EgoRenderer: Rendering Human Avatars from Egocentric Camera Images (Supplementary)

Tao Hu[1]*, Kripasindhu Sarkar[2], Lingjie Liu[2], Matthias Zwicker[1], Christian Theobalt[2]
[1]Department of Computer Science, University of Maryland, College Park
[2]Max Plank Institute for Informatics, Saarland Informatics Campus

This supplementary material provides additional technical details for the main paper.

1. Dataset

More details of the proposed synthetic dataset. To render our synthetic dataset, we animated characters using the SMPL model [3] with around 3000 different motions sampled from the CMU MoCap [1] dataset. More than 600 body textures were randomly chosen from the texture set provided by the SURREAL [5] dataset. In total, we rendered 178,800 images for training.

Dataset Pre-processing After we captured datasets, we first synchronized the egocentric fisheye camera and multi-view cameras to register them in time, and used the pre-trained PointRend [2] for foreground segmentation.

2. More Experimental Results

2.1. Quantitative Comparisons

Single-video comparisons. We also provide the additional L1 distances (Table 1) of each method on single-video datasets.

	Im-Tex	Pix2PixHD [6]	Ex-Tex	Only-Ego	Only-MV	Fea-Net [4]
H1	0.994	0.994	**0.950**	0.995	0.999	0.998
H2	1.191	1.199	1.215	1.233	1.224	**1.175**
H3	**1.448**	1.522	1.562	1.484	1.516	1.511
H4	**0.894**	0.906	0.944	0.911	0.905	0.926

Table 1: L1 distances of single-video training on different datasets. Numbers are multiplied by 10

Multi-video comparisons. The L1 distances of multi-video experiments for indoor (H1) and outdoor scenes (H4) are provided in Table 2, where each method was trained on multiple videos from different viewpoints, 9 multi-view cameras for H1, and 4 multi-view cameras for H4.

	Im-Tex	Pix2PixHD [6]	Ex-Tex	Only-Ego	Only-MV	Fea-Net [4]
H1	0.645	0.650	0.688	0.651	0.646	**0.641**
H2	**0.767**	0.791	0.836	0.790	0.817	0.834

Table 2: L1 distances of multi-video training on H1 outdoor and H4 indoor datasets. Numbers are multiplied by 10

2.2. Local and Global Coordinate System

Our system, EgoRenderer can work in both local (human-centered coordinate) and global coordinate systems (Figure 1). For local system (left), we assume each user-specific viewpoint is relative to the human. For global system (right), we

*Work partly conducted during TH's internship at MPI-INF

synthesize global target poses by integrating local poses estimated by Mo2Cap2 and global tracking by external devices. More results can be found in the video demo.

(a) Renderings at timestamp 1.

(b) Renderings at another timestamp.

Figure 1: EgoRenderer can work in both local (left) and global coordinate systems (right).

References

[1] Carnegie Mellon University Motion Capture Database. http://mocap. cs.cmu.edu/. 1

[2] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9796–9805, 2020. 1

[3] M. Loper, Naureen Mahmood, J. Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015. 1

[4] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[5] G. Varol, J. Romero, X. Martin, Naureen Mahmood, Michael J. Black, I. Laptev, and C. Schmid. Learning from synthetic humans. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4627–4635, 2017. 1

[6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. pages 8798–8807, 06 2018. 1