Neural Capture of Animatble 3D Human from Monocular Video
- Supplemental Material -

Gusi Te[1,2], Xiu Li[2,3], Xiao Li[2], Jinglu Wang[2], Wei Hu[1], and Yan Lu[2]

[1] Peking University
[2] Microsoft Research Asia
[3] Tencent

This supplemental material includes additional results, discussion, and applications to further demonstrate the advantages of our method. We organize this supplemental material in 2 parts: Section 1 includes additional implementation details of our training pipeline, as detailed in the main submission. Section 2 includes additional results of our human modeling, as well as more ablation studies and comparisons. We also provide a video in the supplemental materials to visualize our synthesized results in animation.

1 Implementation details

Training details on each dataset. For one batch, we only select one single image from the training dataset. During training, we sample 6000 rays and 32 points along each ray for each image and utilize a two-stage sampling strategy for better efficiency as in [3]. Similar to [3], we implement two separate networks with different weights: the first is to estimate spatial density, and the second is to sample more intensive points with importance sampling. We hence aggregate 64 points for rendering. For datasets without background masks available, we either apply an off-the-shelf matting algorithm or jointly model the background during training. Specifically, we apply the off-the-shelf image matting method of [2] on the Human3.6M [1] dataset and train our method on the foreground part only. For the Doublefusion [5] dataset, we found the matting results are not good enough and opt to jointly model a single background image for all training frames by a small MLP network of 3 layers during training. The background MLP network takes a 2D pixel position as input and outputs the background image pixel value. We apply positional encoding to the input.

2 Additional experiment results.

More pose synthesis results. We show additional pose synthesis results of our method in Fig. 1. In the top rows the images are rendered from different views, and the bottom images are rendered given novel poses on the Doublefusion dataset. The query embedding allows our model to represent the human body and clothes given arbitrary pose and viewpoint. We refer to our supplemental video for more animated pose synthesis results.

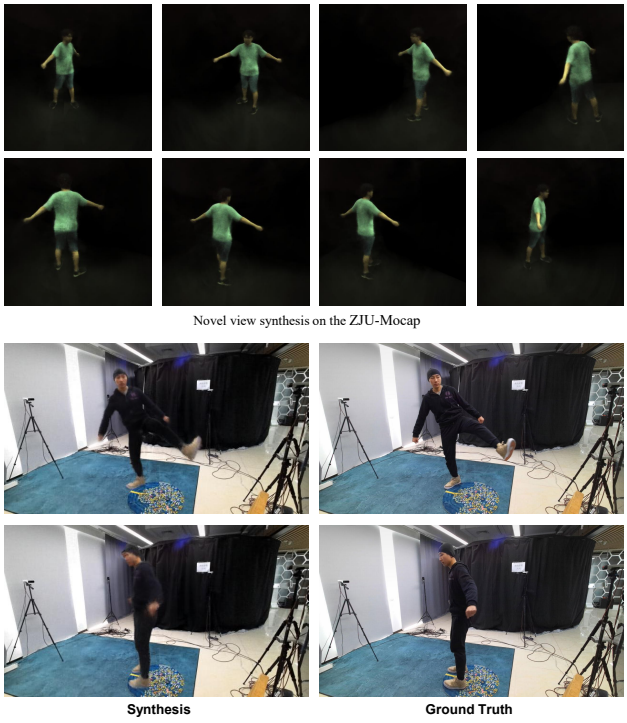

Fig. 1. Additional novel view and pose synthesis results on ZJU-Mocap and Doublefusion dataset.

Ablation - latent code. We test the impact of latent code by removing latent code on the mesh vertices. The comparison is shown in Table 1, where it is observed that the latent codes bring a lot of improvements to the full model.

		full	w/o latent
SSIM	Training View	**0.980**	0.957
	Novel View	**0.973**	0.950
PSNR	Training View	**35.87**	32.16
	Novel View	**34.75**	31.35

Table 1. Impact of latent code.

Comparison with Neural Body (CVPR 2021). Compared with Neural Body [4], our method achieves much better generalization on novel poses. Figure 2 shows a example case on novel pose synthesis. Our model faithfully reconstructs the human under novel pose (right), while NeuralBody completely fails to render the human appearance (left). We also report the quantitative results

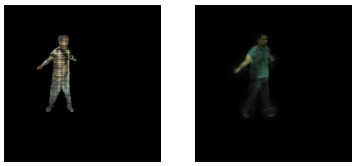

Fig. 2. Rendering results of our method and NeuralBody [4]

with NeuralBody and its successor work AniNeRF in Tab 2. The superiority of AniNeRF over NeuralBody [4] for the generalization ability (i.e., novel pose synthesis) has been also extensively validated in the AniNeRF paper.

	NeuralBody	AniNeRF	Ours
PSNR ↑	27.94	29.11	**34.75**
SSIM ↑	0.930	0.948	**0.973**

Table 2. Quantitative comparison with NeuralBody and AniNeRF on People-Snapshot dataset.

References

1. Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 1
2. Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 1
3. Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1
4. Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2, 3
5. Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7287–7296, 2018. 1