PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image

PRADYUMNA YALANDUR MURALIDHAR*, University of Tuebingen, Germany and Zuse School ELIZA, Germany
YUXUAN XUE*†, University of Tuebingen, Germany
XIANGHUI XIE, University of Tuebingen, Germany and Max Planck Institute for Informatics, Germany
MARGARET KOSTYRKO, University of Tuebingen, Germany
GERARD PONS-MOLL, University of Tuebingen, Germany and Max Planck Institute for Informatics, Germany

Fig. 1. Given a single monocular RGB image containing a human in a complex environment, **PhySIC** reconstructs metrically aligned 3D human and scene geometries as well as a dense vertex-level contact map. Our method jointly optimizes human pose, scene geometry, and global scale to produce a physically plausible human-scene pair, accurately capturing contact and interactions such as sitting and foot-floor adherence, even in the presence of occlusions. Image.

Reconstructing metrically accurate humans and their surrounding scenes from a single image is crucial for virtual reality, robotics, and comprehensive 3D scene understanding. However, existing methods struggle with depth ambiguity, occlusions, and physically inconsistent contacts. To address these challenges, we introduce **PhySIC**, a unified framework for physically plausible Human–Scene Interaction and Contact reconstruction. PhySIC recovers metrically consistent SMPL-X human meshes, dense scene surfaces, and vertex-level contact maps within a shared coordinate frame, all from a single RGB image. Starting from coarse monocular depth and parametric body estimates, PhySIC performs occlusion-aware inpainting, fuses visible depth with unscaled geometry for a robust initial metric scene scaffold, and synthesizes missing support surfaces like floors. A confidence-weighted optimization subsequently refines body pose, camera parameters, and global scale by jointly enforcing depth alignment, contact priors, interpenetration avoidance, and 2D reprojection consistency. Explicit occlusion masking safeguards invisible body regions against implausible configurations. PhySIC is highly efficient, requiring only 9 seconds for a joint human-scene optimization and less than 27 seconds for end-to-end reconstruction process. Moreover, the framework naturally handles multiple humans, enabling reconstruction of diverse human scene interactions. Empirically, PhySIC substantially outperforms single-image baselines, reducing mean per-vertex scene error from 641 mm to 227 mm, halving the pose-aligned mean per-joint position error (PA-MPJPE) to 42 mm, and improving contact F1-score from 0.09 to 0.51. Qualitative results demonstrate that PhySIC yields realistic foot-floor interactions, natural seating postures, and plausible reconstructions of heavily occluded furniture. By converting a single image into a physically plausible 3D human-scene pair, PhySIC advances accessible and scalable 3D scene understanding. Our implementation is publicly available at https://yuxuan-xue.com/physic.

CCS Concepts: • **Computing methodologies** → *Reconstruction*; **Scene understanding**; *Machine learning approaches.*

Additional Key Words and Phrases: Human-Scene Interaction, Digital Human, Reconstruction, 3D Scene Understanding

ACM Reference Format:
Pradyumna Yalandur Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. 2025. PhySIC: Physically Plausible 3D Human-Scene Interaction and Contact from a Single Image. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757377.3763862

*Equal contribution. †Corresponding Author.

Authors' Contact Information: Pradyumna Yalandur Muralidhar, University of Tuebingen, Tuebingen, Germany and Zuse School ELIZA, Darmstadt, Germany, pradyumna.yalandur-muralidhar@student.uni-tuebingen.de; Yuxuan Xue, University of Tuebingen, Tuebingen, Germany, yuxuan.xue@uni-tuebingen.de; Xianghui Xie, University of Tuebingen, Tuebingen, Germany and Max Planck Institute for Informatics, Tuebingen, Germany, xianghui.xie@uni-tuebingen.de; Margaret Kostyrko, University of Tuebingen, Tuebingen, Germany, Margaret.Kostyrko@student.uni-tuebingen.de; Gerard Pons-Moll, University of Tuebingen, Tuebingen, Germany and Max Planck Institute for Informatics, Tuebingen, Germany, gerard.pons-moll@uni-tuebingen.de.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763862

1 Introduction

Holistic 3D understanding of humans and their surrounding environments is essential for emerging technologies such as embodied AI, sports analytics, and augmented reality. These applications require precise scene geometry, accurate localization of humans within scenes, and coherent ground contact estimation. Existing methods, however, usually consider either only static scenes without human [Chen et al. 2019], or only human pose estimation assuming given 3D scene [Hassan et al. 2019a]. Recent method HolisticMesh [Weng and Yeung 2021] can predict both scene and human from single RGB image. Nevertheless, it is limited to a single human interacting with specific indoor furniture categories, which does not scale up to arbitrary scene types. While more recent approaches like HSR [Xue et al. 2024] and HSfM [Müller et al. 2024] achieve holistic human-scene reconstruction, they require video input or multi-view images respectively, limiting their applicability to single-image scenarios.

However, having a general method that can handle diverse scene types and an arbitrary number of humans interacting with a scene is very challenging. The model needs to reason about different scene geometries, intricate human-scene contacts under depth-scale ambiguity, occlusion to both human poses and scene geometry, all from a single RGB image while being fast for practical applications.

Our idea to address these challenges is to simultaneously reason about human and scene, leveraging strong geometry priors from foundation models. During interaction, the scene physically limits possible human poses and human pose provides crucial cues for estimating scene geometry and scale. Based on this observation, we propose **PhySIC**, PhySically plausible human scene Interaction and Contacts from single RGB image. Starting from coarse monocular depth and initial parametric body estimates, our method jointly optimizes these components through an objective that harmonizes *reliable depth alignment*, *realistic contact encouragement*, *interpenetration avoidance*, and *2D reprojection consistency*, yielding coherent 3D human-scene reconstruction. In essence, PhySIC transforms a single RGB image into: (i) a metrically scaled SMPL-X human mesh, (ii) a comprehensive scene representation including dense surfaces and essential support structures like floors, and (iii) a vertex-level dense contact map within a shared metric coordinate system. Our framework is highly efficient and can process one image in *less than 27 seconds*, making it possible to transform everyday images into physically consistent 3D human-scene pairs. This paves the way for scalable single-image 3D understanding.

We evaluated PhySIC on the PROX [Hassan et al. 2019a] and RICH [Huang et al. 2022] datasets. Results show that our method significantly outperforms previous SOTA, HolisticMesh [Weng and Yeung 2021]: on PROX dataset, our method improves the mean joint error of human pose from 77mm to 42mm and contact F1 score from 0.39 to 0.51. Experiments on diverse internet images demonstrate the superior applicability of our approach to various interaction and scene types. Our contributions are summarized as follows:

- We propose PhySIC, the first metric-scale human-scene reconstruction method that can handle multiple humans, diverse scene and interaction types.
- We introduce a robust initialization strategy and occlusion-aware joint optimization, providing valuable insights for human scene reconstruction.
- Our highly efficient reconstruction pipeline will be publicly released, democratizing human scene reconstruction and interaction data collection.

2 Related work

2.1 Single View to 3D Human

Reconstructing the shape and pose of 3D humans from monocular images has seen significant advances, particularly with parametric models such as SMPL [Loper et al. 2015] and its extensions to SMPL-X [Pavlakos et al. 2019], which enables expressive full-body estimation including hands and face. Early methods like SMPLify [Bogo et al. 2016] optimized body parameters to fit 2D joint detections. Subsequent deep learning methods, including HMR [Kanazawa et al. 2018], SPIN [Kolotouros et al. 2019], and PARE [Kocabas et al. 2021] introduced end-to-end regression and attention mechanisms to improve robustness to occlusion and truncation. WHAM [Shin et al. 2024] and TRAM [Wang et al. 2024a] combine human mesh recovery with SLAM-based camera tracking, enabling accurate global localization of SMPL bodies in world coordinates from monocular video. Recently, large-scale learning-based models such as NLF [Sárándi and Pons-Moll 2024] leverage over 25 million annotated frames to directly regress both SMPL-X parameters and global position from a single image, achieving state-of-the-art generalization and accuracy across diverse scenes and poses. Despite these advances, existing methods often lack explicit reasoning about physical interaction or consistency with the surrounding 3D scene, leading to floating, misaligned, or physically implausible human reconstructions. Our work addresses these issues by enabling metrically aligned, physically plausible human recovery that is explicitly consistent with the reconstructed scene.

2.2 Single View to 3D Scene

Early methods for monocular 3D scene reconstruction leveraged geometric and semantic priors to recover layouts, object placements, and meshes from a single RGB image. Notable among these is Total3D [Nie et al. 2020], which jointly infers room layout and object pose. Mesh R-CNN [Gkioxari et al. 2019] and MonoScene [Cao and de Charette 2022] further advance object-centric mesh prediction and semantic scene completion. Recent breakthroughs in monocular depth estimation, such as ZoeDepth [Bhat et al. 2023], Metric3D [Hu et al. 2024], and DepthPro [Bochkovskii et al. 2024], employ large-scale pretraining and transformers to predict sharp, scale-consistent depth, enabling realistic metric point cloud extraction. Gen3DSR [Ardelean et al. 2025] builds on these estimators with category-specific object reconstruction, but omits human modeling and thus cannot reason about physical contact or interaction. In contrast, our method leverages state-of-the-art depth estimation together with explicit human modeling, enabling physically plausible, metrically aligned human-scene reconstruction from a single image, beyond the capabilities of prior object- or scene-centric approaches.

Table 1. Comparison between existing human-scene reconstruction methods and ours. Our method can handle multi-human interaction in both indoors and outdoors, and predicts the full scene at much faster speed.

Method	Multi-human	Scene types	RGB input only	Output	Runtime
PROX	✘	Indoor	✘	Objects	73 sec.
Mover	✘	Indoor	✘	Objects	30 min.
HolisticMesh	✘	Indoor	✔	Objects	5 min.
Ours	✔	In+Outdoor	✔	Full scene	27 sec.

2.3 3D Human-Scene Interaction

Modeling and reconstructing plausible human–scene interactions is central to scene understanding. Early benchmarks addressed interaction detection [Liu et al. 2020], generation [Savva et al. 2016], and pose refinement [Hassan et al. 2019b] with scene constraints. PROX [Hassan et al. 2019a] introduced interpenetration and contact penalties but assumes access to static scene scans. In contrast, our method reconstructs metric-scale scenes from a single RGB image. Several approaches infer scene structure from human motion [Li et al. 2024; Nie et al. 2021; Yi et al. 2022], while large-scale capture works such as EgoBody [Zhang et al. 2022] and HPS [Guzov et al. 2021] provide detailed multi-person and metric pose data using wearable sensors, though they require specialized hardware and do not address single-image reconstruction.

Dynamic tracking and contact estimation approaches, such as CHORE [Xie et al. 2022], InterTrack [Xie et al. 2025], and DECO [Tripathi et al. 2023], can reconstruct articulated humans and contacts, but often rely on incomplete scene geometry. Generative models like ParaHome [Kim et al. 2024] simulate diverse 3D human–object interactions, yet focus on activity synthesis rather than image-based reconstruction. Placement-focused works (e.g., POSA [Hassan et al. 2021], PLACE [Zhang et al. 2020], Putting People in Scenes [Li et al. 2019]) leverage statistical priors, but typically lack dense, metrically accurate scene recovery. Recent methods in holistic reconstruction, such as RICH [Huang et al. 2022], HSR [Xue et al. 2024], HolisticMesh [Weng and Yeung 2021], and the work by Biswas et al. [Biswas et al. 2023] move toward integrated scene understanding but often require controlled environments. In contrast, our method reconstructs metrically accurate, physically plausible humans and diverse scenes with dense contact reasoning directly from a single image, enabling multi-human and in-the-wild scenarios (see Tab. 1).

Our work is most closely related to HSfM [Müller et al. 2024], which reconstructs 3D human-scenes from uncalibrated multi-view images using joint optimization. To our knowledge, PhySIC is the first method to reconstruct both 3D human-scenes and their interactions from a single monocular image: a particularly challenging task due to monocular ambiguity and severe occlusions, yet highly practical given its applicability to internet images. Several additional technical design choices further differentiate PhySIC from HSfM; Please refer to our supplementary material for further details.

3 Method

Given a single RGB image, our method PhySIC predicts metric-scale dense scene point clouds and 3D human mesh with accurate vertex-level contact maps. This is a highly complex problem which requires accurate reasoning of sophisticated human poses and diverse scene geometries under heavy human-scene occlusions. We decompose this problem into separate metric-scale scene estimation (Sec. 3.1) and human reconstruction with alignment to the scene (Sec. 3.2). Human and scene are inherently constrained by each other, which we leverage for a joint optimization to obtain physically plausible human-scene contacts (Sec. 3.3). An overview of our method can be found in Figure 2.

For notational simplicity, we explain our method for single human interaction with scene, but our approach seamlessly handles multiple humans. Specifically, given input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$, PhySIC outputs scene point map $\mathcal{P}_s$ and human mesh vertices $\mathcal{V}_h \in \mathbb{R}^{N\times 3}$ using the SMPL-X body model [Pavlakos et al. 2019].

3.1 Stage 1: Metric-Scale Scene with Detailed Geometry

3.1.1 Scene image inpainting. From monocular image, the human can heavily occlude the background scene, which leads to missing regions if one simply ignores the human when reconstructing the scene, causing false-negative interactions. Instead, we first inpaint the scene to fill in the missing regions and then run scene reconstruction for the complete scene, as shown in Fig. 2. Specifically, we use SAM2 [Ravi et al. 2024] to obtain the human mask and adopt OmniEraser [Wei et al. 2025] to inpaint the human region, yielding an image $\mathcal{I}_s$ with unobstructed view of the scene.

3.1.2 Metric-scale scene points. Our goal is to obtain accurate metric-scale scene points from an image. Existing depth estimators like DepthPro [Bochkovskii et al. 2024] can predict accurate metric-scale depth, however, lack detailed geometry. On the other hand, some models such as MoGe [Wang et al. 2024b] can capture fine-grained details, but the results reside in relative space. We leverage the best of both worlds to obtain metric scene scale with accurate and detailed geometry. Specifically, using the inpainted scene image $\mathcal{I}_s$, we first obtain metric depth map $\mathcal{D}_s$ from DepthPro and unscaled relative point maps $\mathcal{P}_s^{\text{rel}}$ from MoGe. Since MoGe prediction is pixel-aligned, we can align the point maps $\mathcal{P}_s^{\text{rel}}$ with metric depth $\mathcal{D}_s$ by optimizing scale s and translation $\mathbf{t}_z$:

$$(s^*, \mathbf{t}_z^*) = \arg\min_{s,\mathbf{t}_z} \left\| (s \cdot \hat{\mathcal{P}}_s^{\text{rel}} + \mathbf{t}_z) - \pi^{-1}(\mathcal{D}_s, K_{\mathcal{D}}) \right\|_2^2, \tag{1}$$

where π^{-1} is the back-projection function and intrinsic $K_{\mathcal{D}}$ is predicted by DepthPro. We optimize only the depth shift in $\mathbf{t}_z$ and solve this using RANSAC. The metric-scale point-map $\hat{\mathcal{P}}_s$ can then be obtained by: $\hat{\mathcal{P}}_s = s^* \cdot \hat{\mathcal{P}}_s^{rel} + \mathbf{t}_z^*$.

3.1.3 Ground plane fitting. The point map $\hat{\mathcal{P}}_s$ captures accurate local geometry, but can suffer from missing or unreliable floor geometry, which is important for precise human-scene interaction. To this end, we fit a plane to the floor points using normal constraints. Specifically, we adopt SAM2 to obtain a 2D mask of the floor, which is used to segment 3D floor points from $\hat{\mathcal{P}}_s$. We then use RANSAC to fit a plane robustly to the floor points, aligning both normals and positions. We estimate the normal of each point using its two immediate neighboring points defined in a 2D pixel grid.

3.1.4 Combined scene points. We obtain additional floor points $\mathcal{P}_f$ by sampling a 2D grid of points on the plane within the extents of the scene. The final 3D scene, as our initialization for the next step,

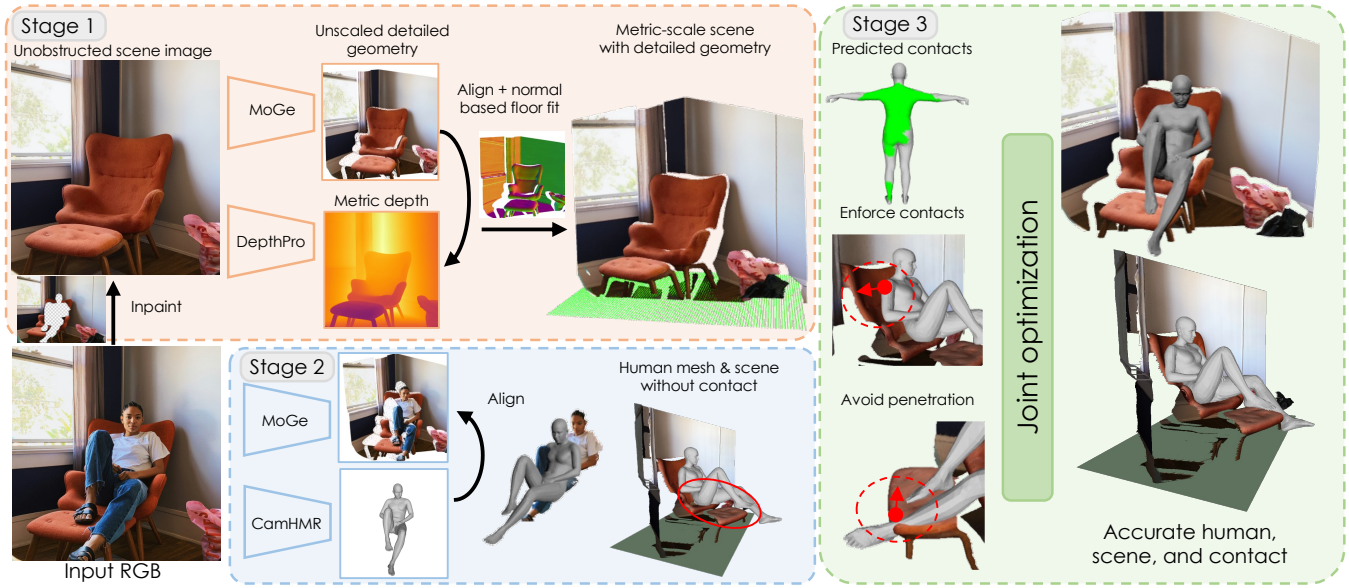

Fig. 2. **Method overview.** Given a single RGB image, we obtain accurate human, scene and contact reconstruction in 3D. We first obtain a complete metric scale scene with detailed geometry (Stage 1, Sec. 3.1) and initialize human mesh which roughly aligns with the scene without contacts (Stage 2, Sec. 3.2). We then jointly optimize human and scene to satisfy contact constraints while avoiding penetrations (Stage 3, Sec. 3.3). Image from Unsplash.

is formed by the union of the refined scene point cloud $\hat{\mathcal{P}}_s$ and the synthesized floor plane points $\mathcal{P}_f$:

$$\mathcal{P}'_s = \hat{\mathcal{P}}_s \cup \mathcal{P}_f. \tag{2}$$

Note that the final scene points $\mathcal{P}_s$ come mainly from MoGe, while the initial camera used in Eq. (1) comes from DepthPro. To ensure better alignment, we recalculate the camera intrinsic for $\mathcal{P}'_s$. Let (u, v) be a 2D pixel and (X, Y, Z) be its corresponding 3D point from $\hat{\mathcal{P}}'_s$, we assume centered principal point [Patel and Black 2024] and derive potential focal lengths: $f_x(u, v) = (u - W/2)\frac{Z}{X}$ and $f_y(u, v) = (v - H/2)\frac{Z}{Y}$. The final focal lengths, f_x and f_y, are robustly set to the median of these respective values. This new intrinsic matrix K is used for all subsequent camera projections.

3.2 Stage 2: Human Reconstruction and Alignment

The previous section masks out the human and only considers the scene. We now reconstruct the human and align it with the predicted scene point cloud $\mathcal{P}'_s$. This consists of two steps: 1) obtain human points $\mathcal{P}_h$ aligned with scene points, and 2) estimate human mesh aligned with the human points $\mathcal{P}_h$, i.e. the underlying scene $\mathcal{P}'_s$.

3.2.1 Metric-scale human points. . From the original input image $\mathcal{I}$, we use MoGe to predict an unscaled point cloud $\hat{\mathcal{P}}_{h+s}$, which contains human points $\hat{\mathcal{P}}_h$ and surrounding scene points $\hat{\mathcal{P}}_s$. We then align this to metric-scale scene points $\mathcal{P}'_s$ by optimizing a scale and depth shift, similar to Eq. (1). Note that we use the human mask to remove $\hat{\mathcal{P}}_h$ from $\hat{\mathcal{P}}_{h+s}$ when performing the alignment. We then apply the optimized scale and shift to human points $\hat{\mathcal{P}}_h$, thus aligning them with the metric-scale scene, denoted as $\mathcal{P}_h$.

3.2.2 Human mesh estimation. . To obtain semantically meaningful contact vertices, we use SMPL-X [Pavlakos et al. 2019] to represent the human. We denote $\mathbf{H}$ as the SMPL-X model which takes body shape $\boldsymbol{\beta}$, hand and full body poses $\boldsymbol{\theta}_h, \boldsymbol{\theta}_b$, and global translation $\boldsymbol{t}_h$ as input, and outputs the human vertices $\mathcal{V}_h = \mathbf{H}(\boldsymbol{\beta}, \boldsymbol{\theta}_h, \boldsymbol{\theta}_b, \boldsymbol{t}_h)$. The initial SMPL-X vertices $\mathcal{V}_h$ are obtained by fusing SMPL [Loper et al. 2015] prediction from CameraHMR [Patel and Black 2024] and hand pose from WiLor [Potamias et al. 2025]. Specifically, we fit SMPL-X into the SMPL mesh predicted by CameraHMR using SMPLFitter [Sárándi and Pons-Moll 2024] and replace the hand parameters with the hand pose predicted by WiLor. This initial estimation does not precisely align with the input image and metric-scale scene, which we address next.

3.2.3 Metric-scale human mesh. . We first optimize the global human translation $\boldsymbol{t}_h$ to improve the pixel-alignment of the estimated SMPL-X vertices using 2D joint projection loss:

$$L_{\text{j2d}} = \left\| \left(\pi(\mathcal{J}(\mathcal{V}_h(\mathbf{t}_h), K) - \hat{J}_h^{2D} \right) \right\|_2^2, \tag{3}$$

where $\mathcal{J} : \mathbb{R}^{N\times 3} \mapsto \mathbb{R}^{J\times 3}$ regresses the 3D body keypoints and $\hat{J}_h^{2D}$ are the 2D keypoints predicted by ViTPose [Xu et al. 2022]. We then align the optimized human vertices with the metric-scale human points $\mathcal{P}_h$ using the Chamfer distance between camera-facing vertices $\mathcal{V}_{\text{cf}}$ and human points $\mathcal{P}_h$:

$$L_{\text{align}} = \lambda_{\text{j2d}} L_{\text{j2d}} + \lambda_{\text{d}} L_{\text{d}}, \text{ where} \tag{4}$$

$$L_{\text{d}} = \sum_{\mathbf{v}\in\mathcal{V}_{\text{cf}}} \min_{\mathbf{p}\in\mathcal{P}_h} \|\mathbf{v} - \mathbf{p}\|_2^2 + \sum_{\mathbf{p}\in\mathcal{P}_h} \min_{\mathbf{v}\in\mathcal{V}_{\text{cf}}} \|\mathbf{p} - \mathbf{v}\|_2^2. \tag{5}$$

We select camera-facing vertices $\mathcal{V}_{cf} \subset \mathcal{V}_h$ as the vertices whose surface normals are at an angle deviating less than 70 degrees from

the camera view direction. This is crucial to avoid aligning the backside vertices with the human points. Note here that we only optimize the global translation $\boldsymbol{t}_h$ parameter.

3.3 Stage 3: Joint Human-Scene Optimization

Even though the human vertices $\mathcal{V}_h$ and metric-scale scene points $\mathcal{P}_s'$, which were obtained from previous steps, reside in the same metric-scale coordinate, they are predicted separately. Hence, physical plausibility is not guaranteed. We further enhance the plausibility by enforcing additional constraints between the human and the scene (Fig. 2, Stage 3). To this end, we formulate a joint optimization objective that adapts principles of contact attraction and interpenetration avoidance [Hassan et al. 2019a; Yi et al. 2022] to our setting of single-image reconstruction with pointmaps. Thus, we additionally introduce the contact and interpenetration loss, together with regularization terms to jointly optimize the human parameters $\boldsymbol{\theta}_b, \boldsymbol{\theta}_h, \boldsymbol{\beta}, \boldsymbol{t}_h$ and a scene scale parameter s_{sc}:

$$L_{\text{total}} = \lambda_{\text{j2d}} L_{\text{j2d}} + \lambda_{\text{d}} L_{\text{d}} + \lambda_c L_c + \lambda_i L_i + \lambda_{\text{reg}} L_{\text{reg}}. \tag{6}$$

Let $\mathcal{P}_s = s_{\text{sc}} \mathcal{P}_s'$ be the scaled scene points; we explain the contact, interpenetration and regularization terms next. The loss weights λ_* are detailed in the supplementary.

3.3.1 Contact loss L_c. This encourages the human vertices in contact with the scene to be close to the scene points $\mathcal{P}_s$. We use DECO [Tripathi et al. 2023], which predicts the human contact vertices $\mathcal{V}_{\text{con}}$ and minimize their distance to the closest scene points. During optimization, we use an active-contact subset by re-evaluating nearest scene distances each iteration and only applying L_c to vertices within ϵ, suppressing spurious long-range contacts:

$$L_{\text{c}} = \sum_{v \in \mathcal{V}_{\text{con}}} \left(\min_{\mathbf{p} \in \mathcal{P}_s} \rho(\|\mathbf{v} - \mathbf{p}\|_2^2) \right) \mathbb{I}\left(\min_{\mathbf{p} \in \mathcal{P}_s} \|\mathbf{v} - \mathbf{p}\|_2^2 < \epsilon \right), \tag{7}$$

where ρ is an adaptive robust loss function [Barron 2019] and the indicator function $\mathbb{I}(\cdot)$ ensures the loss term is only active when the distance to the nearest scene point is less than a threshold ϵ. This hinders penalizing distant false-positive contact predictions or interactions with outlier scene points.

3.3.2 Occlusion aware interpenetration loss L_i. It prevents the human mesh $\mathcal{V}_h$ from unnaturally penetrating the scene geometry $\mathcal{P}_s$. We leverage the estimated per-point normal of $\mathcal{P}_s$ and penalize points lying opposite to the normal direction:

$$L_{\text{i}} = \sum_{\mathbf{v} \in \mathcal{V}_h \setminus \mathcal{V}_{\text{occ}}} \rho \left(\min_{\mathbf{p} \in \mathcal{P}_s} \|\mathbf{v} - \mathbf{p}\|_2^2 \right) \mathbb{I}(\mathbf{n}_{\mathbf{p}} \cdot (\mathbf{v} - \mathbf{p}) < 0). \tag{8}$$

Importantly, we exclude human vertices $\mathcal{V}_{\text{occ}}$ occluded by surrounding objects or by itself. Specifically, we consider human vertices whose 2D projections lie outside the human mask as occluded by object. We divide the vertices into different body parts and consider a part as self-occluded if 30% of its vertices are occluded by other body parts. This prevents the occluded body parts from moving towards unnatural poses due to the penetration loss, as no other signal, like 2D keypoints, is available to regularize the optimization.

3.3.3 Regularization terms L_{reg}. To ensure the optimized human mesh $\mathcal{V}_h$ does not deviate excessively from initial estimates, we apply a mesh regularization loss, treating the initial estimates as a pose prior. This loss penalizes the L2 distance between the current and initial mesh vertices in the root-relative space, constraining the local body pose of the human, while allowing for large updates in global translation of the mesh. We increase the weight of the regularization loss for occluded vertices $\mathcal{V}_{occ}$ since the initial estimates are our best guess for unobserved parts of the human mesh. We further weakly regularize the scene scale s_{sc} and the human translation $\boldsymbol{t}_h$ by preventing large deviations from their initial values.

3.3.4 Contact map extraction. Our joint optimization produces accurate and physically plausible human-scene interactions, which allows us to extract per-vertex contact maps based on proximity. Each human mesh vertex $\mathbf{v}_j \in \mathcal{V}_h$ is labeled as in-contact if its Euclidean distance to the nearest point on the scene surface is less than a predefined threshold ϵ_c. This process yields a binary contact mask over the human mesh vertices, identifying regions of interaction.

3.3.5 Handling multiple humans. The method described above can easily be extended to multiple humans by using another human mask to perform human-scene alignment and joint optimization. Specifically, we use SAM2 to obtain per-instance human masks. We inpaint all humans simultaneously to obtain the scene, then align each human mesh individually with the scene following Sec. 3.2. We then perform one joint optimization between the underlying scene and all humans using Eq. (6).

4 Experiments

4.1 Implementation Details

We implement our optimization framework using PyTorch [Paszke et al. 2019] and batched 3D geometry operations to handle multiple humans with PyTorch3D [Ravi et al. 2020]. During initialization, we perform aggressive outlier point removal using mean k-NN distance to ensure clean scene geometry, where k is adaptively set based on image resolution. For the first optimization (Eq. 3), we perform 30 iterations of gradient descent with Adam [Kingma and Ba 2017]. For the second optimization (Eq. 5), we use two iterations of L-BFGS [Liu and Nocedal 1989]. Our final optimization (Eq. 6) utilizes 100 iterations of gradient descent with Adam. Both gradient descents use a learning rate of $1e$-2, and the L-BFGS optimizer uses a unit learning rate. While the camera-facing mask $\mathcal{V}_{cf}$ remains stable throughout optimization, the self-occlusion state can vary due to pose optimization. Hence, we update $\mathcal{V}_{occ}$ every 30 iterations of the final gradient descent. For more details, please refer supplementary.

A frequent operation used in L_c and L_i is nearest-neighbor search. Despite the varying scene scale during optimization, we leverage the scale-invariant nature of the nearest neighborhood structure to precompute a 128^3 grid of nearest-scene-points, and transform query points to the initial scale. This results in a 15–20× overall speedup, compared to a brute-force implementation. On a NVIDIA H100 GPU, our optimization takes 9 seconds for a 480p image and 12 seconds for a 720p image, yielding end-to-end human-scene reconstruction times of 27 seconds and 36 seconds respectively.

Fig. 3. **Qualitative results on PROX dataset (row 1) and internet images (row 2-3).** We compare the output of PhySIC with PROX [Hassan et al. 2019a] and HolisticMesh [Weng and Yeung 2021]. Note that we run PROX with our estimated scene on internet images as there is no scene scan available. Both PROX and HolisticMesh rely on predefined contact maps, hence are not robust to complex human poses and interactions. Our method reconstructs 3D scene and adapts contact optimization based on input, leading to more coherent reconstruction. Please refer Fig. 5 and Fig. 7 for more results.

Table 2. **Quantitative Comparison on PROX and RICH.** Our method outputs better local pose (PA-MPJPE) and more accurate contacts. Although HolisticMesh shows better number in MPJPE on PROX, we found it is unreliable and cannot robustly reconstruct on in-the-wild data (See Fig. 3).

Method	Human Pose Metrics ↓			Contact Metrics ↑		
	PA-MPJPE	MPJPE	MPVPE	Precision	Recall	F1 score
PROX Quantitative Dataset						
CameraHMR	42.35	997.49	996.20	–	–	–
DECO	–	–	–	0.406	0.349	0.376
PROX	73.31	266.50	266.00	0.260	0.108	0.152
HolisticMesh	77.04	**202.80**	**191.70**	0.373	0.412	0.391
Ours	**41.99**	230.26	227.19	**0.508**	**0.514**	**0.511**
RICH-100 Dataset						
PROX	120.24	706.19	692.07	0.040	0.250	0.069
Ours	**46.50**	**616.27**	**617.33**	**0.310**	**0.689**	**0.428**

4.2 Evaluation Protocol

We evaluate PhySIC against prior arts on the PROX [Hassan et al. 2019b] and RICH [Huang et al. 2022] datasets, both containing humans and static scene scans. The PROX dataset captures a single subject interacting with various objects of a scene in an indoor setting. In contrast, RICH includes videos of two scenes covering both indoor and outdoor settings, each scene captured by 6-8 cameras, resulting in $\sim 125k$ frames, with high redundancy, which makes full evaluation expensive. Hence, we use all 178 images from PROX-quantitative and randomly sample 100 images from RICH, covering all possible cameras, activities and backgrounds. We further provide qualitative results on the PiGraphs dataset [Savva et al. 2016], containing videos of humans interacting with static scenes in indoor environments. Finally, we collect a set of in-the-wild images from the internet to show the generalizability of our approach.

We compare PhySIC with PROX [Hassan et al. 2019b] and HolisticMesh [Weng and Yeung 2021], two state-of-the-art approaches that jointly model human-scene interaction from monocular images. While HolisticMesh estimates human-scenes from a single RGB image, whereas PROX requires a static 3D scene scan for optimization. To enable a fair comparison, and to evaluate PROX on RGB images, we perform two modifications. First, we replace the static scene with an unprojected depth map from DepthPro [Bochkovskii et al. 2024]. Further, we replace PROX's pose prior VPoser [Pavlakos et al. 2019] with SOTA CameraHMR. Specifically, we initialize and regularize the pose optimization using CameraHMR [Patel and Black 2024].

Unlike HolisticMesh, which fits a single static scene to the entire PROX-Quantitative sequence, our method relies only on a single inpainted image from a frame without interactions with thin structures. This avoids the need for per-frame inpainting and allows us to optimize the human and the scene independently at each frame, without relying on sequence-level cues. Despite this lightweight

design, our inpainting generalizes robustly and works well on in-the-wild images directly, even without any sequence information.

4.3 Qualitative Analysis

4.3.1 Human-scene reconstruction. Fig. 3 presents qualitative results of human-scene reconstruction. In contrast to our method, PROX lacks robust occlusion handling and an appropriate distance threshold in its contact loss, resulting in inaccurate poses and mislocalization within the scene. While some interpenetration is expected due to unmodeled scene deformations, PROX exhibits excessive interpenetration beyond these expected discrepancies. Similarly, HolisticMesh also suffers from noticeable interpenetration, and fails to run on certain in-the-wild examples, highlighting limitations with generalizability. In contrast, PhySIC's robust occlusion handling and refined distance thresholding for contacts leads to more accurate poses, better localization, and significantly reduced interpenetration, thereby successfully increasing robustness in complex scenes. For additional results, please refer to Fig. 7 and supplementary.

4.3.2 Contact estimation. We show example comparison with contact estimation method DECO [Tripathi et al. 2023] in Fig. 4. Our joint optimization is guided by the contact estimation from DECO which can be noisy. However, our approach robustly recovers accurate human-scene interactions and further improves contacts, especially in the intricate body parts such as feet and arms. Additional examples can be found in Fig. 6.

4.4 Quantitative Analysis

We quantitatively evaluate our method on both 3D human pose and vertex-level contact metrics. For 3D human pose, we report the Mean Per-Joint Position Error (MPJPE), the average euclidean distance between the camera-relative predicted and GT human joints. We also use Procrustes Aligned MPJPE (PA-MPJPE) which computes MPJPE after global alignment, effectively comparing the root-relative human pose. Additionaly, we report the Mean Per-Vertex Positon Error (MPVPE), the average Euclidean distance between the predicted and GT mesh vertices, which takes into account the predicted human shape $\boldsymbol{\beta}$. For the human-scene contact, we report standard classification metrics (precision, recall, F1 score), calculated using the predicted and GT per-vertex contacts.

The results in Tab. 2 demonstrate that our approach achieves state-of-the-art performance in both human pose and contact estimation. Specifically on the PROX dataset, we significantly reduce PA-MPJPE, by nearly half compared to the PROX and HolisticMesh, even though both methods initialize from CameraHMR. Our method consistently improves upon the state-of-the-art CameraHMR and DECO across all pose and contact metrics. Our method also outperforms HolisticMesh in contact accuracy with a 40% improvement in F1 score. Although HolisticMesh shows a marginally better MPJPE and MPVPE on PROX, it performs poorly on non-PROX in-the-wild images (Fig. 7), and suffers from severe interpenetration and inaccurate local poses, as indicated by its PA-MPJPE.

On the RICH dataset, HolisticMesh could not be evaluated, as it is trained only for indoor living environments occupied with limited object categories, while RICH was captured both indoors and outdoors, beyond HolisticMesh categories. Our method outperforms PROX on both pose and contact metrics.

Table 3. Ablating the impact of different loss terms on joint human-scene optimization. Depth loss L_d is important to ensure good global alignment (MPJPE) and occlusion-aware interpenetration loss L_i improves local pose accuracy (PA-MPJPE).

Ablation	Human Pose Metrics ↓			Contact Metrics ↑		
	PA-MPJPE	MPJPE	MPVPE	Precision	Recall	F1 score
Init. (CHMR)	42.35	997.49	996.20	–	–	–
Init. (DECO)	–	–	–	0.406	0.349	0.376
$L_{reg} + L_{j2d}$	76.79	643.02	641.31	0.288	0.052	0.088
$+L_c$	71.22	637.50	635.98	0.397	0.250	0.307
$+L_i$	69.22	639.16	637.62	0.394	0.228	0.289
$+L_d$	72.64	364.98	358.75	0.490	0.430	0.459
+ occ. aware L_i	**41.91**	238.72	235.75	0.490	**0.550**	**0.518**
+ floor (full model)	41.99	**230.26**	**227.19**	**0.508**	0.514	0.511

4.5 Ablation Study

We investigate the impact of different loss terms on our joint optimization stage and report the results on PROX dataset. Starting from the basic 2D joint reprojection loss L_{j2d} and regularization term, we gradually add more losses defined in Eq. (6) to the optimization process. We report the performance of our initialization approaches, CameraHMR and DECO, and ablation results in Tab. 3. With just $L_{reg} + L_{j2d}$, the human pose metrics degrade compared to the initial estimates. This is due to depth ambiguity in the monocular setting: a perfect 2D fit does not imply accurate 3D pose. This necessitates additional constraints from human-scene interaction losses (L_c and L_i). However, since these losses are applied against nearest scene points, they can misalign with the actual contact regions. To address this, we also include a loss against the human points $\mathcal{P}_h$. The depth alignment loss ($+L_d$) marks a crucial improvement in both pose and contacts, due to improved localization of the human within the scene, which ensures that L_c and L_i act on the correct scene regions and thus become effective. Our occlusion-aware interpenetration loss (L_i with $\mathcal{V}_{\text{occ}}$ excluded) further then delivers the largest PA-MPJPE gain (to 41.91) and achieves the best contact recall and F1 scores. Note that without the occlusion awareness, the local body poses (PA-MPJPE) are even worse than the initialization. This is because the occluded part can be overly penalized due to interpenetration without regularization from the input image, leading to large deviations from correct body poses. Our experiments highlight the significance of occlusion reasoning, which is consistent with the prior works [Xie et al. 2023, 2025]. For detailed qualitative results, please refer to supplementary.

5 Limitations and Future Works

While PhySIC advances the state of the art in physically plausible human-scene reconstruction from a single image, several limitations remain, highlighting directions for future research. (i) Image inpainting. Our approach relies on state-of-the-art inpainting models [Wei et al. 2025] to reconstruct occluded scene regions. These models are imperfect, particularly for thin or intricate structures, leading to erased or deformed geometry. As inpainting methods

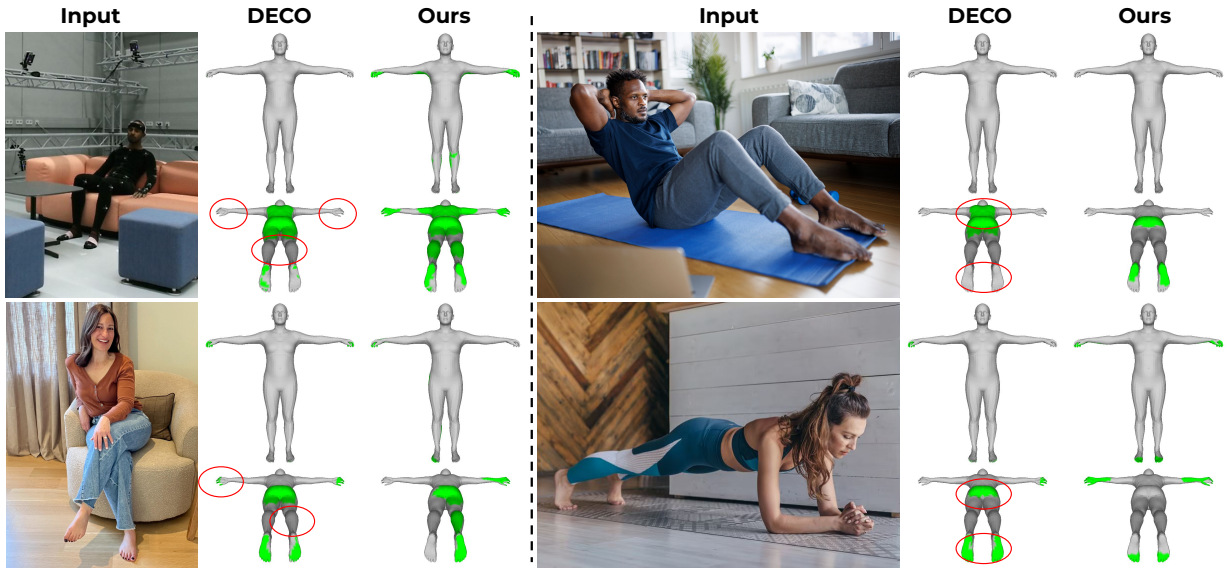

Fig. 4. **Qualitative results for contact estimation.** We compare our approach against the state-of-the-art image-based contact predictor, DECO [Tripathi et al. 2023] in both lab and wild setting. Note how our method improves the nuanced contact on arms and feet. Please refer Fig. 6 for further examples.

improve, we expect direct benefits to our approach. (ii) Scene deformations. We assume static, rigid scene geometry, which may not hold for deformable objects such as cushions or clothing. Extending PhySIC to handle non-rigid scene deformations could enable more realistic human-scene interactions. (iii) Human-object interactions. We focus on human-scene interactions and do not explicitly model fine-grained interactions with small objects, such as grasping or pushing. Future work could integrate off-the-shelf object mesh estimators, align them with reconstructed depth maps, and leverage additional 2D object supervision. (iv) Flat floor assumption. PhySIC assumes that floors are planar to simplify occlusion reasoning and contact estimation. This assumption generally holds, but can fail when no floor points are detected or RANSAC fails to find consensus. In such cases, we skip floor sampling, which may result in false-negative contacts. Maturity of holistic 3D scene reconstruction methods could address this limitation [Roh et al. 2024].

6 Conclusion

We present PhySIC, a framework for physically plausible human-scene interaction and contact reconstruction from a monocular RGB image. By jointly optimizing metrically scaled SMPL-X human meshes and detailed 3D scene geometry, PhySIC enables reconstruction of coherent, physically realistic human-scene pairs in diverse environments. Our method introduces robust initialization strategies, combining metric depth and detailed relative geometry, occlusion-aware refinement, and efficient multi-term optimization that enforces contact, interpenetration, and depth alignment. Extensive experiments on challenging benchmarks demonstrate that PhySIC significantly outperforms prior work in both pose and contact metrics, and generalizes well to multi-human and in-the-wild scenarios. PhySIC provides a scalable, accessible step towards holistic, single-image 3D human-centric scene understanding. We anticipate that continued progress in inpainting, foundation geometry models, and interaction reasoning will further enhance the capabilities and generality of our approach. We will release our code and evaluation scripts to support future research and practical applications.

Acknowledgments

We thank the anonymous reviewers whose feedback helped improve this paper. This work is made possible by funding from the Carl Zeiss Foundation. This work is also funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 409792180 (EmmyNoether Programme, project: Real Virtual Humans) and the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting YX. PYM is supported by the Konrad Zuse School of Excellence in Learning and Intelligent Systems (ELIZA) through the DAAD programme Konrad Zuse School of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. GPM is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645.

PYM and YX contributed equally as the joint first author. YX is the corresponding author. Authors with equal contributions are listed in alphabetical order and allowed to change their orders freely on their resume and website. YX initialized the core idea, organized the project, co-developed the current method, co-supervised the experiments, and wrote the draft. PYM co-initialized the core idea, co-developed the current method, implemented most of the prototypes, conducted experiments, and co-wrote the draft. XX contributed to the draft writing and improving Fig. 2. MK lead the visualization and rendering of presented results in Figs. 1, 5.

Fig. 5. **Additional qualitative results for in-the-wild images.** Please refer to Supp. Mat. for more results.

Fig. 6. **Qualitative results for contact estimation.** We compare our approach against the state-of-the-art image-based contact predictor, DECO.

Fig. 7. **Qualitative results on internet images.** We compare the output of PhySIC with PROX and HolisticMesh. HolisticMesh fails to model scenes with arbitrary surfaces since it estimates per-object geometry.

References

Andreea Ardelean, Mert Özer, and Bernhard Egger. 2025. Generalizable 3D Scene Reconstruction via Divide and Conquer from a Single View. In *International Conference on 3D Vision (3DV)*.

Jonathan T. Barron. 2019. A General and Adaptive Robust Loss Function. *CVPR* (2019).

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. doi:10.48550/ARXIV.2302.12288

Sandika Biswas, Kejie Li, Biplab Banerjee, Subhasis Chaudhuri, and Hamid Rezatofighi. 2023. Physically Plausible 3D Human-Scene Reconstruction from Monocular RGB Image using an Adversarial Learning Approach. arXiv:2307.14570 [cs.CV] https://arxiv.org/abs/2307.14570

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. 2024. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. doi:10.48550/arXiv.2410.02073 arXiv:2410.02073 [cs]

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision (ECCV)*. 561–578.

Anh-Quan Cao and Raoul de Charette. 2022. MonoScene: Monocular 3D Semantic Scene Completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 3981–3991. doi:10.1109/CVPR52688.2022.00396

Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. 2019. Holistic++ Scene Understanding: Single-view 3D Holistic Scene Parsing and Human Pose Estimation with Human-Object Interaction and Physical Commonsense. In *The IEEE International Conference on Computer Vision (ICCV)*.

Georgia Gkioxari, Justin Johnson, and Jitendra Malik. 2019. Mesh R-CNN. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 9784–9794. doi:10.1109/ICCV.2019.00988

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. 2021. Human POSEitioning System (HPS): 3D Human Pose Estimation and Self-localization in Large Scenes from Body-Mounted Sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. 2019a. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. In *International Conference on Computer Vision*. https://prox.is.tue.mpg.de

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. 2019b. Resolving 3D Human Pose Ambiguities with 3D Scene Constraints. doi:10.48550/arXiv.1908.06963 arXiv:1908.06963 [cs]

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. 2021. Populating 3D Scenes by Learning Human-Scene Interaction. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. 2024. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. 2022. Capturing and Inferring Dense Full-Body Human-Scene Contact. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 13274–13285.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. 2018. End-to-end Recovery of Human Shape and Pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7122–7131.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. 2024. ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions. arXiv:2401.10232 [cs.CV]

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] https://arxiv.org/abs/1412.6980

Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. 2021. PARE: Part Attention Regressor for 3D Human Body Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11127–11137.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2252–2261.

Jianan Li, Tao Huang, Qingxu Zhu, and Tien-Tsin Wong. 2024. Physics-based Scene Layout Generation from Human Motion. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 45, 10 pages. doi:10.1145/3641519.3657517

Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. 2019. Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 12368–12376. doi:10.1109/CVPR.2019.01265

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45 (1989), 503–528. https://api.semanticscholar.org/CorpusID:5681609

Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2020. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2020), 2684–2701.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.* 34, 6, Article 248 (Oct. 2015), 16 pages. doi:10.1145/2816795.2818013

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. 2024. Reconstructing People, Places, and Cameras. *arXiv:2412.17806* (2024).

Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. 2021. Pose2Room: Understanding 3D Scenes from Human Activities. *arXiv preprint arXiv:2112.03030* (2021).

Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. 2020. Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes From a Single Image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG] https://arxiv.org/abs/1912.01703

Priyanka Patel and Michael J. Black. 2024. CameraHMR: Aligning People with Perspective. doi:10.48550/arXiv.2411.08128 arXiv:2411.08128 [cs]

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 10967–10977. doi:10.1109/CVPR.2019.01123

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-Wild. doi:10.48550/arXiv.2409.12259 arXiv:2409.12259 [cs]

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714 [cs.CV] https://arxiv.org/abs/2408.00714

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501* (2020).

Wonseok Roh, Hwanhee Jung, Jong Wook Kim, Seunggwan Lee, Innfarn Yoo, Andreas Lugmayr, Seunggeun Chi, Karthik Ramani, and Sangpil Kim. 2024. CAT-Splat: Context-Aware Transformer with Spatial Guidance for Generalizable 3D Gaussian Splatting from A Single-View Image. doi:10.48550/arXiv.2412.12906 arXiv:2412.12906 [cs]

István Sárándi and Gerard Pons-Moll. 2024. Neural Localizer Fields for Continuous 3D Human Pose and Shape Estimation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/fd23a1f3bc89e042d70960b466dc20e8-Abstract-Conference.html

Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. 2016. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)* 35, 4 (2016).

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. 2024. WHAM: Reconstructing World-grounded Humans with Accurate 3D Motion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. 2023. DECO: Dense Estimation of 3D Human-Scene Contact In The Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8001–8013.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. 2024b. MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision. doi:10.48550/arXiv.2410.19115 arXiv:2410.19115 [cs]

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. 2024a. TRAM: Global Trajectory and Motion of 3D Humans from in-the-wild Videos. *arXiv preprint arXiv:2403.17346* (2024).

Runpu Wei, Zijin Yin, Shuo Zhang, Lanxiang Zhou, Xueyi Wang, Chao Ban, Tianwei Cao, Hao Sun, Zhongjiang He, Kongming Liang, and Zhanyu Ma. 2025. OmniEraser: Remove Objects and Their Effects in Images with Paired Video-Frame Data. doi:10.

48550/arXiv.2501.07397 arXiv:2501.07397 [cs]

Zhenzhen Weng and Serena Yeung. 2021. Holistic 3D Human and Scene Mesh Estimation from Single View Images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 334–343. doi:10.1109/CVPR46437.2021.00040

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. 2022. CHORE: Contact, Human and Object Reconstruction from a Single RGB Image. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 13662)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 125–145. doi:10.1007/978-3-031-20086-1_8

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. 2023. Visibility Aware Human-Object Interaction Tracking from Single RGB Camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. 2025. InterTrack: Tracking Human Object Interaction without Object Templates. In *International Conference on 3D Vision (3DV)*.

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. arXiv:2204.12484 [cs.CV] https://arxiv.org/abs/2204.12484

Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Otmar Hilliges. 2024. HSR: Holistic 3D Human-Scene Reconstruction from Monocular Videos. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXII (Lecture Notes in Computer Science, Vol. 15130)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 429–448. doi:10.1007/978-3-031-73220-1_25

Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. 2022. Human-Aware Object Placement for Visual Environment Reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*. 3959–3970.

Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. 2022. Egobody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision*. Springer, 180–200.

Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. 2020. PLACE: Proximity learning of articulation and contact in 3D environments. In *2020 International Conference on 3D Vision (3DV)*. IEEE, 642–651.