

Reconstructing Humans and Objects in Interaction using Large Reconstruction Models

Agniv Chatterjee¹ and Georgios Pavlakos²

University of Texas at Austin

agniv.chatterjee@utexas.edu, pavlakos@cs.utexas.edu

Abstract. Estimation of Human-Object Interactions in 3D (3D HOI) is a fundamental problem in 3D computer vision with applications in AR/VR, robotics, and embodied AI. However, reconstructing these interactions in 3D remains challenging due to depth ambiguities, occlusions, and object shape variability. Existing approaches are primarily concerned with reprojection and contact constraints, fitting parametric human models and object templates to 2D images. In this paper, we explore a different avenue. We present MILO, a framework that leverages the visual capabilities of Large Reconstruction Models (LRMs) to recover detailed 3D human-object interactions from a single image. Our key observation is that LRMs provide a powerful geometric scaffold that preserves relative human-object arrangement and proximity cues. This significantly simplifies the reconstruction procedure, reframing the problem as interpreting the LRM mesh: we segment it into human and object components, fit a parametric body model to the human part, and optionally align an object template to the object part (if such a template is available). MILO achieves strong reconstruction accuracy and outperforms existing baselines across multiple benchmarks and interaction scenarios. Our code is available at <https://ac5113.github.io/MILO>.

Keywords: 3D human-object interactions · Large Reconstruction Models

1 Introduction

Daily human life is characterized by rich 3D interactions with a wide variety of objects, from utensils and smartphones to furniture and tools. Modeling these human-object interactions (HOI) is central to many vision applications, including immersive AR/VR, teleoperation, and robot manipulation. However, HOI reconstruction from a single image is difficult due to diverse object shapes and affordances, multiple interaction modes, occlusions, and depth ambiguities, which make the recovery of human pose and object geometry challenging.

Human parametric models such as SMPL [36], MANO [49], and SMPL-X [43] provide standardized representations which have led to successes for 3D human reconstruction. In contrast, the absence of analogous canonical models for generic objects makes object reconstruction significantly more challenging. Moreover, the generalization capabilities of methods [41, 64] trained on existing HOI



Fig. 1: We present MILO, an approach for reconstructing 3D human-object interactions from a single image. Given an input photo (left), MILO uses a Large Reconstruction Model to predict a combined human-object mesh (center). This mesh is interpreted (right) by identifying human and object parts, fitting a SMPL-H body to the human part, and optionally an object template to the object part. The examples illustrate accurate 3D human mesh together with consistent object pose and shape. Human meshes are shown in **blue** and objects in **red**.

datasets [2, 19] are hindered by the characteristics of these datasets, which are captured in controlled environments with a limited set of objects and scripted interactions. Additionally, approaches [5, 66] that require the retrieval of object meshes further depend on the quality and coverage of available shape repositories. More importantly, accurate placement of retrieved meshes often requires privileged cues such as ground-truth contact or depth [5, 41, 73].

Our insight is that we can circumvent many of these limitations by leveraging the rich visual capabilities of Large Reconstruction Models (LRMs) [17, 21, 57, 69]. LRMs are trained on large-scale data [6, 7] and are capable of reconstructing complex scenes, outputting an image-specific joint reconstruction of the human and the object. This effectively provides a geometric scaffold that encodes relative arrangement and proximity cues. Building on this success, our method proposes to replace the ambiguous image fitting and the contact assumptions with the interpretation of an LRM-generated human-object reconstruction.

Our goal is to systematically assess the extent to which LRMs can be used for HOI reconstruction. While previous work has used LRMs primarily to reconstruct the object [35], our insight is that LRMs can be very powerful for capturing the relative positioning and proximity cues between a human and an object. To this end, we propose MILO; **M**odeling **I**nteractions using **L**arge **R**econstruction **M**odels. MILO first applies an off-the-shelf LRM [57] to the input RGB image to obtain a holistic human-object mesh. This reframes our problem as explaining the LRM interaction mesh. Through multi-view segmentation of the mesh, we can extract the human and the object point cloud. For the human part, we fit the parametric SMPL-H body model to the LRM mesh via predicted 3D keypoints [44, 70]. For the object, we compose the raw LRM reconstruction with the fitted human, or optionally align a template mesh if one is available. In

that case, alignment is supported by establishing semantic correspondences [74] between the template and the LRM mesh. We present example results in Fig. 1.

We quantitatively compare MILO with state-of-the-art methods on multiple datasets, i.e., InterCap [19], HODome [75] and IMHD [77], while also showing qualitative results on in-the-wild images [5]. Importantly, MILO outperforms previous work across the board, without using ground-truth contact information.

Our main contributions are as follows:

1. We demonstrate the power of Large Reconstruction Models for HOI reconstruction. Instead of operating with the high ambiguity of reprojection objectives, we reframe the problem as explaining the LRM mesh results.
2. We design an approach that robustly fits a parametric human model to the LRM mesh and optionally aligns an object template to it.
3. We achieve state-of-the-art performance across multiple benchmarks, all while using weaker information than previous work, i.e., without relying on contact information.

2 Related Works

2.1 Monocular 3D Human Reconstruction

Single-image 3D human reconstruction methods can be broadly categorized into parametric and non-parametric approaches. Parametric methods [8, 25, 28, 29, 50] estimate the parameters of a body model [36, 43, 49, 68], whereas non-parametric methods [26, 31, 32] predict explicit meshes or point clouds. Optimization-based pipelines [3, 27, 43, 68] recover body pose and shape by fitting a parametric body model to 2D cues, e.g., keypoints, silhouettes, or segmentation, via differentiable rendering. In contrast, learning-based approaches directly regress body-model parameters from inputs [9, 22, 29, 30, 39] using deep networks. While these methods have led to robust monocular human reconstruction, they typically focus on the person in isolation and do not explicitly reason about surrounding objects or human-object contact. In our method, we use the HMR2.0 network [14] to obtain initial SMPL parameters from the input RGB image.

2.2 Monocular 3D Object Reconstruction

Monocular 3D object reconstruction has a long history with a wide range of 3D representations explored. Early work employed explicit volumetric [4, 13], point-based [10], or mesh-based [10, 61] representations, while later methods [38, 42] adopted implicit signed-distance or radiance-field representations. More recently, 3D Gaussian splatting [24] has emerged as an alternative to NeRF-style radiance fields, enabling efficient object reconstruction. Diffusion-based image-to-3D approaches [37, 46, 55] and view-conditioned diffusion models [33] further improve visual fidelity and handle complex appearance, but often rely on per-instance optimization or suffer from multi-view inconsistencies across generated views.

Recent work on Large Reconstruction Models [17, 21, 54, 58, 62] has enabled efficient and view-consistent reconstruction of diverse object categories. By leveraging the LRM output, MILO obtains an image-specific joint human-object mesh, which captures accurate geometry and proximity information. For our implementation, we adopt Hunyuan3D-2.0 [57], although we show that our algorithm is not specific to the choice of LRM.

2.3 Monocular 3D Human-Object Reconstruction

Joint reconstruction of humans and objects from a single image has been explored with both learning- and optimization-based approaches. HOI-TG [64] employs a transformer to jointly reason about pose and interaction. LEMON [72] learns 3D human-object interaction relations from 2D images by jointly predicting human contact, object affordance, and spatial relations. CONTHO [41] predicts vertex-level contact maps and refines the human-object mesh poses accordingly. StackFLOW [20] regresses dense human-object offsets and refines both the human and object reconstructions using stacked normalizing flows. CHORE [67] jointly learns a part-correspondence field that links human and object geometry. TriDi [45] supports multiple conditioning modes by introducing a trilateral diffusion model over humans, objects, and interactions.

A common strategy involves reconstructing the object by retrieving the closest object mesh from existing shape repositories [7]. PICO [5] retrieves object candidates, while Wu et al. [66] uses text-to-3D retrieval to obtain an object template. Related to our work, LRMs have also been used to reconstruct objects of interaction directly from images, without relying on explicit CAD templates. For instance, EasyHOI [35] employs InstantMesh [69] to obtain object meshes in hand-object interaction scenarios. Our approach also leverages an LRM, but we propose to use these powerful models for joint human-object reconstruction, simplifying the relative positioning of humans and objects.

Most of the previous methods assume access to object templates and/or GT contact annotations, which constrains scalability and generalization to novel objects and in-the-wild scenes. For our approach, the use of an object template is optional, while we do not rely on explicit contact information.

2.4 Human-Object Contact Estimation

Human-object contact provides a powerful geometric cue that tightly constrains the arrangement of humans and objects. Early work modeled contact at a coarse level [1, 16, 23], over patches [11, 40], or only on the human body surface [48, 52, 76]. More recent approaches [15, 18, 53, 59, 73] estimate dense, vertex-level contact or correspondence maps and exploit them to regularize HOI reconstruction or scene understanding tasks. Several HOI methods integrate contact signals directly into their pipelines. CONTHO [41] uses vertex-level human-object contact maps for mesh refinement. CHORE [67] learns part-to-body correspondence fields that implicitly encode contact patterns. PICO [5] uses paired human-object contact maps to guide joint human-object fitting. A related effort, HULC [51], exploits

body-scene contact to stabilize monocular motion capture. In contrast to them, our method does not require contact annotations. Instead, we show that we can extract competitive contact estimates from our HOI reconstructions.

3 Technical Approach

In this section, we present MILO, our approach that reconstructs detailed human-object interactions from a single RGB image by leveraging an LRM. Sections 3.1 and 3.2 discuss preliminaries, while Sections 3.3 and 3.4 describe how we can fit a parametric human model to the LRM mesh. In Section 3.5, we describe how we can segment the reconstructed mesh into a human and object part. Finally, if an object template is available, Section 3.6 describes how we can align it to the LRM reconstruction. An overview of our pipeline is shown in Fig. 2.

3.1 Human Body Model Preliminaries

We represent the human body with the SMPL-H model [49], parameterized by global orientation $\Phi \in \mathbb{R}^3$, body pose $\Theta \in \mathbb{R}^{21 \times 3}$, hand pose $\Theta_h \in \mathbb{R}^{30 \times 3}$, shape $\beta \in \mathbb{R}^{10}$, and root translation $\Gamma \in \mathbb{R}^3$. Given these parameters, SMPL-H outputs body vertices and kinematic joints as: $[V, J] = \mathcal{M}(\Phi, \Theta, \Theta_h, \beta, \Gamma)$ where $V \in \mathbb{R}^{3 \times 6890}$ are the mesh vertices and $J \in \mathbb{R}^{3 \times 52}$ are the joint locations.

3.2 Human-Object Mesh Reconstruction

One of our key ideas is to use an LRM to directly reconstruct a combined human-object mesh from the input image. Concretely, we employ Hunyuan3D-2.0 [57], which produces a combined mesh capturing both the person and the interacted object. This choice is driven by two key advantages: (1) LRMs can generate a high-quality 3D approximation of the object and interaction present in the image. (2) The combined mesh encodes rich information about human-object spatial relations and proximity cues without requiring explicit contact annotations. To this end, we pass an RGBA image to the LRM to obtain a holistic human-object mesh. The alpha channel represents the combined segmentation maps of the human and the object. We interpret the LRM output mesh as a non-parametric interaction scaffold, segment it into human and object components, and fit the SMPL-H model to the human part to explain how the person interacts with the object. If an object template is available, we can also align it to the LRM mesh.

3.3 3D Keypoint Estimation

The goal of this step is to estimate a sparse set of 3D keypoints on the LRM mesh and use them for subsequent SMPL-H fitting. Specifically, we render the LRM mesh from a set of virtual views $\mathcal{V}$ (set to 60) and run ViTPose [70] and HaMeR [44] on each rendered image to obtain body and hand keypoints, respectively. For each view $v \in \mathcal{V}$, ViTPose predicts the COCO-25 body keypoints

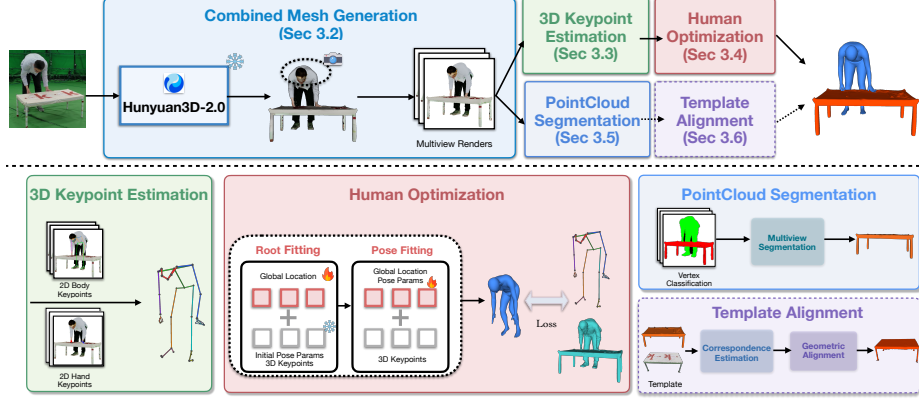


Fig. 2: Overview of MILO, our method for reconstructing 3D human-object interactions from a single image. Given an input RGB image, we first use Hunyuan3D-2.0 [57] to predict a combined human-object mesh (Sec. 3.2). We then render the mesh from multiple virtual views and estimate 2D body and hand keypoints in each view (Sec. 3.3), which are robustly triangulated to obtain 3D keypoints on the LRM reconstruction. Starting from an initial SMPL-H estimate from HMR2.0 [14] and HaMeR [44], we optimize SMPL-H in two stages (Sec. 3.4): (i) *root fitting*, optimizing only global orientation and translation, and (ii) *pose fitting*, additionally optimizing body/hand pose, and shape. Finally, we perform multi-view segmentation of the LRM mesh to extract the object component and compose it with the parametric human mesh for evaluation and visualization (Sec. 3.5). *If an object template is available*, we optionally align it to the segmented object using semantic correspondences (Sec. 3.6).

$x_b^{(v)} \in \mathbb{R}^{25 \times 2}$ with confidence scores $c_b^{(v)} \in [0, 1]^{25}$. Simultaneously, HaMeR predicts hand keypoints $x_{h,s}^{(v)} \in \mathbb{R}^{21 \times 2}$ for $s \in \{L, R\}$ (with confidence set to 1).

For each keypoint, we discard low-confidence detections, triangulate the surviving observations across all pairs of observed views, and select the hypothesis with the largest consensus set under a reprojection-error threshold, followed by nonlinear refinement over the inlier views. This returns the final set of body and hand 3D keypoints $(X, \bar{c})$, with $X \in \mathbb{R}^{67 \times 3}$. The confidence $\bar{c}$ comes from aggregating the per-keypoint confidence over its inlier views. These keypoints are used in the next stage to fit SMPL-H to the LRM mesh. We provide details about the confidence thresholds, triangulation, and aggregation in the SuppMat.

3.4 Human Optimization

We get initial estimates of the SMPL-H parameters using HMR2.0 [14], which predicts human pose and shape in the camera frame. We also initialize the hand pose parameters using HaMeR [44]. We then refine these estimates in two stages: the root fitting, which optimizes for the global orientation and the root translation, followed by the pose fitting stage, which additionally optimizes for the body shape, body, and hand pose. Both stages optimize the SMPL-H mesh, keeping the LRM mesh fixed. Additionally, in the first fitting stage, we optimize a human

scale parameter. This helps align the human to the LRM mesh without devolving to degenerate scales, since the LRM mesh is not guaranteed to produce the combined mesh at the correct metric scale.

Root Fitting. Let $\hat{P} = \{\tilde{\Phi}, \tilde{\Theta}, \tilde{\Theta}^h, \tilde{\beta}, \tilde{\Gamma}\}$ denote the initial SMPL-H prediction for image I as given by HMR2.0 [14] and HaMeR [44], with the parameters expressed in the camera frame. We initialize the body pose (Θ), hand pose (Θ^h) and body shape (β) with these values. In this first stage, we keep Θ, Θ^h, β fixed and only optimize Φ and Γ to align the SMPL-H joints X^S to their corresponding 3D keypoints X estimated in the previous section. The objective is expressed as: $\mathcal{L}_{rf} = \sum_i \rho(X_i^S - X_i)$, where ρ is the Geman-McClure robustifier [12]. This stage runs for 30 iterations and coarsely aligns SMPL-H with the LRM mesh.

Pose Fitting. Next, we optimize the full SMPL-H parameter set, i.e., shape β , body pose Θ , and hand pose Θ^h . To keep the solution feasible, we regularize pose and shape with the VPoser prior [43], the MANO prior [49] and an L_2 shape prior. Let $\zeta_{bp} \in \mathbb{R}^{32}$ be the representation of the body pose parameters Θ in the latent space of VPoser [43], and $\zeta_{hp} \in \mathbb{R}^{30}$ be the representation of the hand pose parameters Θ^h in the latent space of the MANO model [49]. The regularization terms can be expressed as: $\mathcal{L}_{bp} = \|\zeta_{bp}\|^2$, $\mathcal{L}_{hp} = \|\zeta_{hp}\|^2$ and $\mathcal{L}_\beta = \|\beta\|^2$.

Under severe self-occlusion or object occlusion, the LRM may hallucinate missing body parts, which can lead to image-inconsistent pose estimates. While $\mathcal{L}_{bp}$ and $\mathcal{L}_{hp}$ encourage feasible poses, they do not explicitly preserve consistency with the image. To improve robustness in such cases, we add an anchoring loss based on the HMR2.0 [14] pose initialization. We use the initialized latent body pose ζ_{bp}^{hmr} and penalize large deviations from it during pose fitting:

$$\mathcal{L}_{ha} = \sum_j \left[(\zeta_{bp,j} - \zeta_{bp,j}^{\text{hmr}})^2 - \tau \right]_+, \quad (1)$$

where τ is a tolerance margin, set to 2.5. This acts as a soft anchor: small latent refinements are not penalized, while larger deviations are discouraged.

To further improve alignment in occluded regions, we introduce a visible-vertex 3D consistency loss with the LRM reconstruction. Let $P^{\text{obs}} \subset \mathbb{R}^3$ denote a set of 3D points sampled from the observed (human) LRM reconstruction, and let $P^{\text{pred}} \subset \mathbb{R}^3$ denote the set of predicted SMPL-H mesh vertices. Because the reconstruction can be unreliable in occluded areas, we restrict the comparison to a subset $P_{\mathcal{V}}^{\text{pred}} \subseteq P^{\text{pred}}$ containing only those predicted vertices associated with visible joints, where $\mathcal{V}$ is the vertex-visibility set estimated from the rendered views. We then define a one-way robust Chamfer loss from the observed points to these visible predicted vertices:

$$\mathcal{L}_{3D} = \frac{1}{|P^{\text{obs}}|} \sum_{p \in P^{\text{obs}}} \rho \left(\min_{q \in P_{\mathcal{V}}^{\text{pred}}} \|p - q\|_2 \right), \quad (2)$$

where $p \in P^{\text{obs}}$ is an observed 3D point, $q \in P_v^{\text{pred}}$ is a visible predicted vertex, $|P^{\text{obs}}|$ is the number of observed points, and $\rho(\cdot)$ is a robust penalty (Geman–McClure) to reduce sensitivity to outliers. This term improves fitting in both self-occlusion and object-occlusion cases.

We combine these with the root loss to obtain the final fitting objective:

$$\mathcal{L}_{pf} = \lambda_{rf}\mathcal{L}_{rf} + \lambda_{bp}\mathcal{L}_{bp} + \lambda_{hp}\mathcal{L}_{hp} + \lambda_{\beta}\mathcal{L}_{\beta} + \lambda_{ha}\mathcal{L}_{ha} + \lambda_{3D}\mathcal{L}_{3D}, \quad (3)$$

where $\lambda_{rf} = 10$, $\lambda_{bp} = \lambda_{hp} = 0.04$, $\lambda_{\beta} = 0.05$, $\lambda_{ha} = 10$ and $\lambda_{3D} = 50$. We optimize this objective for 60 iterations, yielding a human mesh that is tightly aligned with the LRM reconstruction while remaining anatomically plausible.

3.5 Point Cloud Segmentation

The next step is to extract the object geometry from the LRM mesh to obtain an object point cloud. We use the same multi-view renders from Sec. 3.3 and apply a segmentation model [34,47] to obtain human and object masks. The prompt used is “person.<obj_name>”, indicating the return of one mask for the person and one for the object. The visible vertices in each view are projected onto the image plane and labeled using these masks, followed by multi-view aggregation and filtering. For each vertex, we aggregate the per-viewpoint scores that account for both viewpoint quality and a boundary-closeness weight. The viewpoint quality score balances geometric coverage and segmentation reliability. The boundary-closeness weight handles human-object contact regions. Next, we threshold the aggregated scores to obtain final labels. The vertices labeled as non-human form the raw object point cloud, which we clean via geometric filtering. This procedure is described in detail in the SuppMat. The resulting clean object point cloud is composed with the fitted SMPL-H mesh.

3.6 Template Alignment

MILO can operate without a template at test time. When a dataset provides a template mesh, we can optionally align it to the segmented object component of the LRM reconstruction. Let $M^{\text{tmp}} = \{u_i\}_{i=1}^N$ be the template mesh vertices and let $P^{\text{lrn}} = \{v_j\}_{j=1}^M$ denote the segmented LRM point cloud. We seek a similarity transform: $T(u) = s\mathbf{R}u + \mathbf{t}$, with $s > 0$, $\mathbf{R} \in SO(3)$, and $\mathbf{t} \in \mathbb{R}^3$, aligning the template to the reconstructed object.

Semantic correspondences. Our first step is to establish correspondences between the LRM and the template mesh. We do this by considering multi-view renders of both meshes. For this, we first restrict matching to LRM renders that are most

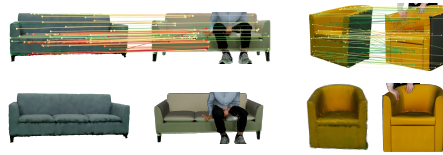


Fig. 3: Template Alignment. Top row: Visualization of semantic correspondences between a render of the template mesh and the LRM mesh. Bottom row: Alignment of the template mesh with the LRM mesh.

consistent with the input image. Using the geometry-aware semantic correspondence model of Zhang et al. [74], we compute a global mutual-nearest-neighbor similarity $S(\cdot, \cdot)$ between the input image and every LRM render (considering only the object foreground), and retain the top- K_{in} views with $K_{\text{in}} = 5$.

Next, for each template render I_q^{tmp} , we evaluate its similarity $S(I_q^{\text{tmp}}, I_r^{\text{lrn}})$ against the retained LRM renders I_r^{lrn} and keep the top- K_{pair} matches ($K_{\text{pair}} = 3$). For each selected pair (q, r) , we compute dense bidirectional pixel correspondences in descriptor space. Let $\Pi_{q \rightarrow r}(x)$ denote the target pixel matched to source pixel x , and let $C_{q \rightarrow r}(x)$ be its cosine similarity.

We lift these 2D matches to 3D correspondences by unprojection. A template vertex u_i is considered only if it is visible in view q , projects inside the foreground mask, and its matched pixel $\Pi_{q \rightarrow r}(\pi_q(u_i))$ lies within a radius τ_{2D} of some visible LRM vertex projection in view r . The matched LRM vertex is:

$$j^*(i, q, r) = \arg \min_{j \in \mathcal{V}_r^{\text{lrn}}} \|\pi_r(v_j) - \Pi_{q \rightarrow r}(\pi_q(u_i))\|_2, \quad (4)$$

subject to: $C_{q \rightarrow r}(\pi_q(u_i)) > \tau_{\text{loc}}$ and $\|\pi_r(v_{j^*}) - \Pi_{q \rightarrow r}(\pi_q(u_i))\|_2 < \tau_{2D}$. We use $\tau_{\text{loc}} = 0.6$ and $\tau_{2D} = 15$ pixels. A template vertex may obtain multiple candidate matches across view pairs. We retain only the highest-scoring one, with score:

$$\omega_{i,q,r} = C_{q \rightarrow r}(\pi_q(u_i)) \cdot S(I_q^{\text{tmp}}, I_r^{\text{lrn}}). \quad (5)$$

This gives a sparse forward map $i \mapsto j_i$ with weight $w_i = \max_{q,r} \omega_{i,q,r}$. To suppress unstable matches, we also compute the reverse map LRM $\rightarrow$ template and apply neighborhood-based cycle consistency. A forward match $u_i \mapsto v_{j_i}$ is retained only if the reverse map sends v_{j_i} back to a template vertex within a local neighborhood of u_i , defined relative to the local template mesh scale. This yields the final sparse correspondence set: $\mathcal{C} = \{(i_k, j_k, w_k)\}_{k=1}^K$, (Fig. 3).

Template alignment. Given sparse weighted correspondences $\mathcal{C} = \{(i_k, j_k, w_k)\}_{k=1}^K$, we first estimate an initial similarity transform $T(u) = s\mathbf{R}u + \mathbf{t}$ by solving a weighted Sim(3) Procrustes problem with a weighted Kabsch fit. We then refine T with an iterative weighted Sim(3) re-estimation that jointly uses $\mathcal{C}$ and nearest-neighbor constraints to the segmented LRM object point cloud P^{lrn} .

At iteration t , we transform template vertices as $\tilde{u}_i^{(t)} = T^{(t)}(u_i)$, assign each $\tilde{u}_i^{(t)}$ its closest point $p_i^{(t)} = \text{NN}(\tilde{u}_i^{(t)}, P^{\text{lrn}})$, and form an inlier set $\mathcal{V}_t = \{i \mid \|\tilde{u}_i^{(t)} - p_i^{(t)}\|_2 < \tau_t\}$ using a threshold τ_t that is progressively tightened across iterations. We then compute $T^{(t+1)}$ by a *single* weighted Kabsch-with-scale fit on the combined pairs $\{(u_{i_k}, v_{j_k})\}_{k=1}^K \cup \{(u_i, p_i^{(t)})\}_{i \in \mathcal{V}_t}$, using weights w_k for semantic pairs and a constant weight of 10 for each nearest-neighbor pair. This procedure retains the semantic alignment captured by $\mathcal{C}$ while improving surface-level agreement with the reconstructed object geometry.

4 Experiments

This section presents our empirical evaluation. Section 4.1 compares MILO with the state-of-the-art on the InterCap [19], HODome [75] and IMHD [77] datasets.

Methods	Type	Needs Template?	Needs Contact?	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
CONTHO [41]	Reg.	✓	✓	8.36	24.30	13.14
HOI-TG [64]	Reg.	✓	✗	8.22	25.05	14.63
PHOSA [73]	Opt.	✓	✓	10.07	23.36	13.38
Open3DHOI [65]	Opt.	✗	✗	<u>6.88</u>	31.18	10.17
PICO [5]	Opt.	✓	✓	7.43	21.85	10.33
Ours (w/ template)	Opt.	✓	✗	6.96	18.97	7.45
Ours (w/o template)	Opt.	✗	✗	6.85	<u>20.74</u>	<u>9.36</u>

Table 1: Quantitative comparison with the state of the art on the Inter-Cap dataset [19]. We report Procrustes-Aligned Chamfer Distance (PA-CD) for the human mesh (PA-CD_h), object mesh (PA-CD_o), and the combined human-object pair (PA-CD_{h+o}); lower is better. Regression (Reg.) and optimization (Opt.) baselines are evaluated under the training protocol from Cseke et al. [5]. Our method, MILO, uses no contact information, yet achieves the best overall reconstruction accuracy.

We evaluate settings with and without access to a template. Sec. 4.2 evaluates the effect of MILO on the downstream task of estimating vertex-level contact on the human body. Section 4.3 studies the effect of the underlying LRM mesh geometry and the dependence of MILO on the choice of LRM. Section 4.4 compares our joint reconstruction scaffold against object-only and individual-reconstruction baselines. Section 4.5 analyzes and ablates the key design choices. Finally, Sec. 4.6 presents qualitative in-the-wild reconstructions on PICO-db [5].

Methods	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
PICO [5]	13.12	14.44	10.07
Ours (w/ template)	9.50	12.97	<u>6.68</u>
Ours (w/o template)	<u>9.71</u>	<u>13.75</u>	6.38

Table 2: Quantitative evaluation on the HODome dataset [75].

Methods	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
PICO [5]	15.81	17.70	13.24
Ours (w/ template)	<u>11.76</u>	<u>13.39</u>	<u>10.10</u>
Ours (w/o template)	8.99	9.71	6.98

Table 3: Quantitative evaluation on the IMHD dataset [77].

4.1 Comparison with the state-of-the-art

InterCap [19] is an indoor dataset with ground-truth human and object meshes, including ten human subjects and ten everyday objects in various interaction scenarios. We evaluate MILO for HOI reconstruction in this controlled setting and compare it with both regression-based [41, 64] and optimization-based methods [5, 65, 73] following the protocol of Cseke et al. [5]. For the learning-based baselines, we use checkpoints trained only on BEHAVE [2], i.e., enforcing an out-of-domain evaluation.

We report Procrustes-Aligned Chamfer Distance (PA-CD) for the human mesh, the object mesh, and the combined human-object pair. The rigid alignment is performed on the combined mesh; after alignment, PA-CD is computed separately for the human and object components as well as for the joint reconstruction. Quantitative results are presented in Tab. 1. MILO outperforms all previous work (even without access to a template), while relying on less privileged information than competing methods.

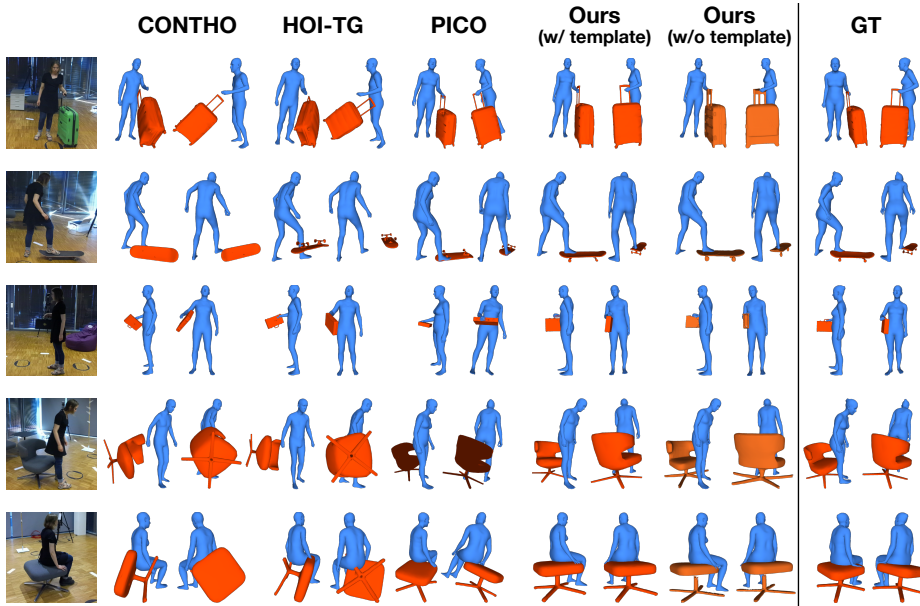


Fig. 4: Qualitative comparison on the InterCap dataset [19]. From left to right: input RGB image, CONTHO [41], HOI-TG [64], PICO-fit [5], MILO reconstructions (with and without template meshes, respectively), and ground-truth human-object meshes. Our method achieves better human-object alignment in 3D. Humans are rendered in **blue**, template objects in **red**, and segmented object meshes in **orange**.

Method	F1↑	Precision↑	Recall↑	geo. (cm)↓
DECO [59]	9.37	12.57	10.28	126.70
Ours	30.00	36.57	39.49	39.57

Table 4: Contact evaluation. Metrics are reported on InterCap. Our 3D geometry-inferred contact outperforms the image-regressed contact by the DECO baseline [59].

Additionally, we evaluate MILO on HODome [75] and IMHD [77] datasets, comparing against the best-performing baseline PICO. HODome and IMHD are also indoor datasets with ground-truth human and object meshes. HODome consists of 10 human subjects and 20 object classes, while IMHD consists of 8 human subjects and 10 object categories with different interactions with each object. As can be seen from Tab. 2 and Tab. 3, MILO performs better than PICO across the board, in both modes.

The improvements from MILO are also evident in the qualitative results on InterCap [19] (Fig. 4), HODome [75], and IMHD [77] (Fig. 5), where MILO achieves better 3D alignment between the human and the object, with more accurate contact and fewer penetrations than prior approaches.

4.2 Contact estimation

MILO does not predict contact, but we can infer contact based on the proximity of the SMPL-H mesh and the aligned object template. We report the results

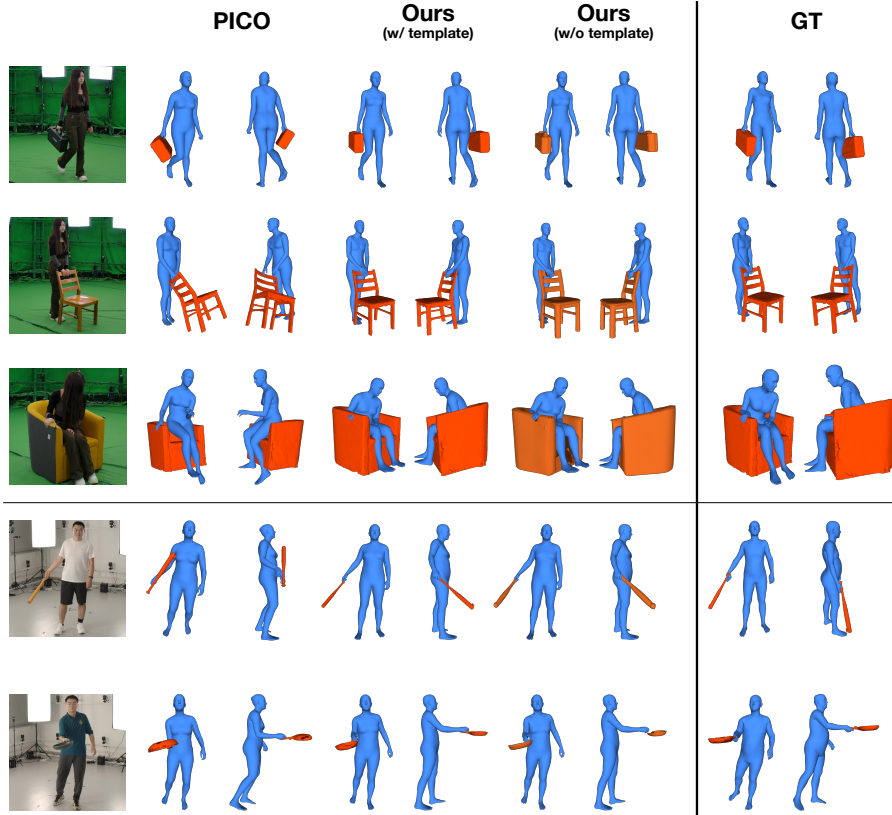


Fig. 5: Qualitative evaluation of PICO-fit [5] and MILO on images from the HODome [75] and IMHD [77] datasets. Our outputs, with or without templates, are more physically plausible and closer to the GT than PICO. Humans are rendered in blue, template objects in red, and segmented object meshes in orange.

of this strategy in Tab. 4. Compared to the DECO RICH checkpoint, MILO achieves higher F1, Precision, and Recall, and lower geometric error on InterCap.

4.3 Effect of mesh geometry

Next, we evaluate the effect of the initial mesh geometry in the MILO pipeline. Table 5 compares our default Hunyuan3D-2.0 choice with alternative LRM backbones, and an oracle mesh obtained from the ground-truth human-object mesh. Replacing Hunyuan3D-2.0 [57] with SAM3D (joint) [56] gives comparable performance, suggesting that MILO is not tied to a specific LRM choice. InstantMesh [69] performs worse, but we acknowledge that it corresponds to an older method. Finally, the oracle result indicates that the remaining error is largely bounded by upstream reconstruction quality.

Mesh Geometry	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
InstantMesh [69]	7.51	33.21	10.84
SAM3D (joint) [56]	7.00	21.99	9.80
Hunyuan3D-2.0 [57]	6.85	20.74	9.36
Oracle	3.28	4.93	3.12

Table 5: Effect of mesh geometry. Results presented on InterCap. We replace the mesh reconstructed by Hunyuan3D-2.0 [57] with meshes produced by SAM3D [56] and InstantMesh [69], as well as the ground-truth geometry (oracle). We observe that SAM3D performs comparably to Hunyuan3D-2.0, which indicates that MILO is not tied to a specific LRM choice.

Baseline	Contact Loss	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
EasyHOI* (full-body)	✗	41.57	77.06	37.21
EasyHOI* (full-body)	✓	9.48	27.93	12.08
SAM3D (independent)	✗	18.18	31.07	15.89
Ours	✗	6.85	20.74	9.36

Table 6: Effect of joint human & object reconstruction scaffold. Results are presented on InterCap. The two EasyHOI-style variants follow an object-only pipeline: the object is reconstructed by the LRM, and its pose is optimized by silhouette reprojection, either alone (first row) or with an additional contact term (second row). SAM3D (independent) instead reconstructs the human and object independently and composes them through MoGe [63] pointmap alignment. The use of a joint human-object scaffold proposed by MILO (last row) outperforms all baselines.

4.4 Comparison with individual reconstruction

MILO leverages the LRM reconstruction as a shared geometric scaffold for HOI reconstruction: it fits a parametric human model to the human region of the reconstructed mesh and extracts the interacting object from the same scaffold. In contrast, other recent methods [35] follow an object-only pipeline, in which an LRM is used to reconstruct the object mesh and the resulting shape is subsequently fitted using reprojection and contact constraints. For example, EasyHOI [35] adopts this strategy for hand-object interaction reconstruction.

Although EasyHOI is not directly comparable to our setting because it considers hand-object rather than full-body interactions, we design a baseline that follows the same object-only paradigm. Specifically, we follow the same pipeline with EasyHOI [35]; first (1) we perform inpainting, then (2) we reconstruct the object with InstantMesh [69] and post-process the output mesh to make it watertight, and finally (3) we optimize the object pose using a silhouette reprojection loss. Additionally, we consider a version where a contact loss is added. We discuss details of contact estimation in the SuppMat.

As shown in Tab. 6, MILO outperforms both versions, with and without contact loss. This result indicates that using a joint reconstruction scaffold is more effective for HOI reconstruction than an object-only reconstruction followed by fitting to the image. Moreover, we observe that contact cues are essential for

Human Fitting	Uses Template	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓	Human Fitting	Uses Template	PA-CD _h (cm) ↓	PA-CD _o (cm) ↓	PA-CD _{h+o} (cm) ↓
No fitting	✗	30.58	23.53	14.10	No fitting	✓	14.97	37.02	14.57
Root Fitting	✗	7.98	22.53	10.15	Root Fitting	✓	7.79	18.63	7.99
Pose Fitting	✗	6.85	20.74	9.36	Pose Fitting	✓	6.96	18.97	7.45

Table 7: Effect of human fitting. Results are presented on the InterCap dataset [19]. **No fitting** computes the CD metrics using the raw LRM mesh (adopting our point-cloud segmentation), without SMPL-H optimization. **Root fitting** optimizes only the global orientation and translation of the human. **Pose fitting** additionally optimizes body/hand pose and shape for SMPL-H. We consider both template and template-free settings. The progressive reductions across metrics highlight the importance of the fitting stages and the necessity of a parametric model for precise 3D HOI reconstruction.

these image-fitting approaches, since otherwise the location and scale of the object are mostly unconstrained (first row of Tab. 6). Removing the reliance on explicit contact estimation is one of the key advantages of MILO.

To further identify whether MILO’s gains come from the joint reconstruction or from the downstream fitting pipeline, we add a controlled baseline that reconstructs the human and the object separately. We follow the composition method of SAM3D [56, 71] and run SAM3D independently on the human and the object regions of the same image. Then, we align the two resulting meshes in a common frame via MoGe [63] pointmap alignment, yielding the “SAM3D (independent)” baseline in Tab. 6. This baseline differs from the “SAM3D (joint)” baseline of Table 5 in that the geometry comes from two independent reconstructions instead of a single joint scaffold, so we can isolate the value of joint reconstruction. As reported in Tab. 6, MILO clearly outperforms this baseline across all metrics. This substantial improvement confirms that the joint scaffold, rather than the fitting procedure, is the main driver of MILO’s accuracy.

4.5 Ablation Experiments

We also evaluate the significance of the two fitting stages of the optimization pipeline. We perform this ablation using both the segmented LRM mesh and the aligned template (Tab. 7). As can be seen, removing the pose fitting, or both fitting stages, consistently degrades performance, confirming that both stages of our optimization are necessary to achieve the best overall reconstruction quality. Additional ablation experiments have been included in the SuppMat.

4.6 Qualitative Evaluation on In-The-Wild Images

To assess MILO’s robustness, we further evaluate it qualitatively on in-the-wild images from PICO-db [5]. As illustrated in Fig. 6, MILO recovers human and object meshes that are better aligned in 3D and exhibit more coherent contact patterns and relative positioning compared to prior work [5].

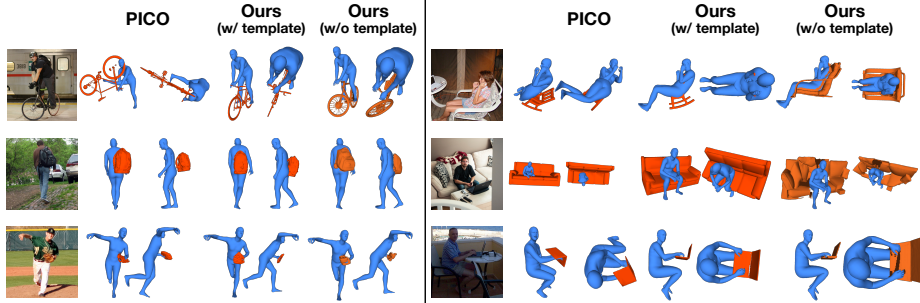


Fig. 6: Qualitative reconstructions on in-the-wild images from PICO-db [5]. MILO recovers 3D human-object geometry from a single image, with coherent arrangement and fewer interpenetrations than the PICO-fit [5] baseline. Humans are rendered in blue, template objects in red, and the segmented object mesh in orange.

5 Conclusion

We introduced MILO, an approach for reconstructing human-object interactions from a single image. MILO leverages the visual capabilities of Large Reconstruction Models to predict a holistic human-object mesh. It then segments the mesh into human and object parts, fits the SMPL-H body to the human part, and optionally the object template to the object part, if such a template is available. By treating the LRM output as a geometric scaffold, our method captures arrangement and proximity cues while remaining agnostic to object category and shape. Quantitative experiments on multiple benchmarks show that MILO surpasses existing HOI reconstruction methods despite using significantly less privileged information. Moreover, qualitative results on in-the-wild images demonstrate the generalization to diverse objects, poses, and interaction contexts. These results highlight the promise of LRMs for scalable, label-efficient HOI estimation.

Limitations and Future Work. One limitation of our approach is that the performance of MILO is bounded by the quality of the underlying LRM. However, given the quick progress in this space [56, 58, 60], we expect that the performance of LRMs will continue to improve. Based on our results, we also observe that our performance is hindered by the accuracy of the point cloud segmentation. We anticipate that with better methods for point-cloud segmentation, our approach will also benefit. Future work includes the refinement of LRM outputs especially for smaller objects, integrating stronger instance-level reasoning and segmentation to extract a better object mesh from the LRM output, and extending the framework to video and multi-person/multi-object interaction settings.

Acknowledgement

This research was supported by NSF-2504906, and 2544200; gifts from Adobe, Google and Nvidia; and computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin.

References

1. Bernardin, K., Ogawara, K., Ikeuchi, K., Dillmann, R.: A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *Transactions on Robotics (T-RO)* **21**(1), 47–57 (2005)
2. Bhatnagar, B.L., Xie, X., Petrov, I.A., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: BEHAVE: Dataset and method for tracking human object interactions. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 15935–15946 (2022)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In: *European Conference on Computer Vision (ECCV)*. vol. 9909, pp. 561–578 (2016)
4. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: *European Conference on Computer Vision (ECCV)*. pp. 628–644. Springer (2016)
5. Cseke, A., Tripathi, S., Dwivedi, S.K., Lakshminpathy, A., Chatterjee, A., Black, M.J., Tzionas, D.: PICO: Reconstructing 3D people in contact with objects. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1783–1794 (June 2025)
6. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., Vanderbilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-XL: A universe of 10m+ 3D objects. *arXiv:2307.05663* (2023)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., Vanderbilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3D objects. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 13142–13153 (2023)
8. Dwivedi, S.K., Schmid, C., Yi, H., Black, M.J., Tzionas, D.: POCO: 3D pose and shape estimation using confidence. In: *International Conference on 3D Vision (3DV)*. pp. 85–95 (2024)
9. Dwivedi, S.K., Sun, Y., Patel, P., Feng, Y., Black, M.J.: TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1323–1333 (2024)
10. Fan, H., Su, H., Guibas, L.J.: A point set generation network for 3d object reconstruction from a single image. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. pp. 2463–2471. IEEE Computer Society (2017)
11. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Learning complex 3D human self-contact. In: *AAAI Conference on Artificial Intelligence*. pp. 1343–1351 (2021)
12. Geman, S., McClure, D.E.: Statistical methods for tomographic image reconstruction. *Bulletin of the International Statistical Institute* **52**(4), 5–21 (1987)
13. Girdhar, R., Fouhey, D., Rodriguez, M.D., Gupta, A.: Learning a predictable and generative vector representation for objects. In: *European Conference on Computer Vision (ECCV)*. vol. 9910, pp. 484–499 (2016)
14. Goel, S., Pavlakos, G., Rajasegaran, J., Kanazawa*, A., Malik*, J.: Humans in 4D: Reconstructing and tracking humans with transformers. In: *International Conference on Computer Vision (ICCV)*. pp. 14737–14748 (2023)
15. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3D scenes by learning human-scene interaction. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 14708–14718 (2021)

16. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 11807–11816 (2019)
17. Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., Tan, H.: LRM: Large reconstruction model for single image to 3D. *arXiv: 2311.04400* (2023)
18. Huang, C.H., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 13264–13275 (2022)
19. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision (IJCV)* **132**(7), 2551–2566 (2024)
20. Huo, C., Shi, Y., Ma, Y., Xu, L., Yu, J., Wang, J.: Stackflow: Monocular human-object reconstruction by stacked normalizing flow with offset. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*. pp. 902–910. *ijcai.org* (2023)
21. Jiang, H., Huang, Q., Pavlakos, G.: Real3d: Towards scaling large reconstruction models with real images. In: *International Conference on Computer Vision (ICCV)*. pp. 5821–5833 (October 2025)
22. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In: *International Conference on 3D Vision (3DV)*. pp. 42–52 (2021)
23. Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., Miura, Y.: Patterns of static prehension in normal hands. *American Journal of Occupational Therapy* **34**(7), 437–445 (1980)
24. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139:1–139:14 (2023)
25. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: PARE: Part attention regressor for 3D human body estimation. In: *International Conference on Computer Vision (ICCV)*. pp. 11127–11137 (2021)
26. Kolotouros, N., Pavlakos, G., Daniilidis, K.: Convolutional mesh regression for single-image human shape reconstruction. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 4496–4505 (2019)
27. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.: Unite the people: Closing the loop between 3D and 2D human representations. In: *Computer Vision and Pattern Recognition (CVPR)* (2017)
28. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 3383–3393 (2021)
29. Li, Z., Liu, J., Zhang, Z., Xu, S., Yan, Y.: CLIFF: Carrying location information in full frames into human pose and shape estimation. In: *European Conference on Computer Vision (ECCV)*. vol. 13665, pp. 590–606 (2022)
30. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3D whole-body mesh recovery with component aware transformer. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 21159–21168 (2023)
31. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 1954–1963 (2021)
32. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: *International Conference on Computer Vision (ICCV)* (2021)

33. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3D object. *International Conference on Computer Vision (ICCV)* (2023)
34. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023)
35. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 7037–7047. *Computer Vision Foundation / IEEE* (2025)
36. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. *Transactions on Graphics (TOG)* **34**(6), 248:1–248:16 (2015)
37. Melas-Kyriazi, L., Laina, I., Rupperecht, C., Vedaldi, A.: Realfusion 360° reconstruction of any object from a single image. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 8446–8455. *IEEE* (2023)
38. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J. (eds.) *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I. Lecture Notes in Computer Science*, vol. 12346, pp. 405–421. *Springer* (2020)
39. Moon, G., Choi, H., Lee, K.M.: Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In: *Computer Vision and Pattern Recognition Workshops (CVPRw)*. pp. 2307–2316. *IEEE* (2022)
40. Müller, L., Osman, A.A.A., Tang, S., Huang, C.H.P., Black, M.J.: On self-contact and human pose. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 9990–9999 (2021)
41. Nam, H., Jung, D.S., Moon, G., Lee, K.M.: Joint reconstruction of 3D human and object via contact-based refinement transformer. In: *Computer Vision and Pattern Recognition (CVPR)* (2024)
42. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 165–174 (2019)
43. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 10975–10985 (2019)
44. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 9826–9836. *IEEE* (2024)
45. Petrov, I.A., Marin, R., Chibane, J., Pons-Moll, G.: Tridi: Trilateral diffusion of 3d humans, objects, and interactions. In: *International Conference on Computer Vision (ICCV)*. pp. 5523–5535 (October 2025)
46. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H., Skorokhodov, I., Wonka, P., Tulyakov, S., Ghanem, B.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In: *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net* (2024)
47. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N.,

- Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024)
48. Rempe, D., Birdal, T., Hertzmann, A., Yang, J., Sridhar, S., Guibas, L.J.: HuMoR: 3D human motion model for robust pose estimation. In: International Conference on Computer Vision (ICCV). pp. 11468–11479 (2021)
 49. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)* **36**(6) (Nov 2017)
 50. Rong, Y., Shiratori, T., Joo, H.: FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration. In: International Conference on Computer Vision Workshops (ICCVw). pp. 1749–1759 (2021)
 51. Shimada, S., Golyanik, V., Li, Z., Pérez, P., Xu, W., Theobalt, C.: HULC: 3D human motion capture with pose manifold sampling and dense contact guidance. In: European Conference on Computer Vision (ECCV). pp. 516–533 (2022)
 52. Shimada, S., Golyanik, V., Xu, W., Theobalt, C.: PhysCap: Physically plausible monocular 3D motion capture in real time. *Transactions on Graphics (TOG)* **39**(6), 235:1–235:16 (2020)
 53. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: European Conference on Computer Vision (ECCV) (2020)
 54. Tang, J., Chen, Z., Chen, X., Wang, T., Zeng, G., Liu, Z.: LGM: large multi-view gaussian model for high-resolution 3d content creation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part IV. Lecture Notes in Computer Science*, vol. 15062, pp. 1–18. Springer (2024)
 55. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from A single image with diffusion prior. In: *Computer Vision and Pattern Recognition (CVPR)*. pp. 22762–22772. IEEE (2023)
 56. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025)
 57. Team, T.H.: Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation (2025)
 58. Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V., Cao, Y.: Triposr: Fast 3d object reconstruction from a single image. *CoRR* **abs/2403.02151** (2024)
 59. Tripathi, S., Chatterjee, A., Passy, J.C., Yi, H., Tzionas, D., Black, M.J.: DECO: Dense estimation of 3D human-scene contact in the wild. In: International Conference on Computer Vision (ICCV). pp. 8001–8013 (2023)
 60. Voleti, V., Yao, C., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: SV3D: novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part I. Lecture Notes in Computer Science*, vol. 15059, pp. 439–457. Springer (2024)
 61. Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.: Pixel2mesh: Generating 3d mesh models from single RGB images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *Computer Vision - ECCV 2018 - 15th European Conference*,

- Munich, Germany, September 8-14, 2018, Proceedings, Part XI. Lecture Notes in Computer Science, vol. 11215, pp. 55–71. Springer (2018)
62. Wang, P., Tan, H., Bi, S., Xu, Y., Luan, F., Sunkavalli, K., Wang, W., Xu, Z., Zhang, K.: PF-LRM: pose-free large reconstruction model for joint pose and shape prediction. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024)
 63. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Computer Vision and Pattern Recognition (CVPR). pp. 5261–5271. Computer Vision Foundation / IEEE (2025)
 64. Wang, Z., Zheng, Q., Ma, S., Ye, M., Zhan, Y., Li, D.: End-to-end HOI reconstruction transformer with graph-based encoding. In: Computer Vision and Pattern Recognition (CVPR). pp. 27706–27715. Computer Vision Foundation / IEEE (2025)
 65. Wen, B., Huang, D., Zhang, Z., Zhou, J., Deng, J., Gong, J., Chen, Y., Ma, L., Li, Y.L.: Reconstructing in-the-wild open-vocabulary human-object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 17426–17436 (2025)
 66. Wu, J., Pavlakos, G., Gkioxari, G., Malik, J.: Reconstructing hand-held objects in 3d. arXiv preprint arXiv:2404.06507 (2024)
 67. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: CHORE: contact, human and object reconstruction from a single RGB image. In: European Conference on Computer Vision (ECCV). vol. 13662, pp. 125–145 (2022)
 68. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: Computer Vision and Pattern Recognition (CVPR). pp. 6183–6192 (2020)
 69. Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., Shan, Y.: Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. CoRR **abs/2404.07191** (2024)
 70. Xu, Y., Zhang, J., Zhang, Q., Tao, D.: Vitpose: Simple vision transformer baselines for human pose estimation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022)
 71. Yang, X., Kukreja, D., Pinkus, D., Fan, T., Park, J., Shin, S., Cao, J., Liu, J.W., Ugrinovic, N., Sagar, A., et al.: Sam 3d body: Robust full-body human mesh recovery. In: Computer Vision and Pattern Recognition (CVPR). pp. 7209–7219 (2026)
 72. Yang, Y., Zhai, W., Luo, H., Cao, Y., Zha, Z.J.: LEMON: Learning 3D human-object interaction relation from 2D images. In: Computer Vision and Pattern Recognition (CVPR) (2024)
 73. Zhang, J.Y., Pepose, S., Joo, H., Ramanan, D., Malik, J., Kanazawa, A.: Perceiving 3D human-object spatial arrangements from a single image in the wild. In: European Conference on Computer Vision (ECCV). vol. 12357, pp. 34–51 (2020)
 74. Zhang, J., Herrmann, C., Hur, J., Chen, E., Jampani, V., Sun, D., Yang, M.H.: Telling left from right: Identifying geometry-aware semantic correspondence. In: Computer Vision and Pattern Recognition (CVPR) (2024)
 75. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: Neurdome: A neural modeling pipeline on multi-view human-object interactions. In: Computer Vision and Pattern Recognition (CVPR) (2023)
 76. Zhang, S., Zhang, Y., Bogo, F., Pollefeys, M., Tang, S.: Learning motion priors for 4D human body capture in 3D scenes. In: International Conference on Computer Vision (ICCV). pp. 11343–11353 (2021)

77. Zhao, C., Zhang, J., Du, J., Shan, Z., Wang, J., Yu, J., Wang, J., Xu, L.: I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In: Computer Vision and Pattern Recognition (CVPR). pp. 729–741 (2024)