Human-Aware Object Placement for Visual Environment Reconstruction

Hongwei Yi[1] Chun-Hao P. Huang[1] Dimitrios Tzionas[1] Muhammed Kocabas[1,2]
Mohamed Hassan[1] Siyu Tang[2] Justus Thies[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tubingen, Germany [2] ETH Zurich
{firstname.lastname}@{tuebingen.mpg.de,inf.ethz.ch}

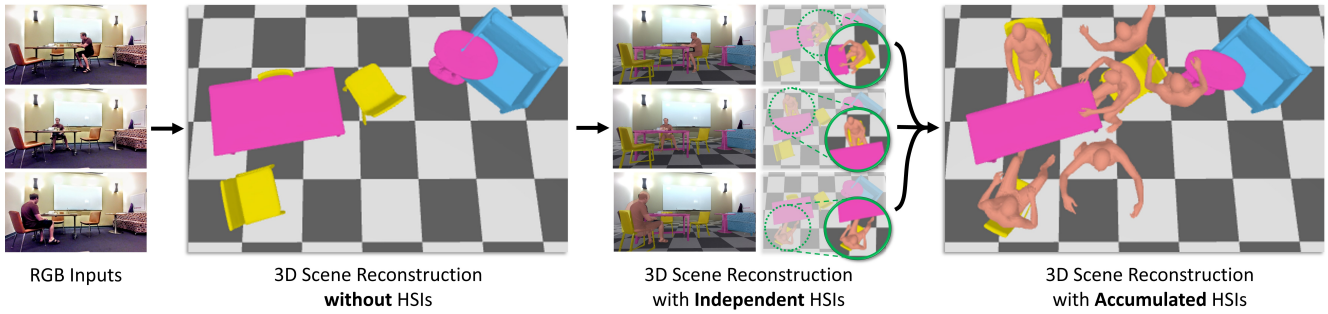


Figure 1. From a monocular video sequence, MOVER reconstructs a 3D scene that best affords humans interacting with it. Existing methods for monocular 3D scene reconstruction ignore people and produce non-functional scenes. MOVER takes as input: (1) several images of human-scene interaction (HSI) from a static camera, (2) a rough estimate of 3D object shape and placement in 3D space [61], and (3) estimated 3D human bodies interacting with the scene [43, 63]. Each frame contains valuable information about humans, objects, and the proximal relationship between them. MOVER accumulates this information across frames, to optimize for a physically plausible and functional 3D scene. The final 3D scene is more accurate than the input and enables reasoning about human-scene contact.

Abstract

Humans are in constant contact with the world as they move through it and interact with it. This contact is a vital source of information for understanding 3D humans, 3D scenes, and the interactions between them. In fact, we demonstrate that these human-scene interactions (HSIs) can be leveraged to improve the 3D reconstruction of a scene from a monocular RGB video. Our key idea is that, as a person moves through a scene and interacts with it, we accumulate HSIs across multiple input images, and optimize the 3D scene to reconstruct a consistent, physically plausible and functional 3D scene layout. Our optimization-based approach exploits three types of HSI constraints: (1) humans that move in a scene are occluded or occlude objects, thus, defining the depth ordering of the objects, (2) humans move through free space and do not interpenetrate objects, (3) when humans and objects are in contact, the contact surfaces occupy the same place in space. Using these constraints in an optimization formulation across all observations, we significantly improve the 3D scene layout reconstruction. Furthermore, we show that our scene reconstruction can be used to refine the initial 3D human pose and shape (HPS) estimation. We evaluate the 3D scene layout reconstruction and HPS estimation qualitatively and quantitatively using the PROX and PiGraphs datasets. The code and data are available for research purposes at https://mover.is.tue.mpg.de/.

1. Introduction

Human behavior, and the interaction of humans with their environment, is fundamentally about the 3D world. Hence, 3D reconstruction of both the human and scene can facilitate human behavior analysis. Where and how the human interacts with a scene can be used to predict future motions and interactions for human-centered AI and robots, or to synthesize these for AR/VR and other computer-graphics applications.

Tremendous progress has been made in reconstructing 3D human bodies [24, 36, 38, 41, 43, 44, 64, 65] and 3D scenes [6, 16, 30, 61, 88] from monocular images or videos, typically in isolation from each other. In real life, though, humans always interact with scenes. Consequently, humans (partially) occlude the scene, and the scene (partially) occludes humans. Strong human-scene occlusion can cause problems for both scene and human reconstruction.

In contrast, recent work on human-scene interaction (HSI), estimates humans and scenes together [10, 26, 83].

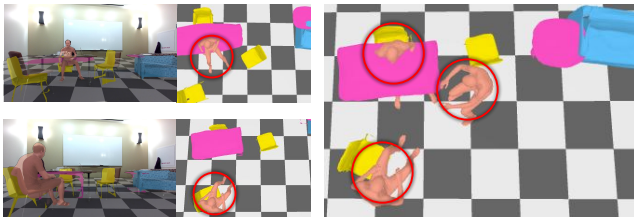
PROX [26] demonstrates how HSI can be used to constrain 3D human pose estimation, but it requires a 3D scan of the full scene to be known a priori. This is often unrealistic and cumbersome, as it requires one to conduct offline 3D reconstruction by walking around the scene with a depth sensor [95] to observe it from many view points.

What we need, instead, is a method that estimates the scene and humans from images of a single color camera. This is challenging, as the lack of depth information causes the scale and placement of objects to be inconsistent w.r.t. the humans interacting with them. This leads to physically implausible results, like humans penetrating objects, or lacking physical contact when walking, sitting, or lying down, causing bodies to "hover" in the air (see Fig. 2). Methods that reconstruct 3D humans from single views leverage statistical body models [37, 54, 64, 85] as priors on the body shape and pose. However, the same tools do not exist for the collective space of 3D scene layouts. This is due to the enormous space of possible object arrangements in indoor 3D scenes, the large number of different object classes, and the huge inter-class (e.g., chairs and desks) and intra-class (e.g., desk chair and club chair) shape variability.

To address the above issues, we present MOVER, which stands for "human Motion driven Object placement for Visual Environment Reconstruction". MOVER leverages information across several HSI frames to estimate both a plausible 3D scene and a moving human that interacts with the scene. Fig. 1 provides a high-level overview. MOVER takes as input: (1) a set of color frames from a static monocular camera, (2) a 3D human mesh inferred for each frame [43, 63], and (3) a 3D shape inferred for each object detected in the scene [40, 61]. As output, MOVER produces a refined 3D scene, comprised of repositioned input objects, so that it is consistent with the estimated 3D human; i.e., it satisfies the expected contacts on the body [27], while preventing interpenetration. MOVER uses a novel efficient optimization scheme, that jointly optimizes over camera pose, ground-plane pose, and the size and position of 3D objects, while being constrained by various HSI constraints.

MOVER takes three types of HSI constraints into account: (1) humans that move in a scene are occluded or occlude objects, thus, defining the depth ordering of the objects, (2) humans move in free space that is not occupied by objects and do not interpenetrate objects, (3) contact between humans and objects means that the contacting parts of their surfaces occupy the same place in space. Thus, we leverage both explicit (i.e., contact) and implicit (i.e., free space, no penetrations) HSI cues. MOVER is able to use these because it employs detailed meshes for both the scene and the moving human. In contrast, the few attempts that have been made towards a similar goal either use oversimplified shapes [10], i.e., 3D bounding boxes for objects and skeletons for humans, work only for static humans that contact a single object [89], or do not integrate information across several interaction frames [10, 83, 89].

(a) 3D scene reconstruction [61] and HPS [63] in isolation.



(b) HolisticMesh [83]. L: single-image results. R: multiple-images result.

Figure 2. Where existing methods struggle: (a) humans in estimated scenes penetrate objects or lack contact with objects and "hover" in the air when estimated in isolation [61, 63] (b) humans interpenetrate objects, even, when the 3D scenes and humans are jointly optimized with single (left) or sequential images (right) [83]. In contrast, we leverage human scene interaction constraints in a global optimization across all input frames, to get a scene that is coherent with the human motions (see Fig. 1).

Comparisons against the state of the art on the PROX [26] and PiGraphs [74] datasets show, that MOVER estimates more accurate and realistic 3D scene layouts that satisfy the expected contacts, while minimizing penetrations, w.r.t. the moving humans. Interestingly, we find that MOVER's estimated 3D scene can be used to refine the human poses, with a PROX-like method [26]. While estimating 3D scenes and 3D humans from a monocular camera is challenging, our results suggest that they are synergistic tasks that benefit from each other.

2. Related Work

Single-view 3D Human Pose in "Isolation": Estimating human pose from an image is a long standing problem [59, 73]. Typically, this is cast as estimating 2D or 3D joints of body [2, 57, 67, 78, 79] or whole-body [7, 35, 82] skeletons. Recently, there has been a significant shift in research interest towards reconstructing the 3D human body surface which, in contrast to the joints, interacts directly with objects and can be observed by commodity cameras. To this end, many non-parametric methods [20, 46, 70, 71, 77, 80, 94] have been developed, estimating either depth maps [20, 77], 3D voxels [80, 94], 3D distance fields [70, 71], or free-form 3D meshes [46]. While these methods can reconstruct bodies with details like hair and clothing, they miss semantic information and correspondence information. In contrast, parametric statistical 3D shape models for the body [3, 25, 54] and whole body [37, 64, 68, 85] provide this information and allow re-posing. Since parametric models represent the shape and pose in a low-dimensional space, they are

Method	GDI	Cam.	C-HOI	N-HOI	FGC
PHOSA [89]	✓	✗	✓	✗	✗
Holistic++ [10]	✗	✗	✗	✗	✓
HolisticMesh [83]	✓	✓	✓	✗	✓
Ours	✓	✓	✓	✓	✓

Table 1. Comparison of the most relevant methods. GDI: Geometric Detailed Interaction. C-HOI: Contact-Human-Object Interaction. N-HOI: Exploiting free space constraints with no object contact. FGC: Feet-Ground Contact. Cam.: Camera orientation and ground-plane are refined with humans or not.

a powerful tool to estimate the surface from incomplete data (e.g., 2D images with occlusions) through optimization [5, 37, 64, 84], regression [12, 38, 42, 45], or hybrid approaches [36]. However, all the above methods reason about the human in "isolation", i.e. without taking the surrounding objects and scenes into account. Thus, they struggle to reconstruct details like contact with objects, and often fail due to occlusions (e.g., bodies standing behind furniture). PARE [43] addresses this by leveraging localized features and attention, gaining robustness to occlusions. We initialize our approach with [43] to refine the 3D scene layout.

Single-view 3D Scene in Isolation: 3D reconstruction from single views has been addressed in several recent works that leverage learned geometrical priors for specific object classes or entire scenes. Shapes from single views are reconstructed using generative models for specific object classes [13, 23, 58, 76, 81]. The methods differ in the underlying representation, which ranges from volumetric representations like occupancy fields [58] and implicit surface functions [53, 62], to explicit surface representations like triangular meshes [21, 81]. To reconstruct scenes, single objects can be detected [28] and reconstructed in isolation. Mesh-RCNN [21] detects the objects in an RGB image, and predicts geometry for each object individually. Instead of a generative mesh model, Izadinia et al. [32] and Kuo et al. [47] propose to retrieve individual CAD models for the detected objects in the scene. Bansal et al. [4] predict a normal map from the input image that is used to align a retrieved CAD model. Instead of predicting normal maps from the input image, there is a series of methods that estimate depth maps [19, 22, 49, 75], or pixel-aligned implicit functions for objects [69, 72] and scenes [16, 18]. Joint estimation of the room layout and objects with scene context information has been proposed by [11, 30, 31, 61, 88, 92, 93]. However, these methods only consider an isolated 3D scene without a human in it.

Note that there are also methods that predict room layouts with 3D bounding boxes [17, 29, 50, 56]. In contrast, we reconstruct the detailed object geometry to leverage explicit contact point constraints based on the human scene interactions, while optimizing for the scene layout.

3D Human-Scene Interaction (HSI): Humans inhabit 3D scenes. Several methods model this and learn to populate a 3D scene [27, 51, 90, 91]. In contrast, our work reasons about the human and its interaction with the 3D scene from RGB observations. There are several methods that explore different kinds of human scene interaction; these can be divided into three categories based on the granularity of the interaction between human and scenes: (1) Hand-Object [8, 9, 33, 48, 52, 86]. (2) Body-Object [15, 89]. (3) Body-Scene [10, 60, 83].

Our proposed method focuses on reconstructing 3D scenes composed of objects and structural elements like the floor plane, using accumulated human scene interactions (body-objects and body-scene). Table 1, overviews the most related work that operates on single-view RGB images/videos. PHOSA [89] infers humans and objects together when they are in contact. They do not consider the fact that humans do not need to contact an object to constrain its location; their movement through free space constrains object placement. Sminki et al. [87] only consider feet-ground contact. iMapper [60] maps RGB videos to dynamic "interaction snapshots", by learning "scenelets" from PiGraphs data and fitting them to videos. However, the estimated scene is not aligned with the 2D image, and consists of predefined CAD templates with fixed shape and size. Holistic++ [10] takes learned 3D HOI (Human Object Interaction) into account, to jointly reason about the arrangement of bodies and objects. Both [60] and [10] do not model geometrically detailed human-scene interaction, due to their simplified representation of the scene and bodies. Weng et al. [83] jointly optimize the reconstructed mesh-based 3D scene and bodies, which are initialized from [61] and [64]. The approach only considers interpenetration between objects and the human, and does not model the explicit human-scene contact. Additionally, both [10, 83] do not model the coherence of human-scene interactions across frames from monocular video. In contrast to the prior work, our contribution lies in incorporating multiple human-scene interactions collectively, such that we can reconstruct a more accurate and consistent scene, with physically plausible human-scene interactions.

3. Method

MOVER is an optimization-based approach that reconstructs a physically plausible 3D scene that is consistent with predicted human-scene interactions over time (see Fig. 3). Specifically, our method takes a single RGB video or multiple images $\{I_t\}_{t=1}^{T}$ as input and reconstructs the human bodies at each time step t as well as the numerous static scene objects, both of which reside in a common 3D space and are supported by a ground plane. In our experiments, we consider large objects in indoor scenes that humans frequently interact with, i.e., chairs, beds, sofas, and tables.

We initialize our approach using separate estimates for the 3D human poses [43, 63], the 3D scene [61], and the ground

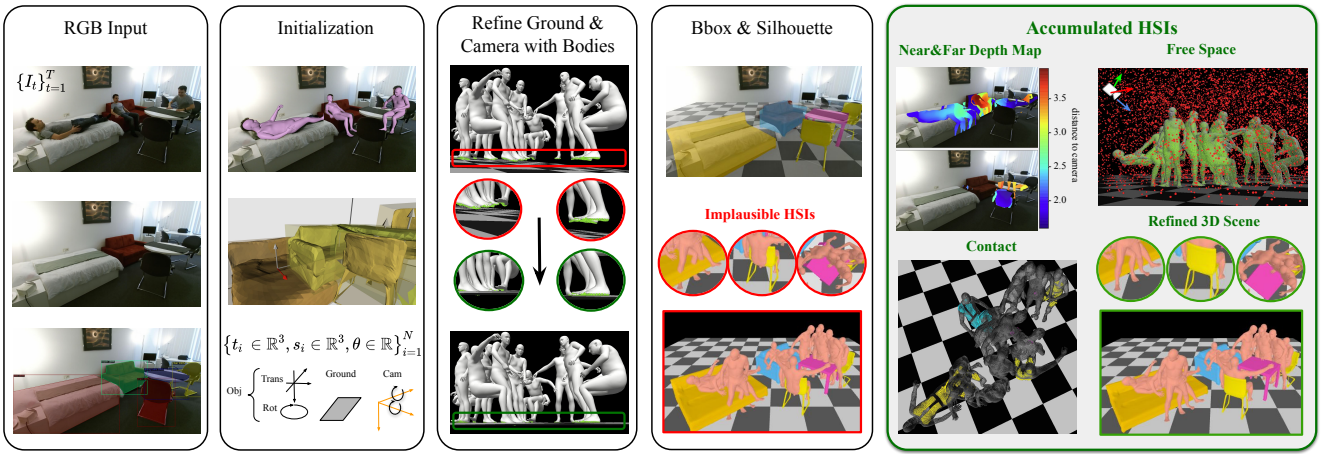


Figure 3. Overview of MOVER. Given a video/multiple images, the initialization involves using [61] to reconstruct a 3D scene from labeled or detected 2D instance segmentation masks [40], estimating the 3D human poses and shape [43, 63], and extracting the expected contact vertices on the estimated bodies using POSA [27]. The first step then refines the camera orientation and ground plane using the human bodies and their foot contact. Then we optimize the object layout based on 2D bounding boxes and silhouettes to remove interpenetration between people and objects, e.g., the human sits through the chair, stands into a table, and the legs are in a bed. Finally, incorporating multiple HSIs collectively from the whole video, we can improve the 3D scene further such that the bodies perform more realistic HSIs.

plane. Using the estimated body poses, we predict contact vertices $\mathcal{C}$ for all bodies using POSA [27], which predicts likely contact vertices on the body conditioned on pose. We further divide these vertices into foot contacts $\mathcal{C}^{\text{feet}}$ and other body part contacts $\mathcal{C}^{\text{body}}$. The explicit foot contact points $\mathcal{C}^{\text{feet}}$ are used as constraints to refine the camera orientation and ground plane prediction. Based on this initialization, we optimize the alignment of the objects by minimizing an objective function based on multiple human-scene interactions (HSIs) across the entire input data.

3.1. 3D Scene Layout Optimization

Our method leverages multiple HSIs to refine the 3D scene. Recall that these HSIs provide the following constraints: (1) humans that move in a scene are occluded or occlude objects, thus, defining the depth ordering of the objects (depth order constraint), (2) humans move through free space and do not interpenetrate objects (collision constraint), (3) when humans and objects are in contact, the contact surfaces occupy the same place in space (contact constraint). Using these constraints, our objective $L_{\text{scene-human}}$ is:

$$\mathcal{L}_{\text{scene-human}} = \lambda_1 \mathcal{L}_{\text{bbox}} + \lambda_2 \mathcal{L}_{\text{occ-sil}} + \lambda_3 L_{\text{scale}} + \lambda_4 \mathcal{L}_{\text{depth}} + \lambda_5 \mathcal{L}_{\text{collision}} + \lambda_6 \mathcal{L}_{\text{contact}}. \quad (1)$$

We apply an occlusion-aware silhouette term $\mathcal{L}_{\text{occ-sil}}$ from [89], a 2D bounding box projection term $\mathcal{L}_{\text{bbox}}$ that constrains the top-left corner and the width of the bounding boxes of the objects, and L_{scale}, an ℓ_2-based regularizer to constraint the variation of the object scales, see more details in Sup.Mat.

Depth Order Constraint $\mathcal{L}_{\text{depth}}$. The occlusion between humans and objects can provide clues about the object's depth. We assume the human's depth is accurate. If a human occludes an object, then the far side of the person sets a limit on how close the object can be. Alternatively, if the object occludes the person, then the visible side of the person sets a maximum distance for the object. This is summarized in Fig. 4. In this way, human-object occlusion provides constraints on scene layout even when there is no human-object contact.

Directly applying the ordinal depth loss proposed by Jiang et al. [34] for each image is inefficient, because the required memory increases with the number of images. In contrast, we accumulate all single depth ordering maps into one far depth range map $\hat{D}_{\text{far}}$ and one near depth range map $\hat{D}_{\text{near}}$ as:

$$\hat{D}_{\text{far}}(p) = \min\left(D^1_{\text{far}}(p), ..., D^T_{\text{far}}(p)\right),$$
$$\hat{D}_{\text{near}}(p) = \max\left(D^1_{\text{near}}(p), ..., D^T_{\text{near}}(p)\right),$$

where the pixel p is in the overlapping region between the human bodies and the objects. Using these accumulated depth range maps, we constrain the depth $D_i(q)$ of a projected pixel q from object i to lie in the corresponding range:

$$\mathcal{L}_{\text{depth}} = \sum_i \sum_{q \in Sil_i \cap M_i} [\text{ReLU}(D_i(q) - \hat{D}_{\text{far}}(q)) + \text{ReLU}(\hat{D}_{\text{near}}(q) - D_i(q))],$$

where Sil_i is the rendered silhouette of the object i, M_i is the 2D segmentation mask of *it*, and $D_i(q)$ is the depth of the object i at the pixel q. See more details in Sup.Mat.

Collision Constraint $\mathcal{L}_{\text{collision}}$. To penalize all interpenetrating vertices of objects and bodies in the scene, we use

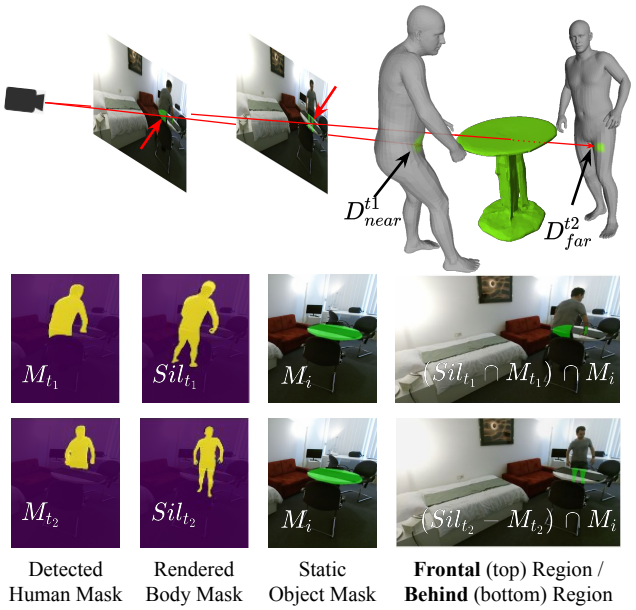


Figure 4. Computing depth range maps for the depth order constraint $\mathcal{L}_{\text{depth}}$. Given a detected human mask M_t and a rendered body mask Sil_t, for each object i, we compute the overlap region between M_i and $Sil_t \cap M_t$ as the frontal region and extract the depth of the backside surface of the body as near depth range D^t_{near} of the object i. Similarly, we compute $(Sil_t - M_t) \cap M_i$, which defines the far depth range D^t_{far} of the object.

the signed distance field (SDF) of all reconstructed bodies. Specifically, we calculate a signed distance field volume V_j for each body j in a shared 3D world space, and accumulate them into a global SDF volume as $\hat{V} = \min(V_1, ..., V_j, ...)$. The SDF $\hat{V}$ is stored in a volumetric grid of size 256^3, which spans a padded bounding box of all bodies. For a vertex u_i of an object O_i, we compute the voxel coordinate $f(u_i) = (p(u_i), q(u_i), k(u_i))$ in the global SDF volume, where p, q, k denote the indices of the grid in it, and retrieve the corresponding SDF value $\hat{V}_{f(u_i)}$.

Based on the SDF values of all vertices of all N objects, we resolve the scene-body inter-penetration by penalizing vertices with a negative SDF value:

$$\mathcal{L}_{\text{collision}} = \sum_i \sum_u \|\hat{V}_{f(u_i)}\|_2^2, \quad \hat{V}_{f(u_i)} < 0.$$

Contact Constraint $\mathcal{L}_{\text{contact}}$. When humans and objects are in contact, the contact surfaces occupy the same place in space. We propose a contact constraint to minimize the distance between the contacted body parts and its assigned corresponding contacted object. PHOSA [89] proposes a loss in which they assign a whole body to only one object, whereas humans sometimes interact with multiple objects; e.g., a person sits on a chair and puts their hand on a table. In contrast, we directly assign the contacted body vertices $\mathcal{C}_i^{body}$ of each body to different objects, based on the overlap between the 2D projection of the vertices and the detected object masks, and based on the 3D distances between them. We consider the vertices of sofa and chair backs and seat bottoms as contactable regions, see more details in Sup.Mat.

We minimize the distance between the contacted bodies and the contacted object parts:

$$\mathcal{L}_{\text{contact}} = \sum_i \sum_{v \in \mathcal{C}^{\text{body}}} \mathbb{I}(v, O_i)[\text{CD}(v^y, \mathcal{C}(O_i)^y) + \text{CD}(v^{\perp y}, \mathcal{C}(O_i)^{\perp y})],$$

where $\mathcal{C}(O_i)^{\perp y}$ and $\mathcal{C}(O_i)^y$ denote the back and the bottom seat contact part of an object i, respectively. $\mathbb{I}(v, O_i)$ is an indicator function (1 only if the contact vertex v is assigned to the contacted object O_i, 0 else). CD denotes the one-directional ChamferDistance (CD), i.e., from bodies to objects, because for large furniture like a bed or a sofa, a human only contacts a small region of the object. In contrast, PHOSA [89] uses a bi-directional CD, which tends to shrink the object to match the contacted body parts.

3.2. Optimization

We optimize Eq. (1) for a specific scene w.r.t. the parameters $\mathbf{s}_i$ (scale), θ_i (rotation), $\mathbf{t}_i$ (translation) of the objects $\{i = 1...N\}$, with the Adam optimizer [39]. In the following, we detail the initialization of the 3D scene and the HPS.

Initial 3D Scene. We extract a representative 2D image $\mathbf{I}$ from the input data without any human-object occlusion. For this image, we label or compute 2D bounding boxes B_i and an instance masks M_i of all N objects in the scene using PointRend [40]. We use [61] to get an initial 3D scene $\mathbf{S}_0$, consisting of a ground plane $y = y_p$ and multiple object meshes $\{O_i\}_{i=1}^N$, and a perspective camera with $yaw, pitch$ orientation. Each object i has a translation $\mathbf{t}_i \in \mathbb{R}^3$, scale $\mathbf{s}_i \in \mathbb{R}^3$, and a rotation along the y-axis $\theta_i^y \in [0, 2\pi)$ parameters. Since the predicted meshes of [61] are incomplete and have holes, we use Occupancy Networks [58] and Marching Cubes [55] to transform each object mesh into a water-tight mesh. Based on this preparation, we first optimize the objective function without considering the HSIs:

$$\mathcal{L}_{\text{scene}} = \mathcal{L}_{\text{occ-sil}} + \lambda_1 \mathcal{L}_{\text{bbox}} + \lambda_2 \mathcal{L}_{\text{scale}}.$$

Initialization the ground and the camera. As shown in the third column of Fig. 3, the estimated ground plane and camera orientation from [61] violates the reconstructed bodies (e.g., people float in the air). Previous methods either fix the camera orientation and only optimize the ground plane and humans [10], or estimate them independently per image [83], which generates inconsistent camera orientation and ground planes throughout a video. However, the camera orientation and ground plane are essential for producing plausible HSIs.

Thus, we jointly estimate the ground, camera and multiple humans together, by applying $\mathcal{L}_{\text{feet}}$:

$$\mathcal{L}_{\text{feet}}(R,p) = \rho_1(R^\top \sum_t \mathcal{C}_t^{\text{feet}} - [0, y_p, 0]^\top),$$

where R is the camera rotation matrix calculated from *pitch*, *yaw*, and ρ denotes a robust Geman-McClure error function [14] for down-weighting outliers.

Initial Estimate of 3D Bodies. As an initial shape and body pose estimate for the input images $\{I_t\}_{t=1}^T$, where a human interacts with a 3D scene, we use OpenPose [7] and SMPLify-X [63]. Specifically, we use a perspective camera projection and estimate the pose parameters θ_t of SMPL-X for each frame with shared body shape parameters β. SMPLify-X requires a good initialization and, for this, we use PARE [43] because it is robust to occlusion and our scenes involve significant occlusion. PARE outputs SMPL, which we convert to SMPL-X, and use the resulting 3D joints to initialize SMPLify-X, see more details in Sup.Mat.

We then optimize all SMPL-X parameters to minimize an objective function E_{Body} of multiple terms, as described in SMPLify-X [63] (see $E_{\text{SMPLify-X}}$) :

$$E_{\text{Body}} = \sum_{t=1}^{T} \left(E_{\text{SMPLify-X}}\left(t\right)\right) + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}}.$$

To reduce the jitter, we add a constant-velocity motion smoothing term on 3D joints J and their 2D projections J^{Proj}:

$$\mathcal{L}_{\text{smooth}} = \sum_{t=1}^{T-1} \rho_2 \left(\|J_{t-1} + J_{t+1} - 2 \times J_t\|\right) + \rho_3 \left(\|J_{t-1}^{\text{Proj}} + J_{t+1}^{\text{Proj}} - 2 \times J_t^{\text{Proj}}\|\right).$$

To avoid noisy and unreliable body poses, and therefore, wrong human-scene interactions during optimization, we also apply a filter based on the constant-velocity assumption. We calculate the pelvis acceleration ν_t and local joints' acceleration α_t of a person in frame t to describe the global body translation and local pose articulations of the body. We filter out those bodies with either large pelvis translation or incorrect human pose with a large ν or a large α respectively: $\{j : \nu_j < \tau_{\text{pelvis}} \cap \alpha_j < \tau_{\text{local}}, j \in \{1...T\}\}$, where τ_{pel}, τ_{local} are the thresholds for the pelvis acceleration and the local pose acceleration, respectively.

4. Experiments

To evaluate the influence of accumulated HSIs on the optimized 3D scene layout, we use two different datasets, PiGraphs [74] and PROX [26] (see Sup. Mat.). In comparison to [61] and [83], we achieve state-of-the-art 3D scene layout reconstruction, both quantitatively (see Sec. 4.1) and qualitatively (see Sec. 4.3). On the PROX *quantitative* dataset, we find that our 3D scene reconstructions lead to more accurate human shape and pose estimations than our baselines. In Sec. 4.2, we analyze the different energy terms and how they contribute to our final results. Qualitative results are shown in Fig. 5 and in the suppl. video.

4.1. Quantitative Analysis

We perform several experiments to investigate the effectiveness of our proposed method in three parts: 3D scene reconstruction, human-scene interaction (HSI) reconstruction, and human pose and shape (HPS) estimation. The results are listed in Tab. 2.

3D Scene Reconstruction. Following [10, 31, 61, 83], we compute the 3D IoU and 2D IoU of object bounding boxes to evaluate the 3D scene reconstruction and the consistency between the 3D world and 2D image on PROX and PiGraphs. However, the 3D IoU is coarse and does not capture the error in an object's orientation, which is quite important for physically plausible HSI, e.g., a human can not sit on an armed chair with the wrong orientation. Therefore, we introduce the *point2surface distance:* $p2s$ to measure the distance from a cropped object mesh to the estimated 3D object mesh. It enables 3D scene reconstruction evaluation with more geometric details including orientation and shape. Given 2D labeled or detected [40] bounding boxes and masks, our methods improves the input [61] significantly, and outperforms [83] on all scene-reconstruction metrics and different datasets, as shown in Tab. 2 and Tab. 3.

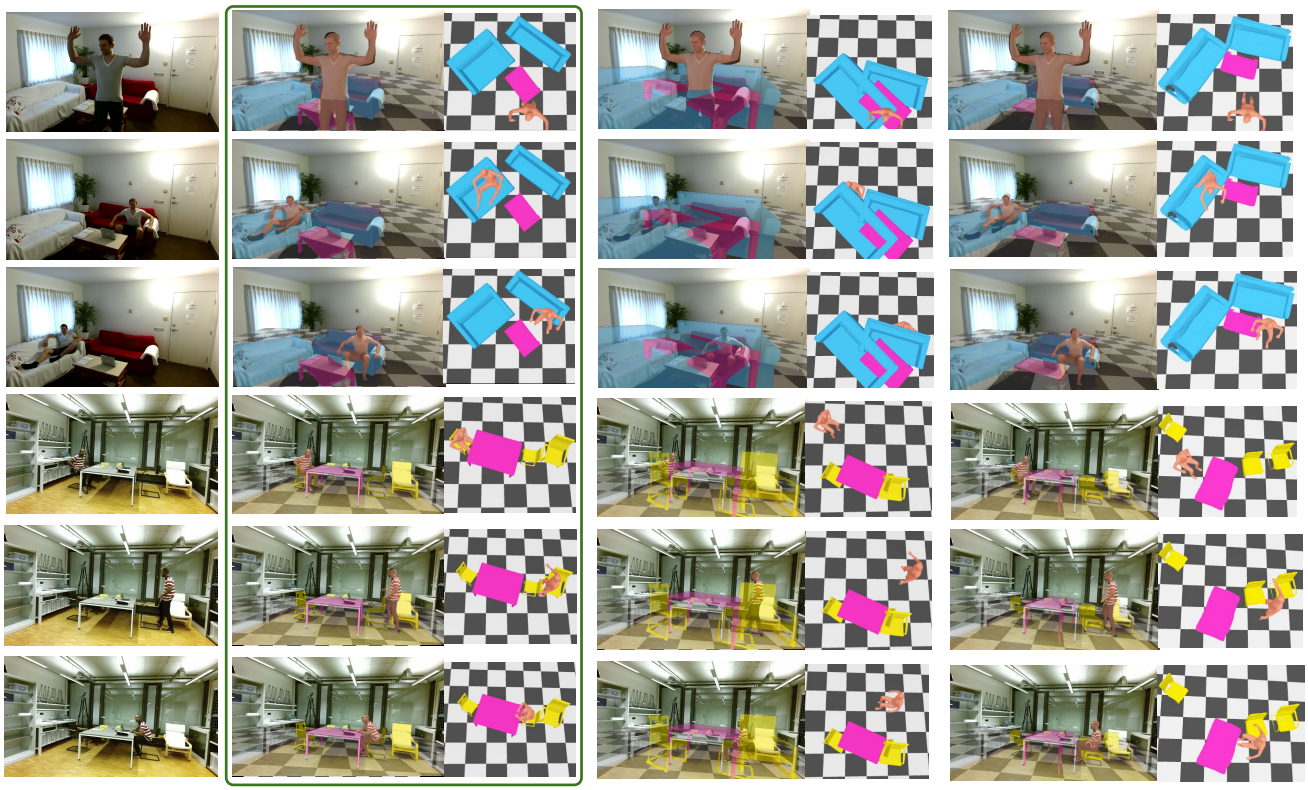
Furthermore, we also evaluate the error of the camera orientation and ground plane penetration [66] using the estimated foot contact vertices (see Tab. 4). We find that that jointly optimizing the camera orientation and the ground plane using foot contact significantly improves accuracy compared to the initial estimate from [61].

Human-scene Interaction Reconstruction. To evaluate the estimated HSI or functionality of the scene (denoting how well the estimated scene can support human motion), we compute the metrics as [27, 90, 91]. Specifically, for each reconstructed body and 3D scene, we calculate (1) the *non-collision score* to measure the ratio of body mesh vertices without collision with the estimated 3D scene, divided by the number of all body mesh vertices, and (2) the *contact score* to denote whether the body is in contact with the 3D scene or not. The *contact score* is 1, if at least one vertex of a body interpenetrates with the 3D scene. We report the mean *non-collision score* and mean *contact scores* among all videos and all bodies. In Tab. 2, MOVER achieves the best balance between non-collision and contact.

The estimated scenes with detected 2D boxes and masks [40] provide lower HSI scores than with 2D GT. It is mainly

Methods	Setting					Scene Recon.			HSI	
	BBOX&Mask	Cam.	Contact	Depth	Colli.	IoU_{3D} ↑	P2S↓	IoU_{2D} ↑	Non-Col ↑	Cont. ↑
HolisticMesh [83]	PointRend					0.211	0.410	0.648	0.990	0.369
Total3D [61]	PointRend					0.246	0.319	0.522	0.974	0.510
Ours	PointRend	✓	✓	✓	✓	0.309	0.221	0.777	0.977	0.612
HolisticMesh [83]	2D GT					0.267	0.237	0.745	0.988	0.491
Total3D [61]	2D GT					0.196	0.369	0.227	0.963	0.440
Ours	2D GT	✓	✓	✓	✓	0.383	0.199	0.898	0.986	0.673
Ablation Study	2D GT	✓				0.374	0.206	0.859	0.979	0.738
		✓	✓	✓		0.389	0.199	0.904	0.983	0.697
		✓	✓			0.381	0.205	0.904	0.980	0.773
		✓		✓		0.393	0.194	0.907	0.983	0.638
		✓			✓	0.383	0.199	0.903	0.984	0.674

Table 2. Quantitative results for 3D scene understanding (3D object detection) and human-scene interaction on the PROX *qualitative* dataset. P2S, Non-Col and Cont denote *point2surface distance*, Non-Collision and Contactness respectively. In each column, red is the best result among methods that take 2D labeled masks as input; blue is the second best.



RGB Input **Ours** Total3D [61] HolisticMesh [83]

Figure 5. Qualitative results on PiGraphs (top) and PROX. Our method recovers better 3D scenes and HPS, which supports more plausible HSIs, compared with our baseline [61] (Separated Composition) and another single-image baseline (Sequentially Joint Optimize) [83].

because of the mis-detected objects from [40]. Since the reconstructed scenes of [83] do not support human-scene contact well, e.g., a sitting body often floats, due to the lack of explicit human-scene contact modeling, it has a better non-collision score but a lower contact score.

Human Pose and Shape (HPS) Estimation. Can we use the estimated 3D scene to, in turn, improve 3D HPS? Here we follow PROX but replace the scanned 3D scene of PROX with our estimated 3D scene. In Tab. 5, we evaluate the HPS estimation on PROX *quantitative* using the metrics as [26]. Specifically, we report (1) the mean per-joint error (PJE) and (2) the mean vertex-to-vertex distance (V2V). For completeness, we also compute these metrics on the Procrustes-aligned predictions (denoted as p.PJE and p.V2V, respectively). But note that the metrics w./o. Procrustes alignment (PJE and V2V) are more meaningful, since we

Methods	IoU_{2D} ↑	IoU_{3D} ↑
Cooperative [30]	68.6	21.4
Holistic++ [10]	75.1	24.9
HolisticMesh [83]	75.6	26.3
Ours	**79.2**	**27.8**

Table 3. Quantitative results for 3D scene understanding (3D object detection) on *PiGraphs* dataset [74].

	Cam. Orien.			Ground Pen	
Methods	pitch ↓	roll ↓	mean ↓	Freq. ↓	Dist. ↓
Total3D [61]	0.059	**0.031**	0.045	0.316	0.167
Ours	**0.042**	0.034	**0.038**	**0.100**	**0.112**

Table 4. Errors in the camera orientation and the ground penetration using foot contact on the PROX *qualitative* dataset.

want to evaluate the translation, rotation and scaling of the human body. As shown in Tab. 5, with estimated camera orientation and ground plane constraints (+CamGP), the PJE and V2V are both improved by a significant margin $+43.21$ and $+42.41$ respectively, w.r.t. our baseline. We also see that our refined scene can further refine our estimated bodies by applying the SDF loss (+SDF) and the contact loss (+Contact) from [26]. Our final body estimation outperforms HolisticMesh [83] and is similar to PROX, without having access to a scanned 3D scanned scene.

4.2. Ablation Study

To analyze the contribution of the accumulated HSIs and the influences of the different constraints, we conducted multiple ablation studies; see Tab. 2. All three proposed HSI constraints (depth, contact, and collision) help improve 3D scene reconstruction in different ways. The *contact* constraint produces the highest human-scene contact scores, but decreases the non-collision score. The *collision* and *depth* both contribute to the non-collision score. However, using only the *depth* achieves a slightly better 3D scene evaluation than our full model, but leads to worse human-scene contact scores. By applying all constraints, our method can generate a more accurate 3D scene, which supports more physically plausible HSI.

4.3. Qualitative Analysis

In Fig. 5, we show reconstructed 3D scenes and humans along with RGB videos, to demonstrate the effectiveness and generality of our approach on different datasets (PROX and PiGraphs). MOVER recovers better 3D scenes and HPS compared to our baseline [61] (Separated Composition) and another single-image baseline [83] (Sequentially Joint Optimize). See Sup.Mat. for more examples.

5. Discussion

Based on single-view inputs, our proposed method optimizes the 3D alignment of objects in a *static scene*. However, humans also move objects, resulting in a dynamical scene layout. While our approach uses individual mesh models for each object, we assume a static scene. Nevertheless, we believe that our proposed constraints based on HSIs will be beneficial for future work on the reconstruction of dynamic scenes. Besides optimizing the 3D scene layout, we do not change the initial shape estimate of an object. A more flexible and adjustable geometry representation, e.g., an implicit representation, would be needed, since the initial mesh could have a wrong topology.

Human motion reconstruction and 2D instance segmentation struggle with severe occlusions in the input, which leads to poor estimations of HSIs, and, thus, influence our 3D scene layout prediction. While not the scope of our work, the robustness and accuracy of human motion estimation can be improved by incorporating human motion priors. Also, jointly predicting human motion and the 3D scene with HSIs in a probabilistic framework can be another interesting direction for future work.

With **G.T Captured 3D Scene Scans**				
Methods	PJE↓	V2V ↓	p.PJE ↓	p.V2V↓
RGB [26]	220.27	218.06	73.24	60.80
PROX [26]	167.08	166.51	71.97	61.14
With **Image2Mesh Models**				
HolisticMesh [83]	190.78	192.21	**72.72**	**61.01**
baseline*	219.62	222.50	75.92	68.34
+CamGP	176.41	180.09	73.41	67.33
+CamGP+SDF	175.98	179.98	73.96	68.29
Ours	**174.37**	**178.31**	73.60	67.89

Table 5. Quantitative results for human pose estimation on PROX *quantitative* dataset (baseline* denotes batch-wise SMPLify-X, **Ours**: +CamGP+SDF+Contact.)

6. Conclusion

We have introduced MOVER, which reconstructs a 3D scene by exploiting 3D humans interacting with it. We have demonstrated that accumulated HSIs, computed from a monocular video, can be leveraged to improve the 3D reconstruction of a scene. The reconstructed scene, in turn, can be used to improve 3D human pose estimation. In contrast to the state of the art, MOVER can reconstruct a consistent, physically plausible 3D scene layout.

Acknowledgments. We thank Yixin Chen, Yuliang Xiu for their fruitful feedback and discussions, Yao Feng, Partha Ghosh and Maria Paola Forte for proof-reading, and Benjamin Pellkofer for IT support. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B.

Disclosure. MJB has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. MJB has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH.

References

[1] https://github.com/vchoutas/smplx/tree/master/transfer_model. 13

[2] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5167–5176, 2018. 2

[3] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of PEople. *Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 2

[4] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr revisited: 2D-3D alignment via surface normal prediction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5965–5974, 2016. 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, volume 9909, pages 561–578, 2016. 3

[6] Aljaž Božič, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. TransformerFusion: Monocular RGB scene reconstruction using transformers. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1

[7] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2021. 2, 6

[8] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[10] Yixin Chen, Siyuan Huang, Tao Yuan, Yixin Zhu, Siyuan Qi, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3D holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, pages 8647–8656, 2019. 1, 2, 3, 5, 6, 8

[11] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Understanding indoor scenes using 3D geometric phrases. In *Computer Vision and Pattern Recognition (CVPR)*, pages 33–40, 2013. 3

[12] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision (ECCV)*, volume 12355, pages 20–40, 2020. 3

[13] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *European Conference on Computer Vision (ECCV)*, volume 9912, pages 628–644, 2016. 3

[14] M Comer, C Bouman, and J Simmons. Statistical methods for image segmentation and tomography reconstruction. *Microscopy and Microanalysis*, 16:1852 – 1853, 2010. 6

[15] Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. Gravity-aware monocular 3d human-object reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[16] Manuel Dahnert, Ji Hou, , Matthias Nießner, and Angela Dai. Panoptic 3D scene reconstruction from a single RGB image. *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3

[17] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. DeLay: Robust spatial layout estimation for cluttered indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 616–624, 2016. 3

[18] Maximilian Denninger and Rudolph Triebel. 3d scene reconstruction from a single viewport. In *European Conference on Computer Vision (ECCV)*, pages 51–67. Springer, 2020. 3

[19] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018. 3

[20] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding Humans: Non-parametric 3D human shape estimation from single images. In *International Conference on Computer Vision (ICCV)*, pages 2232–2241, 2019. 2

[21] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 9785–9795, 2019. 3

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017. 3

[23] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. Atlasnet: A papier-mâché approach to learning 3d surface generation. *CoRR*, abs/1802.05384, 2018. 3

[24] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10884–10894, 2019. 1

[25] Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and Hans-Peter Seidel. A statistical model of human pose and body shape. *Computer Graphics Forum*, 28(2):337–346, 2009. 2

[26] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 1, 2, 6, 7, 8, 15

[27] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 4, 6

[28] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 3

[29] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *International Conference on Computer Vision (ICCV)*, pages 1849–1856, 2009. 3

[30] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 207–218, 2018. 1, 3, 8

[31] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3D scene parsing and reconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, volume 11211, pages 194–211, 2018. 3, 6

[32] Hamid Izadinia, Qi Shan, and Steven M Seitz. IM2CAD. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5134–5143, 2017. 3

[33] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[34] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588, 2020. 4

[35] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12354, pages 196–214, 2020. 2

[36] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In *International Conference on 3D Vision (3DV)*, 2020. 1, 3

[37] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8320–8329, 2018. 2, 3, 13

[38] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131, 2018. 1, 3

[39] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2015. 5

[40] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. PointRend: Image segmentation as rendering. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9799–9808, 2020. 2, 4, 5, 6, 7

[41] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5253–5263, 2020. 1

[42] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5252–5262, 2020. 3

[43] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11127–11137, 2021. 1, 2, 3, 4, 6, 13

[44] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, 2019. 1

[45] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019. 3

[46] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4496–4505, 2019. 2

[47] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2CAD: 3D shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision (ECCV)*, volume 12348, pages 260–277, 2020. 3

[48] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[49] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2016. 3

[50] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2136–2143, 2009. 3

[51] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3D indoor environments. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12368–12376, 2019. 3

[52] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[53] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2019–2028, 2020. 3

[54] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 2

[55] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1987. 5

[56] Arun Mallya and Svetlana Lazebnik. Learning informative edge maps for indoor scene layout prediction. In *International Conference on Computer Vision (ICCV)*, pages 936–944, 2015. 3

[57] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose

estimation. In *International Conference on Computer Vision (ICCV)*, pages 2659–2668, 2017. 2

[58] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, 2019. 3, 5

[59] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding (CVIU)*, 104(2):90–126, 2006. 2

[60] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J Mitra. iMapper: interaction-guided scene mapping from monocular videos. *Transactions on Graphics (TOG)*, 38(4):92:1–92:15, 2019. 3

[61] 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. In *Computer Vision and Pattern Recognition (CVPR)*, pages 55–64, 2020. 1, 2, 3, 4, 5, 6, 7, 8, 13, 15

[62] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 3

[63] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3, 4, 6, 13

[64] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3

[65] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *International Conference on Computer Vision (ICCV)*, pages 803–812, 2019. 1

[66] Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 6

[67] Grégory Rogez and Cordelia Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3108–3116, 2016. 2

[68] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6):245:1–245:17, 2017. 2

[69] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[70] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, pages 2304–2314, 2019. 2

[71] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, pages 84–93, 2020. 2

[72] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[73] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding (CVIU)*, 152:1–20, 2016. 2

[74] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. *Transactions on Graphics (TOG)*, 35(4):139:1–139:12, 2016. 2, 6, 8, 13, 15

[75] Daeyun Shin, Zhile Ren, Erik B Sudderth, and Charless C Fowlkes. 3d scene reconstruction with multi-layer depth and epipolar transformers. In *International Conference on Computer Vision (ICCV)*, pages 2172–2182, 2019. 3

[76] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 3

[77] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. FACSIMILE: Fast and accurate scans from an image in less than a second. In *International Conference on Computer Vision (ICCV)*, pages 5329–5338, 2019. 2

[78] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3D human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016. 2

[79] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698, 2017. 2

[80] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *European Conference on Computer Vision (ECCV)*, pages 20–38, 2018. 2

[81] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018. 3

[82] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. DOPE: distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12371, pages 380–397, 2020. 2

[83] Zhenzhen Weng and Serena Yeung. Holistic 3D human and scene mesh estimation from single view images. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 5, 6, 7,

8, 13, 15

[84] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10957–10966, 2019. 3

[85] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6183–6192, 2020. 2

[86] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[87] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes - the importance of multiple scene constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2148–2157, 2018. 3

[88] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3D scene understanding from a single image with implicit representation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 8833–8842, 2021. 1, 3

[89] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, volume 12357, pages 34–51, 2020. 2, 3, 4, 5

[90] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. PLACE: Proximity learning of articulation and contact in 3D environments. In *International Conference on 3D Vision (3DV)*, 2020. 3, 6

[91] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204, 2020. 3, 6

[92] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5287–5295, 2017. 3

[93] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3119–3126, 2013. 3

[94] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3D human reconstruction from a single image. In *International Conference on Computer Vision (ICCV)*, pages 7738–7748, 2019. 2

[95] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3D reconstruction with RGB-D cameras. *Computer Graphics Forum (CGF)*, 37(2):625–652, 2018. 2

[96] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834, 2021. 15

Appendices

In this supplemental document, we provide additional information about the dataset, implementation details, extended sensitivity analysis, failure cases, additional qualitative results and discussion of potential misuse.

A. Dataset

PiGraphs. PiGraphs [74] consists of 60 RGB-D videos of 30 scenes. The dataset is recorded with a *Microsoft Kinect One*, and is designed to capture human and object arrangements in different kinds of interaction. Each video recording is about 2-minute long with 5 fps. It contains labeled 3D bounding boxes of objects in the scene and human poses represented as 3D skeletons. We use this dataset to evaluate the scene reconstruction and compare with [61, 83]. Note that the provided human poses are noisy and not suitable for an evaluation of 3D human shape and pose estimation.

PROX Qualitative. PROX *qualitative* contains 61 RGB-D videos at 30 fps of human motion/interaction in 12 scanned static 3D scenes. The data has been recorded using the *Microsoft Kinect One* and *StructureIO* sensor. To enable 3D scene reconstruction evaluation on this dataset, we segment and label each object with its 3D bounding box. Since there are two scenes (i.e., "BasementSittingBooth" and "N0SittingBooth") containing an inseparable object (see Fig. 6), we evaluate all methods on the remaining 10 scenes using the corresponding 51 videos as input.

PROX Quantitative. PROX *quantitative* captures a sequence of human-scene interaction RGB-D frames within a synchronized *Vicon* marker-based motion capturing system. In total, the dataset contains 178 frames and provides groundtruth body meshes, which accounts for human pose and shape (HPS) evaluation. For fair evaluation on HPS, we input all images into HolisticMesh [83] and ours to get a refined scene and use a refined scene to get refined bodies. In addition, we also label this scene for 3D scene reconstruction evaluation, see Fig. 6.

B. Implementation Details

Loss Terms. The *2D bounding box term* $\mathcal{L}_{\text{bbox}}$ is a ℓ_1 norm between the object's projected 3D bounding box $Proj_i$ and its corresponding detected 2D bounding box Det_i, expressed with the top-left corner coordinate x_{min}, y_{min} and $width$ value.

$$\mathcal{L}_{\text{bbox}} = \sum_i \|Proj_i^{\alpha} - Det_i^{\alpha}\|, \quad \alpha \in \{x_{min}, y_{min}, width\}.$$

The *scale term* prevents object scales s deviating far from the initial estimates s^{init} from Total3D [60]:

$$\mathcal{L}_{\text{scale}} = \sum_i \|\frac{s_i}{s_i^{init}} - 1.0\|_2.$$

Initial Estimate of 3D Bodies. We use PARE [43] to initialize the body poses and shape (shape β, pose θ, scale s). Since our approach uses the SMPL-X [63] model, we apply [1] to convert the SMPL parameter estimated from PARE. In addition, we use the calibrated camera intrinsic parameters K provided by the datasets (PiGraph and PROX). To convert the estimations of PARE which uses a weak perspective camera model, we compute the corresponding translation t^{body} by:

$$\Pi_{K_0}\left(s(R_\theta(J(\beta))\right) = \Pi_K\left((R_\theta(J(\beta)) + t^{body}\right),$$

where K_0 denotes the camera intrinsic parameters of the weak perspective camera model with focal length 5000.

Contact Regions of Objects. We automatically calculate the contact regions of objects based on the normal of the vertices. Specifically, the vertices, whose normals are along y-axis, are the bottom or top part of the objects, while the vertices with along z-axis normal are the back part of the objects. We term that sofas and chairs have two contact regions, i.e., bottom and back parts, while beds and tables only have top part as the contact region, shown in Fig. 7.

Optimization. We use the Adam optimizer [37] to optimize the final energy term with a step size of 0.002 and 3000 iterations. We set $\lambda_1, \lambda_2, \lambda_3$ as $1000, 0.3, 1000$ respectively, for 2D bounding box term, occlusion-aware term and scale term. The weights of our proposed depth order constraint, collision constraint, and contact constraint are set to $\lambda_4 = 8, \lambda_5 = 1000$, and $\lambda_6 = 1e5$, respectively. We use two robust Geman-McClure error fuctions, ρ_1, ρ_2 with parameter 0.1 on 3D joints, and one ρ_3 with parameter 100 on 2D projection of 3D joints.

Our method takes around 30 minutes for 3000 iterations to optimize a 3D scene with accumulated HSIs constraints. In comparison, HolisticMesh [83] which jointly optimizes human and a 3D scene for one single image, directly trains the parameters of the network in Total3DUnderstanding [61] to regress the 3D scene, which is time-consuming and costs around 40 minutes. For the human optimization, it runs twice (5 minutes), i.e., one is a HPS initialization used to refine the scenes, and the second pass is done using the refined scenes. In total, HolisticMesh takes 45 minutes for one single image. Our method takes almost the same time for a scene (around 10 objects) regardless how many frames in the input video. The number of frames in a video only influences the time of calculating the depth map, the SDF volume and the contact information of each body. However this can be done once and is easily processed in parallel before the optimization. In contrast, HolisticMesh [83] processes a video sequentially, i.e., one frame after another. Therefore, the optimization time increases w.r.t. the number of frames in a video.

(A) PROX *qualitative* dataset (B) PROX *quantitative* dataset

Figure 6. We crop out each object separately and label the corresponding 3D bounding box for 10 scenes in PROX *qualitative* dataset and one scene in PROX *quantitative* dataset.

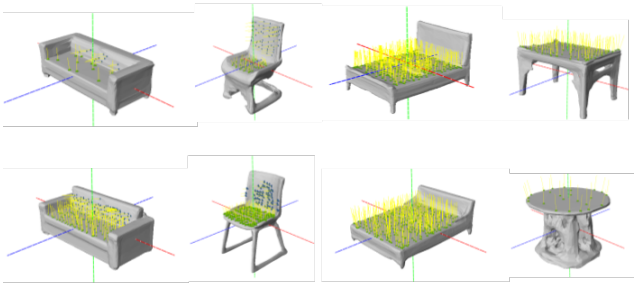


Figure 7. Contact regions of different objects.

C. Sensitivity Analysis.

Our approach uses HSIs observed in a video. A longer video potentially has more HSIs, which results in more constraints for our objective function. In Tab. 6, we analyze how different video lengths influence scene reconstruction, by reporting the 3D intersection-over-union (IoU) metric. Specifically, we use 10 sequences of the PROX qualitative dataset (one sequence per scene) and randomly sample 10 segments of 10s, 20s, 30s length from each sequence. We observe that longer sequences result in better performance, i.e., higher IoU and lower standard deviation. We will add this experiment and clarify that performance depends on the *number of HSIs* and not the video length, i.e., a short video with many HSIs results in a better reconstruction than a long video with a few unique HSIs.

	10s	20s	30s	entire videos (51s)
3D IoU mean ↑	0.389	0.395	0.407	0.424
3D IoU std. ↓	0.018	0.015	0.010	-

Table 6. Ablation study on different length of videos as input. The average length of entire videos is 51s.

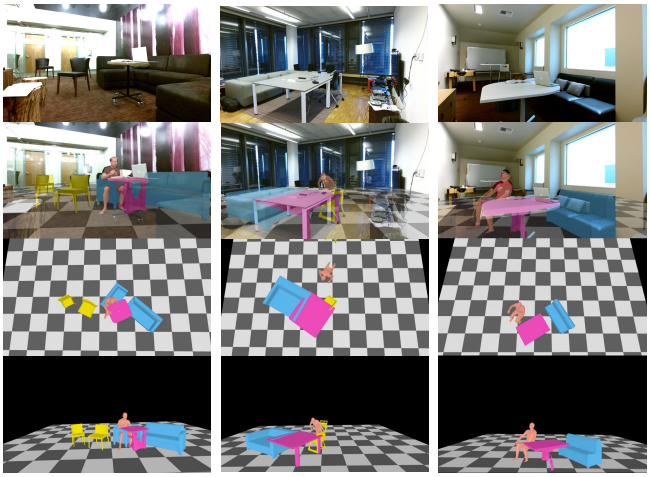
We also did a sensitivity study w.r.t. noise in the initialization. In Tab. 7, we add uniform noise on the initial scale, translation and orientation of objects predicted by Total3D [60], and report the 3D IoU. MOVER is robust to noisy orientation and translation estimates from Total3D [60], but sensitive to the scale variation. This is because we cur-

scale noise	± 25%	± 15%	± 0.05%
3D IoU ↑	0.345	0.3805	0.4105
transl.	± 30cm	± 20cm	±10m
3D IoU ↑	0.4175	0.416	0.415
orien.	±45°	±30°	±15°
3D IoU ↑	0.4205	0.418	0.4205

Table 7. Sensitivity analysis on scene reconstr. with uniform noise on input scale, translation and orientation from Total3D [60] (*Werkraum_03301_01* video). Scene w/o noise has 0.417 3D IoU.

Methods	Scene Recon.			HSI	
	IoU_{3D} ↑	P2S↓	IoU_{2D} ↑	Non-Col ↑	Cont. ↑
HolisticMesh [83]	0.239	0.133	0.533	0.948	0.951
Total3D [61]	0.063	0.409	0.342	0.940	0.436
Ours	**0.390**	**0.095**	**0.862**	**0.972**	**0.934**

Table 8. Quantitative results for 3D scene understanding (3D object detection) and human-scene interaction on the PROX *quantitative* dataset. P2S, Non-Col and Cont denote *point2surface distance*, Non-Collision and Contactness respectively.



(A) (B) (C)

Figure 8. Failure cases. (A) The estimated sofa has arms, which does not match the unarmed sofa in the input image. (B) The half bottom body is occluded, that leads to a wrong pose estimation as well as HSI observation. (C) The body is sitting "in the air", where the chair is missing.

rently regularize the optimization to the initial scale relatively strongly; i.e., we cannot deviate much from a noisy estimate to "correct" it. Relaxing $\mathcal{L}_{\text{scale}}$ easily resolves this.

D. More Evaluation Results on PROX Quantitative Dataset.

We also evaluate 3D scene reconstruction and human-scene interaction on PROX *quantitative*, as shown in Tab. 8. Our method improves our input baseline [61] significantly and outperforms the previous method [83] with a big margin in both 3D scene reconstruction metrics and human-scene interaction metrics.

E. Failure Cases

In this section, we discuss and show the failure cases of our method. Besides optimizing the 3D scene layout, we do not change the initial shape estimate of an object. Thus, wrong estimated geometry shape can still violate human's interaction, as shown in (A) in Fig. 8. A more flexible and adjustable geometry representation, e.g., an implicit representation, would be needed. Human motion reconstruction struggle with severe occlusions in the input, that leads to wrong body poses as well as poor estimations of HSIs, and, thus, influence our 3D scene layout prediction, see (B) in Fig. 8. While not the scope of our work, the robustness and accuracy of human motion estimation can be improved by incorporating human motion priors or learning-based probabilistic human pose and estimation network. Severe occlusion can also causes missing objects in the scene like the chair in Fig. 8(C).

In our pipeline, we currently consider the contact between *detected* objects and bodies. As a potential future extension of our method, one can also leverage the information from 2D learning-based human-object interaction (HOI) detection network [96], by using contacted bodies to discover missing objects; or learn a model that jointly regress human-object interaction and their shape.

F. Additional Qualitative Results

In Fig. 9 and Fig. 10, we present additional qualitative results on PROX [26] qualitative and PiGraphs [74] dataset respectively. As can be seen, our method performs well on a variety of different scenes and predicts a physically plausible and functional scene layout. We also refer to the suppl. video for results.

G. Discussion of Potential Misuse

Our approach is not intended for any surveillance application. Our goal is to understand how humans interact and move in scenes from videos (e.g., from TV sitcoms), to this end both the scene geometry and the human pose need to be reconstructed. Our method could be misused in potential surveillance applications that curtail human rights and civil liberties, but we will restrict the usage of our method in a legal way.

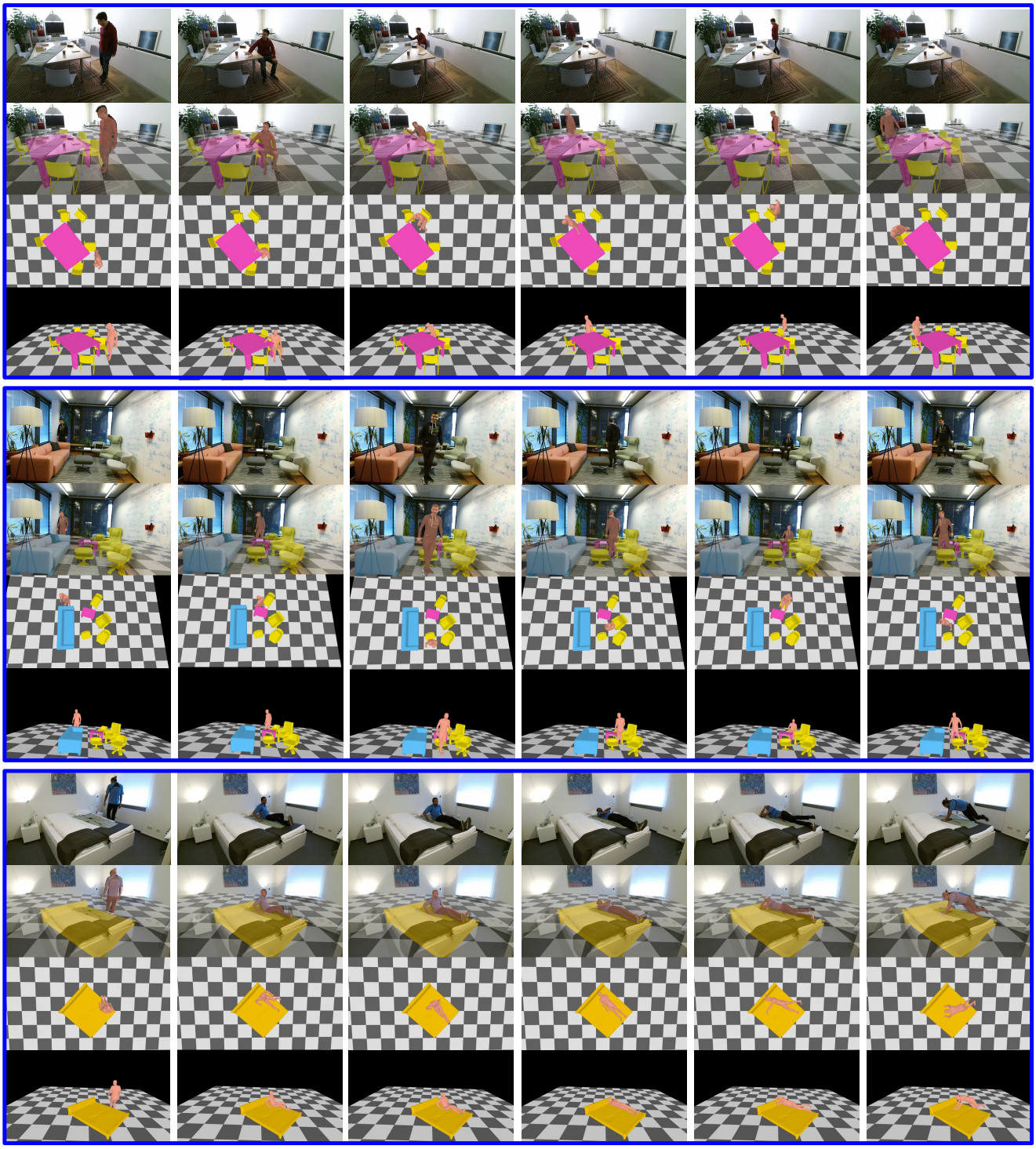


Figure 9. More qualitative results on PROX qualitative dataset.

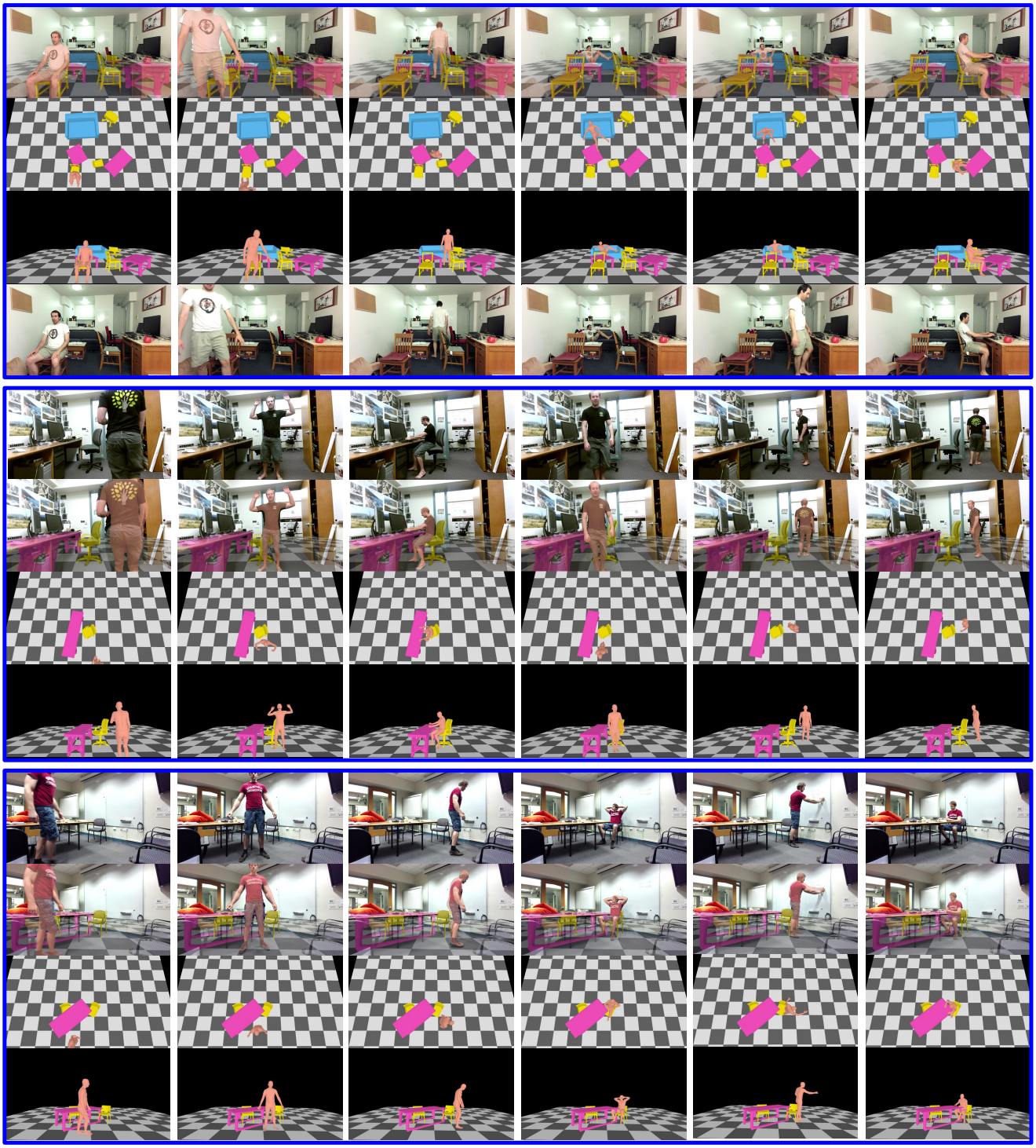


Figure 10. More qualitative results on PiGraphs dataset.