RHINO: Reconstructing Human Interactions with Novel Objects from Monocular Videos

Lixin Xue[1] Chengwei Zheng[1,2] Georgios Paschalidis[3]
Chen Guo[1] Manuel Kaufmann[1] Juan Zarate[1] Dimitrios Tzionas[3,4]

[1]ETH Zürich [2]The University of Tokyo [3]University of Amsterdam [4]Aristotle University of Thessaloniki

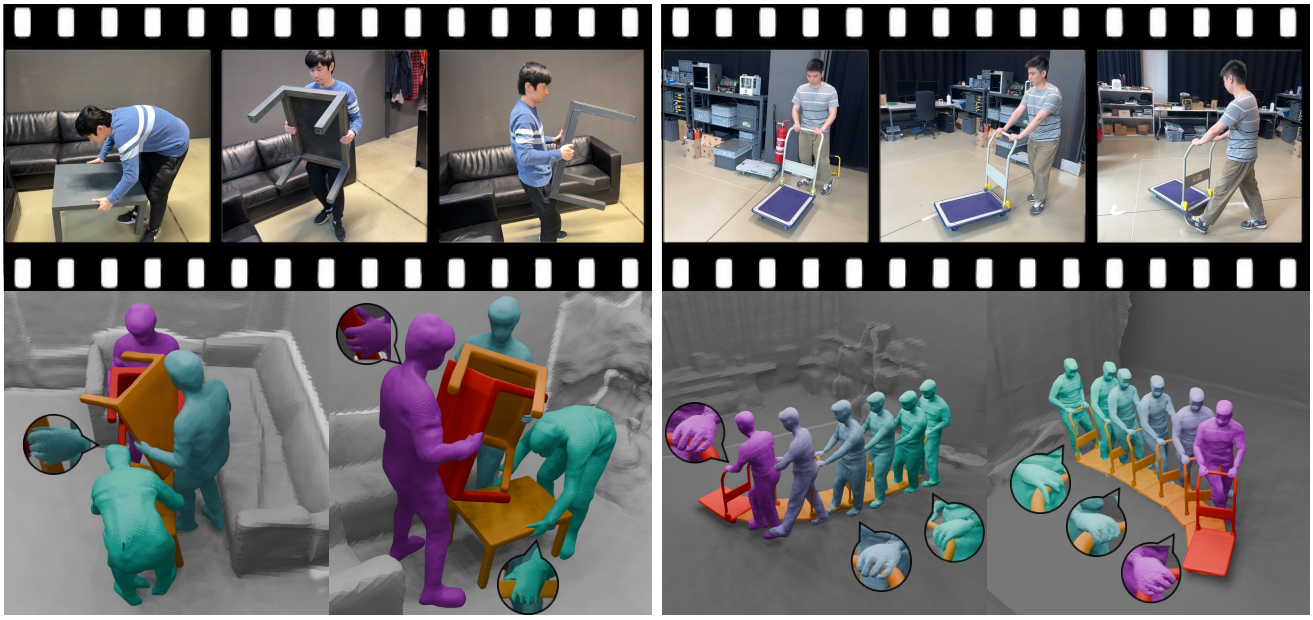


Figure 1. We develop RHINO, a novel framework that reconstructs detailed (dynamic) 3D human-object interactions (HOI) and the surrounding scene within a common world frame from a monocular RGB video with a moving viewpoint. RHINO uses per-component neural SDFs to: (i) capture shape details, and (ii) encourage contact via a differentiable distance term. The "zoom-in insets" highlight plausible contacts. RHINO requires neither a pre-scanned object template nor prior knowledge of the object, unlike most existing work.

Abstract

Reconstructing people, objects, and their interactions in 3D is a long-standing and fundamental goal for intelligent systems. Often the input is RGB video from a moving camera, making the task ill-posed; depth is ambiguous, humans and objects occlude each other, and camera and object motion entangle to create apparent motion. Most prior work addresses humans or objects in isolation, ignoring their interplay, or assumes known 3D shapes or cameras, which is impractical for real-world applications. We develop RHINO (Reconstructing Human Interactions with Novel Objects), a novel three-step framework that recovers in 3D a human, novel (unseen) manipulated object, and static scene in a common world frame from a monocular RGB video. First, we leverage 3D-aware foundation models to obtain cues that stabilize Structure-from-Motion (SfM) even for low-texture regions; this yields a coarse shape and apparent motion of a manipulated object from foreground pixels, and a coarse scene shape and camera motion from background pixels. Second, we estimate a human in the camera frame via an off-the-shelf method, and subtract the camera motion from apparent motion to extract the object motion; this registers the human, object, and coarse scene shapes into a common world frame. Third, we refine shapes using a compositional neural field with per-component signed-distance fields. The latter further enables differentiable contact priors that attract surfaces while penalizing interpenetration, improving the physical plausibility of the final reconstruction. For evaluation, we capture a new dataset of handheld monocular videos synchronized with a volumetric 4D capture stage, providing ground-truth shape and camera motion. RHINO outperforms state-of-the-art baselines on novel-view synthesis and 4D reconstruction. Ablations show that each stage contributes substantially. Code and data are available at https://lxxue.github.io/RHINO.

1. Introduction

Humans constantly interact with and manipulate objects in their surroundings. Enabling computers to perceive these interactions in 3D from a single RGB video benefits assistive robotics, AR/VR, healthcare, media, and learning from internet-scale videos. In this work, we reconstruct a 4D scene with dynamic human-object interactions (HOI) from such a monocular video.

This task is challenging due to depth ambiguities and human-object occlusions. Moreover, camera motion produces "apparent motion" that entangles camera and object motion. Finally, while 3D human recovery is now tractable, reconstructing manipulated objects remains highly challenging because objects vary widely in shape and appearance and are often low-texture or symmetric, making feature detection and tracking difficult.

Due to these challenges, prior methods address only parts of the problem. Most methods reconstruct human-free scenes [4, 54, 59, 60] or humans in isolation [12, 16, 17, 45]. A few methods jointly reconstruct humans and scenes in a common world frame, but still fail to capture manipulation-induced object motion [68, 76]; see Fig. 2. Moreover, many methods assume known object/scene shapes [2, 20, 21, 24] or calibrated cameras [2, 23], which is often impractical.

To our knowledge, no method provides a framework for recovering a 4D HOI scene in a world frame from a moving-view monocular RGB video. We address this gap with RHINO (*Reconstructing Human Interactions with Novel Objects*), a three-stage framework; see its results in Fig. 1.

First, we estimate the shape and motion of a novel (unseen) object. Prior work [22, 52] tackles this via Structure-from-Motion (SfM) and feature correspondences. However, everyday objects in full-body videos are often low-texture and occupy a small region of the frame, making sparse features [10] unstable and dense correspondences [51] inconsistent across frames. We leverage recent 3D-aware foundation models [32] to produce dense, robust correspondences that enable more reliable SfM for these objects.

However, the camera also moves, so apparent motion entangles camera and object motion. To disentangle the two motions, we first estimate the camera motion from static background regions via SfM, and scale and align it with the apparent motion. Then, we "subtract" this camera-induced component from the apparent motion to obtain object motion. We also estimate an initial human shape and motion in the camera frame [53]. Overall, this stage registers the human, object, and scene into a common world frame.

The initial human, object, and scene shapes are coarse. We refine them using a compositional neural field with per-component signed-distance and appearance fields. We optimize the fields for photometric and mask consistency, with geometric regularization. This yields detailed 3D human, object, and scene shapes aligned with the image cues. Crucially, operating in a world frame—rather than individual camera frames—enables multi-frame optimization over all components, mitigating per-frame initialization errors.

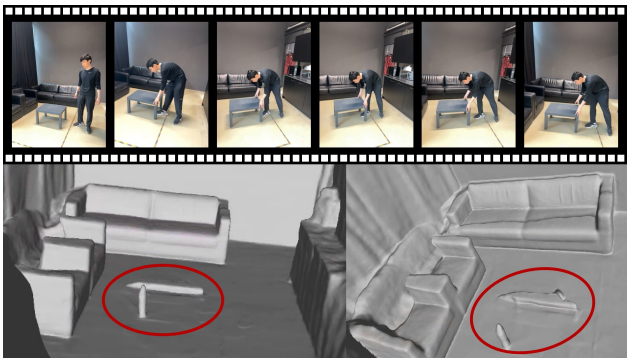


Figure 2. **Existing work**, such as the HSR [68] SotA method, can faithfully reconstruct the 3D shape of a static scene and of a person moving in it, but struggles when people manipulate objects. As illustrated, when a person pushes a table (top), the static part of the scene is reconstructed well (bottom row, two views), but the table's reconstruction is degenerate (see the red highlight). Here we do not show the reconstructed person to reduce occlusions.

However, reconstructed 3D shapes often lack physically-plausible contact; hands frequently hover above objects or visibly interpenetrate them. Our key insight is that neural signed distance fields (neural SDFs) not only encode geometry, but also provide continuous, differentiable distances to object surfaces—an ideal signal for reasoning about contact. We therefore repurpose per-component neural SDFs to formulate a contact-aware loss that simultaneously attracts surfaces and penalizes inter-penetrations, improving physical plausibility even under occlusions. We show that using neural SDFs to jointly encode shape and reason about contact is effective for monocular HOI reconstruction.

For evaluation, we capture a new dataset of 7 sequences using a handheld camera synchronized with a 4D capture stage to obtain monocular RGB video paired with ground-truth 4D HOI geometry and camera motion. We evaluate on novel-view synthesis and 4D reconstruction; RHINO outperforms state-of-the-art (SotA) baselines. Ablations show that each of RHINO's three stages contributes significantly.

Overall, our main contributions are the following:

1. We address the problem of reconstructing a 4D HOI scene in a world frame from monocular RGB video, a setting not tackled by prior work.
2. We use a compositional neural field framework with per-component appearance fields and SDFs; the latter encode both shape and human-object contact cues.
3. We use 3D-aware foundation models to estimate object motion, while decoupling it from camera motion.
4. We collect a new evaluation dataset with a handheld, moving RGB camera paired with 4D ground truth.

Code and data are public at https://lxxue.github.io/RHINO.

2. Related Work

Object Pose Estimation. Template-based pose estimation methods [3, 11, 34, 38, 58] achieve strong performance by tracking a known object template, but the need for a pre-acquired template limits their applicability in the wild. Template-free methods [1, 31] avoid this assumption, yet current approaches still require depth input [63] or a static reference video to build an object model [22, 52], limiting their use in the wild. Furthermore, these approaches rely on traditional 2D feature-matching techniques [10, 51] that lack multi-view consistency and often produce unreliable correspondences. In contrast, we leverage 3D foundation models [32] for dense, geometrically consistent correspondences, enabling robust pose initialization without pre-scanned templates. We further reconstruct objects in 3D, and refine object poses using photometric cues, achieving accurate pose estimation and high-fidelity reconstruction.

3D Human-Object Reconstruction. Most work on 3D HOI relies on object templates that are known [18, 30, 40, 64, 65, 67, 77] or retrieved from databases [8, 13, 73], limiting applicability. ProciGen [66], trained on a large synthetic dataset, enables template-free HOI reconstruction. InterTrack [67] extends this to coherently track HOI in 3D over videos. Most existing methods are evaluated on datasets such as BEHAVE [2] and InterCap [24], focusing on minimally clothed bodies—neglecting detailed surface geometry that serves as a critical contact cue (*e.g.*, shoes). Recent methods [15, 26, 71, 72] reconstruct detailed interacting humans and objects with separable representations, but rely on multi-view input, which is rarely available in practice. Hand-only tracking and contact cues have been used to reconstruct unseen objects [43, 56], but interactions are simple. Recently, HOLD [14] attains detailed hand–object reconstructions from a single video, yet it does not tackle full-body interactions. Unlike CHORE and InterTrack, which operate in the camera frame, RHINO reconstructs full-body human–object interactions in a world frame, including the scene, directly from monocular RGB video—without requiring prior object knowledge.

3D Human-Scene Reconstruction. Existing methods recover 3D human skeletons [5, 19, 37, 49] or meshes [9, 20, 21, 23, 28, 33, 35, 36, 39, 70, 75] within static scenes. However, these methods typically rely on pre-scanned scene geometry. Recent feedforward models [6, 47] relax this dependency by inferring camera and human parameters from images, though at reduced accuracy and without explicit human modeling. Several works [50, 68, 69, 76] seek joint human-scene reconstruction from monocular inputs. HSR [68] reconstructs humans within static scenes but cannot handle moving objects. Going further, RHINO tackles the challenging task of additionally reconstructing the dynamically manipulated object—whose shape, pose, and motion are all unknown—in the same world frame.

3. Method

We consider an input video containing a human interacting with an object, captured with a single, moving RGB camera. Our goal is to recover the detailed 3D shapes and appearances of a human, a manipulated object, and their surrounding environment in a common world frame.

We build a three-stage framework (Figs. 3 and 4). First, we estimate an initial scene, object, and human (Sec. 3.1), and align them into a common world frame (Sec. 3.2). Then, we recover details via compositional per-component neural SDFs (Sec. 3.3). Lastly, we use these neural SDFs to encourage contact and avoid inter-penetration (Sec. 3.4).

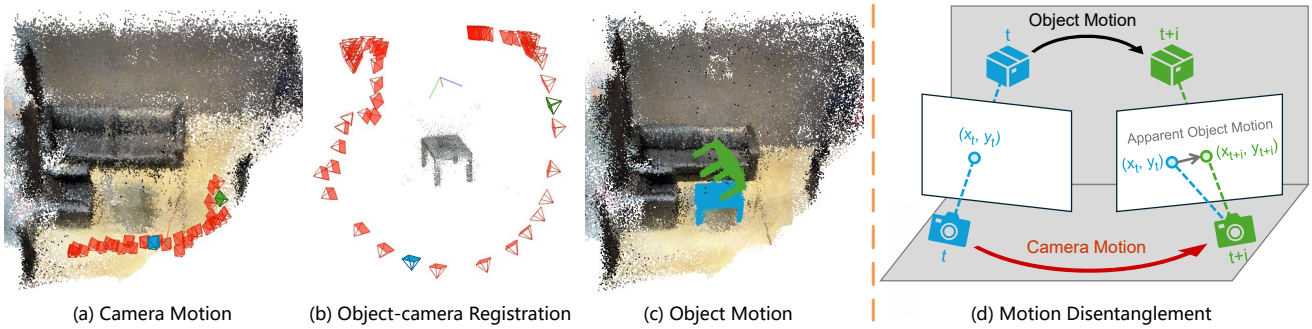
3.1. Initialization

We estimate coarse shape for the static scene, and coarse shape and motion for the moving object and human in the camera frame. We also estimate camera motion. This lets us later align the human, object, and scene into a common world frame (Sec. 3.2), and jointly refine these (Sec. 3.3).

Camera Motion & Scene Initialization. We estimate the camera motion under the assumption that the scene background is static and defined in the world frame. Consequently, any apparent motion between the scene and the camera is caused only by the camera motion. However, the input video also contains the dynamic human and object. To isolate the static background, we exploit SAM2 [46] to segment the video into a scene-only video, masking out the human and the object. Based on these background pixels, we perform SfM [32, 42] and estimate camera motion for all frames $C_{\text{scn}} = \{C^i_{\text{scn}}\}_{i=1}^{N}$ where C^i_{scn} is the camera pose (extrinsics) for frame i and N is the video length, and a rough 3D scene point cloud, PCL_{scn}, as shown in Fig. 3 (a).

Object Pose Initialization. We segment the moving object from the video and establish feature correspondences across adjacent frames on the object pixels. However, the object usually occupies a small area in images, and it might be occluded, low-texture, or symmetric. This makes standard keypoint detectors (*e.g.*, SuperPoint [10, 14]) or keypoint-free feature matching (*e.g.*, LoFTR [22, 51]) yield sparse or non-distinctive matches, which challenges Structure-from-Motion (SfM). We empirically find that recent 3D-aware foundational models remain surprisingly effective. Thus, we adopt MASt3R [32] to establish correspondences across neighboring frames. MASt3R casts this as a 3D task based on pointmap regression rather than a 2D problem in image space, resulting in feature correspondences that are more accurate and robust, even in low-texture regions. Based on these reliable matches, we perform triangulation and obtain a composed camera trajectory, C_{obj}, as if the object was static, as shown in Fig. 3 (b).

Human Initialization. We infer SMPL-X [44] bodies in the camera frame via the AiOS [53] model applied on a per-frame basis. These bodies have a reasonable but rough



(a) Camera Motion (b) Object-camera Registration (c) Object Motion (d) Motion Disentanglement

Figure 3. **Camera & Object motion (Sec. 3.1, Sec. 3.2).** When both the camera and object move, their motion entangles into "apparent" motion. To disentangle them, we (a) estimate camera motion in the world frame via SfM on scene-only pixels, (b) estimate apparent motion via SfM on object-only pixels, (c–d) estimate object motion in the world frame by "removing" the camera motion from the apparent one.

pose, so their pixel-alignment can be improved. Thus, we refine poses to align to Sapiens [29] 2D keypoints, while regularizing poses with a temporal smoothness loss, and interpolating keypoints in the case of occlusions by objects.

3.2. Aligning into World Frame

Section 3.1 defines two camera trajectories, C_{scn} and C_{obj}, in the scene (or world) and object frame, respectively. But in reality we have one camera, so we need to "unify" these.

Object to World. We consider as real the camera trajectory estimated from the static scene, $\mathbf{C}_{\text{scn}}$, so we need to align C_{obj} to C_{scn}. Let $\mathbf{S}$ be a scaling and $\mathbf{T} = [\mathbf{R}, \mathbf{t}; \mathbf{0}, 1]$ a rigid transformation, and $\mathbf{P}_{\text{obj}} = \{\mathbf{P}^i_{\text{obj}}\}_{i=1}^N$ be the object motion (sequence of poses) in the world frame. Then:

$$\mathbf{T} \cdot \mathbf{S} \cdot \mathbf{C}_{\text{obj}} = \mathbf{C}_{\text{scn}} \cdot \mathbf{P}_{\text{obj}}, \tag{1}$$

where we need to estimate $\mathbf{S}, \mathbf{T}$. We identify time frames i' where the object is static, using RANSAC to find a similarity transform between C_{obj} and C_{scn}; the frames that provide consensus are static-object frames. For these frames, $P^{i'}_{obj} = I$, so Eq. (1) simplifies to:

$$\mathbf{T} \cdot \mathbf{S} \cdot \mathbf{C}_{\text{obj}} = \mathbf{C}_{\text{scn}}, \tag{2}$$

which helps solve for scale, $\mathbf{S}$, rotation, $\mathbf{R}$, and translation, $\mathbf{t}$, through the Umeyama least-squares algorithm [57]:

$$\min_{\mathbf{s},\mathbf{R},\mathbf{t}} \sum_{i'=1}^{n} \|\mathbf{s}\mathbf{R}\mathbf{c}^i_{\text{obj}} + \mathbf{t} - \mathbf{c}^i_{\text{scn}}\|^2, \tag{3}$$

where $\mathbf{c}^i_{\text{obj}}, \mathbf{c}^i_{\text{scn}}$ are the camera centers of $\mathbf{C}_{\text{obj}}, \mathbf{C}_{\text{scn}}$. We then solve for object pose in the world frame (Fig. 3 (c–d)):

$$\mathbf{P}_{\text{obj}} = \mathbf{C}^{-1}_{\text{scn}} \cdot \mathbf{T} \cdot \mathbf{S} \cdot \mathbf{C}_{\text{obj}}, \tag{4}$$

by removing the real-camera motion, $\mathbf{C}_{\text{scn}}$, from the apparent motion, $\mathbf{C}_{\text{obj}}$, after morphing $\mathbf{C}_{\text{obj}}$ via Procrustes ($\mathbf{S}$, $\mathbf{T}$).

Human to World. The per-frame bodies of Sec. 3.1 live in the camera frame, under a weak-perspective assumption. We recover the body's trajectory in the world frame under a perspective camera by exploiting 2D projection constraints along with 3D contact constraints with a ground estimated by applying RANSAC on PCL_{scn}, as in [25, 68].

3.3. Joint Optimization

Given initial scene, object, and human estimates in a common world frame (Sec. 3.2), we refine these to recover details via a compositional neural field that has per-component neural Signed-Distance Fields (SDF) and appearance fields. Joint optimization in a shared world frame enforces multi-view consistency and mitigates initialization imperfections.

3D Shape Representation. Extending the modeling paradigm in [16, 68], we represent the shape of a human (H), object (O), and scene (S) as neural SDFs:

$$f^H_{\text{sdf}} : \mathbb{R}^{3+n_{\theta_b}} \to \mathbb{R}^{1+n_z};\ (\mathbf{x}^H, \boldsymbol{\theta}_b) \mapsto (\xi^H, \mathbf{z}^H), \tag{5}$$

$$f^O_{\text{sdf}} : \mathbb{R}^3 \to \mathbb{R}^{1+n_z};\ \mathbf{x}^O \mapsto (\xi^O, \mathbf{z}^O), \tag{6}$$

$$f^S_{\text{sdf}} : \mathbb{R}^3 \to \mathbb{R}^{1+n_z};\ \mathbf{x}^S \mapsto (\xi^S, \mathbf{z}^S). \tag{7}$$

where $f^{(\cdot)}_{\text{sdf}}$ maps a 3D point $\mathbf{x}^{(\cdot)}$ to a signed distance value, $\xi^{(\cdot)}$, and a geometric feature, $\mathbf{z}^{(\cdot)}$. To capture pose-dependent deformations (*e.g.*, clothing wrinkles), f^H_{sdf} conditions on body articulation $\boldsymbol{\theta}_b$, excluding global orientation and translation. The human and object fields operate in canonical space; for mapping to the world frame see below.

Appearance Representation. To model the appearance of a human (H), object (O), and scene (S) we employ three neural fields that predict RGB colors from 3D points:

$$f^H_{\text{rgb}} : \mathbb{R}^{3+n_{\theta_b}+n_z+3} \to \mathbb{R}^3;\ (\mathbf{x}^H, \boldsymbol{\theta}_b, \mathbf{z}^H, \mathbf{n}^H) \mapsto \mathbf{c}^H, \tag{8}$$

$$f^O_{\text{rgb}} : \mathbb{R}^{3+3+n_z+3} \to \mathbb{R}^3;\ (\mathbf{x}^O, \mathbf{v}, \mathbf{z}^O, \mathbf{n}^O) \mapsto \mathbf{c}^O, \tag{9}$$

$$f^S_{\text{rgb}} : \mathbb{R}^{3+3+n_z+3} \to \mathbb{R}^3;\ (\mathbf{x}^S, \mathbf{v}, \mathbf{z}^S, \mathbf{n}^S) \mapsto \mathbf{c}^S. \tag{10}$$

All color fields condition on shape features, $\mathbf{z}^{(\cdot)}$, and normals, $\mathbf{n}^{(\cdot)}$. The latter encourages disentanglement of shape and appearance, and is obtained by computing the gradient of the respective SDF. Note that f^H_{rgb} conditions on $\boldsymbol{\theta}_b$, while f^O_{rgb} and f^S_{rgb} condition on the viewing direction, $\mathbf{v}$. For the object and the scene model, we optimize a per-frame latent code to capture effects like shadows and highlights.

Mapping Canonical to World Frame. We model the human and object in respective canonical, pose-independent

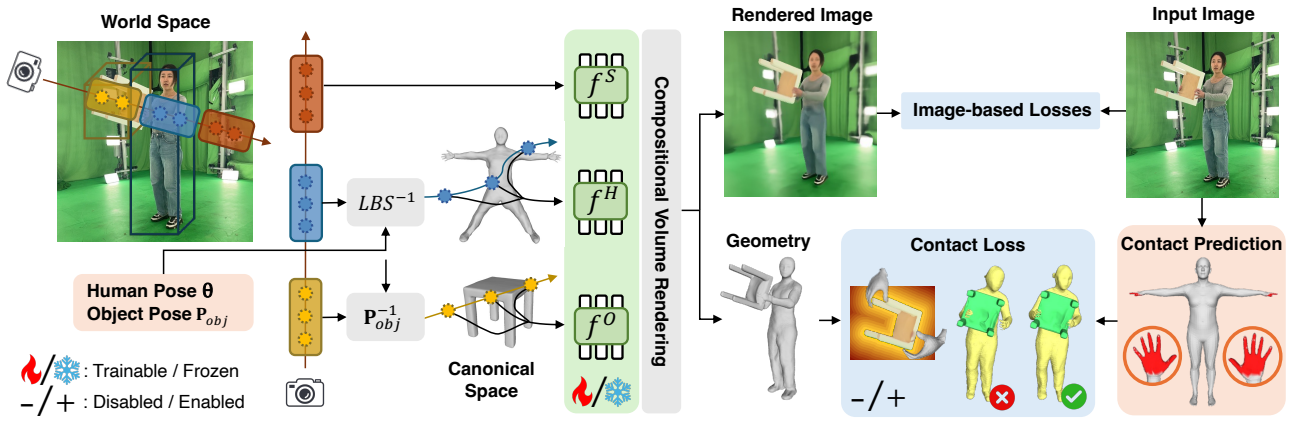


Figure 4. **Method Overview.** Starting with initialized global human and decoupled object poses (Sec. 3.1, Sec. 3.2), we sample points along the camera ray for the human, object and static scene. To enable consistent representation, sampled points are warped into canonical space using inverse LBS for the human and the estimated rigid transformation for the object. All components are then rendered holistically via compositional volume rendering. A global optimization (Sec. 3.3) helps learn the 3D representation of all elements and refine the initial poses via a photometric loss, while encouraging physically plausible contact by leveraging differentiable contact priors (Sec. 3.4).

frames and find maps from these to the world frame. For the human field, we use skeletal deformations and linear blend skinning (LBS) with bone transformations extracted from $\boldsymbol{\theta}$ as in [16, 68]. To do so: $\mathbf{x}'^H = LBS(\mathbf{x}^H, \boldsymbol{\theta})$ and $\mathbf{x}^H = LBS^{-1}(\mathbf{x}'^H, \boldsymbol{\theta})$, where $\mathbf{x}'$ lives in the world space. For the object, we simply apply the object's pose to move between the spaces: $\mathbf{x}'^O = \mathbf{P}_{obj}\mathbf{x}^O$ and $\mathbf{x}^O = \mathbf{P}_{obj}^{-1}\mathbf{x}'^O$.

Compositional Volume Rendering. We perform a compositional volume rendering for each camera ray, r, to render an image and model occlusions between scene elements. Specifically, we sample N 3D points within the bounding box of each component in the world frame, and then sort based on their distance to the camera. The color value is:

$$C(r) = \sum_{i=1}^{3N} \tau_i \mathbf{c}^{(\cdot)}(\mathbf{x}^i), \tag{11}$$

where τ_i is an opacity value defined in [68]. Similarly, we can render depth, surface normals, and masks.

Joint Optimization. We train the human, scene, and object neural fields jointly to recover details, via global optimization over all K frames using a per-pixel RGB loss and auxiliary losses (mask, depth, and normal losses) as in [68].

In our work, we model close human-object interactions, which often feature severe mutual occlusions (so reconstructed bodies seem truncated), and require detailed hand reconstruction. To address this, we sample points within SMPL-X [44] human bodies (with negative SDF values) and penalize for points outside these (with positive SDF). Due to hand dexterity and inaccurate hand pose estimation, naive reconstruction often leads to clumpy hand geometry. We resolve this via a hand-specific SDF loss that is guided by the SMPL-X body mesh. For details, see Supp. Mat.

3.4. Improving Physical Plausibility

The recovered scene, object, and human align well to pixels (Sec. 3.3), but may be misaligned w.r.t. each other due to depth ambiguity; *e.g.*, hands might hover over or penetrate an object. We tackle this via contact and collision losses with a two-stage alternating refinement framework.

Contact Estimation. We leverage the recent image-based InteractVLM [13] model to estimate 3D body contact points from each frame. However, this sometimes yields false positives and temporally-inconsistent detections. We tackle this by exploiting object motion. Since objects move only under manipulation, any frame containing object motion is labeled as a "contact frame," suppressing false-positives on frames where no interaction occurs. Moreover, we apply a temporal filter on raw contact predictions to improve their temporal consistency; for details see Supp. Mat.

Then, we use the object's neural SDF of Eq. (6) to define a differentiable term $\xi_{x_c}^O = f_{sdf}^O(x_c)$ that attracts the body-contact points, x_c, onto the object. We also apply contact and collision losses for physical plausibility as in [55]:

$$\mathcal{L}_{\text{contact}} = \alpha_1 \tanh\left(\xi_{x_c}^O / \alpha_2\right)^2 \quad \text{if } \xi_{x_c}^O \geq 0, \text{and} \tag{12}$$

$$\mathcal{L}_{\text{collision}} = \beta_1 \tanh\left(\xi_{x_c}^O / \beta_2\right)^2 \quad \text{if } \xi_{x_c}^O < 0. \tag{13}$$

Pose Refinement via Contact. We use the physical losses of Eq. (12) and Eq. (13) in an optimization-based framework. However, applying these from the start harms object shape due to inter-penetrations. To tackle this, we take a two-stage approach. In the first stage, we optimize everything using the losses of Sec. 3.3 (RGB, mask, and shape cues); the physical losses are omitted. In the second stage, we freeze the shape networks (Eq. (5) – Eq. (7)) and appearance networks (Eq. (8) – Eq. (10)) and optimize only the human and object poses; all losses are used, including physical losses. We alternate between these two stages throughout the optimization; in this way the physical losses refine poses without corrupting the object's shape.

Method	Setup			Chamfer Distance [cm] ↓			Hausdorff Distance [cm] ↓			F1 Score @2 cm [%] ↑		
	H	O	S	H	O	H+O	H	O	H+O	H	O	H+O
HSR [68]	✓	✗	✓	2.69	—	—	22.28	—	—	55.17	—	—
HOLD [14]	✗	✓	✗	—	4.41	—	—	11.92	—	—	33.64	—
InterTrack [67]	✓	✓	✗	4.66	11.16	7.18	20.58	33.28	30.97	29.41	16.81	25.02
RHINO (Ours)	✓	✓	✓	**2.65**	**1.21**	**2.42**	**15.64**	**10.80**	**14.90**	**56.16**	**90.42**	**56.51**

Table 1. **Evaluation on shape reconstruction.** We evaluate on all BenchRHINO sequences, using standard metrics. Columns denote the human (H), object (O), or scene (S). The "Setup" columns indicate whether each model (row) estimates shape for the human, object, or scene.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
HSR [68]	22.65	0.791	0.246
HOLD [14]	17.92	0.646	0.513
RHINO (Ours)	**25.80**	**0.832**	**0.212**

Table 2. **Evaluation on novel-view synthesis.** We evaluate on all BenchRHINO sequences. Our method (RHINO) provides substantially better view synthesis quality, outperforming both baselines across all metrics.

4. Evaluation

First, we discuss the dataset and metrics. Next, we evaluate RHINO against SotA methods on shape reconstruction and novel-view synthesis. Last, we ablate our design choices.

4.1. Dataset – BenchRHINO

Most HOI datasets have been captured with static cameras [2, 24, 71, 72, 77]. No existing dataset captures full-body and object interactions with a moving camera, similarly to cameras of Internet videos. We fill this gap by capturing BenchRHINO, a new benchmark dataset captured with a hand-held camera moving within a volumetric 4D capture studio. This lets us capture monocular, moving-viewpoint RGB video, with frames paired with 3D ground truth (GT).

BenchRHINO: Setup & Statistics. We use a capture studio comprising 106 synchronized cameras (53 RGB and 53 IR cameras) at 12 MP resolution and 30 FPS. We capture 7 sequences, featuring 4 subjects manipulating 6 objects.

Shape GT. Shape is reconstructed from the raw images via a commercial software [7]. To separate this shape into a human mesh and object mesh, we first detect respective masks (paired with labels) in RGB images via SAM2 [46], and then back-project masks from all views while applying majority voting to factor-out noise in SAM2 mask labels.

Camera Motion GT. We use an iPhone to capture video with a moving viewpoint. To recover its GT motion (sequence of 6D poses), we take two steps. First, we attach an AprilTag [41] to get a rough initial pose trajectory within the capture stage. Then, we utilize the mask loss between the projected mesh mask and the segmentation masks obtained with SAM2 [46], similar to the MultiPly method [27].

WildRHINO: In-the-wild videos. Since our annotated dataset is captured in lab settings, we also capture videos in natural indoor environments for qualitative evaluation. We capture 5 sequences of 3 subjects manipulating 4 objects.

4.2. Evaluation Protocol

Baselines. No existing method can reconstruct detailed human-object interactions in a world frame from monocular RGB videos. We compare to the following closest related work: HOLD [14] reconstructs hand-object interactions only in the camera frame, and struggles with textureless objects and rapid pose changes. HSR [68] reconstructs a human and a scene in a world frame, but cannot handle dynamic human-object interactions. InterTrack [67] predicts sparse human and object point clouds from an RGB video, but struggles generalizing for previously unseen objects, and cannot produce a detailed reconstruction.

Evaluation Metrics. Shape reconstruction is evaluated via Chamfer distance (CD), Hausdorff distance (HD), and F1 score at 2 cm. Baselines perform reconstruction in the camera frame, so we perform ICP to align their results to the GT mesh before computing the metrics. Novel-view synthesis is evaluated via PSNR, SSIM [61], and LPIPS [74].

4.3. Task 1: Shape Reconstruction

We compare to HSR [68], HOLD [14], and InterTrack [67] on shape reconstruction on the BenchRHINO dataset. As shown in Tab. 1, our method clearly outperforms all baselines [14, 67, 68]. The difference becomes more evident in the qualitative comparisons presented in Figs. 5 and 6. HOLD [14] struggles reconstructing good shape when the object pose is noisy (Fig. 5, 2nd row and 4th row). Equipping HOLD with accurate object pose estimation leads to significantly better object reconstruction quality in Fig. 6. However, it still struggles modeling hand-object interaction, as the reconstructed shapes look aligned to image cues, but are misaligned w.r.t. each other in 3D space. InterTrack [67] reconstructs reasonable shape for objects similar to its training data (*e.g.*, chair, suitcase), but struggles modeling the relative spatial configuration of the human and object. For out-of-distribution objects, InterTrack struggles significantly more, as shown in Fig. 6. In contrast, our method robustly reconstructs unseen objects and HOIs that closely resemble the ground truth.

4.4. Task 2: Novel–View Synthesis

We compare to HSR [68] and HOLD [14] quantitatively on novel-view synthesis on BenchRHINO, even though this task is not our primary goal. InterTrack [67] is excluded from this comparison as it only outputs point clouds and does not model appearance. As shown in Tab. 2, our method produces better rendering quality than the baselines. For visual comparisons, please see Supp. Mat.

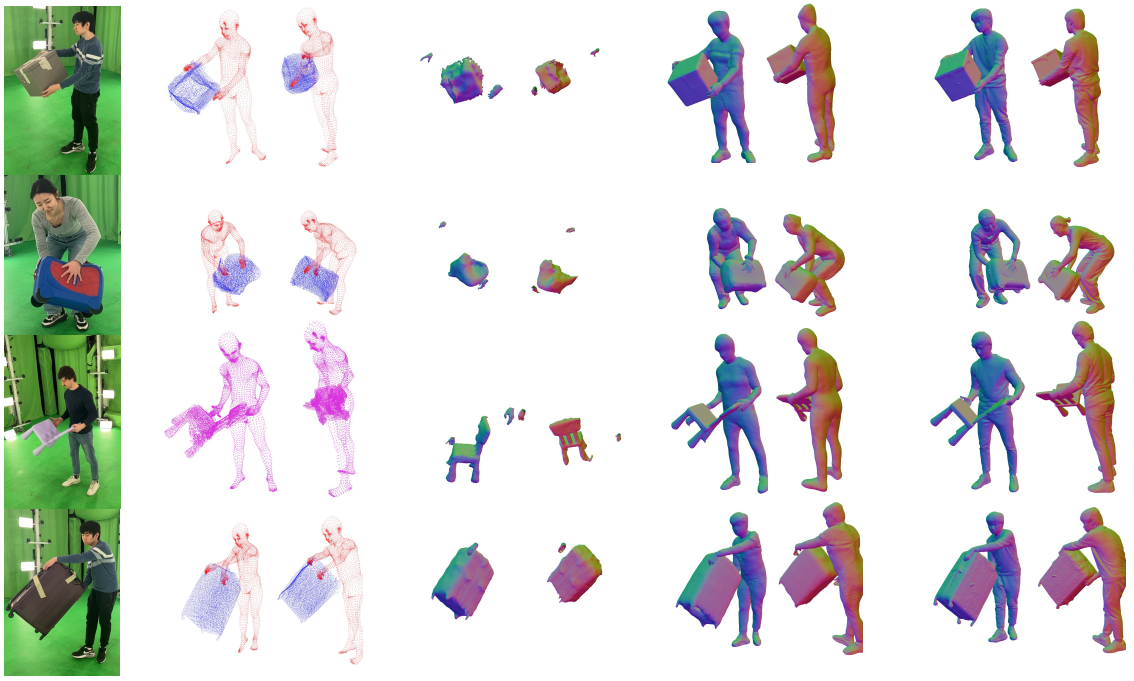


Figure 5. **Evaluation on shape reconstruction (Sec. 4.3)** on our BenchRHINO dataset (Sec. 4.1). HOLD [14] struggles with noisy object poses (rows 2, 4) and fails to model interaction. InterTrack [67] recovers reasonable object shape but fails to model the interaction due to large human and object pose errors. Our method (RHINO) faithfully recovers interactions, which lie closer to the ground truth (GT).

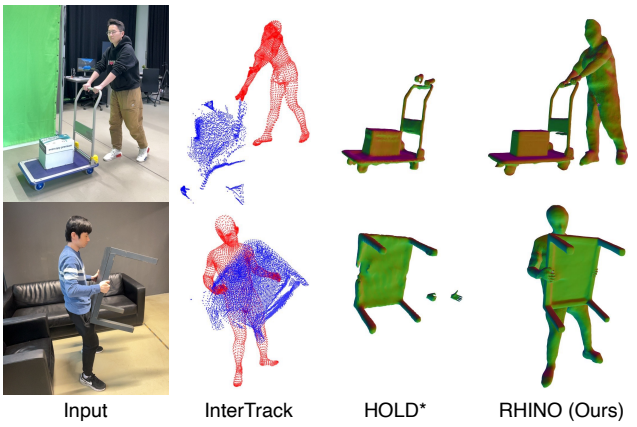


Figure 6. **Evaluation on shape reconstruction** on WildRHINO. InterTrack [67] fails on these out-of-distribution objects. While HOLD* (vanilla HOLD [14] using our method's object poses) reconstructs good object shape, it fails to model interactions. RHINO yields reasonable reconstruction reflecting the interaction.

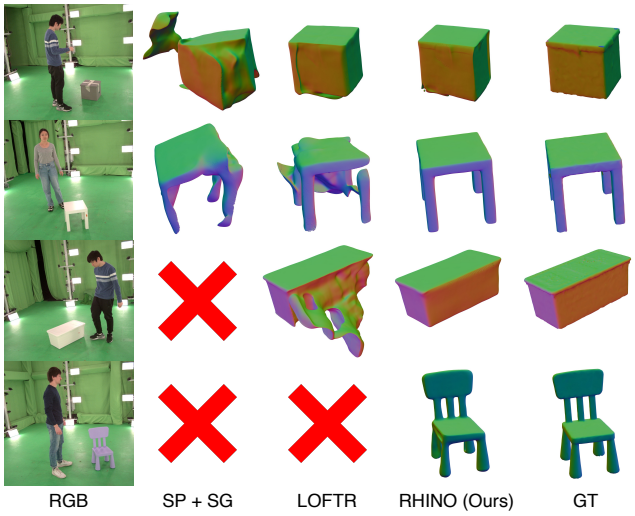


Figure 7. **Evaluation on object pose estimation (Sec. 4.5).** We compare to SP+SG, a HOLD-inspired [14] baseline that uses SuperPoints [10] and SuperGlue [48], and one that uses LoFTR [51]. × denotes failed reconstruction due to wrong object poses.

4.5. Ablation Study

Object Pose Estimation. We compare our object pose estimation, which uses MASt3R [32] features, against traditional pipelines. These include the SuperPoint [10] + SuperGlue [48] pipeline ("SP+SG"), as used in HOLD [14] and OnePose [52], a strong 2D feature matching baseline. We also compare against the dense feature LoFTR [51], used in BundleSDF [62] and OnePose++ [22]. As shown in Fig. 7 and Tab. 3, object poses estimated with our method enable significantly more robust and accurate object reconstruction. The qualitative results in Fig. 7 visually explain this performance gap. The SP+SG pipeline tends to fail for large textureless objects (Fig. 7, 3rd row) due to

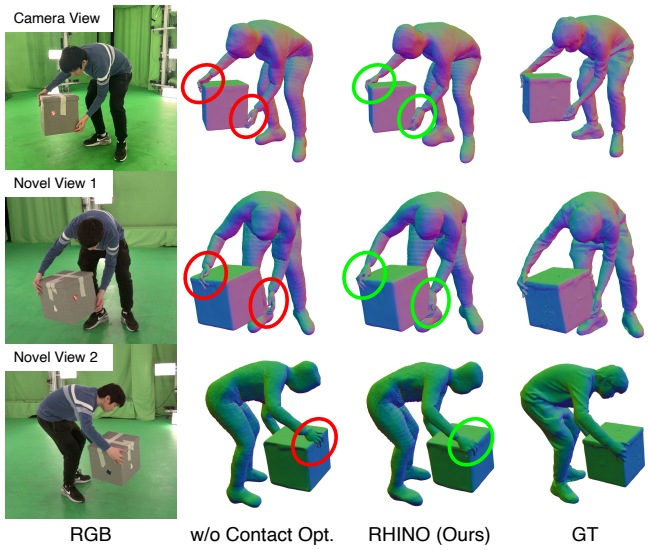


Figure 8. **Effects of contact.** We show reconstructions of our framework with ("`RHINO (Ours)`") and without ("`w/o Contact Opt`") the physical losses of Eq. (12) and Eq. (13). For a bigger version with zoom-in impressions, see Supp. Mat.

non-repeatable keypoint detection, leading to degenerate reconstruction. Similarly, LoFTR [51] is not accurate enough to find correspondences for objects under complex motions (Fig. 7, 4th row). In contrast, our method proves robust in these challenging scenarios. More visual comparisons on feature matching are shown in the video on our website.

Motion Disentanglement (MD). To evaluate our camera–object motion disentanglement (Sec. 3.2), we compare against a variant that uses apparent object motion without removing the camera component. As shown in Tab. 4, removing MD ("`w/o MD`") drastically worsens all metrics, showing that MD is crucial for world-frame reconstruction.

Contact Refinement. To investigate the importance of our contact-based pose refinement, we compare our full model ("`RHINO (Ours)`") to a version without this step ("`w/o Contact Opt`"), and show results in Fig. 8. As discussed in Sec. 3.4, while initial reconstructions may align well with pixels (Fig. 8, 1st row), they often suffer physical implausibility due to depth ambiguity and occlusions, causing hands to hover over or penetrate an object. Figure 8 clearly shows these artifacts in novel views (red circles of the 2nd and 3rd row). On the other hand, our full RHINO model leverages contact priors and produces reconstructions in which the hands establish firm, realistic contact with the object, closely matching the ground truth. We quantitatively evaluate in Tab. 5 by reporting penetration depth (PD), contact precision, recall, and F1. Contact refinement halves the PD and nearly triples the recall, which shows its importance for physical plausibility.

	CD [cm] ↓	HD [cm] ↓	F1 Score [%] ↑
SP [10] + SG [48]	4.25	25.43	60.06
LoFTR [51]	3.97	20.19	62.80
RHINO (Ours)	**1.09**	**10.94**	**91.38**

Table 3. **Evaluation on object pose estimation (Sec. 4.5).** We compare to `SP+SG`, a HOLD-inspired [14] baseline that uses SuperPoint [10] and SuperGlue [48], and one that uses LoFTR [51]. Results are reported for BenchRHINO sequences for which all baselines do not fail; see the list of sequences in Supp. Mat.

Method	CD [cm] ↓	F1@2cm [%] ↑	PSNR ↑	LPIPS ↓
w/o MD	10.21	26.32	22.89	0.306
RHINO (Ours)	**2.65**	**56.16**	**25.80**	**0.212**

Table 4. **Ablation on motion disentanglement (MD) (Sec. 4.5).** Removing motion disentanglement leads to a large drop across all metrics, confirming it is essential for world-frame reconstruction.

Method	PD [cm] ↓	Precision [%] ↑	Recall [%] ↑	F1 [%] ↑
w/o Contact Opt.	1.088	24.06	18.39	20.84
RHINO (Ours)	**0.477**	**25.67**	**63.57**	**36.57**

Table 5. **Ablation on contact refinement (Sec. 4.5).** Refining pose via contact reduces penetration depth (PD) and increases contact recall and F1, showing its importance for physical plausibility.

5. Conclusion

In this paper, we present RHINO, a novel framework that reconstructs 4D human-object interaction scenes from a single monocular video with a moving viewpoint. First, we develop a robust 3D-aware methodology that estimates an initial object and scene shape, human motion, and disentangled object and camera motion. Then, we refine these using a compositional neural field with per-component SDFs. This not only captures shape details, but also enables defining a contact loss with differentiable SDF-based distances, ensuring physically-plausible interactions. For evaluation, we capture BenchRHINO, a new video dataset of human-object interactions with 4D ground truth. RHINO is able to reconstruct high-quality human-object interactions in challenging scenarios such as occlusions and complex motions.

Future Work. Our current framework is designed for interactions involving a single person and a single rigid object. Extending it to handle more complex scenes with multiple interacting humans or objects, remains a significant challenge. Moreover, our framework assumes object rigidity. A valuable direction for future work would be to incorporate the reconstruction of non-rigid objects, such as articulated objects, to capture a wider range of real-world interactions. Finally, our current optimization process is slow. Speeding it up to enable fast 4D capture would be crucial for applications in AR/VR and robotics. We also see potential for improving robustness to sparse observations, as our method currently performs best with good object coverage.

Acknowledgements

We thank all participants of the captured dataset. The ETH part of the team was partly supported by the Swiss SERI Consolidation Grant AI-PERCEIVE. Manuel Kaufmann was partly supported by ETH AI center. Compute was partly performed on the ETH Zurich Euler cluster The UvA part of the team was partly supported by the ERC Starting Grant (project STRIPES, 101165317, PI: D. Tzionas), by a research gift from Google, by the NVIDIA Academic Grant Program, and by EuroHPC JU via access to the supercomputers LEONARDO (project ID EHPC-AI-2024A06-077), hosted by CINECA in Italy, and JUPITER (project ID e-reg-2025r02-393), hosted by JSC in Germany.

References

[1] Dimitrije Antić, Georgios Paschalidis, Shashank Tripathi, Theo Gevers, Sai Kumar Dwivedi, and Dimitrios Tzionas. SDFit: 3D object pose and shape by fitting a morphable SDF to a single image. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6

[3] Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. FreeZe: Training-free zero-shot 6D pose estimation with geometric and vision foundation models. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[4] Weirong Chen, Ganlin Zhang, Felix Wimbauer, Rui Wang, Nikita Araslanov, Andrea Vedaldi, and Daniel Cremers. Back on track: Bundle adjustment for dynamic scene reconstruction. In *International Conference on Computer Vision (ICCV)*, 2025. 2

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

[6] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Pons-Moll Gerard. Human3R: Everyone everywhere all at once. arXiv:2510.06219, 2025. 3

[7] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *Transactions on Graphics (TOG)*, 34(4), 2015. 6

[8] Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[9] Yudi Dai, Yitai Lin, Chenglu Wen, Siqi Shen, Lan Xu, Jingyi Yu, Yuexin Ma, and Cheng Wang. HSC4D: Human-centered 4D scene capture in large-scale indoor-outdoor space using wearable imus and lidar. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2018. 2, 3, 7, 8

[11] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. SO-Pose: Exploiting self-occlusion for direct 6D pose estimation. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[12] Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J. Black, and Dimitrios Tzionas. POCO: 3D pose and shape estimation using confidence. In *International Conference on 3D Vision (3DV)*, 2024. 2

[13] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 5

[14] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6, 7, 8

[15] Suhas Gopal, Rishabh Dabral, Vladislav Golyanik, and Christian Theobalt. Betsu-Betsu: Multi-view separable 3D reconstruction of two interacting objects. In *International Conference on 3D Vision (3DV)*, 2025. 3

[16] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 5

[17] Chen Guo, Junxuan Li, Yash Kant, Yaser Sheikh, Shunsuke Saito, and Chen Cao. Vid2Avatar-Pro: Authentic avatar from videos in the wild via universal prior. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[18] Zichao Guo, Xiangyu Zhang, Haoyuan Mu, Wen Heng, Zechun Liu, Yichen Wei, and Jian Sun. Single path one-shot neural architecture search with uniform sampling. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[19] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3D scene geometry to human workspace. In *Computer Vision and Pattern Recognition (CVPR)*, 2011. 3

[20] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3D human pose estimation and self-localization in large scenes from body-mounted sensors. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3

[21] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[22] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-free one-shot object pose estimation without CAD models. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 7

[23] Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel

Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[24] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. *International Journal of Computer Vision (IJCV)*, 2024. 2, 3, 6

[25] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *European Conference on Computer Vision (ECCV)*, 2022. 4

[26] Yuheng Jiang, Suyi Jiang, Guoxing Sun, Zhuo Su, Kaiwen Guo, Minye Wu, Jingyi Yu, and Lan Xu. NeuralHOFusion: Neural volumetric rendering under human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[27] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. MultiPly: Reconstruction of multiple people from monocular video in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[28] Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tang, Juan José Zárate, and Otmar Hilliges. EMDB: The electromagnetic database of global 3D human pose and shape in the wild. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[29] Rawal Khirodkar, Timur M. Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *European Conference on Computer Vision (ECCV)*, 2024. 4

[30] Hedvig Kjellström, Danica Kragic, and Michael J. Black. Tracking people interacting with objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2010. 3

[31] Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6D: Model-free 6D pose estimation of novel objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[32] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 7

[33] Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. MoCapDeform: Monocular 3D human motion capture in deformable scenes. In *International Conference on 3D Vision (3DV)*, 2022. 3

[34] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment anything model meets zero-shot 6D object pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[35] Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4D human-scene reconstruction in the wild. arXiv:2501.02158, 2025. 3

[36] Diogo C. Luvizon, Marc Habermann, Vladislav Golyanik, Adam Kortylewski, and Christian Theobalt. Scene-aware 3D multi-human motion capture from a single camera. In *Computer Graphics Forum (CGF)*, 2023. 3

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

[38] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. GenFlow: Generalizable recurrent flow for 6D pose refinement of novel objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[39] Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[40] Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3D human and object via contact-based refinement transformer. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[41] Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *International Conference on Robotics and Automation (ICRA)*, 2011. 6

[42] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes Lutz Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[43] Pantelis Panteleris, Nikolaos Kyriazis, and Antonis A. Argyros. 3D tracking of human hands in interaction with unknown objects. In *British Machine Vision Conference (BMVC)*, 2015. 3

[44] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5

[45] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[46] 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, and Ross Girshick et. el. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. 3, 6

[47] Sara Rojas, Matthieu Armando, Bernard Ghamen, Philippe Weinzaepfel, Vincent Leroy, and Gregory Rogez. HAMSt3R: Human-aware multi-view stereo 3D reconstruction. In *International Conference on Computer Vision (ICCV)*, 2025. 3

[48] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 8

[49] Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning interaction snapshots from observations. 35(4):139:1–139:12, 2016. 3

[50] Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. Total-Recon: Deformable scene

reconstruction for embodied view synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[51] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 7, 8

[52] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 7

[53] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-one-stage expressive human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[54] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[55] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4713–4725, 2023. 5

[56] Dimitrios Tzionas and Juergen Gall. 3D object reconstruction from hand-object interactions. In *International Conference on Computer Vision (ICCV)*, 2015. 3

[57] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1991. 4

[58] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[59] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

[60] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[61] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 6

[62] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[63] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

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

[65] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. CHORE: Contact, human and object reconstruction from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[66] Xianghui Xie, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Template free reconstruction of human-object interaction with procedural interaction generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[67] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)*, 2025. 3, 6, 7

[68] Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Hilliges Otmar. HSR: Holistic 3D human-scene reconstruction from monocular videos. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 5, 6

[69] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. PPR: Physically plausible reconstruction from monocular videos. In *International Conference on Computer Vision (ICCV)*, 2023. 3

[70] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[71] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. NeuralDome: A neural modeling pipeline on multi-view human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 6

[72] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M3: Capture multiple humans and objects interaction within contextual environment. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6

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

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[75] Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. EgoBody: Human body shape and motion of interacting people from head-mounted devices. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[76] Zetong Zhang, Manuel Kaufmann, Lixin Xue, Jie Song, and Martin R. Oswald. ODHSR: Online dense 3D reconstruction of humans and scenes from monocular videos. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3

[77] Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'M HOI: Inertia-aware monocular capture of 3D human-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 6