Supplementary Document for
LPSNet: End-to-End Human Pose and Shape Estimation with Lensless Imaging

Haoyang Ge[1,†], Qiao Feng[1,†], Hailong Jia[1], Xiongzheng Li[1], Xiangjun Yin[1],
You Zhou[1], Jingyu Yang[1], Kun Li[1,*]
[1]Tianjin University, China [2]Nanjing University, China
{ghy0623,fengqiao,jhl,lxz,yinxiangjun,yjy,lik}@tju.edu.cn zhouyou@nju.edu.cn

In this document, we provide the following supplementary contents:

- Training Loss;
- Evaluation;
- Experimental Results on Simulated Dataset;
- Failure Cases and Limitations.

We also provide a demo video on our project page.

1. Training Loss

We introduce multiple training losses as supervision for LPSNet: the reconstruction loss, the Kullback-Leibler divergence loss, and the environmental constraint loss. Formally, the entire training loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{das}} . \tag{1}$$

- **Regressor Loss $\mathcal{L}_{\text{reg}}$.** This loss is formulated as:

$$\mathcal{L}_{\text{reg}} = \lambda_{\text{2D}} \mathcal{L}_{\text{2D}} + \lambda_{\text{3D}} \mathcal{L}_{\text{3D}} + \lambda_{\text{para}} \mathcal{L}_{\text{para}} , \tag{2}$$

where $\mathcal{L}_{\text{2D}}$, $\mathcal{L}_{\text{2D}}$ and $\lambda_{\text{para}} \mathcal{L}_{\text{para}}$ denote the 2D keypoints loss, the 3D keypoints loss and the SMPL parameter reconstruction loss, respectively.

- **Double-Head Auxiliary Supervision Loss $\mathcal{L}_{\text{das}}$.** This loss is formulated as:

$$\mathcal{L}_{\text{das}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{den}} , \tag{3}$$

where $\mathcal{L}_{\text{sc}}$ and $\mathcal{L}_{\text{den}}$ denote the IUV Supervision loss and Keypoints Supervision loss:

$$\mathcal{L}_{sc} = \lambda_{xy} \left(\text{KL-Loss}(x, \hat{x}) + \text{KL-Loss}(y, \hat{y}) \right) , \tag{4}$$

$$\begin{aligned} \mathcal{L}_{den} = \lambda_{pi} & \text{ CrossEntropy } (P, \hat{P}) \\ & + \lambda_{uv} \text{ SmoothL1}(U, \hat{U}) \\ & + \lambda_{uv} \text{ SmoothL1}(V, \hat{V}). \end{aligned} \tag{5}$$

2. Evaluation Metrics

Here, we describe the evaluation metrics we used in our experiments. First, we report the widely used MPJPE (mean per joint position error), which is calculated as the mean of the Euclidean distances between the ground truth and the predicted joint positions after centering the pelvis joint on the ground truth location (as is common practice):

$$MPJPE = \frac{1}{N} \sum_{i=1}^{N} \|p_i - q_i\|_2 . \tag{6}$$

Also, we report PA-MPJPE (Procrustes Aligned MPJPE), which is calculated similarly to MPJPE but after a rigid alignment of the predicted pose to the ground truth pose,

$$PA-MPJPE = \frac{1}{N} \sum_{i=1}^{N} \|p_i - q_i\|_2 . \tag{7}$$

Furthermore, we calculate Per-Vertex-Error (PVE), which is denoted by the Euclidean distance between the ground truth and predicted mesh vertices that are the output of SMPL layer:

$$PVE = \frac{1}{M} \sum_{j=1}^{M} \|v_j - w_j\|_2 . \tag{8}$$

3. Experimental Results on Simulated Dataset

We also create a simulation dataset; through the simulation dataset, the feasibility of our experiment can be further verified. We create a simulation application based on the principle of lensless imaging to model the imaging process of lensless imaging. We design the simulation application by referring to the imaging model of PhlatCam [1]. A real-world 2D scene $\mathbf{i}(x, y; z)$ at a distance z can be assumed to be made up of incoherent point sources. Each point source will produce a shifted version of PSF $p_z(x, y)$, and since the sources are incoherent to each other, the shifted PSF will

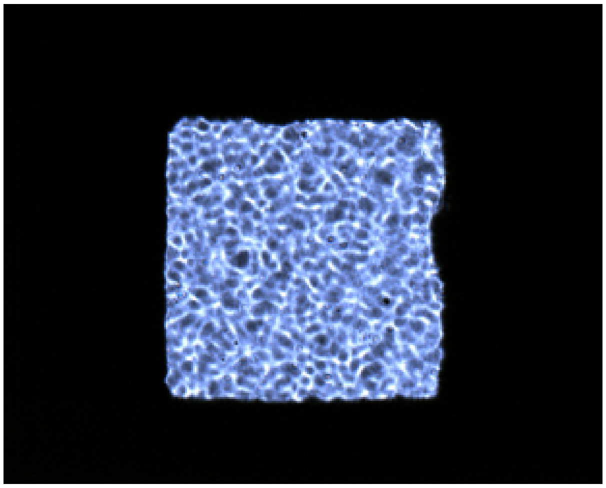

Figure 1. The PSF acquired by our imaging system.

add linearly in intensity [2] at the sensor. By Property 1 of the PSF, we can write the imaging model as the following convolution model:

$$\mathbf{b}(x,y) = p_z(x,y) * \mathbf{i}(x,y;z). \tag{9}$$

Here, b is the sensor's capture, and * denotes 2D convolution over (x,y). We capture the PSF of our system as shown in Fig. 1, and we use the PSF and the existing human posture dataset to get the simulation dataset based on the imaging modality. We perform simulations on the Human3.6M dataset [3]. Fig. 2 and Fig. 3 show the results of LPSNet on the simulated dataset.

4. Failure Cases and Limatations

There are some limitations and some difficult cases that we have not solved very well. Fig. 4, Fig. 5, Fig. 6 shows, some failure cases of our approach. When human movement is more complex, the human pose and shape, as estimated by LPSNet, will have a large deviation. As shown in Fig. 5, the error in human pose and shape estimation also becomes larger when the human body proper is obscured.

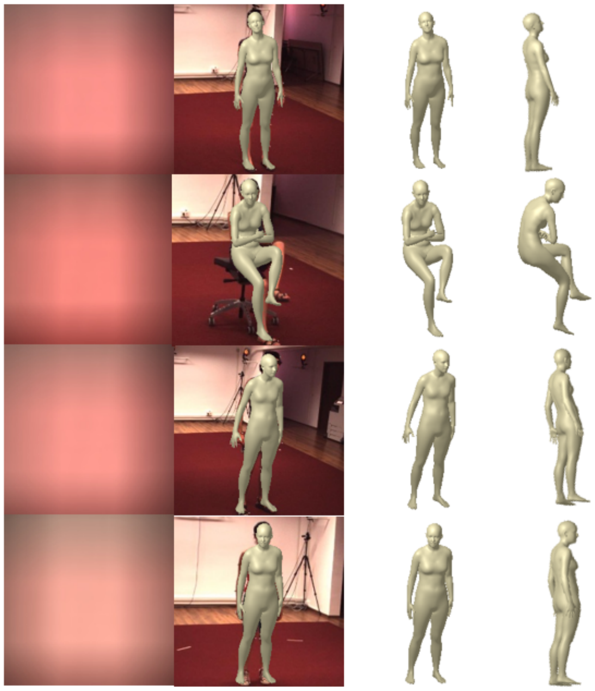

Figure 2. From left to right: lensless measurements, alignment of the estimated body mesh with the original scene, and reconstruction results of LPSNet on the simulated dataset.

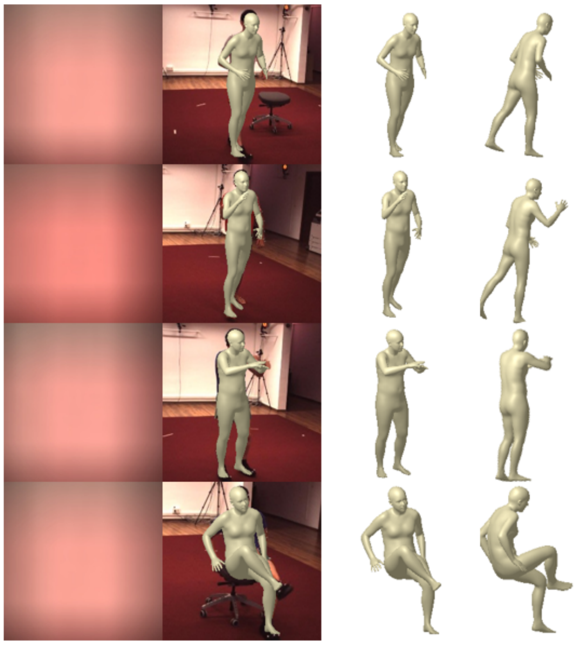

Figure 3. From left to right: lensless measurements, alignment of the estimated body mesh with the original scene, reconstruction results of LPSNet on the simulated dataset.

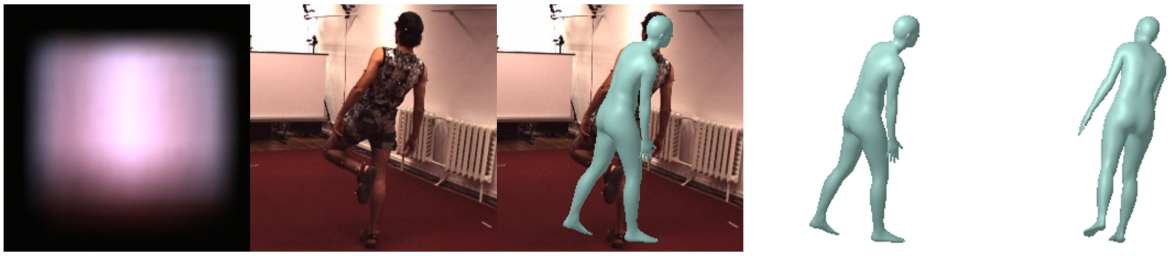

Figure 4. **Failure results for LPSNet.** From left to right: lensless measurements, real scene, alignment of the estimated body mesh with the original scene, reconstruction results of LPSNet. The poses of the characters in the scene are more complex and difficult to reconstruct.

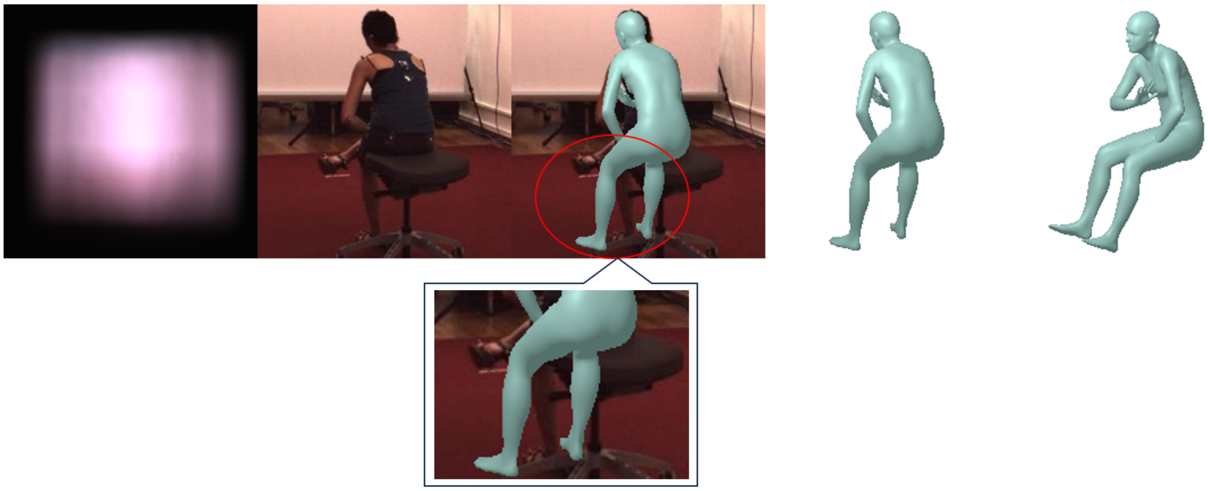

Figure 5. **Failure results for LPSNet.** From left to right: lensless measurements, real scene, alignment of the estimated body mesh with the original scene, reconstruction results of LPSNet. The enlarged image shows that the reconstruction of the obscured part fails.

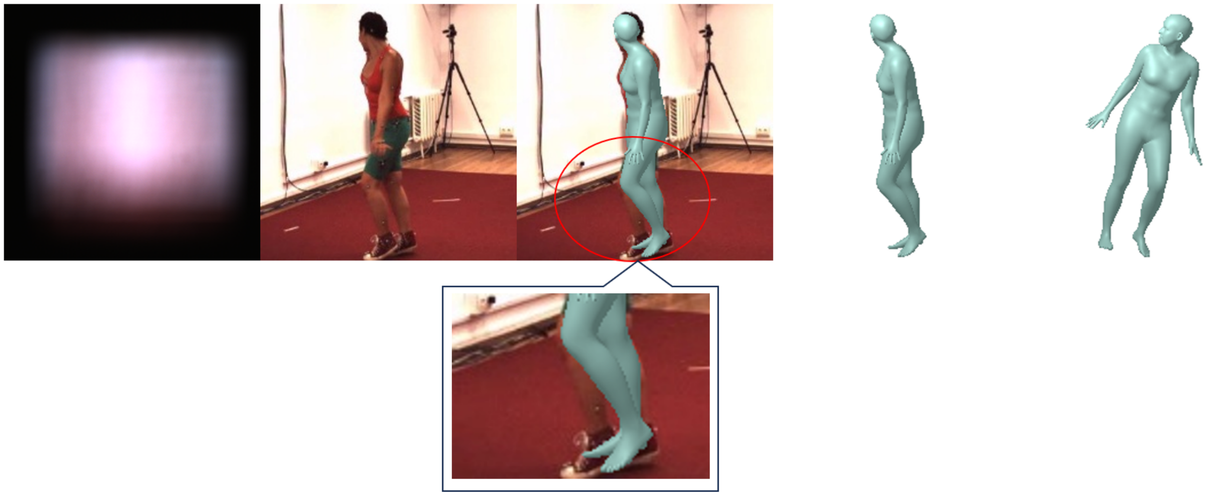

Figure 6. **Failure results for LPSNet.** From left to right: lensless measurements, real scene, alignment of the estimated body mesh with the original scene, reconstruction results of LPSNet. Poses overlap heavily with large reconstruction errors.

References

[1] Vivek Boominathan, Jesse K Adams, Jacob T Robinson, and Ashok Veeraraghavan. Phlatcam: Designed Phase-Mask Based Thin Lensless Camera. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(7):1618–1629, 2020. 1

[2] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company publishers, 2005. 2

[3] 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 Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 2