian parameter prediction network. Given the mean position x_c, it outputs the rotation r_c, scale s_c, opacity o_c, and color c_c in the initialized space. The left segment of the deformation MLPs, introduced in both Sec.3.1 and Sec.3.2, delineates the motion process from the initialized space to the canonical space and ultimately to the deformed space, in terms of the mean position. Conversely, The right segment of deformation MLPs, detailed in Sec. 3.2, facilitates the deformation of the remaining Gaussian parameters from the initialized space to the deformed space.

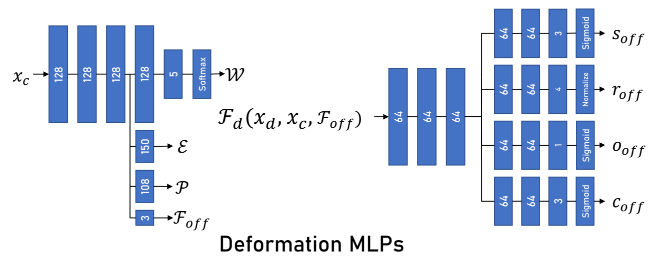

Fig. 15. **The Deformation MLPs** In the left segment of the deformation MLP, each linear layer is succeeded by weight normalization, and the activation function utilized is the Softplus, with the exception of the final layer. Conversely, in the right segment of the deformation MLP, each linear layer is succeeded by weight normalization, and the ReLU function serves as the activation function, except for the final layer.

B.3 Training Details

We show the loss weights as follows: we choose $\lambda_{\text{RGB}} = 1$, $\lambda_{\text{D-SSIM}} = 0.25$, $\lambda_{\text{flame}} = 1$, and $\lambda_{\text{vgg}} = 0.1$ for all of our experiments. For the flame loss, we set $\lambda_{\text{e}} = 1000$, $\lambda_{\text{p}} = 1000$, $\lambda_{\text{w}} = 1$. The training process is optimized using the Adam optimizer with a learning rate of $lr = 1e^{-4}$ and $\beta = (0.9, 0.999)$. Additionally, we implement a learning rate decay at the 80th and 100th epoch, employing a decay factor of 0.5. Moreover, we implement a decay of flame regularization at the 20th, 30th, 50th, and 70th epoch, employing a decay factor of 0.5.

B.4 Evaluation Details

Consistent with the approach employed in NHA [Grassal et al. 2022], we undertake fine-tuning of pre-tracked FLMAE [Li et al. 2017] expression and pose parameters both in the training and evaluation phases. The detailed loss weights during training are outlined in Sec. B.3 of the Supp. Mat. In the evaluation process, we exclusively employ the RGB loss.

B.5 Data Preprocessing

We adhere to the identical data preprocessing pipeline as employed in PointAvatar [Zheng et al. 2023a], which is derived from IMavatar [Zheng et al. 2022]. Additionally, we employ consistent camera and FLAME parameters across all methods. This ensures a fair comparison of head avatar methods, eliminating variations introduced by different face-tracking schemes during data preprocessing.

For the three human subjects captured by us, the initial preprocessing involves cropping the images to a square shape and resizing them to dimensions of both 512×512 and 1024×1024. Subsequently, we apply the data preprocessing pipeline mentioned above to further process the images

B.6 Ethics

We conducted experiments by capturing images of three human subjects using smartphones and additionally utilized data from three human subjects obtained from other datasets. For the 3 subjects captured by us, written consent was obtained from all subjects for the use of the captured images in this project. The data will be made publicly available for research purposes.