Supplementary Document for: SHERF: Generalizable Human NeRF from a Single Image

Shoukang Hu[1*] Fangzhou Hong[1*] Liang Pan[1] Haiyi Mei[2] Lei Yang[2] Ziwei Liu[1]
[1]S-Lab, Nanyang Technological University [2]Sensetime Research

1. Implementation Details

1.1. More Implementation Details of SHERF

SHERF model is trained with images from different actors at the same time. For example, when sampling data pairs from THuman (90 subjects $\times$ 20 poses $\times$ 24 views), we randomly sample one input and one target image from the same subject. To render the target image during training and evaluation, we randomly sample an input image from given camera views and sample 48 points for each the ray belong to the human region bound box part at the target space. During the optimization, we use the Adam [2] optimizer. We set the initial learning rate as 2×10^{-3} and decay the learning rate by a factor of 0.5 for every epoch. The maximum iteration number is set as 5 epochs.

1.2. Novel Pose Synthesis of NHP

As discussed in the main paper, there lacks a clear framework to synthesize novel poses in NHP [3] as it models the neural radiance field in the canonical space. In this work, we synthesis novel pose results of NHP by using the Linear Blend Skinning of SMPL algorithm. Specifically, we transform the 3D sampled points from the target space to observation space and query the corresponding features. Then queried features, along with the coordinates of 3D sampled points and ray directions in the target space, are fed into the NeRF decoder to predict density $\boldsymbol{\sigma}$ and RGB $\boldsymbol{c}$ values.

2. Analysis on SMPL and Camera Parameters Estimated from a 2D Input Image.

Current human NeRF methods, including multi-view images or monocular video settings, rely on accurate SMPL parameters. In our experiments, we also use accurate SMPL and camera parameters provided by the datasets. Recently, single-view SMPL estimation methods have made great progress and are reliable. To verify the effectiveness of our proposed SHERF in real-world scenarios with only one single 2D image available and no accurate SMPL and camera parameters available, we use SMPL and camera parameters predicted by CLIFF [4] to evaluate the performance on the RenderPeople dataset. As shown in the Tab. 1 and Fig. 1, SHERF produces plausible results and surpasses baseline methods.

Table 1: Performance (PSNR, SSIM and LPIPS) comparison with SMPL and camera parameters estimated from a 2D input image among NHP, MPS-NeRF and our SHERF method on the RenderPeople dataset.

Method	Novel View			Novel Pose		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NHP [3]	18.04	0.72	0.31	17.59	0.70	0.33
MPS-NeRF [1]	17.81	0.74	0.30	17.33	0.71	0.32
SHERF (Ours)	**19.64**	**0.79**	**0.22**	**19.22**	**0.78**	**0.24**

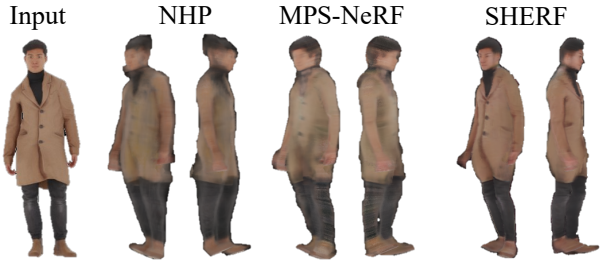

Figure 1: Visualization results with SMPL and camera parameters estimated from a 2D input image among NHP, MPS-NeRF and our SHERF method on the RenderPeople dataset.

3. More Qualitative Results

3.1. Models Trained with Free View Inputs

More qualitative results with different viewing angles as inputs on test subjects of THuman are shown in Fig. 2 - Fig. 3. The models are trained with free viewing angles as inputs on training subjects of THuman. Two main trends can be observed. 1) NHP [3] tends to render images with smoothed effects in face and cloth, failing to produce realistic image details. MPS-NeRF [1] can somehow produce image details, but still suffers from recovering face details. Thanks to the

bank of hierarchical features, our SHERF can render more realistic images with details in face and cloth when compared with NHP and MPS-NeRF. 2) When given the front viewing angle input, NHP and MPS-NeRF overfit to the cloth patterns of the front view input image when synthesizing the back view output image while our SHERF can learn to synthesize images with more acceptable results. 3) When given the back viewing angle input, NHP and MPS-NeRF fails to render images with reasonable face details especially for the front viewing angle output, while our SHERF can generate results with acceptable image quality. For more qualitative results in RenderPeople data set, please refer to our demo video.

3.2. Models Trained with Front View Inputs

In the analysis part, we show that models trained with front view inputs are not suitable for the real-world scenarios where human images are captured individually from a random camera viewing angle. To further support our claim, we show qualitative results with different viewing angles as inputs on models trained only with front view inputs of THuman. As shown in Fig. 4 - Fig. 5, although all three methods can produce good results with front view inputs, the image quality degrades significantly when other free viewing angle inputs are provided. For example, when given the back viewing angle input, almost no reasonable results can be produced. Even in the front view input setting, SHERF still produces better results when compared with two SOTA baseline methods.

References

[1] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-nerf: Generalizable 3d human rendering from multiview images. *arXiv preprint arXiv:2203.16875*, 2022. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[3] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34:24741–24752, 2021. 1

[4] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 1

Figure 2: More qualitative results with different viewing angles as inputs on test subjects of THuman.

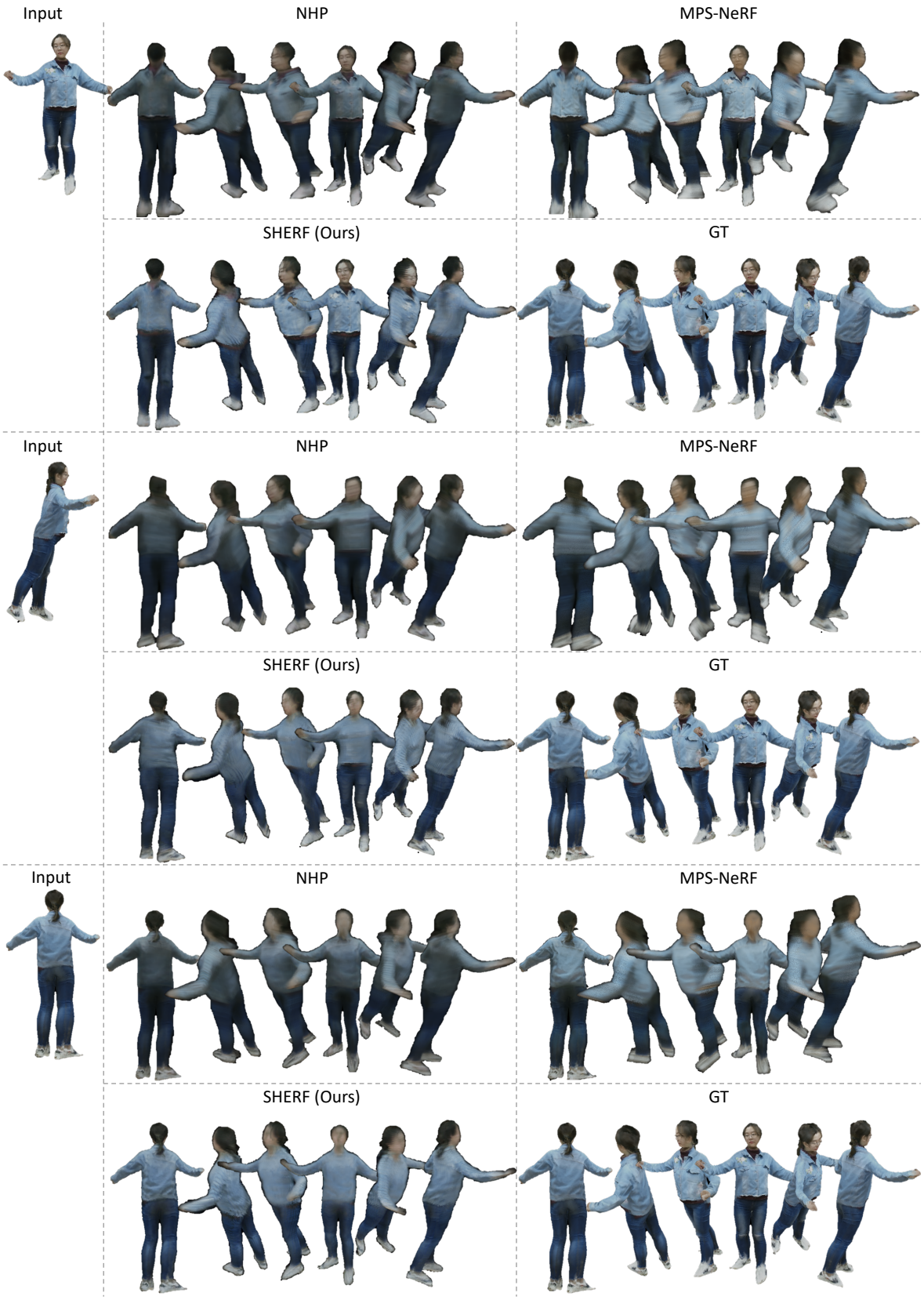

Figure 3: More qualitative results with different viewing angles as inputs on test subjects of THuman.

Figure 4: More qualitative results with different viewing angles as inputs on models trained only with front view inputs of THuman.

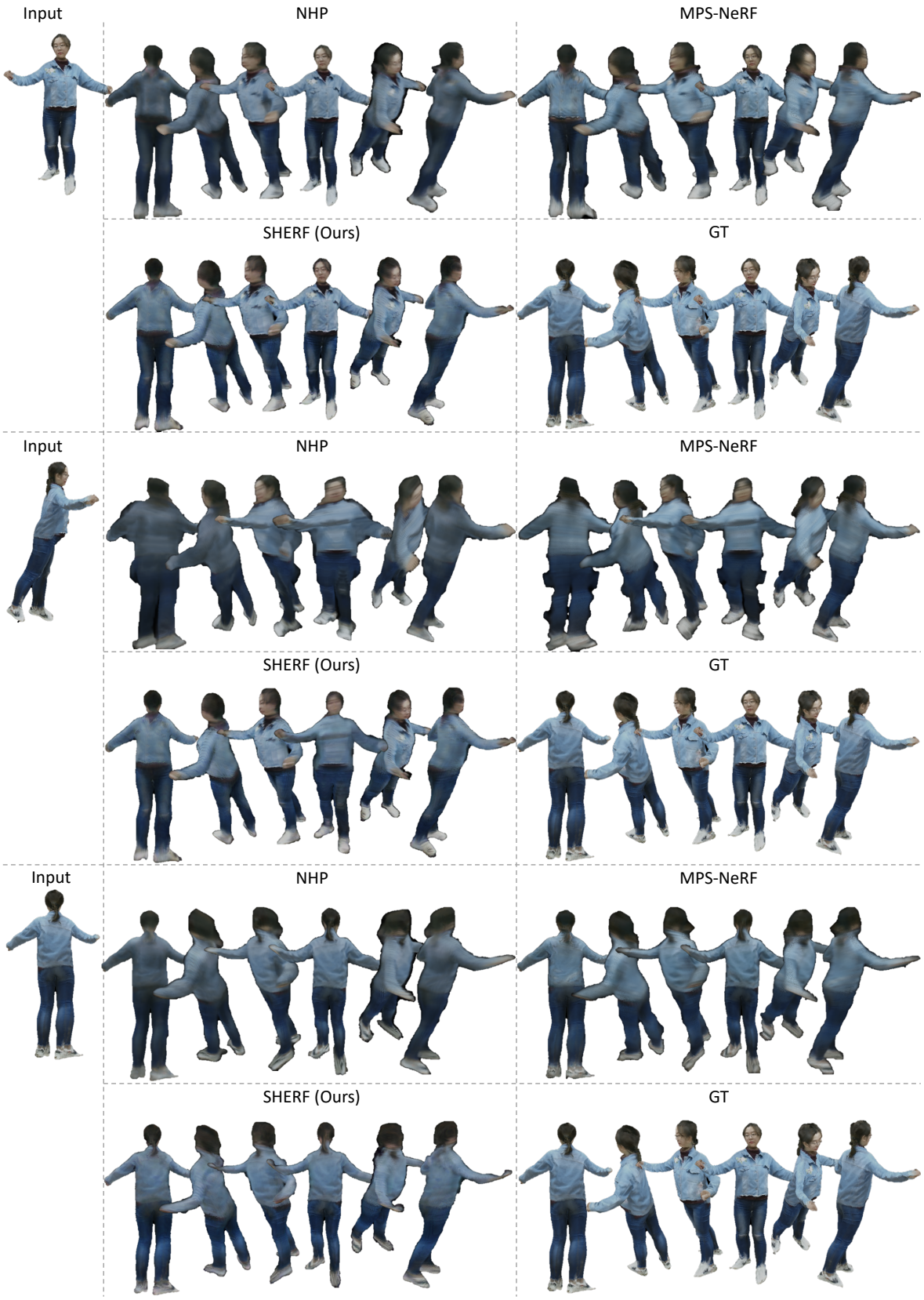

Figure 5: More qualitative results with different viewing angles as inputs on models trained only with front view inputs of THuman.