GAF: Gaussian Avatar Reconstruction from Monocular Videos via Multi-view Diffusion

Supplementary Material

In this supplementary material, we provide additional information about the dataset in Sec. A. Subsequently, we present more detailed explanations about method implementations in Sec. B, including parametric head tracking and multi-view head diffusion. Following that, we showcase the results of our multi-view head diffusion results in Sec C. Next, we provide additional comparisons in Sec D, including novel view synthesis, self-/cross-reenactment, and robustness analysis. Finally, we discuss the ethical considerations and potential negative impacts in Section E.

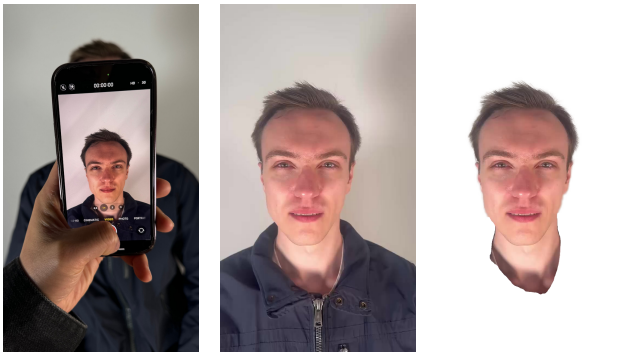
A. Dataset

Smartphone Video Capture. We capture monocular video sequences using an iPhone 14 Pro. The subject is seated in a chair, and the room lights are turned on during the recording, providing adequate illumination. The duration of the recording is about 10-15 seconds, at 30 frames per second. The image resolution is 1280×720.

Data preprocessing. To simplify the optimization process for animatable Gaussian splats, we integrate two preprocessing steps on raw images extracted from monocular videos. Firstly, we leverage the image matting techniques proposed in [5, 6] to remove the background. More specifically, we use [6] for our smartphone video capture, while we adopt [5] for the NeRSemble [2] dataset, where the initial background image is provided. Secondly, we utilize face segmentation maps acquired from BiSeNet [13] to isolate and crop out the torso portion, thus concentrating solely on head reconstruction. An example of our image preprocessing pipeline is illustrated in Fig. 1.

Train-Test Split. In the NeRSemble dataset, we use monocular videos from the 8-th camera as the input, only capturing the head from the front view. we evaluate head avatar reconstruction and animation quality in two settings: 1) *novel view synthesis*: driving a reconstructed avatar with seen head poses and expressions during training, and rendering it from 15 hold-out viewpoints. 2) *novel expression synthesis*: driving a reconstructed avatar with unseen head poses and expressions during training, rendering it from 5 nearby hold-out views, *i.e.* cameras 6–10. In Tab. 1, we provide detailed statistics about the used sequences in Nersemble and train/val/test split.

The statistics of the monocular video data are summarized in Tab. 2. Since monocular videos captured on commodity devices lack corresponding ground truths for novel view renderings, we only evaluate avatar animation performance in the quantitative comparisons, by applying pose and expression parameters from those unseen frames during training.



(a) Video capture (b) Raw image (c) Processed image

Figure 1. **Data capture and processing of monocular videos from smartphones.** We capture a short video using an iPhone 14 Pro. From the raw images, we remove the background using image matting techniques and segment out the torso to focus on the head region.

#Sequence Names	#timesteps × #cams		
	Train	Novel view	Novel expression
017 EXP-5	208×1	208×15	52×5
037 EXP-8	203×1	203×15	50×5
055 EMO-4	105×1	105×15	26×5
074 EXP-5	179×1	179×15	44×5
134 EMO-1	79×1	79×15	19×5
165 EXP-8	189×1	189×15	47×5
221 EXP-5	147×1	147×15	36×5
251 EMO-1	51×1	51×15	12×5
264 EMO-1	75×1	75×15	18×5
304 EMO-1	127×1	127×15	31×5
417 EMO-4	124×1	124×15	31×5
460 EXP-4	124×1	124×15	30×5

Table 1. **Statistics of the train/val/test splits used for NeRSemble sequences.** For each sequence, we use 80% of timesteps for the training and validation datasets. We select the 8th camera (front-facing) for the train split, while all remaining cameras are used for novel-view evaluation (validation set). The novel-expression evaluation is conducted by selecting 5 nearby cameras for the remaining 20% of timesteps (test set).

#Sequence Names	#timesteps	
	Train	Test
wojteck-1	760	2678
person0004	450	1050
subject1	229	48
subject2	312	83
subject3	139	34
subject4	440	154

Table 2. **Statistics of the train/test splits used for the Monocular Video dataset.** To effectively evaluate the ability of our method to represent unseen regions of the head, we select training frames with limited head rotation. The remaining frames, which contain unseen poses and expressions, are used as the test set.

B. Implementations

B.1. Monocular Head Tracking

We track the FLAME [4] parameters using the VHAP-tracker [1] proposed in [8]. Given a monocular video we optimize both shared parameters (shape, albedo map, diffuse light) and per-timestep parameters (pose, translation, expression). The tracking algorithm is divided into three stages: (i) initialization stage; (ii) sequential optimization stage; (iii) global optimization stage. The tracking process begins with an initialization stage, performed on the first frame of the video, which sets up all the aforementioned parameters. Following this, a sequential optimization stage is applied to each successive frame of the video. In this stage, the parameters of each frame are optimized for 50 iterations, using the previous timestep as initialization. Finally, the tracking parameters are refined through a global optimization stage, where a random frame is sampled at each iteration, for a total of 30 epochs.

The tracking is performed by minimizing a combination of multiple energy terms: (i) a photometric energy term, computed between the rendered image and the ground-truth one; (ii) a landmarks energy term, which computes the distance between the projected 2D FLAME [4] landmarks and the 2D landmarks predicted by an off-the-shelf detector [15]; (iii) temporal energy terms, applied on the per–timestep parameters, which ensure smoothness over time; (iv) regularization energy terms, applied on all FLAME [4] parameters. We revised the loss weights for the smoothness terms as: $\lambda_{smooth,transl} = 3e4$, $\lambda_{smooth,rot} = 3e3$, $\lambda_{smooth,jaw} = 4.0$, $\lambda_{smooth,eyes} = 1.0$, $\lambda_{smooth,expr} = 0.5$. For all the remaining hyper-parameters we refer to the original work [1].

We use NVDiffRast [3] as the differentiable mesh renderer and the FLAME 2023 version [4] with the additional 168 triangles to represent the teeth, as proposed by [8].

B.2. Multi-view Latent Head Diffusion

In Fig. 2, we show the network architecture details of our multi-view head latent diffusion. The denoiser network is based on a 2D U-Net [9] with attention blocks [11]. The U-Net comprises four Down Blocks, one Middle Block, and four Up Blocks. Each Down Block contains a Residual block, a 3D Attention block, and a Downsampling layer. The Middle Block is composed of a Residual block and a 3D Attention block. The Up Block mirrors the Down blocks but with Upsampling layers.

B.3. Gaussian Regularizations

The position regularization term ensures that Gaussians remain close to their attached triangles during optimization through:

$$\mathcal{L}_{pos} = \|max(\mu, \epsilon_{pos})\|_2 \tag{1}$$

where $\epsilon_{pos} = 1$ serves as the threshold, allowing small positional errors within the scaling of the attached triangle. The scale regularization term mitigates the formation of large Gaussians, which could lead to jittering problems due to small rotations of triangles.

$$\mathcal{L}_{scale} = \|max(\mathbf{s}, \epsilon_{scale})\|_2 \tag{2}$$

It will be disabled when the local scale of the Gaussian w.r.t the attached triangle is less than $\epsilon_{pos} = 0.6$.

B.4. Ablation Studies of Gaussian Avatars Fusion

In the main paper, we evaluate various diffusion priors for novel view constraints, demonstrating the effectiveness of our face-specific multi-view diffusion priors for 3D Gaussian head reconstruction from monocular videos. Here, we provide additional details on the implementation of our ablation studies. We use six sequences from the NeRSemble dataset, including '055 EXP-5', '098 EMO-1', '134 EMO-1', '165 EMO-1', '221 EXP-8', and '417 EMO-4'.

No diffusion. This variant does not apply any priors to constrain novel view renderings. It is implemented using GA [8] with SH =0.

Pretrained Stable Diffusion. This variant randomly renders a novel view at each iteration and refines it using Stable Diffusion 2.1, guided by ControlNet [14] with normal maps. The iteratively denoised images serve as pseudo-ground truths.

Personalized Stable Diffusion. Instead of using a pretrained model, this variant employs DreamBooth [10] to fine-tune the U-Net and text encoder with a learning rate of 5e−6 for 500 iterations. The iteratively denoised images are used as pseudo-supervision.

Pose-conditioned multi-view diffusion. This variant uses pose embedding-conditioned multi-view diffusion models to generate pseudo-ground truths.

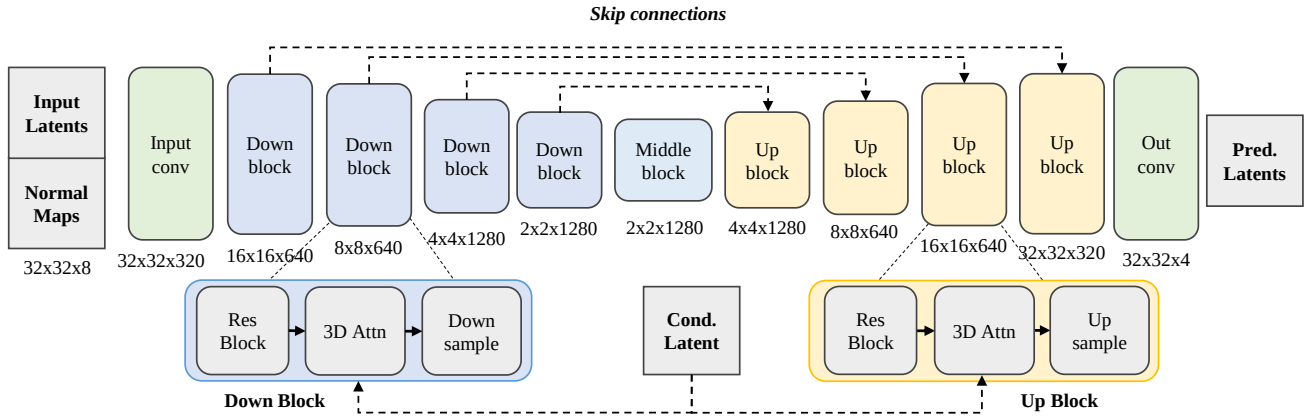


Figure 2. **Network architecture of our multi-view head latent diffusion model.** The denoiser network is built on a 2D U-Net architecture with attention blocks. The input consists of multi-view image latents concatenated with VAE latents of normal maps rendered from the FLAME mesh. The 3D Attention block enforces 3D consistency by applying cross-attention across all views. It also incorporates the input image latent into the denoising process, effectively preserving the identity and appearance details of the input portrait.

Raymap-conditioned multi-view diffusion. This variant uses ray map-conditioned multi-view diffusion models to generate pseudo-ground truths.

Our multi-view diffusion using Score Distillation Sampling (SDS) loss. Instead of using iteratively denoised images as pseudo-ground truths, this variant employs SDS loss [7] based on single-step denoising.

Ours without latent upsampler $\times 2$. We remove the pretrained latent upsampler $\times 2$. Then the resolution of pseudo ground truths is 256×256.

Ours without 3D-aware denoising. The pseudo-ground truths generated by diffusion models rely on multi-view renderings of 3D Gaussian avatars, embedding 3D awareness into the diffusion process. To eliminate this 3D awareness, we set the time step to 0 when adding noise. Comparisons in Tab. 3 demonstrate the benefits of 3D-aware denoising.

Method	Novel Views			Novel Expressions		
	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
Ours, w/o 3D-aware	0.119	21.52	89.71	**0.079**	25.20	91.99
Ours final	**0.118**	**21.82**	**89.87**	**0.079**	**25.39**	**92.02**

Table 3. Ablation study of 3D-aware denoising.

C. Results of Multi-view Head Diffusion

In Fig. 3, we showcase the sampling results from our multi-view head diffusion model. The model generates four view-consistent images from a single input image while effectively preserving facial identity and appearance. This demonstrates the model's capability to synthesize coherent and identity-preserving novel views.

D. Additional Comparisons

D.1. NeRSemble Dataset

In Fig. 4, we provide additional qualitative comparisons on dynamic head avatar reconstruction from monocular videos sampled from NerSemble dataset [2].

D.2. Monocular Videos on Commodity Devices

In Fig. 5, we provide additional qualitative comparisons against INSTA [16], FlashAvatar [12], and GA [8] on monocular videos captured by commodity devices.

D.3. Self- & Cross-Reenactment

We show the self and cross reenactment results of our method and Gaussian Avatars in Fig. 6 and 7.

D.4. Robustness Analysis

To demonstrate the robustness of our method with sparse input data, we evaluate reconstruction performance across different frame numbers in the input video. We use the '104 EMO-1' sequence from the NeRSemble dataset, which contains 56 frames in the input. To reduce the frame count, we sample keyframes at uniform intervals. For instance, for an 8-frame input, we select frames at timesteps 0, 7, 14, 21, 28, 35, 42, and 55; for a single-frame input, we use only the 28th frame. As shown in Fig. 9 , our method can maintain stable quantitative performance with as few as 8 frames, while GaussianAvatars drops dramatically. This highlights the resilience of our method to limited observations.

In Fig. 9, we present qualitative results from a robustness analysis conducted with varying numbers of frames in the input monocular videos. Our approach consistently achieves photorealistic novel view rendering across various sequence lengths, even with only 8 frames as input.

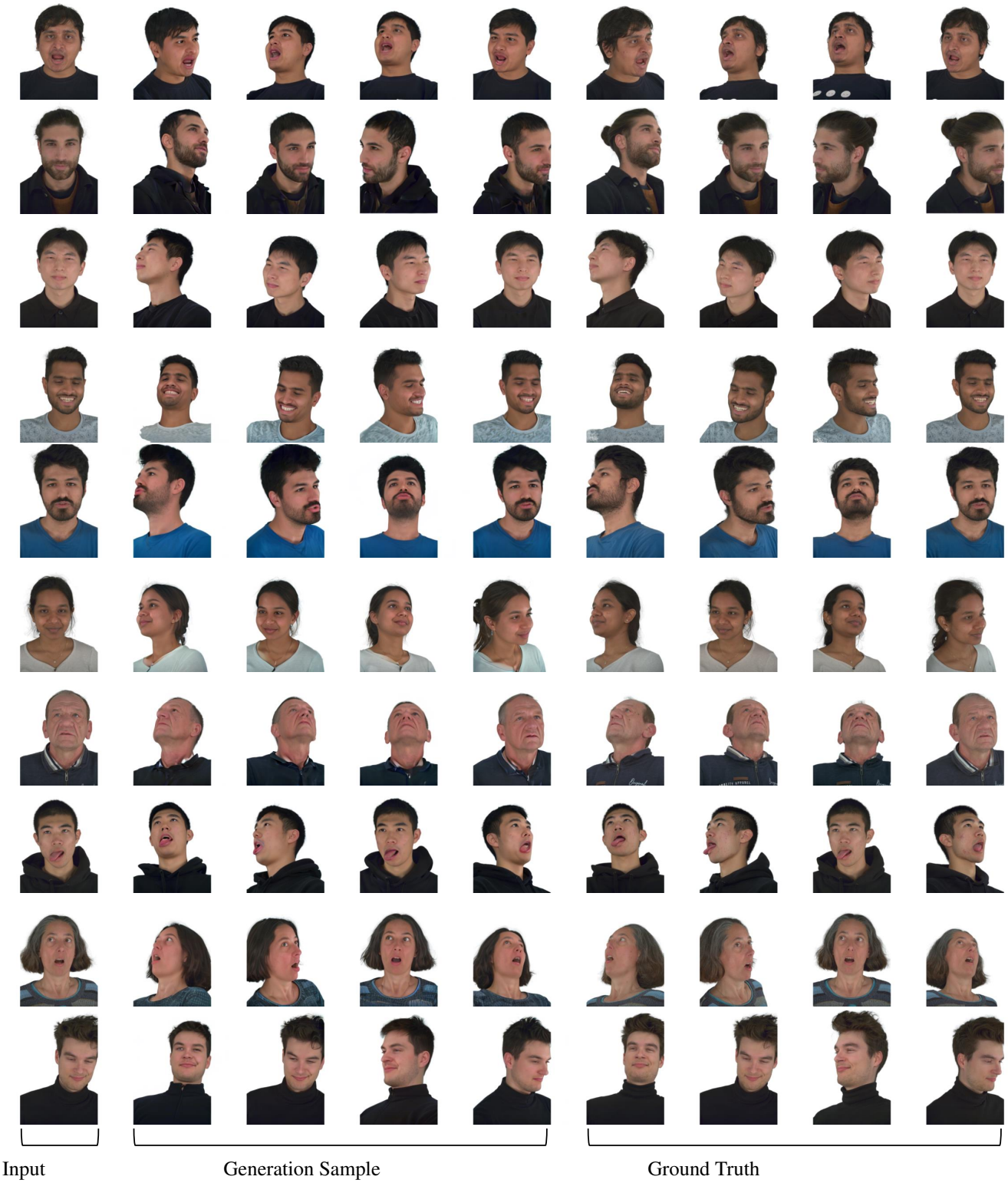


Figure 3. **Generation sample results of the multi-view head latent diffusion model.** Given a single image as input, our method can generate identity-preserved, view-consistent multi-view portrait images.

Figure 4. **Additional results on novel view synthesis from monocular videos from the NeRSemble dataset.** Our method demonstrates robust reconstruction of less observed regions (e.g., side facial areas), maintains facial identities across viewpoints, and consistently produces more plausible and view-consistent renderings from hold-out views.

Figure 5. **Additional comparisons of head avatar reconstruction from monocular videos on commodity devices.** We present novel expression animation results using unseen frames from the monocular videos during training. Additionally, we display the fitting results for the input frame (top right) and novel view renderings of the same frame (bottom right). While all methods accurately fit observed frames in front-facing sequences with limited head poses, baseline methods fail to generalize to novel views and poses due to the absence of effective priors for less unobserved regions.

E. Ethical Discussion and Negative Impacts

The creation of photorealistic and animatable head avatars from an input video poses several ethical challenges and significant risks related to the possible malevolent usage of this technology. One major concern is the potential for misuse in creating deepfakes, which are highly realistic but fake videos that can be used to spread misinformation, manipulate public opinion, or damage reputations. Additionally, this technology can lead to privacy violations, as individuals' likenesses can be replicated without their consent, lead-

(a) Driving Expression (Ground truth) (b) Gaussian Avatars (c) Ours

Figure 6. **Self-reenactment from monocular videos on the NeRSemble dataset.** We use the tracked FLAME pose and expressions of a driving sequence to animate the reconstructed Gaussians. We show three novel view renderings for each reenactment result. Our method demonstrates more plausible head animations through more detailed face reconstruction, such as wrinkles, and faithfully produces view-consistent head renderings from different novel viewpoints.

(a) Driving Expression (Reference) (b) Gaussian Avatars (c) Ours

Figure 7. **Cross-reenactment from monocular videos on the Nersemble dataset.** We show three novel view renderings for each reenactment result. Our method outperforms Gaussian Avatars by showcasing more vivid expression transfers and more plausible renderings around the mouth.

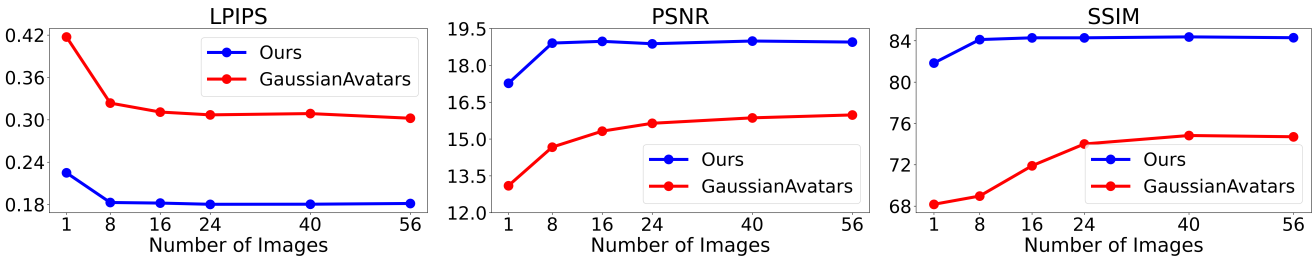


Figure 8. **Robustness analysis to the number of frames in the input monocular video.**

Figure 9. **Robustness analysis to the number of frames in the input monocular video.** (a) Input view (top) & Ground truth (bottom); (b) 56 frames; (c) 40 frames; (d) 24 frames; (e) 16 frames; (f) 8 frames; (g) 1 frame. Compared to GaussianAvatars [8], our method demonstrates robust reconstruction of novel view synthesis even with as few as eight frames, highlighting its robustness to limited observations.

ing to unauthorized use in various contexts. There is also the risk of identity theft, where malicious actors could use these avatars to impersonate others for fraudulent activities. Moreover, the psychological impact on individuals who see their digital likeness used inappropriately can be profound, causing distress and harm. Our commitment is to promote the responsible and ethical use of this technology, and we are firmly against any malicious usage that aims to harm individuals or communities.

References

[1] Versatile head alignment with adaptive appearance priors. https://github.com/ShenhanQian/VHAP. 2

[2] 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, 3

[3] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics*, 39(6), 2020. 2

[4] 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. 2

[5] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 1

[6] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 1

[7] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[8] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. *arXiv preprint arXiv:2312.02069*, 2023. 2, 3, 7

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 2

[10] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 2

[11] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[12] Jun Xiang, Xuan Gao, Yudong Guo, and Juyong Zhang. Flashavatar: High-fidelity digital avatar rendering at 300fps. *arXiv preprint arXiv:2312.02214*, 2023. 3

[13] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 1

[14] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[15] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15475–15484, 2023. 2

[16] 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. 3