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

Zhenzhen Weng, Serena Yeung
Stanford University
{zzweng, syyeung}@stanford.edu

arXiv:2012.01591v2 [cs.CV] 16 Apr 2021

Abstract

The 3D world limits the human body pose and the human body pose conveys information about the surrounding objects. Indeed, from a single image of a person placed in an indoor scene, we as humans are adept at resolving ambiguities of the human pose and room layout through our knowledge of the physical laws and prior perception of the plausible object and human poses. However, few computer vision models fully leverage this fact. In this work, we propose a holistically trainable model that perceives the 3D scene from a single RGB image, estimates the camera pose and the room layout, and reconstructs both human body and object meshes. By imposing a set of comprehensive and sophisticated losses on all aspects of the estimations, we show that our model outperforms existing human body mesh methods and indoor scene reconstruction methods. To the best of our knowledge, this is the first model that outputs both object and human predictions at the mesh level, and performs joint optimization on the scene and human poses.

1. Introduction

Holistic scene perception is key to our human ability to accurately interpret and interact with the 3D world. The human visual system naturally integrates context from actors, objects, and scene layout to infer realistic, robust estimations of the world. Suppose a human is partially included in an image because they are positioned behind a desk. We can still effortlessly extract rich information from the static scene to resolve ambiguities due to the occlusion. Likewise, the appearance of humans also provides useful information about scenes, such as the ground plane and depth of surrounding objects. Humans and objects in scenes jointly manifest spatial occupancies that constrain their relative positions. For computer vision systems to achieve high accuracy in recognizing and interpreting complex scenes, it is therefore important to develop approaches for holistic scene perception and reasoning.

In recent years, holistic scene understanding from single view images has gained increasing interest from computer vision researchers. [35] [14] proposed methods for joint reasoning over inanimate scenes, and recovered room layout and 3D object bounding boxes using consistency losses such as a constraint for objects to be enclosed within the room bounding box. [4] additionally discouraged intersection between object bounding box estimations, and was the first model to bring 3D human pose estimation into the holistic scene understanding problem. It incorporated human-object interaction priors to reason about approximate relations between humans and objects. However all of these works still operate at the relatively coarser level of bounding boxes and joint key points, and are therefore limited in their ability to use precise shapes, surfaces, and physical occupancies to design holistic scene constraints and improve estimation accuracy.

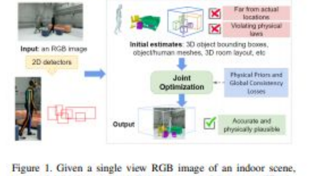

Figure 1. Given a single view RGB image of an indoor scene, our model is able to (i) predict all aspects of the scene (3D object bounding boxes, object and human meshes, 3D room layout, camera pose), and (ii) jointly optimize over a comprehensive set of global consistency losses. The final result is more physically plausible and accurate.

In this work we propose the first single-view, holistic scene understanding method that jointly optimizes over all aspects of 3D human pose, objects, and room layout at the mesh level, to produce state-of-the-art mesh estimations of the scene. Our approach builds on recent advances in mesh prediction. [5] [7] [30] proposed methods for reconstructing the individual object meshes with varying topological structures. [28] builds on [30] and proposed the first holistic 3D scene understanding method with mesh reconstruction at the instance level, however they did not consider humans. Recently, [11] introduced a method for 3D mesh-based human pose estimation, that utilizes physical occupancy information of the static scene to discourage body penetration into the scene. However, [11] requires the ground truth 3D scans of the scene, and does not perform joint human and scene estimation.

Given a single RGB image, our method simultaneously reconstructs the human body mesh and multiple aspects of the scene – 3D object meshes and bounding boxes, room layout, and camera pose – all in 3D (Figure 1). Our approach outputs the SMPL-X (SMPL eXpressive) [31] human mesh model, which fully parameterizes the 3D surface of the human body. It also leverages a variant of the Topology Modification Network (TMN) [5], proposed in [28], as the base model for static object mesh and scene reconstruction. Importantly, we introduce a joint optimization process that incorporates a comprehensive set of physical constraints and priors including 2D/3D reprojection constraints, object-object mesh constraints, object-human mesh constraints, and object/human - room layout constraints, to obtain robust, physically plausible predictions. We perform experimental evaluation on the PiGraphs [34] and PROX [11] datasets and demonstrate that our model outperforms state-of-the-art methods on either 3D scene understanding or 3D human pose estimation.

In summary, our contributions are the following:

- We propose a holistic trainable model for jointly reconstructing 3D human body meshes and static scene elements (3D object meshes and bounding boxes, room layout, and camera pose) from monocular RGB images. To the best of our knowledge, we are the first to jointly estimate this rich scene understanding at the mesh level.
- Our model does not require any ground truth annotations of the 3D scene or the human poses, and can be directly used on any indoor dataset to produce high quality mesh reconstructions.
- Through our joint optimization process that incorporates a comprehensive set of physical constraints and priors, we show that our model outperforms prior state-of-the-art methods on either 3D scene understanding or 3D human pose estimation, on the PiGraphs and PROX Quantitative datasets.

2. Related Work

Single View 3D Human Pose Estimation. Previous 3D pose estimation methods from single view RGB images can be divided into two types: (i) directly learning 3D human keypoints from 2D image features [39], and (ii) 2D pose estimation with subsequent separate lifting of the 2D coordinates to 3D via deep neural networks [32] [24]. Although these works have showed impressive results on in-the-wild images with relatively clean backgrounds, estimating 3D poses with cluttered background and partial occlusions is still very challenging. Recent works in human body models [22] [31] and single view body mesh reconstruction methods [2] [19] have pushed the richness of body details available for reasoning, and provide opportunities for bringing novel constraints to the training stage. Recently, [11] proposed the first 3D human body mesh reconstruction method that takes the static scene into consideration; however they rely on ground truth 3-D scene scans. Our work builds on these directions and is the first to leverage mesh representations of both human and scene in performing holistic estimation of 3D human body and scene meshes jointly.

Holistic Scene Understanding. The 3D holistic scene understanding problem, in particular 3D scene reconstruction from single view images, has received increasing attention over the past few years. While most of these works have focused on coarser bounding boxes and keypoints as opposed to meshes, methods have differed in model outputs and constraint formulations [14][28][4]. Works such as [14] have focused on the static scene; [14] proposed an end-to-end model that learns the 3D room layout, camera pose and 3D object bounding boxes. Drawing insight from the camera projection process and and physical commonsense, [14] encourages projected 3D bounding boxes to be close to their 2D locations on the image plane, and forces object bounding boxes to be within the room layout bounding box.

Some works have attempted to incorporate scene/object information in human pose estimation [43] [11] [26] [44] and/or vice-versa [8]. [44] relies on mesh exemplars with annotated contact points, and does not perform full layout/scene reconstruction. [26] uses a database of "scenelets" and works with human skeletons, [11] utilizes ground truth scene scans. In contrast to these, we consider the more challenging setting of directly estimating scene and human meshes (in general indoor settings), whereas joint mesh estimation is beyond the scope of these works. [4] jointly tackles two tasks from a single-view image: (i) 3D estimations of object bounding boxes, camera pose, and room layout; and (ii) 3D human keypoints estimation. They used an energy-based inference optimization process that refines direct 3D outputs by jointly reasoning across aspects of the objects and human keypoints. However, their constraint formulations based on 3D bounding boxes and human keypoints are still lacking in precision. Additionally, energy-based models have the disadvantage of an expensive inference step compared to feed-forward models, and [4]'s MAP estimation method searches over a discrete set of object locations which may give sub-optimal results. In contrast, we impose precise physical constraints at the *mesh level* in our joint optimization procedure and directly back-

propagate the underlying neural networks.

Holistic Scene Mesh Reconstruction. An emerging line of work attempts to reconstruct richer information about objects in scenes such as depth [36], voxel [21] [40], or mesh representations [7] [28]. Meshes contain much richer 3D shape information about the objects, but are generally harder to reconstruct due to the diverse topology of the shapes. Mesh-retrieval methods [16] [15] [17] retrieve 3D models from a large 3D model repository, however the size of these repositories remain a bottleneck. Object-wise mesh reconstruction methods [5] [42] [7] [30] take a different approach using end-to-end prediction and refinement of the target mesh of individual objects. Recently, [28] incorporated an object-wise mesh reconstruction module in their holistic 3D understanding model for static scenes. However, they did not take advantage of the rich information about object shapes that comes with the meshes, and their reconstructed scene meshes are often physically implausible. Although a recent 3D human mesh estimation method [11] takes advantage of precise object shapes in their constraint formulation, they use ground truth 3D scene scans. In contrast, we estimate *both humans and the static scene* jointly from single view images.

3. Model

We introduce a two-stage approach for joint 3D human and scene mesh estimation. In Stage I, we separately parse and reconstruct the human meshes and the 3D scene – 3D object bounding boxes and meshes, camera pose, and 3D room layout – to obtain initial estimates. In this stage, holistic reasoning is limited to encouraging physical plausibility within the human only and within the static scene only. Then in Stage II, we jointly minimize global consistency losses across humans and the static scene together, which extends the holistic reasoning to simultaneously improve performance of all sub-tasks.

An overview of our method is illustrated in Figure 2. In Section 3.1, we first define our notation and representation of the 3D scene and our human body mesh model. In Section 3.2, we describe the model architectures we use for producing each part of the body and scene estimations. Based on these, in Section 3.3, we present our joint optimization process that incorporates a comprehensive set of physical rules and priors – including reprojection constraints, object-object mesh constraints, object-human mesh constraints, and object/human - room layout constraints – to perform holistic estimation of both human and scene meshes.

3.1. Representation

3D Scene. The input to our model is a 2D image $I \in \mathbb{R}^{(h,w,3)}$. We use a pre-trained Faster R-CNN [33] to obtain initial 2D bounding box estimates $b \in \mathbb{R}^{(4,2)}$ for each of the n_{obj} objects in the scene. The 2D bounding box centers are represented as $c \in \mathbb{R}^2$. Our representation for the camera pose, room layout, and 3D object bounding boxes and meshes in a scene follows the notation used in [14][28]. The camera pose is a 3×3 rotation matrix defined by the pitch and roll angles of the camera system relative to the world system. In the world system, an object bounding box is represented by a 3D bounding box $X \in \mathbb{R}^{(8,3)}$, which can be determined from its 3D center $C \in \mathbb{R}^3$, spatial size $s \in \mathbb{R}^3$, and orientation angle $\theta \in [-\pi, \pi)$. The cuboid room layout is also represented by a 3D box $X^L \in \mathbb{R}^{(8,3)}$, and is parameterized in the same manner as an object bounding box. The triangular mesh for object i in the image is represented by its vertices and faces $M_i = (V_i, F_i)$, where $V_i \in \mathbb{R}^{(N_i,3)}$. N_i is the number of vertices and F_i defines the triangular faces of the mesh. M_i is normalized to fit in a unit cube, and the vertices of the mesh can be converted to the 3D camera coordinate system by translation and rotation as specified by the 3D bounding box parameters.

Human Body Model. We represent the human body using SMPL-X (SMPL eXpressive) [31], a generative model that captures how the human body shape varies across a human population, learned from a corpus of registered 3D body, face and hand scans of people of different sizes, genders and nationalities in various poses. SMPL-X extends the SMPL model [22] with fully articulated hands and an expressive face. It is essentially a differentiable function parameterized by shape β_b, pose θ_b, facial expressions ψ and translation γ of the body. The output of SMPL-X is a 3D triangular mesh $M_b = (V_b, F_b)$ that contains $N_b = 10475$ vertices $V_b \in \mathbb{R}^{(N_b,3)}$ and triangular faces F_b.

3.2. Model Architecture

Body Model. Since the SMPL-X [31] body model is a fully differentiable function, we simply compute the body loss terms (Section 3.3) that are formulated in terms of the vertices and faces of the output human body mesh, and back-propogate the SMPL-X model to find the optimal set of parameters such as the shape and pose of the human body. As in [31], the parameters of the SMPL-X model are regularized with a set of body priors including a VAE-based body pose prior, and L_2 priors on hand pose, facial pose, body shape and facial expressions, penalizing deviation from the neutral state.

Scene Models. We use three sub-modules to predict 3D object boxes, camera pose and 3D room layout, and 3D object meshes in the scene, respectively. Specifically, we adopt the Object Detection Network (ODN), Layout Estimation Network (LEN), and Mesh Generation Network (MGN) from [28]. For 3D object box prediction, the ODN first takes 2D detections of a Faster R-CNN model trained on LVIS [10], extracts appearance features in an object-wise fashion using ResNet-34 [12], and encodes the relative position and size between 2D object boxes into geometry fea-

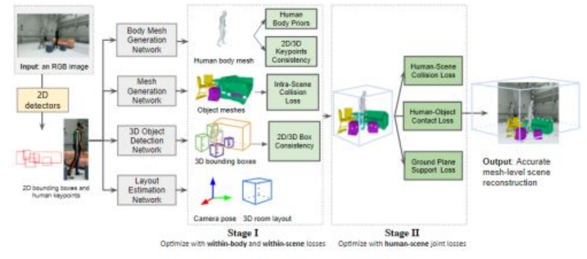

Figure 2. Overview of our model. Given a single RGB image, we first use off-the-shelf 2D detectors to predict the 2D human keypoints and 2D bounding boxes of the objects in the scene. Then, the body mesh network reconstructs a SMPL-X body mesh model through the human keypoints re-projection loss and 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. In **Stage I**, the individual modules are optimized with 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.

tures using the method in [13]. For each target object, an "attention sum" is then computed using relational features to other objects [13]. Finally, each set of box parameters is regressed using a two-layer MLP. The LEN consists of a ResNet-34 feature extractor and two separate branches with fully-connected layers, one for predicting the camera pose and the other for predicting the 3D room bounding box attributes. Finally, for 3D object mesh prediction, the MGN takes a 2D detection of an object as input and uses ResNet-18 to extract 2D appearance features. Then, the image features concatenated with the one-hot LVIS [10] object category encoding are fed into the decoder of AtlasNet [9], which performs mesh deformation from a template sphere mesh. An edge classifier is trained to remove redundant edges from the deformed mesh and a boundary refinement module [30] is used to refine the smoothness of boundary edges and output the final mesh. We pre-trained on SUN RGB-D [37] to initialize the scene models. However, no ground truth annotations are required when training our model on a new dataset.

3.3. Loss Functions and Optimization

We optimize a comprehensive set of losses based on physically plausible constraints and priors, across two stages of training, to perform holistic estimation of 3D human and scene meshes. These losses can be organized as within-body losses (Stage I), within-scene losses (Stage I), and global human-scene losses (Stage II).

Within-body losses As part of Stage I of our approach, we first utilize within-body constraints to generate an initial human mesh estimation. Following [11] [2] [31], we formulate fitting SMPL-X to monocular images as an optimization problem, and seek to minimize the loss function

$$
\begin{aligned}
\mathcal{L}_{\text{body}} &= E(\beta, \theta, \psi, \gamma) \\
&= E_J + \lambda_{\theta_b} E_{\theta_b} + \lambda_{\theta_f} E_{\theta_f} + \lambda_{\theta_h} E_{\theta_h} + \\
&\quad \lambda_{\mathcal{E}} E_{\mathcal{E}} + \lambda_{\beta} E_{\beta} + \lambda_{\alpha} E_{\alpha} + \lambda_{\mathcal{P}_{self}} E_{\mathcal{P}_{self}}
\end{aligned} \tag{1}
$$

Here E_J is the re-projection loss that we use to minimize the weighted robust distance between 2D joints estimated from the RGB image I and the 2D projection of the corresponding 3D joints of SMPL-X. $\theta_b, \theta_f, \theta_h$ are the pose vectors for the body, face (neck, jaw) and the two hands respectively. The terms E_{θ_f}, E_{θ_h}, $E_{\mathcal{E}}$ and E_β are L_2 priors for the hand pose, facial pose, facial expressions and body shape, penalizing deviations from the neutral state. E_β is a VAE-based body pose prior called VPoser introduced in [31]. E_α is a prior penalizing extreme bending only for elbows and knees. The terms $E_J, E_{\theta_b}, E_{\theta_h}, E_\alpha, E_\beta$ are as described in [31]. $E_{\mathcal{P}_{self}}$ is a penetration penalty for self-penetrations (e.g. hand intersecting knee). The λ's are the weights for the terms.

Our formulation is closest to that in [11], which performs human mesh estimation and was built upon [31] with the addition of scene contact ($E_\mathcal{C}$) and penetration ($E_\mathcal{P}$) terms by assuming access to ground truth scene scans. There are several differences between their full loss function and our

formulation in Eq. 1. First, we do not include any depth related terms, because we wish to perform estimation using solely RGB images whereas [11] propose model variants leveraging RGB-D information. Second, since we are performing joint estimation of the 3D scene from a monocular RGB image, we are not yet able to reason on scene contact or penetration after only human mesh estimation. So we include only a body self-penetration term in Eq. 1, which is computed following the approach in [1] [31][41], and will consider human-scene constraints instead during our global optimization stage.

Within-scene losses In Stage I of our approach, we also utilize within-scene constraints to generate an initial static scene estimation. Specifically, we design two within-scene constraints, one for encouraging 2D/3D consistency of the predicted object bounding boxes and the other one for penalizing the collision between the object meshes.

For the first constraint, we utilize the fact that based on the camera projection model, if we project predicted 3D bounding boxes onto the 2D image plane, the projected corners should be close to the 2D bounding box corners. This constraint therefore optimizes both camera pose and 3D bounding boxes. [28] imposes a similar loss where they penalize the deviation of the 2D projections of predicted 3D bounding box corners from ground truth 3D bounding box corners for both object bounding boxes and the room bounding box. However, since our model does not rely on any ground truth annotations in our described optimization process, we propose to use our detected 2D bounding boxes as a pseudo ground truth. We show the effectiveness of this loss term in Section 4. The formal definition of this term can be written as

$$
\mathcal{L}^{J}_{\text{scene}} = \frac{1}{n_{\text{obj}}} \sum_{i=1}^{n_{\text{obj}}} \text{SmoothL}_1(f(X_i(s_i, C_i, \theta_i)), b_i) \tag{2}
$$

where s_i, C_i θ_i are the size, centroid and orientation of the object i. b_i is the 2D bounding box estimate for object i, and f is a differentiable projection function that projects the corners of a 3D bounding box to a 2D image plane. Like [28], we use a smooth L_1 loss function comprised of a squared term if the absolute element-wise error falls below a threshold and an L_1 term otherwise.

Our second constraint is a loss term that penalize the collision between reconstructed object meshes. Although some pose estimation works [11] [18] have incorporated body collision losses, prior works in scene understanding have not explored this loss, because they either did not have the object shape information necessary to calculate the precise collision [14] [4], or did not take advantage of the object shape information that comes with the meshes [28]. We notice that inter-object collision is common in the output of these works. We detect collision using the signed distance field (SDF) of each object. For each object mesh, we voxelize its 3D bounding box into a grid, where for each grid cell center, we calculate its signed distance to the nearest point in the rest of the object meshes in the scene. A negative distance means that this cell center is inside the nearest scene object and denotes penetration. We use a squared sum term of the signed distances of each penetrating grid cell. Formally,

$$
\mathcal{L}^{\mathcal{P}}_{\text{scene}} = \frac{1}{n_{\text{obj}}} \sum_{i=1}^{n_{\text{obj}}} \sum_{c_j \in V_i} ||d(c_j, M_{-i}) \mathbb{1}(d(c_j, M_{-i}) < 0)||_2^2 \tag{3}
$$

where c_j is the center of the j_{th} cell in the voxel grid V_i for object i. $d(c_j, M_{-i})$ is the signed distance between the cell center c_i and the scene mesh composed of all object meshes except for object i. $\mathbb{1}$ is an indicator function.

Global human-scene losses In Stage II of our approach, we jointly fine-tune the human and scene estimation components by imposing additional human-scene losses across the reconstructed human mesh and scene mesh. We consider four types of human-scene losses here.

First, observing that indoor furniture are very likely to be on the floor, we penalize the absolute distance between the object bounding boxes and the ground plane as estimated by the Layout Estimation Network. In the camera coordinate system that we use, $+y$ axis is perpendicular to the ground plane and pointing upward. Hence, we can write this term formally as

$$
\mathcal{L}^{\text{obj-ground}}_{\text{joint}} = \frac{1}{n_{obj}} \sum_{i=1}^{n_{\text{obj}}} d(y_{min}(X^L), y_{min}(X_i))) \tag{4}
$$

where $y_{\min}(X)$ returns the minimum y coordinate values of the 3D bounding box $X \in \mathbb{R}^{(8,3)}$.

Second, like objects in the room, humans need a supporting plane to counteract the gravity. Therefore, we penalize the distance between the lowest point in the human body mesh and the room ground plane. We denote this term as $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$.

Third, we include the contact term $E_\mathcal{C}$ from [11], although[11] utlized ground truth scene scans. The intuition is that when humans interact with the scene, they come in contact with it. Thus, [11] annotates a set of candidate contact vertices $V_\mathcal{C} \subset V_b$ across the whole body that come frequently in contact with the world, focusing on the actions of sitting and touching with hands. Formally,

$$
\mathcal{L}^{\mathcal{C}}_{\text{joint}} = \sum_{v_\mathcal{C} \in V_\mathcal{C}} \rho_\mathcal{C}(\min_{v_s \in V_s} ||v_\mathcal{C} - v_s||) \tag{5}
$$

where $\rho_\mathcal{C}$ denotes a robust Geman-McClure error function [6] for down-weighting vertices in $V_\mathcal{C}$ that are far from the nearest vertices the 3D scene mesh M_s which consists of all the meshes in the scene. Note that since we do not have access to (or reconstruct) a floor mesh as in [11], we leave out [11]'s body-floor contact terms; instead, our loss term $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$ encourages contact between the feet and the floor.

Finally, we penalize any collisions between the body mesh and object meshes in the scene. The formulation is similar to Eq. 3. We call this term $\mathcal{L}^{\mathcal{P}}_{joint}$.

To summarize,our model's total loss is

$$
\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{body}} + \mathcal{L}_{\text{scene}} + \mathcal{L}_{\text{joint}} \tag{6}
$$

where

$$
\mathcal{L}_{\text{scene}} = \lambda_1 \mathcal{L}^{J}_{\text{scene}} + \lambda_2 \mathcal{L}^{\mathcal{P}}_{\text{scene}} \tag{7}
$$

$$
\begin{aligned}
\mathcal{L}_{\text{joint}} &= \lambda_3 \mathcal{L}^{\text{obj-ground}}_{\text{joint}} + \lambda_4 \mathcal{L}^{\text{body-ground}}_{\text{joint}} \\
&\quad + \lambda_5 \mathcal{L}^{\mathcal{C}}_{\text{joint}} + \lambda_6 \mathcal{L}^{\mathcal{P}}_{\text{joint}}
\end{aligned} \tag{8}
$$

In Stage I, only within-body ($\mathcal{L}_{\text{body}}$) and within-scene ($\mathcal{L}_{\text{scene}}$) constraints are used. In Stage II, we add global consistency losses ($\mathcal{L}_{\text{joint}}$) across humans and the static scene together, and continuously fine-tune the modules to simultaneously improve performance of all sub-tasks.

4. Experiments

In this section, we evaluate the performance of our method. Since we are the first to jointly predict and reconstruct both 3D human poses and objects at the mesh level, we compare our model with the state-of-the-art methods for each task. Specifically, we compare with [11] on human body mesh prediction, [25][4] for 3D human keypoints estimation, and [14][4] for 3D bounding box estimation.

4.1. Datasets

Pigraphs [34]. PiGraphs contains 30 3D scene scans and 63 video recordings of five human subjects with skeletal tracking provided by Kinect v2 devices. The dataset contains annotations for 3D human keypoints and 3D object bounding boxes in the scenes. We will perform quantitative evaluation on both of these prediction tasks.

PROX Quantitative and Qualitative [11]. PROX Quantitative has 180 static RGB-D frames and was captured using Vicon and MoSH markers. [11] placed everyday furniture and objects into the scene to mimic a living room, and performed 3D reconstruction of the scene. The ground truth human body mesh annotations were obtained by placing markers on the body and the fingers, and then using MoSh++ [23] to convert MoCap data into realistic 3D human meshes represented by a rigged body model. To the best of our knowledge, this is the only available dataset that has both real furniture in a cuboid room as well as a human subject actively interacting with the scene, which makes it ideal for our task. Since PROX Quantitative does not provide ground truth object-level meshes and therefore does not support scene estimation task, we will quantitatively evaluate our model only on the human mesh estimation task. PROX Qualitative [11] provides 100K synchronized and spatially calibrated RGB-D recordings of humans in 12 indoor scenes. While it was released together with PROX Quantitative, it does not have ground truth human mesh annotations. We perform additional qualitative evaluation on this dataset.

Object Detection			Pose Estimation		
Methods	2D IoU	3D IoU	Methods	2D (pix)	3D (m)
[14]	68.6	21.4	[25]	63.9	0.732
[4]	75.1	24.9	[4]	15.9	0.472
w/o joint	74.2	25.2	w/o joint	15.9	0.469
Ours	**75.6**	**26.3**	Ours	**15.8**	**0.460**

Table 1. **Left**: Quantitative results for 3D scene reconstruction on Pigraphs. Higher IoU values indicate better performance. **Right**: Quantitative results for human keypoints estimation on Pigraphs. For both 2D (pix) and 3D (m) metrics, lower values are better. "w/o joint" is the performance of our model without joint optimization.

4.2. Implementation Details

Given an RGB image of an indoor scene as the input to the model, we first use off-the-shelf 2D detectors to estimate 2D object bounding boxes and 2D human keypoints. For 2D object detections, we use Faster R-CNN [33] trained on the LVIS [31] dataset; for 2D keypoint detections we use OpenPose [3]. ODN, LEN, and MGN are pretrained on the SUN RGB-D dataset [37] and Pix3D [38], following prior work for our task.

In Stage I, we optimize the SMPL-X body model using only the within-body ($\mathcal{L}_{\text{body}}$) losses. We use L-BFGS optmizer [29] with learning rate 1e−3. For the scene model, we freeze the MGN and the feature extractors components of ODN and LEN, and use Adam [20] optimizer with learning rate 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.

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 [29] and 5e−5 for Adam).

4.3. Quantitative Results

3D Object and Human Pose Estimation. To show the efficacy of our method in holistic scene understanding, we quantitatively evaluate 3D object detection and 3D human pose estimation on PiGraphs. No prior works for holistic scene understanding have attempted mesh level reconstruction of the scene and human body; both [14] and [4] outputs 3D bounding boxes of objects, and [4] additionally outputs 3D human keypoints. Thus, we evaluate on the same tasks as these baselines. Since our approach is fully based on physical constraints from externally available mesh models, we do not use any of the 3D annotations in PiGraphs for training, as [14] does. However, we are still able to outperform both (Table 1), showing the power of leveraging the rich shape information available through meshes.

Following [14], for object detection evaluation, we report mean 3D bounding box IoU, as well as 2D IoU between the 2D projections of the 3D object bounding boxes and the ground-truth 2D boxes. For 3D human keypoints evaluation, we extract the 144 body joints from the fitted SMPL-X model and only keep the ones used in [25] [4], which is a subset of the SMPL-X joints. As in [4], we compute the Euclidean distance between the estimated 3D joints and the ground-truth, and average over all joints. For 2D evaluation, we project the estimated 3D keypoints back to the 2D image plane and compute pixel distance to ground truth.

The quantitative results for both tasks in Table 1 show that our model outperforms both [14] and [4] on the 3D object detection task, and [25] [4] on the 3D pose estimation task, which illustrates the effectiveness of our method. The boost in 3D performance is significant, because a large source of error of the baseline models come from inaccurate depth estimation of the objects or the humans. Depth estimation from single view images is generally a difficult problem because 2D visual features are limited in suggesting the depth information. We show that the constraints in our joint optimization help to disambiguate the depth information. The improvement on the object bounding box IoUs suggests that applying fine-grained constraints at the mesh level helps with refining coarser details of the objects.

Human Mesh Estimation We quantitatively evaluate our human body mesh estimation results on PROX Quantitative [11] (Table 2). We follow the evaluation of [11], and report the mean per-joint error without/with procrustes alignment (noted as "PJE" / "p.PJE"), and the mean vertex-to-vertex error (noted as "V2V" / "p.V2V"). Procrustes alignment is a common trick to adjust the predicted 3D vertices for errors in translation, rotation, and scaling. We include the procustes aligned numbers for completion, but note that since our method optimizes all aspects of the human body including translation, rotation and scaling, V2V and PJE are more meaningful quantitative metrics in evaluating the overall quality of the predicted 3D vertices of the mesh.

We compare our body mesh reconstruction method with [11], the state-of-the-art human body mesh reconstruction method on PROX Quantitative. [11] shares the same body loss ($\mathcal{L}_{\text{body}}$) as us; however it imposes contact ($E_\mathcal{C}$) and collision ($E_\mathcal{P}$) constraints between the human mesh and the ground truth 3D scene scans. In our method, we consider an estimated scene mesh in formulating our losses instead. Therefore, in Table 2, we include quantitative performance of [11]'s models using ground truth 3D scene scans for reference, and additionally including the following three baselines models for a fair comparison with our model:

- [11] (body terms only): [11], without using scene terms (since these utilize a ground truth scene).
- [11] + estimated scene: [11] with their contact ($E_\mathcal{C}$) and collision ($E_\mathcal{P}$) terms calculated using the 3D scene mesh predicted by [28] (our base scene model).
- [11] + w/in-scene losses: [11] with their contact ($E_\mathcal{C}$) and collision ($E_\mathcal{P}$) terms calculated using an optimized scene mesh (the base scene model [28], plus our within-scene losses $\mathcal{L}_{\text{scene}}$).

with ground truth 3D scene scans				
	V2V	PJE	p.V2V	p.PJE
[11] (including $E_\mathcal{C}$)	208.03	208.57	72.76	60.95
[11] (including $E_\mathcal{P}$)	190.07	190.38	73.73	62.38
Full [11] ($E_\mathcal{C}$ + $E_\mathcal{P}$)	167.08	166.51	71.97	61.14
without ground truth 3D scene scans				
[11] (body terms only)	220.27	218.06	73.24	**60.80**
[11] + estimated scene	224.53	220.47	73.49	61.32
[11] + w/in-scene losses	212.48	209.67	73.13	62.06
Ours	**192.21**	**190.78**	**72.72**	61.01

Table 2. Quantitative results for human mesh estimation on PROX Quantitative. Top half of the table contains the performance of [11]'s models that use ground truth 3D scene scans in optimizing the human body model. Bottom half of the table contains the baseline models that are most comparable to ours, because no ground truth 3D scene scans are used during training. We highlight the best numbers among the models that do not require ground truth scene scans.

Our model outperforms all three baselines that do not use ground truth scene scans (bottom half of Table 2), and is competitive to [11]'s models using ground truth scene scans (top half). This shows the effectiveness of our scene mesh estimation in refining the human meshes, and that simply adding estimated scenes to [11] is not sufficient. The gap between [11] + w/in-scene losses and Ours highlights the utility of our joint optimization process.

4.4. Ablation Analysis

To analyze the contributions of different losses, we compare variants of our proposed full model. In Tables 3 and 4, we compare quantitative results on the human body mesh prediction and 3D object detection tasks as we take out each one of the losses in Eqs. 7 and 8, except for the essential body loss ($\mathcal{L}_{\text{body}}$) and box re-projection loss ($\mathcal{L}^{J}_{\text{scene}}$). We observe that all of the losses are essential in improving both the scene estimation and body estimation tasks. The joint losses $\mathcal{L}^{\text{object-ground}}_{\text{joint}}$, $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$, and $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$ play an essential role in jointly improving the global consistency, which boosts the performance of human body mesh reconstruction task. In particular, $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$ seems to be the most important term in refining the body meshes. The $\mathcal{L}^{\text{object-ground}}_{\text{joint}}$ and $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$ terms improves the ground plane estimation, which helps the 3D object detection task significantly.

Figure 3. **Left half**: Qualitative results on PROX Quantitative and Qualitative datasets. The left frames is from PROX Quantitative. The right frame is from PROX Qualitative. **Right half**: Qualitative results on Pigraphs dataset. From top to bottom are the RGB input, the direct output from the scene and body mesh without any optimization, and the final mesh with the joint optimization.

4.5. Qualitative Results

Figure 3 shows qualitative results of our models on the PROX Quantitative and Qualitative, and PiGraphs datasets. We observe that the direct output of the scene model (pre-trained on SUN-RGBD and Pix3D) without our holistic optimization contains inaccurate object attributes. Our proposed joint optimization method improves the overall accuracy of the predictions by constraining the orientations, positions and the sizes of the objects to be realistic with respect to each other. Also, human pose estimation task helps the optimization of the scene - the chair that the human sits on tend to have more accurate orientations than the other two chairs (column 2). Besides, the initially estimated ground plane could be very inaccurate (column 3), and our joint optimization process helps adjust the ground plane and improve the location of all objects at the same time. Although not obvious from the qualitative results in Figure 3, the estimated scene mesh helps refining the 3D locations of the human body mesh vertices through the joint losses, which is supported by our quantitative results in Tables 2 and 3. Finally, we show additional qualitative results in Section 2 of the Supplementary, and we discuss limitations and failure cases in Section 3 of the Supplementary.

Metrics	V2V	PJE	p. V2V	p. PJE
w/o $\mathcal{L}^{\mathcal{P}}_{\text{scene}}$	200.43	194.27	73.20	62.76
w/o $\mathcal{L}^{\text{body-ground}}_{\text{joint}}$	192.18	190.84	72.21	62.39
w/o $\mathcal{L}^{\text{object-ground}}_{\text{joint}}$	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

Table 3. Ablations for human mesh estimation on PROX Quant.

Tasks	Object Detection		Pose Estimation	
Metrics	IoU_{2D}	IoU_{3D}	2D (pix)	3D (m)
w/o $\mathcal{L}^{\mathcal{P}}_{\text{scene}}$	58.1	19.1	16.5	0.472
w/o $\mathcal{L}^{\text{body-grnd}}_{\text{joint}}$	52.6	10.3	16.3	0.463
w/o $\mathcal{L}^{\text{obj.-grnd}}_{\text{joint}}$	49.3	11.2	17.9	0.523
w/o $\mathcal{L}^{\mathcal{C}}_{\text{joint}}$	74.6	26.4	18.4	0.493
w/o $\mathcal{L}^{\mathcal{P}}_{\text{joint}}$	73.2	24.7	21.6	0.540
Full model	75.6	26.3	15.8	0.460

Table 4. Ablation results on PiGraphs. For the IoU metrics, higher values indicate better performance. For the pose estimation metrics (2D (pix) and 3D (m)), lower values are better.

5. Conclusion

In this work, we focus on the challenging problem of single view holistic reconstruction and joint optimization of human pose together with static scene. We propose the first holistically trainable model for reconstructing and jointly estimating both 3D human pose and 3D scene at the mesh level. Through a joint optimization process that incorporates a comprehensive set of physical plausibility and priors, we show that our model outperforms state-of-the-art methods on either 3D scene understanding or 3D human pose estimation, on the PiGraphs and PROX Quantitative datasets.

Acknowledgements This material is based upon work supported by the National Science Foundation under Grant No. 2026498, as well as a seed grant from the Institute for Human-Centered Artificial Intelligence (HAI) at Stanford University.