Supplementary Material for Holistic 3D Human and Scene Mesh Estimation from Single View Images

Zhenzhen Weng
Stanford University
zzweng@stanford.edu

Serena Yeung
Stanford University
syyeung@stanford.edu

1. Training Details

We direct the readers to [4] for camera/world system setting and details on the network architecture of ODN, LEN and MGN. Here we elaborate on the training details.

Stage I In Stage I, we optimize the SMPL-X body model using only the within-body ($\mathcal{L}_{\text{body}}$) losses. We instantiate a body model for each human in the frame, and use L-BFGS optmizer [5] with learning rate $1e-3$ to learn the optimal body parameters (e.g. body shape, pose, translation). First, the translation vector of the body model is optimized for 20 iterations with only the human keypoints re-projection loss. This step is used to roughly position the body model in the camera coordinate system. Then, all the within-body loss terms are considered and the entire body model is optimized for 80 iterations.

For the scene model, we freeze the MGN and the feature extractors components of ODN and LEN, and use Adam [3] optimizer with learning rate $1e-4$ with weight decay $1e-4$ to back-propagate the linear layers for predicting object bounding box attributes (eg. centroid, orientation), camera pose and 3D room layout. For this part, only the within-scene ($\mathcal{L}_{\text{scene}}$) losses are used. For each frame, the scene model is optimized for 150 iterations.

Stage II In Stage II, we add the global consistency losses ($\mathcal{L}_{\text{joint}}$), and continue fine-tuning of all modules. In this stage, we additionally fix the orientation of the 3D object and room bounding boxes and the camera pose. We train the linear layers for predicting the centroid and the size of the object and room boxes to further refine the 3D location of the objects and the ground plane of the scene. We use the same optimizers as Stage I but with reduced learning rates ($1e-4$ for L-BFGS and $5e-5$ for Adam). The body model and scene model are optimized alternately for 20 iterations. The hyperparamters used are $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 10$, $\lambda_4 = 20$, $\lambda_5 = 1e3$, $\lambda_6 = 1e2$.

2. Additional Qualitative Results

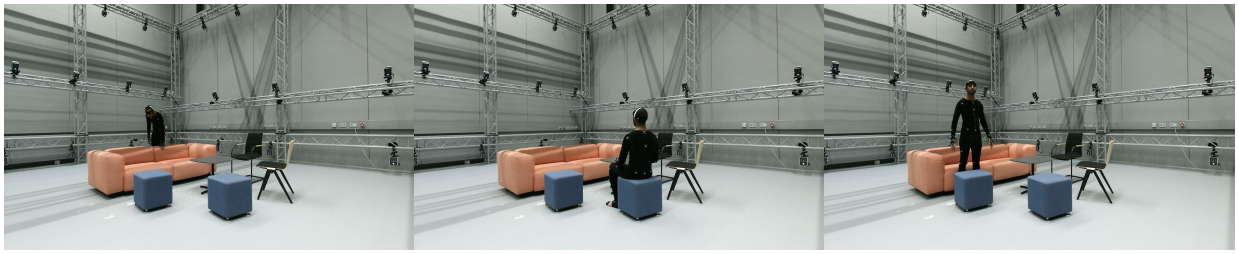
In Figure 1 we show qualitative examples in PROX Quantitative [2] where the scene estimation task significantly helps the body estimation task. These are complementary examples to those in the main paper, which showed that the human body estimation task helps the scene estimation.

From Figure 1 we can see that the initial body meshes are either not physically plausible (column 1), or are intersecting with the scene (column 2, 3). Using the human-scene joint optimization method proposed in our paper, the final body meshes are much more realistic. Note that since we are overlaying the meshes on the 2D images, we can still see the legs behind the furniture after the joint optimization. However, there is no mesh intersection in the 3D coordinate system.

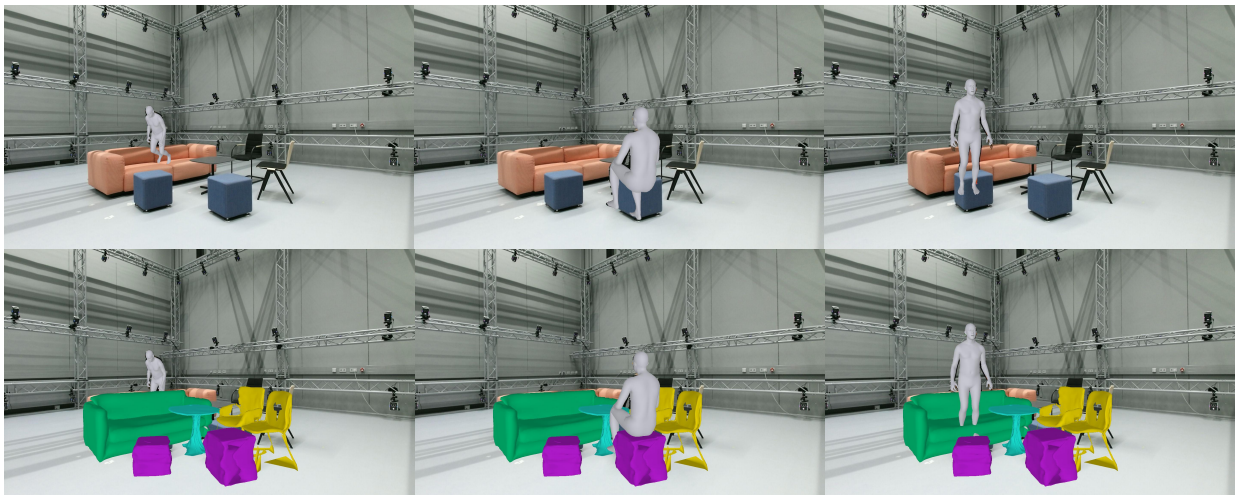
In Figure 2 we show similar results on PiGraphs and PROX Qualitative. We show that the final body meshes in both examples improve through the human-scene optimization stage. In the PiGraphs example, the human body is lifted to reduce the intersection with the sofa. In the PROX Qualitative example, the right hand of the human is occluded so the 2D keypoints predicted by OpenPose [1] do not include the keypoints on the right hand. As a result, the initial hand pose is far from the ground truth. However, through the human-scene optimization that encourages contact between the scene and the hands, the hand pose ended up closer to ground truth.

3. Limitations and Failure Cases

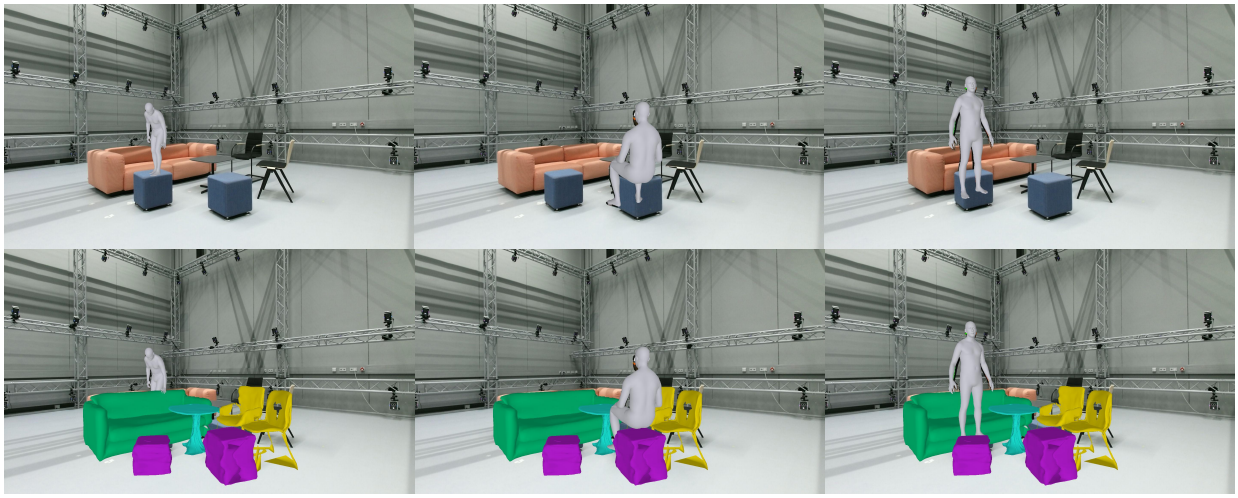
Our method is limited by the performance of the 2D detectors and the mesh generation network. Examples are in rows 1 (missing desk) and 2 (missing coffee table) of Figure 3. Since during joint optimization, the base object mesh structures are not altered, the mesh generation network decides the quality of the generated meshes. Another failure case is due to difficulty or lack of useful physical hints from the scene. When objects and humans are sparsely allocated, the designed losses are not helpful in adjusting their posi-



RGB input



Initial body mesh (shown together with final scene mesh in lower row)



Final body mesh (shown together with final scene mesh in lower row)

Figure 1. Qualitative Results on PROX Quantitative. Each column shows the result on one frame from the PROX Quantitative recordings. From top to bottom are: (1) the RGB input, (2) the initial body mesh (shown together with the final scene mesh in lower row), (3) the final body mesh (shown together with the final scene mesh in lower row).

tions. For instance, Figure 3, row 2 shows the incorrect orientation of the chair in the back. In the right columns of Figure 3 in the paper, where the scenes have more occlusion, the ground plane estimation has a small shift away from the actual ground plane.

4. Code

We will publicly release our code on Github.

RGB input

Initial body mesh (shown together with final scene mesh in lower row)

Final body mesh (shown together with final scene mesh in lower row)

Figure 2. Left: Qualitative example on PiGraphs. Right: Qualitative example on PROX Qualitative. From top to bottom are (1) the RGB input, (2) the initial body mesh (shown together with the final scene mesh in lower row), (3) the final body mesh (shown together with the final scene mesh in lower row).

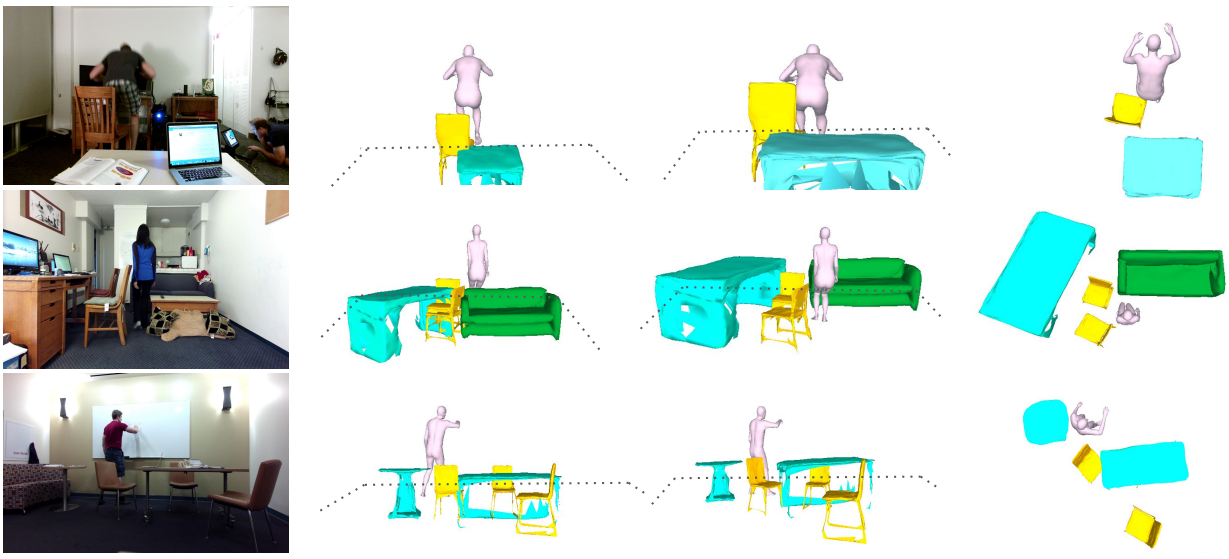


Figure 3. Additional qualitative results with two viewpoints. From left to right are: RGB input, scene w/o optimization, scene w/ optimization (front view and top view).

References

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018. 1

[2] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2282–2292, 2019. 1

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

[4] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image supplementary material. 1

[5] Jorge Nocedal and Stephen J Wright. Nonlinear equations. *Numerical Optimization*, pages 270–302, 2006. 1