ROMEO: Revisiting Optimization Methods for Reconstructing 3D Human-Object Interaction Models From Images

Alexey Gavryushin[1], Yifei Liu[2], Daoji Huang[1], Yen-Ling Kuo[3], Julien Valentin[4], Luc van Gool[1,5,6], Otmar Hilliges[1], and Xi Wang[1]

[1] ETH Zurich, Raemistrasse 101, 8092 Zurich, Switzerland
[2] Univ. of Zurich, Raemistrasse 71, 8006 Zurich, Switzerland
[3] Univ. of Virginia, 1827 Univ. Avenue, Charlottesville, VA 22903, USA
[4] Microsoft Research, Talstrasse 9, 8001 Zurich, Switzerland
[5] KU Leuven, Oude Markt 13, 3000 Leuven, Belgium
[6] INSAIT, 111R Tsarigradsko Shose Blvd., 1784 Sofia, Bulgaria

Abstract. We present ROMEO, a method for reconstructing 3D human-object interaction models from images. Depth-size ambiguities caused by unknown object and human sizes make the joint reconstruction of humans and objects into a plausible configuration matching the observed image a difficult task. Data-driven methods struggle with reconstructing 3D human-object interaction models when it comes to unseen object categories or object shapes, due to the difficulty of obtaining sufficient and diverse 3D training data, and often even of acquiring object meshes for training. To address these challenges, we propose ROMEO, a novel method that does not require any manual human-object contact annotations or 3D data supervision. ROMEO integrates the flexibility of optimization-based methods and the effectiveness of foundation models with large modeling capacity in a plug-and-play fashion. It further incorporates a novel depth-based loss term and largely simplifies the optimization objective of previous methods, eliminating the requirement for manual annotations of contacts and object scales and rendering object-category–specific parameter finetuning unnecessary. We quantify the improvement of ROMEO over existing state-of-the-art methods on two human-object interaction datasets, BEHAVE and InterCap, both quantitatively and qualitatively. We further demonstrate the generalization ability of ROMEO on in-the-wild images.

1 Introduction

Accurate modeling of human-object interactions is an important task, as humans constantly engage with objects in their immediate environment. Gaining insights into human-object interactions provides numerous benefits, including interaction modeling in virtual reality, object-centric learning for embodied agents executing daily tasks, and enhanced human-robot interactions within the 3D world. The prospect of using an approach capable of reconstructing 3D models from commonly available 2D images holds significant promise, as it offers an accessible

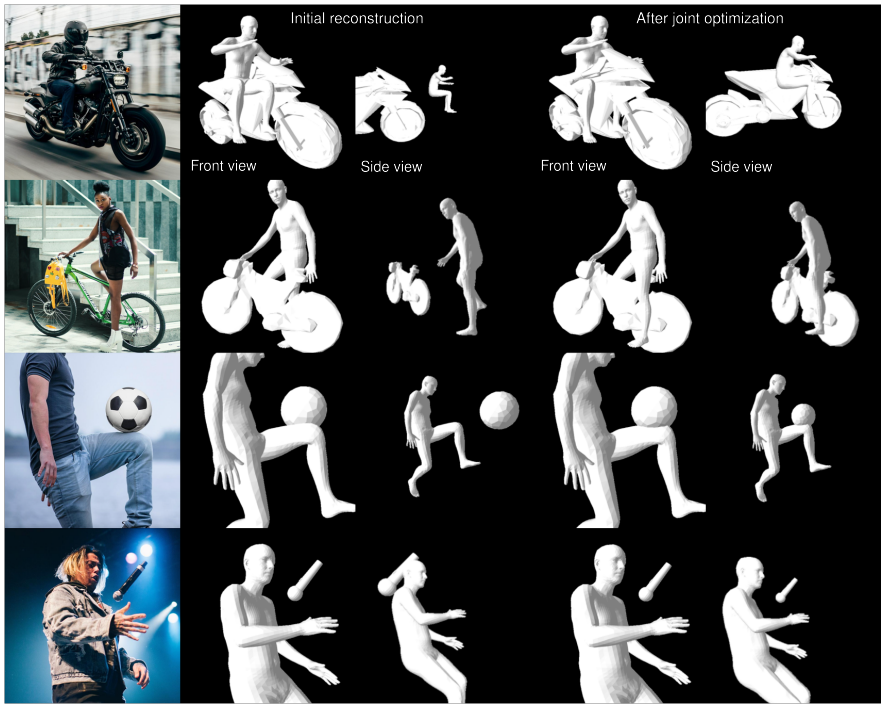


Fig. 1: Reconstruction of 3D human-object interaction models from in-the-wild images. We propose ROMEO, a method that reconstructs 3D interaction models from images using a simplified yet effective optimization objective using estimated depth maps. Our method does not require any human-object contact annotations or prior knowledge of object scales or object categories. We visualize input images, initial scene reconstructions showing strong depth-size ambiguities when viewed from different viewpoints, and improved final reconstructions after our joint optimization step reasoning about the human-object relationships using depth estimation.

and generalizable means to obtain authentic and diverse 3D models of humans interacting with objects.

Reconstructing 3D human-object interaction models from generic images is a challenging task. Merely reconstructing 3D models from 2D images is already difficult, demanding accurate detection and segmentation, all without prior knowledge about the scene and in the presence of object occlusions. Interaction modeling faces a bigger challenge: it requires not only identifying which concrete object in the scene a human is interacting with but also reconstructing detailed 3D humans and objects in a manner consistent with the image observation and conforming to plausible configurations. Examples of implausible configurations include a person suspended in mid-air or a flying basketball located at an unnatural angle relative to the thrower's hand. The task is further complicated by natural occlusions resulting from interactions, such as a human sitting on a chair, alongside depth ambiguities arising from only observing the scene from a single perspective. Lastly, diverse object orientations emerging from different

object affordances [5], for instance *carrying* a keyboard as opposed to *typing* on one, greatly expand the object pose search space.

Recent advancements in approaching this task involve leveraging 3D data, obtained via multiview capturing systems [2,8,9,34], to develop learning methods for reconstructing 3D human-object interaction models. Nonetheless, collecting high-fidelity 3D data typically requires specialized hardware systems, and extensive effort. The effectiveness of current learning models [9, 17, 26–28, 34] is hindered by both the scarcity of available 3D data and their limited ability to generalize to out-of-distribution scenarios.

On the other hand, existing image-based interaction modeling approaches [30, 33] rely on off-the-shelf 3D human reconstruction models [15, 32] and shape reconstruction methods from images [3, 4, 11] for initialization, and jointly optimize the obtained 3D humans and objects to leverage interaction constraints and reduce ambiguity. However, one of the inherent challenges faced by these image-based reconstruction methods is their difficulty in handling occlusions. Furthermore, many existing approaches encounter scalability issues, as they rely on human annotations to identify contact regions between bodies and objects. These annotated regions are then used in a joint optimization step to bring humans and objects in contact. This process of joint optimization often involves complex loss formulations, necessitating the tuning of weight parameters specific to each object category.

Meanwhile, the recent advancements in foundation models for various core computer vision tasks such as visual grounding [16], segmentation [13] and depth estimation [1,19], showcasing remarkable performance, have opened up new possibilities for human-object interaction modeling. Building upon this progress, we aim to revisit existing image-based optimization approaches and gain insights that can further guide the development of 3D interaction modeling techniques. Our primary objective is to investigate the extent to which 3D information can be extracted from images without relying on 3D data supervision.

To accomplish this, we make the following key contributions: First, we identify a major bottleneck in existing approaches, which lies in the reconstruction quality of objects following initialization. Previous approaches have overlooked the complexity associated with the initialization step, which encompasses both shape reconstruction and pose estimation. Each of these components is still an active research field in its own right. To address this challenge, we use foundation models in a plug-and-play fashion to improve both the segmentation quality and the detection rate of the objects being interacted with. Subsequently, we propose a new joint optimization step that only uses a silhouette loss [33] and a novel relative depth loss. Surprisingly, we find that many of the previously used 2D image-based and 3D geometry-based losses can potentially be substituted with a depth-based loss, in particular when the initialized 3D models are reasonably accurate. Our simplified approach focused on optimizing the depth ambiguity between humans and objects is highly effective, reducing the need for multiple complex loss terms and the fine-tuning of weight parameters specific to each object category [33].

Building upon our findings, we propose a revised optimization-based reconstruction approach that delivers significantly better reconstruction results while simplifying the optimization objectives. We conduct quantitative and qualitative evaluations on two datasets: BEHAVE [2], which consists of full-body human-object interaction frame sequences with ground-truth 3D human and object models, and InterCap [8], a large-scale interaction dataset featuring both object categories found in BEHAVE as well as novel objects with substantial differences. Our approach ROMEO demonstrates significant improvements over the state-of-the-art optimization-based method [33] and greatly reduces the distance to the state-of-the-art learning-based model [27] in terms of reconstruction quality. We further showcase the generalization ability of our approach on in-the-wild images, highlighting its robustness in real-world scenarios where training data is missing and a large variety of possible object categories and instances exist.

2 Related Work

Human-object interaction modeling. Efficiently modeling 3D human-object interactions (HOIs) requires reasoning about the human bodies and objects jointly instead of performing the reconstruction separately. In particular, several existing works show that human-object contact estimation plays a crucial role in interaction modeling. These methods explicitly consider the contacts between human bodies and objects when inferring their shapes and 3D spatial arrangements from a single RGB image. PHOSA [33] proposes an optimization-based pipeline that minimizes the distances between pairs of manually segmented human and object parts deemed likely to be in contact if the human and the object are close. HolisticMesh [26] jointly reconstructs human and object meshes by minimizing a set of human-object interaction losses concerning contacts/collisions and the human-ground distance. However, the requirement of manual annotations of contact regions [6, 23, 26, 33] makes it hard to generalize these methods to images in the wild as well as unusual object interactions, such as sitting on a table [2]. The BEHAVE [2] dataset provides a large multi-view RGB-D dataset with annotated human-object contacts. CHORE [27] builds a neural reconstruction model that learns to predict contact points from input images while reconstructing humans, objects, and their interactions. RICH [7] also learns 3D contacts, but merely predicts points on the human in contact with any unidentified point in the scene. VisTracker [28] and InterDiff [29] reason about human-object interactions across time rather than focusing on single-image reconstruction, exploiting temporal cues unavailable in our task setting to boost performance.

In addition to the learning-based approaches, Wang et al. [25] leverage large language models (LLMs) to generate estimations of contacts based on the recognized actions. Our approach also aims to remove the requirement of human annotation. Unlike previous methods, we propose to leverage the alignment of the estimated human and object depths as an informative prior for reasoning about 3D arrangements. This yields a principled and much simpler optimization objective and shows strong performance against the SOTA.

H-O interaction modeling with depth information. Depth ambiguities are one of the challenges for directly recovering 3D human-object interactions. Optimization based on predefined contact regions presents merely one way to resolve such ambiguities. Other approaches attempt to resolve depth ambiguities by including depth information in pipelines. NeuralDome [34] optimizes contacts based on the tracked humans and objects in multiviews. The multiview data implicitly contains the depth alignment of humans and objects, recoverable from different views. MOVER [30] makes use of depth constraints derived from human-scene occlusions and human movement trajectories. However, these approaches use multiple images, either multiview or from motion sequences, to construct depth information. We do not have access to such data when modeling interactions from single images. For a single image, the relative depths between humans and objects can be estimated using any off-the-shelf depth estimation model. We hypothesize this relative depth should be preserved in the rendered humans and objects when the reconstruction matches the single RGB input image. Our depth loss follows this intuition, and we demonstrate that it indeed presents a strong prior for human-object interaction modeling.
Generalizing H-O interaction to large object category domains. Many H-O interaction reconstruction approaches specialize in only a limited set of shapes or perform category-specific optimization. To improve generalizability across object categories, AutoSDF [18] and 3DILG [31] model the distribution over 3D shapes to generate multiple plausible outputs. Wang et al. [25] use LLMs to generalize contact labels to new categories. In this work, we leverage the Grounding DINO [16] detection model combined with Segment-Anything [12] to perform open-vocabulary instance segmentation. Together with the simplified loss, we succeed in performing reconstructions for open vocabulary categories.

3 Method

This section describes our approach to reconstructing a 3D representation that captures the interaction between the human and the object depicted in a given image significantly simpler than previous work. Further implementation details of our approach are provided in subsection 4.4.

3.1 Optimization Method Preliminaries

Existing optimization frameworks [25, 26, 33] leverage holistic context cues by considering the interactive relationship between humans and objects. These frameworks typically consist of a separate initialization stage followed by a joint human-object optimization stage. We follow the same two-stage pipeline in our approach (see Figure 2).
Stage I. First, humans and objects are independently reconstructed. For 3D human reconstruction, state-of-the-art methods such as PARE [15] and FrankMocap [10,21] are commonly used to initialize the 3D human meshes. These methods work reasonably well even under occlusion. The reconstructed meshes are

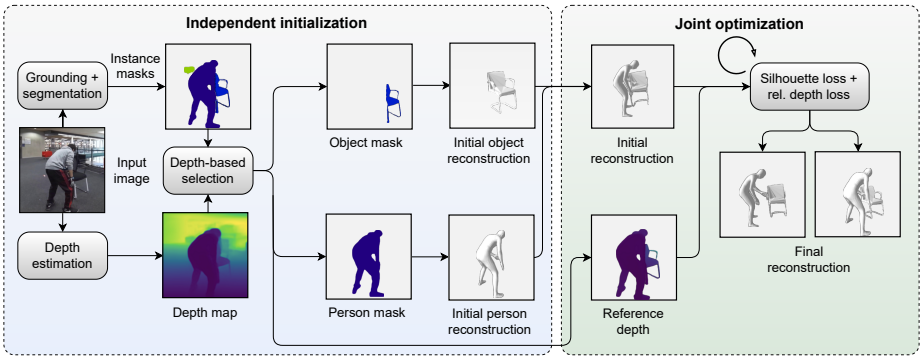


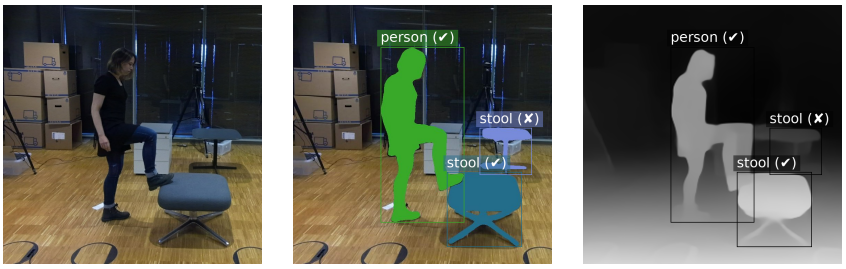
Fig. 2: Method overview. Our method reconstructs 3D human-object interaction models from images. In the initialization step, we reconstruct 3D humans and objects independently. An instance segmentation mask and a depth map, estimated by existing foundational models, are used to obtain better segmentation masks of the object of interest and subsequently improve its initial estimation. We then jointly reason about the 3D arrangement of both humans and objects using a novel combination of relative depth loss and a silhouette loss.

represented by the SMPL model $\mathcal{M}(\beta, \theta)$, where $\beta \in \mathbb{R}^{10}$ represents the shape parameters and $\theta \in \mathbb{R}^{3\times 24}$ represents the pose parameters for the 24 joints. On the other hand, reconstructing 3D objects poses a greater challenge due to the diversity of object categories and shapes within each category. Current solutions rely on template matching with observed instance segmentation masks and employ differentiable renderers [3, 22, 24] to estimate object poses. However, the object reconstruction performance is heavily dependent on the quality of the segmentation masks and is also hindered by occlusions from human interactions.

Stage II. The arrangements between 3D humans and objects are jointly estimated in this optimization step to match the image observations while obtaining plausible interaction models. The main idea of this step is to leverage the interactions between humans and objects, which are commonly formulated as 3D geometry-based losses such as contact loss and penetration loss [33] to ensure correct contacts between human and object while avoiding mesh interpenetration. Previous approaches of modeling human-object and human-scene interactions rely on manual annotations to establish contact between humans and objects [6,30,33]. In more recent work, text prompts are employed to extract contact labels from large language models [25], providing improved generalization abilities. However, this method still requires a known 3D object segmentation to apply the semantic contact information in the 3D space. Notably, such a joint optimization step is also used in learning-based methods such as CHORE [27] and VisTracker [28].

Challenges. While current state-of-the-art learning methods can estimate the distance fields and reconstruct 3D interaction models from images, they frequently fail to obtain satisfactory results without nearly perfect segmentation masks or when objects are occluded. Often, the models have limited general-

ization ability and struggle to handle out-of-distribution objects not seen in the training data. Optimization-based methods, on the other hand, do not require training data and show better generalizability to new object categories and shapes. Their central idea is to perform joint optimization by leveraging the interactions between humans and objects. However, existing approaches rely on manual annotations of contact regions for each object category. Such label collection is not only expensive but also fails to capture diverse interactions within each object category. In addition, joint optimization frequently requires category-specific parameter tuning to work with various object sizes and shapes. In this paper, we address these challenges and improve the performance of optimization methods over the current state of the art in terms of accuracy, robustness, and generalization ability while boosting computational speed.



(a) Input image **(b)** Detected instances **(c)** Estimated depth

Fig. 3: Depth-based object selection. Two instances of the "stool" class are detected (3b), and the estimated depth (3c) is used to select the object with the closest estimated proximity to the human for subsequent reconstruction.

3.2 Object Initialization

Reconstructing 3D shapes from images involves two challenging tasks: object shape reconstruction and pose estimation. These tasks become particularly challenging when objects are occluded by humans during interactions. We observe that the quality of the initial object reconstruction has a direct influence on the final reconstruction models. Surprisingly, this crucial aspect has been largely overlooked by previous approaches [25, 27, 33].

Object segmentation. We employ Grounding DINO [16] together with the Segment Anything Model (SAM) [12] to obtain the object segmentation masks in an open-vocabulary setting. Specifically, we run Grounding DINO on images using the category of the object currently being interacted with as a prompt, and for each obtained bounding box, we infer the object's mask through a mask prompt to SAM. The open-vocabulary approach made possible by using foundation models alleviates the need to adapt the segmentation pipeline to dataset-specific vocabularies, providing flexibility in handling diverse object categories.

Object selection. In order to reconstruct the interaction pair, we need to first identify the pair of a human and an object that is involved in the interaction.

Unlike prior works that rely on carefully engineered heuristics, e.g. overlaps between the detected humans and objects, we introduce a novel, simple depth-based selection rule. This rule aims to determine the object that exhibits the closest proximity to the detected human. Specifically, we use an off-the-shelf depth estimator [19, 20] and intersect its output with each previously obtained SAM mask to determine the average depth for each instance. We then select the object closest to the human in terms of depth. See Figure 3 for a visualization of the object selection procedure. Note that while in the following we focus on the interactions of the human that is closest to the camera based on the estimate, our framework still allows for reconstructing multiple interactions (see Figure 5).
Object pose initialization. To estimate the 6-degree-of-freedom object poses, we use a differentiable renderer [4]. This process involves fitting the projected shapes of the object templates to the segmentation masks obtained from the input images. Note that object pose estimation can be challenging, particularly in the presence of occlusion. Similar to PHOSA [33], we minimize a silhouette loss $\mathcal{L}_{\text{silhouette}}$ combined with a Chamfer loss for a fixed number of iterations during the initialization. See [33] for details regarding these two losses. Nevertheless, we significantly speed up the pose initialization process by using a more performant renderer and reducing the number of vertices in our object templates. More details can be found in the supplementary material.

3.3 Joint Optimization

Inspired by previous works [6, 27, 33], we formulate a joint optimization process to reason about spatial arrangements between humans and objects in 3D.

We hypothesize that the quality of the reconstructed objects in the initial step could be significantly improved by using foundational models, and no contact loss or penetration loss [33] is needed for the joint optimization. Specifically, we consider a novel combination of silhouette loss and relative depth loss and optimize for the object pose, which involves rigid transformations. This leads to a much simpler objective function compared to prior works. This simplicity allows us to obtain meaningful reconstructions, without the need for fine-tuning weight parameters for multiple loss terms and for each object category.
Silhouette loss. We use the same silhouette loss $\mathcal{L}_{\text{silhouette}}$ as in the object pose initialization and minimize the mismatch between the masks of the rendered 3D object's silhouette and the corresponding instance segmentation mask.
Relative depth loss. We observe that the image-based optimization objective often leads to reconstructions that look plausible from the camera view, yet have obvious mismatches when viewed from the side, such as the bicycle in Figure 1. Motivated by this, we propose a relative depth loss term that adds constraints in the camera viewing (depth) direction. During each iteration of the joint optimization, we obtain a rendered object depth map $\tilde{D}_O$ as well as a person depth map $\tilde{D}_P$ using a differentiable renderer. We then compare them to the reference 2D object depth map D_O and person depth map D_P (see Figure 2, Reference depth). Here, D_O and D_P are obtained from an off-the-shelf depth estimator [1], or alternatively a depth sensor, if available.

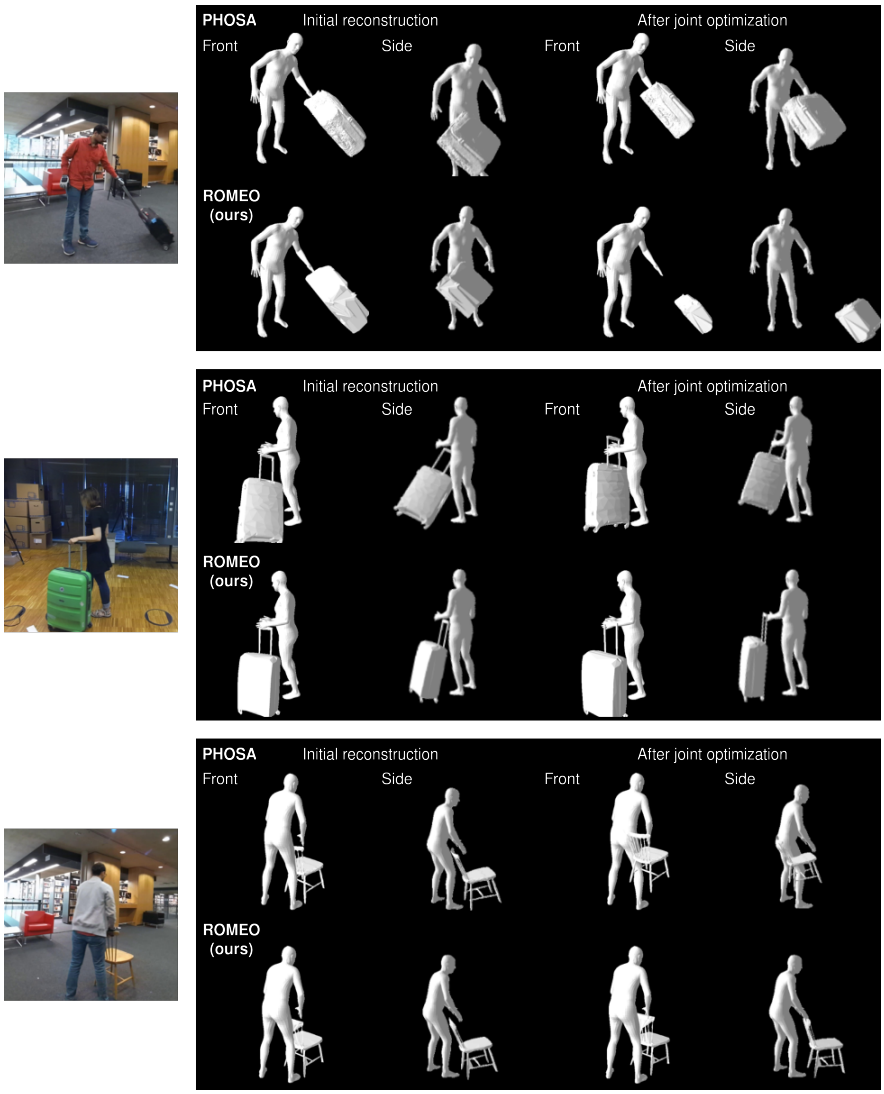


Fig. 4: Qualitative results of ROMEO. We compare the results of our proposed method ROMEO with PHOSA [33] on the BEHAVE dataset. We show front and side views of the initial results and the final results after joint optimization of each method. Our joint optimization improves the initial estimations in recovering plausible contacts with more plausible depth relationships between humans and objects.

We use a simple yet effective depth loss by computing, for both the reference depth map D and rendered depth map $\tilde{D}$, the ratios r resp. $\tilde{r}$ between each depth map's object mean depth and person mean depth, averaged on the pixels of their respective segmentation masks, and enforcing that this ratio matches

between the reference and the rendered depth maps, as in Eq. 2:

$$r = \frac{\sum\limits_{i \in M_O} D(i)/|M_O|}{\varepsilon + \sum\limits_{i \in M_P} D(i)/|M_P|}, \; \tilde{r} = \frac{\sum\limits_{i \in M_O \cap \tilde{M}_O} \tilde{D}(i)/|M_O \cap \tilde{M}_O|}{\varepsilon + \sum\limits_{i \in M_P \cap \tilde{M}_P} \tilde{D}(i)/|M_P \cap \tilde{M}_P|} \tag{1}$$

$$\mathcal{L}_{\text{depth}} = \left(1 - \tilde{r}/\left(\epsilon + r\right)\right)^2, \tag{2}$$

where we sum depth values over the pixels of the reference and rendered object and person masks (M_O, M_P), $(\tilde{M}_O, \tilde{M}_P)$, and ε is a small positive constant to avoid dividing by 0. We assume $D_O(i), D_P(i) > 0 \; \forall i$. This depth loss term $\mathcal{L}_{\text{depth}}$ ensures consistent depth values between the reconstructed human and object and their corresponding depth maps, while also establishing scale invariance due to the inherent cancellation of the depth scale through division, allowing us to put into relation depth maps of varying scales from different sources such as the mesh renderer and the depth estimator as long as the camera center is assumed to be at depth 0, as is the case for metric depth maps [1].

Our final optimization objective is

$$\mathcal{L} = \alpha \lambda_1 \mathcal{L}_{\text{depth}} + \lambda_2 \mathcal{L}_{\text{silhouette}} \tag{3}$$

where $\alpha = \mathcal{L}_{\text{silhouette, 0}}/\mathcal{L}_{\text{depth, 0}}$ is a scaling factor based on the ratio of the silhouette loss and the depth loss during the first joint optimization iteration. This factor acts to counter the possibly large disparity between $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{silhouette}}$ arising from the ratio $\tilde{r}/r$ and causing one of the loss terms to be neglected.

4 Experiments and Analysis

In this section, we investigate variations of our framework to evaluate key design decisions (Sec. 4.1); present quantitative and qualitative comparisons with state-of-the-art methods, including an optimization-based method and a learning-based approach (Sec. 4.2); and assess the generalization ability of our approach on in-the-wild images (Sec. 4.3).

Datasets. We evaluate on the **BEHAVE** [2] and **InterCap** [8] datasets. Both datasets capture full-body human-object interactions. They consist of multi-view RGB-D video frames of people interacting with objects in diverse ways. The corresponding 3D SMPL/SMPL-H body models, object shapes and poses are provided in both datasets. BEHAVE [2] contains about 15k frames where humans interact with 20 common objects, with multiple interactions included for many object categories. For instance, for the *chair* category, interaction types include sitting, holding, lifting, standing, and touching. We evaluate our proposed method on the images from camera view 1 and recording date 3 for BEHAVE, consistent with [27]. For InterCap, we evaluate on 1.1K keyframes employed in a previous study [28]. Using the available ground-truth masks provided by BEHAVE, we filter out images that are $> 70\%$ occluded during the evaluation, following the setup in CHORE [27]. This results in about 4.1K samples on

Category	PointRend		Ours	
	Det. Rate (%)	IoU (%)	Det. Rate (%)	IoU (%)
Chairwood	39.6	37.4	**93.3**	**57.1**
Chairblack	32.1	31.9	**93.9**	**61.1**
Suitcase	62.8	62.7	**90.4**	**78.7**
Backpack	22.7	19.4	**90.7**	**68.4**
Yogaball	47.2	39.9	**97.2**	**72.6**
Basketball	5.8	4.3	**80.2**	**62.7**

Table 1: Comparison of object mask IoUs and object detection rates on BEHAVE between the PointRend [14]-based strategy employed by previous work [33] and our Grounding DINO [16] + Segment Anything [13] strategy.

Instance Selection	Det. Rate (%)	IoU (%)
Ground truth	75.5	66.8
Heuristic	74.8	65.7
Depth-based	**75.1**	**66.3**

Table 2: Comparison of object instance selection methods on BEHAVE, with object detection rates and IoUs between selected object mask and GT object mask. Category-wise results are provided in the supplementary material.

which we evaluate for BEHAVE. InterCap [8] contains about 67k frames where humans interact with 10 objects of various sizes and affordances. Similar to other work [28], data from subjects 9 (female) and 10 (male) are used as the test set.
Evaluation metrics. We evaluate the reconstruction accuracy by computing the two-way Chamfer distances [2, 25, 27] for humans and objects independently. We use Procrustes analysis to align the estimated SMPL model $\mathcal{H}$ and the 3D object $\mathcal{O}$ to the ground truth. Note that this alignment is performed on the combined human-object meshes as in previous works [27, 28]. All Chamfer distances are reported in centimeters.

4.1 Object Initialization

Acquiring the object segmentation masks is the first step of the pipeline, and serves as the foundation for all following processes. Prior approaches, such as PHOSA [33], rely on detectors trained for distinct categories, limiting their efficacy on unfamiliar categories. In contrast, our method integrates Grounding DINO [16] and SAM [12], enabling robust performance across a diverse range of categories. In Table 1, we only use categories where PointRend [14], utilized by PHOSA [33], can successfully detect objects. For other categories in BEHAVE [2], PHOSA's detection abilities are limited as the category is neither seen during training nor exists in the pre-defined classes. As evidenced in Table 1, our detection strategy outperforms PHOSA by a large margin in terms of both detection rates and IoUs. In the BEHAVE dataset, the "suitcase" category is most frequently encountered during the training of PointRend, resulting in the highest detection rate and IoU. Nevertheless, our detection strategy still surpasses it by a notable margin.

Method	Mask	Depth	Stage	BEHAVE		InterCap	
				$\mathcal{H}\downarrow$	$\mathcal{O}\downarrow$	$\mathcal{H}\downarrow$	$\mathcal{O}\downarrow$
CHORE [27]	GT	-	Final	5.58 ± 2.11	10.66 ± 7.71	7.12	12.59
PHOSA [33]		-	Final	12.17 ± 11.13	26.62 ± 21.87	11.20	20.57
ROMEO (ours)		GT	Initial	8.54 ± 3.17	20.13 ± 16.34	10.23 ± 6.41	22.67 ± 14.82
ROMEO (ours)		GT	Final	**8.23** ± 3.00	**16.91** ± 11.65	**9.16** ± 5.70	**18.47** ± 12.81
CHORE [27]	Est	-	Final	7.44 ± 4.41	25.17 ± 28.02	8.96 ± 2.85 *	24.83 ± 15.48 *
PHOSA [33]		-	Final	**11.01** ± 6.29	34.53 ± 24.24	**8.77** ± 3.67	26.90 ± 21.45
ROMEO (ours)		Est	Initial	11.52 ± 7.03	31.77 ± 26.29	9.60 ± 5.04	22.49 ± 14.67
ROMEO (ours)		Est	Final	11.33 ± 7.53	**28.43** ± 25.24	9.61 ± 5.69	**20.34** ± 13.60

Table 3: Comparison with the state of the art on the BEHAVE [2] and InterCap [8] datasets. We compare our approach with the state-of-the-art optimization-based method PHOSA, and use state-of-art learning-based model CHORE (in gray) as a reference. The table is split into two parts, using ground-truth (GT) and estimated (Est) object masks, respectively. Our optimization-based approach ROMEO achieves significant improvements in reconstruction quality over the state-of-the-art optimization-based method [33] while reducing the gap to the learning-based method [27] and outperforming its BEHAVE-trained model (*) using a cross-dataset evaluation on InterCap [2], showing the better generalizability of ROMEO.

Method	Mask	Depth	Stage	BEHAVE	
				$\mathcal{H}\downarrow$	$\mathcal{O}\downarrow$
ROMEO	GT	Est	Initial	8.53 ± 3.17	20.13 ± 16.34
ROMEO	GT	Est	Final	8.51 ± 3.20	20.00 ± 16.04
ROMEO	Est	GT	Initial	11.15 ± 6.79	29.95 ± 26.21
ROMEO	Est	GT	Final	10.53 ± 6.95	23.71 ± 23.71

Table 4: Ablations of joint optimization using mixed estimated and GT data on the BEHAVE test set. We observe a strong improvement of joint optimization in terms of Chamfer distance by using ground-truth depth maps, yet no notable difference in performance by using estimated depth on reconstructions using GT masks. Depth map quality is thus crucial for the joint objective.

Next, we assess the performance of our depth-based object instance selection approach. We evaluate three selection techniques: (1) GT selection, which computes the IoU between each candidate object mask and GT object mask, and chooses the one with the highest IoU. Note that the GT object mask is not accessible in real-world applications. (2) Heuristic selection, which computes the IoU between each candidate object mask and the person mask, and chooses the one maximizing this IoU, and (3) our depth-based selection, as discussed in Sec. 3.2. Table 2 demonstrates that our depth-based selection enhances the detection rate by 0.3% and IoU by 0.6%, approaching the performance achieved with ground truth selection, thereby underscoring the efficacy of our method.

4.2 Joint Optimization

Baselines. We compare our approach with the state-of-the-art 3D human-object interaction reconstruction methods PHOSA [33] and CHORE [27], which reconstruct 3D models of humans and objects through joint optimization or using learned implicit distance fields, with less emphasis on generalizability.

When evaluating using PHOSA [33], we manually annotate each object category with contact and scale priors identically to previous work [27] for a fair

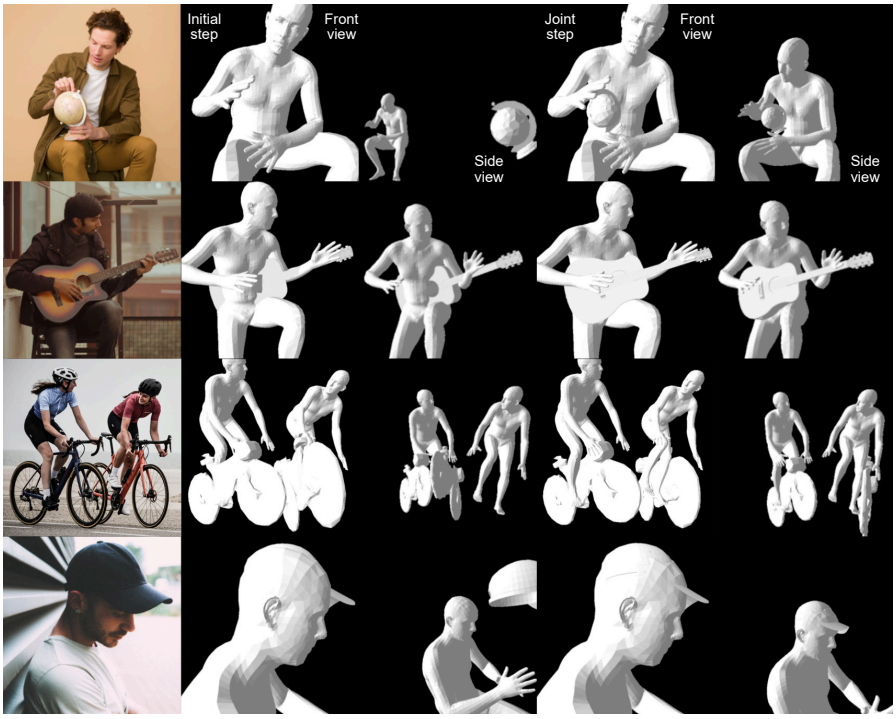


Fig. 5: More examples of reconstructed 3D interaction models from in-the-wild images. From left to right: input image, front and side views of our initial reconstruction, and front and side views of our final reconstruction after joint optimization. Our depth-based selection allows our method to reconstruct scenes with multiple people and objects.

comparison between all approaches. We use PARE [15] to reconstruct 3D humans for the same reason.

Quantitative results. Table 3 shows the reconstruction performance of all methods on the test sets of BEHAVE and InterCap. We consider a setup where ground-truth masks and depth maps are available, as well as a setup with access to estimated masks and depth maps only. Note that not all samples are used due to object detector failure. Results show that ROMEO outperforms PHOSA and reduces the gap to learning-based CHORE, most importantly outperforming it when evaluating its BEHAVE-trained model on InterCap in a generalization experiment, proving the better generalizability of our method. The reconstruction model of CHORE is trained with 3D data supervision. It is therefore especially sensitive to the quality of the input segmentation masks, and can not perform as well when object masks do not possess the noise-free quality of ground-truth data. PHOSA relies on manual contact annotations, thereby implicitly assuming a fixed type of interaction per object. It thus struggles with diverse interactions

captured in BEHAVE. In contrast, we leverage the information provided by foundational models to achieve significantly better results, in particular for objects.
Qualitative results. We show qualitative comparisons of ROMEO and PHOSA in Figure 4. We note that our method reconstructs more plausible models, without using any contact labels or fine-tuned category-specific parameters. We further see that our joint optimization stage can effectively improve the reconstructions, even when the initial estimates are already reasonable.
Ablations. To demonstrate the effect of the quality of the segmentation masks and depth maps, we conduct an ablation study using mixed estimated and GT data. The results are shown in Table 4. Contrasting the "estimated masks, GT depth maps" result with the estimation-only result in Table 3 suggests that the quality of the depth maps determines the performance of our relative depth loss, and that ROMEO can benefit from future improvements in depth estimation.

4.3 Reconstruction Generalizability

We evaluate ROMEO on several in-the-wild images featuring both object categories that are included in BEHAVE and InterCap, as well as novel object categories. Results are visualized in Figure 1 and Figure 5. Due to having no requirement for training data, our method is able to obtain good 3D reconstruction results even for unusual objects.

4.4 Implementation Details

During initialization, we use the Kaolin renderer [4] to render the depth maps and silhouettes, which is considerably faster than the Neural Mesh Renderer [11] used in previous work [33]. After optimizing each of 2000 randomly sampled object rotations and translations using the silhouette loss and Chamfer loss for 70 steps, we pick the candidate with the lowest combined loss. We obtain the human reconstruction from PARE [15]. Next, during the joint optimization, we use the silhouette loss and relative depth loss to refine the object pose for at least 500 iterations, after which we further optimize until the total loss does not decrease for 100 consecutive iterations. See Supp. Mat. for further details.

5 Discussion and Conclusion

We present ROMEO, a method to reconstruct 3D human-object interaction models from RGB images. Our method benefits from foundational models on several vision tasks and uses a novel yet simple optimization objective, eliminating the need for manual contact annotations, 3D data supervision, category-specific parameter fine-tuning and prior knowledge of object scales. Experiments show that ROMEO produces robust and high-quality reconstructions, outperforming the SOTA optimization method and reducing the gap to the learning-based SOTA method, and showing the best generalization on in-the-wild images. In conclusion, our findings provide compelling evidence that reconstructing 3D human-object interaction models does not necessarily require explicit 3D supervision.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-Shot Transfer by Combining Relative and Metric Depth. arXiv preprint arXiv:2302.12288 (2023)
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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15935–15946 (2022)
3. Chen, W., Ling, H., Gao, J., Smith, E., Lehtinen, J., Jacobson, A., Fidler, S.: Learning to Predict 3D Objects with an Interpolation-Based Differentiable Renderer. In: NeurIPS. pp. 9605–9616 (2019)
4. Fuji Tsang, C., Shugrina, M., Lafleche, J.F., Takikawa, T., Wang, J., Loop, C., Chen, W., Jatavallabhula, K.M., Smith, E., Rozantsev, A., Perel, O., Shen, T., Gao, J., Fidler, S., State, G., Gorski, J., Xiang, T., Li, J., Li, M., Lebaredian, R.: Kaolin: A Pytorch Library for Accelerating 3D Deep Learning Research. `https://github.com/NVIDIAGameWorks/kaolin` (2022)
5. Gibson, J.J.: The Ecological Approach to the Visual Perception of Pictures. Leonardo **11**(3), 227–235 (1978)
6. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019)
7. Huang, C.H.P., 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: Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 13274–13285 (Jun 2022)
8. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction. In: German Conference on Pattern Recognition (GCPR). Lecture Notes in Computer Science, vol. 13485, pp. 281–299. Springer (2022)
9. Jiang, N., Liu, T., Cao, Z., Cui, J., Chen, Y., Wang, H., Zhu, Y., Huang, S.: CHAIRS: Towards Full-Body Articulated Human-Object Interaction. arXiv preprint arXiv:2212.10621 (2022)
10. Joo, H., Neverova, N., Vedaldi, A.: Exemplar Fine-Tuning for 3D Human Pose Fitting Towards In-the-Wild 3D Human Pose Estimation. 3DV (2021)
11. Kato, H., Ushiku, Y., Harada, T.: Neural 3D Mesh Renderer. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment Anything. arXiv:2304.02643 (2023)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment Anything. arXiv preprint arXiv:2304.02643 (2023)
14. Kirillov, A., Wu, Y., He, K., Girshick, R.: PointRend: Image Segmentation as Rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9799–9808 (2020)
15. Kocabas, M., Huang, C.H.P., Hilliges, O., Black, M.J.: Pare: Part attention regressor for 3d human body estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11127–11137 (2021)

16. 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)
17. Mirmohammadi, M., Saremi, P., Kuo, Y.L., Wang, X.: Reconstruction of 3d interaction models from images using shape prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2141–2147 (2023)
18. Mittal, P., Cheng, Y.C., Singh, M., Tulsiani, S.: AutoSDF: Shape Priors for 3D Completion, Reconstruction and Generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 306–315 (2022)
19. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. ArXiv preprint (2021)
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-dataset Transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
21. Rong, Y., Shiratori, T., Joo, H.: FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration. In: IEEE International Conference on Computer Vision Workshops (2021)
22. Shugurov, I., Pavlov, I., Zakharov, S., Ilic, S.: Multi-View Object Pose Refinement With Differentiable Renderer. IEEE Robotics and Automation Letters **6**(2), 2579–2586 (2021)
23. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8001–8013 (October 2023)
24. Wang, G., Manhardt, F., Shao, J., Ji, X., Navab, N., Tombari, F.: Self6D: Self-Supervised Monocular 6D Object Pose Estimation. In: The European Conference on Computer Vision (ECCV) (August 2020)
25. Wang, X., Li, G., Kuo, Y., Kocabas, M., Aksan, E., Hilliges, O.: Reconstructing action-conditioned human-object interactions using commonsense knowledge priors. In: 2022 International Conference on 3D Vision (3DV). pp. 353–362 (2022)
26. Weng, Z., Yeung, S.: Holistic 3d human and scene mesh estimation from single view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 334–343 (2021)
27. 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). Springer (October 2022)
28. Xie, X., Bhatnagar, B.L., Pons-Moll, G.: Visibility Aware Human-Object Interaction Tracking from Single RGB Camera. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
29. Xu, S., Li, Z., Wang, Y.X., Gui, L.Y.: InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion. In: ICCV (2023)
30. Yi, H., Huang, C.H.P., Tzionas, D., Kocabas, M., Hassan, M., Tang, S., Thies, J., Black, M.J.: Human-Aware Object Placement for Visual Environment Reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3959–3970 (2022)
31. Zhang, B., Nießner, M., Wonka, P.: 3DILG: Irregular Latent Grids for 3D Generative Modeling. arXiv preprint arXiv:2205.13914 (2022)
32. Zhang, H., Tian, Y., Zhou, X., Ouyang, W., Liu, Y., Wang, L., Sun, Z.: Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop.

In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11446–11456 (2021)
33. 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) (2020)
34. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: NeuralDome: A Neural Modeling Pipeline on Multi-View Human-Object Interactions (2022). https://doi.org/10.48550/ARXIV.2212.07626, https://arxiv.org/abs/2212.07626