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

Zhenzhen Weng, Serena Yeung

Stanford University

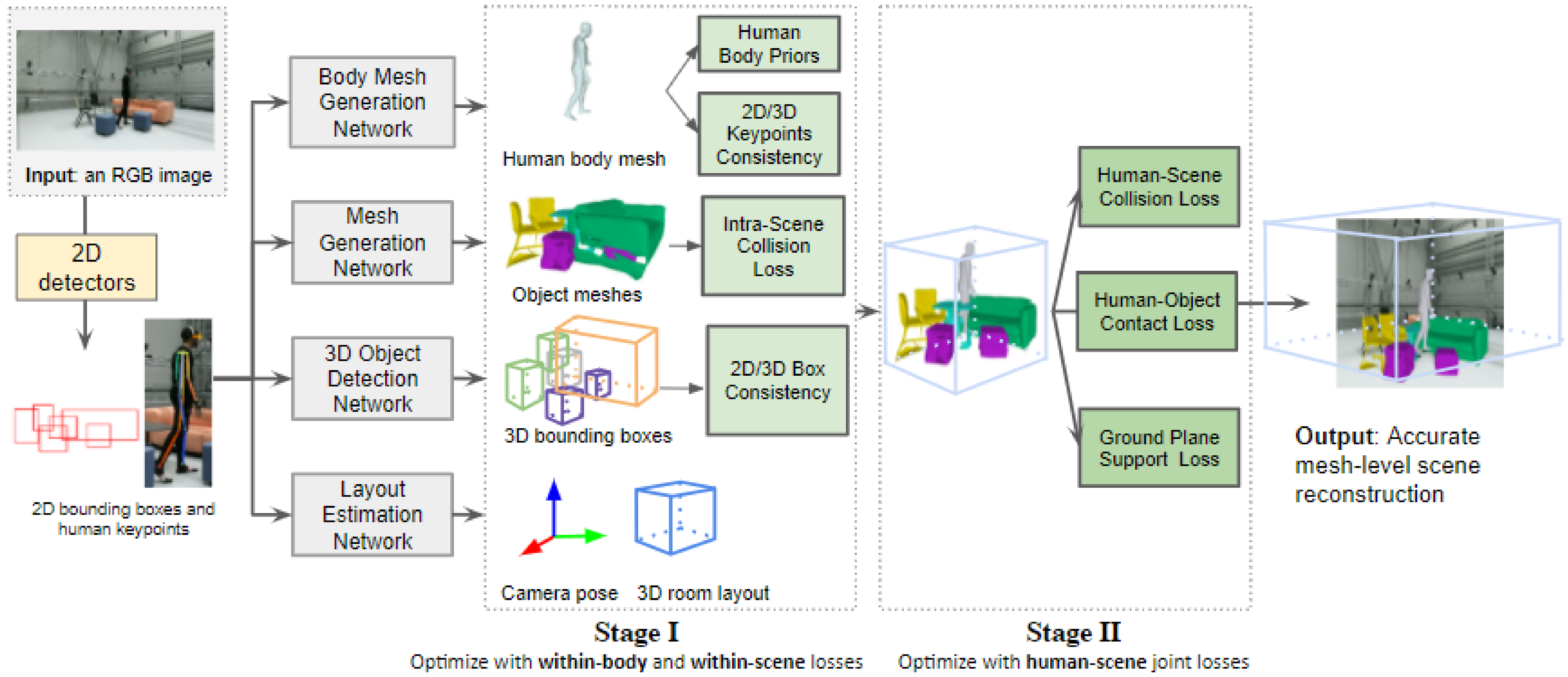
Motivation

- Holistic scene perception is key to our human ability to accurately interpret and interact with the 3D world. This work proposes a holistic trainable model for jointly reconstructing 3D human body meshes and static scene elements from monocular RGB images.
- Our insight is that the 3D world limits the human body pose and the human body pose conveys information about the surrounding, and therefore through **joint estimation and optimization** of the scene mesh and human pose, we can get **more accurate and physically plausible** results.

Overview

- Given a single RGB image, we first use off-the-shelf 2D detectors to predict the 2D human keypoints and 2D bounding boxes of the objects in the scene. Then, the body mesh network reconstructs a SMPL-X body mesh model through the human keypoints re-projection loss and the human body prior losses. The Mesh Generation Network (MGN) reconstructs the object-wise meshes. 3D Object Detection Network (ODN) predicts the 3D bounding boxes of the objects. Layout Estimation Network (LEN) predicts the camera pose and the 3D room bounding box.
- We use a **two-stage** optimization strategy. In Stage I, the human body and the scene are considered separately, and individual modules are optimized with only within-body and within-scene losses. In Stage II, the modules fine-tune with the additional human-scene joint losses to achieve consistency and physical plausibility across all aspects of the output.

Method



$\mathcal{L}_{\text{body}}$ Body loss consisting of the body prior loss and keypoint reprojection loss

$\mathcal{L}_{\text{scene}}$

- $\mathcal{L}^{\mathcal{P}}_{\text{scene}}$ Penetration loss between the scene object meshes
- $\mathcal{L}^{\mathcal{J}}_{\text{scene}}$ Reprojection loss of the scene object 3D bounding boxes

$\mathcal{L}_{\text{joint}}$

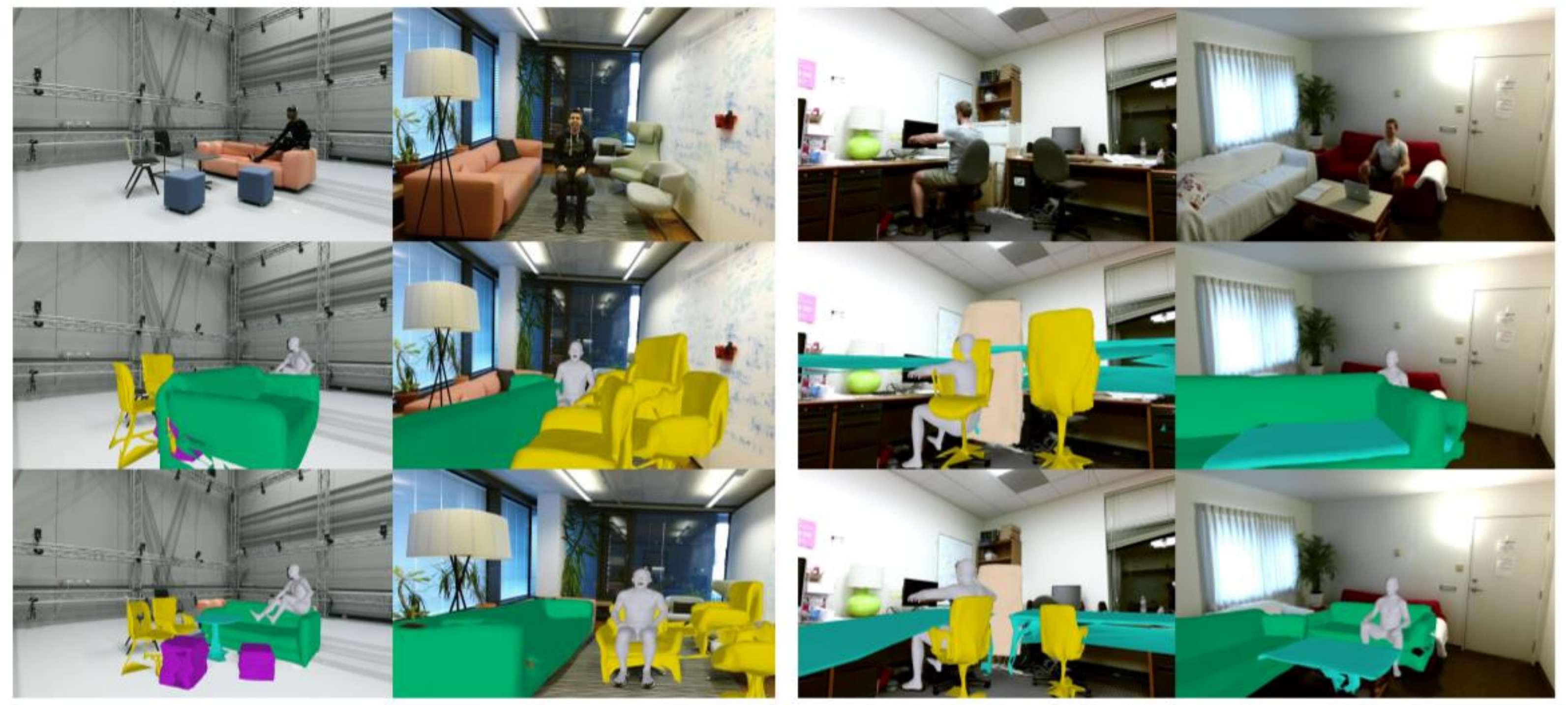
- $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$ $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$ Penetration and contact loss between the body mesh and scene object meshes
- $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$ Loss that minimizes the distance between the feet and the floor
- $\mathcal{L}^{\text{obj-ground}}_{\text{joint}}$ Loss that minimizes the distance between the objects and the floor

Loss function: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{body}} + \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{joint}}$

Input Image

Direct output without optimization

Final mesh with joint optimization



Experiments

We use PROX Quantitative [1] to evaluate the human mesh reconstruction quality and report the vertex-to-vertex error and per-joint error. We compare our method with the baseline methods that do not use scene or use ground truth scene information.

Without ground truth scene scan				
	pje	v2v	p. pje	p. v2v
[1] (body terms only)	220.27	218.06	73.24	60.80
[1] (+ estimated scene)	224.53	220.47	73.49	61.32
[1] (+ w/in scene losses)	212.48	209.67	73.13	62.06

With ground truth scene scan				
	pje	v2v	p. pje	p. v2v
[1] (with $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$)	208.03	208.57	72.76	60.95
[1] (with $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$)	190.07	190.38	73.73	62.38
[1] (with $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$ and $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$)	167.08	166.51	71.97	61.14

Ours	192.21	190.78	72.72	61.01

Ablation study to test the effect of each loss term.

	pje	v2v	p. pje	p. v2v
w/o $\mathcal{L}^{\mathcal{P}}_{\text{scene}}$	200.43	194.28	73.20	62.76
w/o $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$	192.18	190.84	72.21	62.39
w/o $\mathcal{L}^{\text{obj-ground}}_{\text{joint}}$	196.32	193.43	72.47	62.00
w/o $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$	196.48	194.32	73.24	62.96
w/o $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$	212.24	213.26	73.64	62.90
Full model	192.21	190.78	72.72	61.01

On PiGraphs we evaluate the 2D/3D object detection IoU and human keypoints errors (See paper).

Limitations

Our method is limited by the performance of the 2D detectors and the capability of the mesh generation network. Another failure case is due to lack of useful physical hints from the scene. When objects and humans are sparsely allocated, the designed losses are not helpful in adjusting their positions. (See examples in Suppl.)

References

[1] 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.