Animatable Gaussians: Learning Pose-dependent Gaussian Maps for High-fidelity Human Avatar Modeling

Supplementary Material

In this supplemental document, we will show implementation & experiment details, more results and additional experiments.

A. Implementation Details

Template Reconstruction. We optimize an SDF and color field represented by an MLP consisting of intermediate layers with (512, 256, 256, 256, 256, 256) neurons. Given a posed point, we find accurate correspondence in the canonical space by root finding. Following ARAH [12], we initialize the correspondence as the canonical position that is computed by inverse skinning based on blending weights of the closest SMPL vertex. Different from SNARF [2] and ARAH [12] that utilize the Broyden's method [1] to solve Eq. 3, we employ the Gauss-Newton method by implementing a customized CUDA kernel. The training loss of template reconstruction involves an RGB loss, a mask loss and an Eikonal loss [3].

Network Architecture. The network in our avatar representation is composed of StyleUNet [11], a conditional StyleGAN-based [6] generator. Differently, we adapt the original StyleUNet by incorporating two decoders to predict both front & back Gaussian maps. The resolution of the input position map is 512×512, and the resolution of the output Gaussian maps is 1024×1024. Specifically, we utilize three different StyleUNets to output color (3-channel), position (3-channel) and other Gaussian attributes (8-channel). In the color StyleUNet, we modulate the color output with a view direction map to model view-dependent effects. Each pixel on the view direction map indicates the angle between the view direction and the template normal. The view direction map is encoded through a tiny CNN, then the encoded feature map is injected into an intermediate decoder layer of the color StyleUNet.

Training. We adopt the Adam optimizer [7] for training the StyleUNet with a learning rate of 5×10^{-4}. The loss weights are set as: $\lambda_{\text{perceptual}} = 0.1$, $\lambda_{\text{reg}} = 0.005$. The batch size is 1, the total iteration number is 500k, and the training procedure takes about two days on one RTX 4090.

B. Experiment Details

Metric Evaluation. We utilize PSNR, SSIM [13], LPIPS [14] and FID [4] as the metrics for quantitative evaluations. PSNR and SSIM are computed on the entire image at the original resolution, while LPIPS and FID are computed on the cropped minimal square that covers the human body.

Table A. **Quantitative ablation study on the parametric template.**

	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
Parametric Template	**31.2183**	**0.9858**	**0.0344**	**36.9905**
SMPL-X	30.5241	0.9842	0.0401	47.5066

Table B. **Quantitative comparison between representations with different backbones.**

	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
StyleUNet [11]	**29.3127**	**0.9664**	**0.0378**	**27.3143**
U-Net [10]	26.4255	0.9435	0.0507	31.3838
MLP	26.8961	0.9497	0.0650	87.0793

Table C. **Quantitative ablation study on pose projection.**

	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
w Pose Proj.	**24.9932**	**0.9285**	**0.0685**	**45.6266**
w/o Pose Proj.	23.5594	0.9189	0.0792	59.9083

Comparison with Body-only Avatars. The quantitative comparison (Tab. 1 in the main paper) is conducted on the "subject00" sequence of THuman4.0 dataset [15]. The first 2000 frames are utilized as the training dataset, and the numerical results of Tab. 1 are evaluated on the rest 500 frames and the "cam18" camera view.

Comparison with Full-body Avatars. The quantitative comparison (Tab. 2 in the main paper) is conducted on the "avatarrex_zzr" sequence of AvatarReX dataset [16]. The numerical results of Tab. 2 are evaluated on the first 500 frames and the "22010710" camera view.

C. Additional Sequential Results

Fig. A shows additional sequential results animated by challenging out-of-distribution pose sequences from the AMASS dataset [8], including basketball, football and dancing. It demonstrates that our method can also generate realistic and reasonable dynamic details under out-of-distribution poses thanks to the effective avatar representation and pose projection strategy.

D. Additional Experiments

Quantitative Ablation Study on Parametric Template. We conduct a qualitative ablation study on the parametric template in Fig. 7 of the main paper. We also quantitatively compare the reconstructed parametric template with the naked SMPL-X model [9] on the animation accuracy

Figure A. **Example sequential animation results by our method.** Each row is an animation sequence involving 3 subjects. Our method can generate realistic and reasonable dynamic details even under novel poses from the AMASS dataset [8].

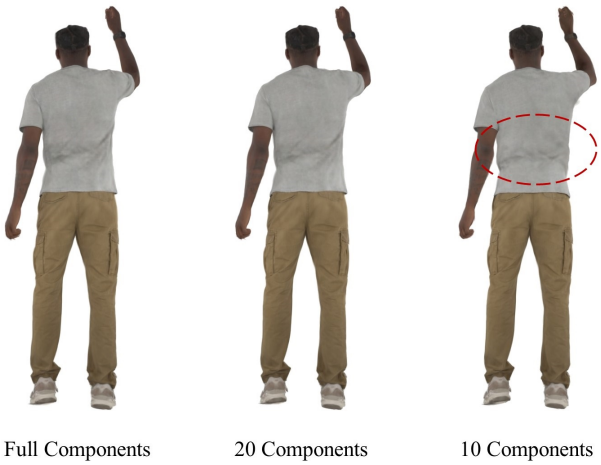

Figure B. **Ablation study on the component number in the pose projection.**

in Tab. A. The numerical results are computed on the first 500 frames and the "22010708" camera view in the "avatarrex_lbn1" sequence from AvatarReX dataset [16].

Quantitative Ablation Study on Backbones. We additionally report quantitative results of StyleUNet [11], UNet [10] and MLPs in Tab. B to further prove the effectiveness of the introduction of 2D CNNs and 2D parameterization. The numerical results are computed on the 60-560 frames and the "Cam127" camera view in the "Actor02/Sequence1" from ActorsHQ dataset [5].

Quantitative Ablation Study on Pose Projection. We quantitatively ablate pose projection in Tab. C. It shows pose projection realizes more accurate animation under novel poses. The numerical results are computed on the 1109-1119, 1447, 1515 and 1578 frames where the testing poses are out of the distribution of training poses.

Number of Principal Components in Pose Projection. Fig. B shows the animation results with different numbers of principal components in the pose projection strategy. It demonstrates that although PCA can project a novel pose into the distribution of the training poses for better pose generalization as shown in Fig. 9 of the main paper, too few principal components may lose some fine-grained garment details. We empirically found that setting the number of principal components to 20 could produce both detailed and generalized animation.

View Number. We quantitatively and qualitatively show the animation results trained with 3 views, 6 views and 14 views in Tab. D and Fig. C. They demonstrate that our method also supports sparse-view input and can realize comparable high-fidelity results. The numerical results are evaluated on the first 500 frames and the "22070932" camera view of the "avatarrex_zzr" sequence from AvatarReX dataset [16].

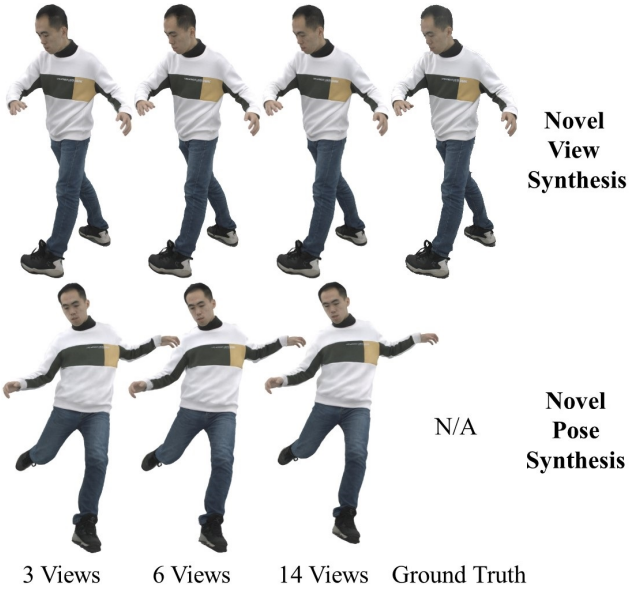

Figure C. **Animation results trained with different numbers of views.**

Table D. **Quantitative evaluation on different view numbers.**

	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
3 Views	30.6123	0.9807	0.0306	11.3066
6 Views	30.3565	0.9803	0.0310	10.9966
14 Views	30.7622	0.9816	0.0297	10.6744

Figure D. **Failure cases.** (a,c) Animation results by our method, (b,d) ground-truth images. Our method fails to model the motion of hairs.

E. Failure Cases

Our method cannot model the physical motion of components that are not driven by the body joints, e.g., the hairs, as illustrated in Fig. D, since we model the whole body including clothes, hands and hairs as an entangled Gaussian representation. We leave for future work a disentangled and compositional representation for modeling the dynamics of different components of the character.

References

[1] Charles G Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of computation*, 19 (92):577–593, 1965. 1

[2] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, pages 11594–11604, 2021. 1

[3] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, pages 3789–3799. PMLR, 2020. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30, 2017. 1

[5] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *TOG*, 42(4):1–12, 2023. 3

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[8] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *ICCV*, pages 5442–5451, 2019. 1, 2

[9] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 1

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1, 3

[11] Lizhen Wang, Xiaochen Zhao, Jingxiang Sun, Yuxiang Zhang, Hongwen Zhang, Tao Yu, and Yebin Liu. Styleavatar: Real-time photo-realistic portrait avatar from a single video. In *SIGGRAPH Conference Proceedings*, 2023. 1, 3

[12] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *ECCV*, pages 1–19. Springer, 2022. 1

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE T-IP*, 13(4):600–612, 2004. 1

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 1

[15] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, pages 15893–15903, 2022. 1

[16] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *TOG*, 42(4), 2023. 1, 3