arXiv:2411.14280v4 [cs.CV] 29 Apr 2025

EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild

Yumeng Liu[1,2] Xiaoxiao Long[3,*] Zemin Yang[2] Yuan Liu[3,4] Marc Habermann[5]
Christian Theobalt[5] Yuexin Ma[2,*] Wenping Wang[6]
[1] The University of Hong Kong , [2]ShanghaiTech University,
[3]Hong Kong University of Science and Technology, [4]Nanyang Technological University,
[5]Max Planck Institute for Informatics, [6]Texas A&M University

Figure 1. We leverage large models and prior-guided optimization to reconstruct physically plausible hand-object interactions from single-view in-the-wild images, effectively handling occlusions and diverse hand-object configurations.

Abstract

Our work aims to reconstruct hand-object interactions from a single-view image, which is a fundamental but ill-posed task. Unlike methods that reconstruct from videos, multi-view images, or predefined 3D templates, single-view reconstruction faces significant challenges due to inherent ambiguities and occlusions. These challenges are further amplified by the diverse nature of hand poses and the vast variety of object shapes and sizes. Our key insight is that current foundational models for segmentation, inpainting, and 3D reconstruction robustly generalize to in-the-wild images, which could provide strong visual and geometric priors for reconstructing hand-object interactions. Specifically, given a single image, we first design a novel pipeline to estimate the underlying hand pose and object shape using off-the-shelf large models. Furthermore, with the initial reconstruction, we employ a prior-guided optimization scheme, which optimizes hand pose to comply with 3D physical constraints and the 2D input image content. We perform experiments across several datasets and show that our method consistently outperforms baselines and faithfully reconstructs a diverse set of hand-object interactions. Here is the link of our project page: https://lym29.github.io/EasyHOI-page/.

* indicates the corresponding authors.

This work was finalized during the first author's visit to ShanghaiTech University.

1. Introduction

Reconstructing **H**and-**O**bject **I**nteractions (HOI) from a single image is an essential task in computer vision and graphics, which plays important roles in many applications, such as human behavior understanding, Augmented Reality, teleoperation, and robotic grasping. This task aims to not only reconstruct the shapes of objects and hands but also to recover fine-grained hand-object interactions satisfying physical constraints.

Single-view HOI reconstruction is quite challenging, since the monocular setting is naturally ill-posed and the interactions contain severe mutual and self-occlusions. Many existing works [18, 23, 60] typically simplify the HOI task by assuming the object is known. However, such 3D template assumption is too strong and prevents these systems from generalizing to unknown objects. Some other works [7, 23, 24] directly train neural networks for joint object and hand recovery on public datasets in an end-to-end manner. However, since acquiring 3D annotations of HOI data is significantly difficult and expensive, the public datasets are low-quality and limited. The lack of high-quality HOI data proves to be the key challenge to boost the development of this task [5]. Overall, these methods unavoidably suffer from poor robustness and could not generalize to unseen scenarios.

Recent advancements in large models have shown promising potential in reconstructing both objects and hands. Notable studies in single-view 3D reconstruction models [26, 33, 39, 40, 61, 66] have demonstrated remarkable capabilities in understanding complex geometries from limited visual input. Similarly, advances in hand estimation [46, 55, 67] have shown significant progress in inferring intricate hand poses in complex scenarios. These parallel developments offer encouraging prospects for addressing the HOI task.

A seemingly straightforward approach is independently reconstructing and merging hands and objects using pretrained large models. However, this approach encounters three critical limitations: 1) **Inconsistent Coordinate Systems:** Object reconstruction methods generally rely on a canonical coordinate system, while hand reconstruction methods often adopt a camera-based system that aligns with the input viewpoint. This incompatibility leads to misaligned results when merged directly. 2) **Estimation Inaccuracies:** The reconstruction methods for, both, objects and hands frequently exhibit significant inaccuracies, often resulting in spatial misalignments that fail to reflect the actual observed interactions. 3) **Impact of Occlusions:** Inferring geometry and interactions for occluded regions from a single view poses significant challenges. The inability to effectively reconstruct these unobserved areas leads to physically implausible results and unrealistic interaction estimates.

To achieve robust and high-quality Hand-Object Interaction (HOI) estimation, we propose leveraging the strong priors of multiple large models to reconstruct both the object and hand. Additionally, we introduce a novel prior-guided optimization framework to jointly optimize these reconstructions, ensuring physically plausible HOI results. Our proposed optimization framework comprises three key stages: 1) **Camera System Setup:** Using differentiable rendering techniques, we align the reconstructed object and hand within a unified coordinate system, guided by segmentation priors to ensure coherent positioning. 2) **HOI Contact Alignment:** We identify potential 3D interaction regions between the object and hand, applying the Iterative Closest Point (ICP) algorithm to register their shapes. This step yields an initial, approximate alignment that captures the interaction. 3) **Hand Parameter Refinement:** To enhance physical plausibility, we refine hand parameters using a combination of loss functions, including segmentation mask loss, penetration loss, contact loss, and regularization loss. This multi-loss approach optimizes the hand's interaction with the object for a realistic and accurate result with limited visual observation.

Compared to the prior works, our system presents strong generalization and robustness on diverse in-the-wild images and could produce plausible HOI reconstructions with only single image as input. Extensive experiments have been conducted to validate the strong generalization and robustness of our method.

2. Related Works

2.1. Hand-Object Interactions Reconstruction

Reconstructing hand-and-object interactions is a challenging problem that has attracted significant attention in recent years. Many reconstruction methods rely on richer input sources, such as videos [16, 25, 27, 70] and object templates [4, 5, 9, 12, 18–20, 22, 22, 23, 37, 47, 49, 52, 57, 60, 60, 62, 63, 63, 73], restricting their use in real-world applications. To achieve model-free reconstruction from single view, some methods deform a sphere mesh [24] or learn an implicit shape field for objects [10, 11, 29, 69, 72]. However, these approaches are often trained on limited datasets, leading to reduced effectiveness on objects that differ from the training data. In contrast, our method leverages advanced large-scale models to significantly enhance generalization capabilities.

2.2. Single-view 3D Reconstruction

Single-view 3D reconstruction has been a long-standing focus in computer vision and graphics, serving as a foundational step for numerous applications. Early approaches [6, 14, 17, 21, 30, 34, 44, 51, 59] applied neural networks to reconstruct 3D shapes from single-view images through regression [34], retrieval [59], or NeRF-based GANs [44].

Recent trends leverage diffusion models using strategies such as Score-Distillation Sampling [35, 36, 48] and multiview diffusion [36, 39, 40, 54]. Additionally, recent Large Reconstruction Models (LRMs) [26, 58, 64, 65] employ transformers to reconstruct object meshes directly from input images. Our method incorporates an LRM model as well but focuses on single-view reconstruction within the context of human-object interactions (HOI).

2.3. Image-based Hand Pose Estimation

Image-based hand pose estimation aims to recover hand pose, including keypoints, parametric shape parameters, or geometries, from input images. Early methods [28, 41, 42, 45] primarily focused on estimating 2D or 3D hand keypoints. Recent single-view hand pose estimation studies have benefited from advancements in human mesh recovery, with the MANO parametric hand model [50] emerging as a powerful tool for extracting hand pose and shape parameters from RGB images [2, 3, 56, 74, 75]. These methods either directly regress MANO parameters [2, 3] or optimize shape fitting to the images [56, 74, 75]. Our approach also involves hand pose estimation, but we address it within the context of human-object interaction, where reconstructed objects introduce additional constraints to refine the hand meshes.

3. Methodology

3.1. Problem Formulation

Given an input image I depicting hand-object interaction, our goal is to reconstruct the 3D object shape and determine its relative pose with respect to the hand. This task requires optimizing several interrelated components. First, we obtain initial reconstructions of both the object mesh Ω_o and hand mesh Ω_h. For hand representation, we employ the MANO parametric model [50], which is characterized by two key components: the global 6D pose $\phi_h \in \mathbb{R}^6$ and the articulated pose parameters $\theta_h \in \mathbb{R}^{45}$. The subsequent optimization involves estimating camera parameters, including intrinsic matrix K and extrinsic parameters $[R|t]$ for object-image alignment, and refining hand parameters (ϕ_h, θ_h).

3.2. Pipeline Overview

As shown in Fig. 2, our pipeline mainly consists of two stages: 1) **Initial Reconstruction of Hand and Object.** We first employ LISA [32] to segment hand and object masks. A diffusion model [71] then removes and inpaints the hand region, followed by SAM [31] segmenting the complete object. Finally, InstantMesh [66] reconstructs the object mesh while HaMeR [46] simultaneously reconstructs the hand mesh. 2) **Hand-Object Interaction Optimization.** Since the initial hand and object reconstructions are obtained separately, they may be inconsistent with the input image. To overcome this challenge, we formulate a prior-guided optimization framework for HOI tasks that progressively refines the reconstruction in a coarse-to-fine manner.

3.3. Initial Reconstruction of Hand and Object

Hand-Object Interaction Reasoning. We first segment the regions of interest - the hand and its interacting object - from the input image, which serves as the foundation for subsequent reconstruction. To handle images that may contain multiple objects, we employ LISA [32], a vision-language model for segmentation task, to obtain semantic segmentation masks for the relevant hand-object interaction.

While LISA effectively segments hands and objects in most cases, it sometimes produces redundant masks for visually similar objects that are not actually being interacted with. To address this, we propose a contour-guided filtering strategy that discards objects whose contours are not neighboring the hand's. Finally we obtain accurate segmentation masks for the interacting objects, denoted as M_o. Additional details are presented in the supplementary material.

Hand Reconstruction. We employ a recent transformer-based approach, HaMeR [46] to recover the hand pose from the input image. First, ViTPose [67] is utilized to estimate the 2D keypoints of the hands and compute the bounding box based on these keypoints, which helps crop the regions of interest (ROI) from the whole input image. Afterward, the hand parameters (ϕ_h, θ_h) are predicted from the bounding box via HaMeR.

Object Reconstruction. Due to the occlusion caused by hand, the interacted object is often visually incomplete in the input image and therefore could only produce distorted 3D geometry of the object. To address this problem, as the example shown in Fig. 3, our reconstruction pipeline first employs a diffusion model [43, 71] to recover the complete object appearance by removing occlusions. As a by-product, we could obtain the segmentation mask of the inpainted object with complete appearance, denoted as $\hat{M}_o$.

Finally we apply a large reconstruction model, InstantMesh [66], to reconstruct the object's full geometry and conduct a post-processing to make the generated mesh watertight. A detailed explanation can be found in the supplementary material.

3.4. Hand-Object Interaction Optimization

Although initial reconstructions of the hand and object are obtained, their separate processing and differing coordinate systems result in unrealistic spatial relationships, such as excessive distance or interpenetration. Furthermore, depth and scale ambiguities in monocular images exacerbate these issues. To resolve this, we propose a three-step optimization scheme to reconstruct a realistic interaction.

1) Camera System Setup. The problem of aligning the

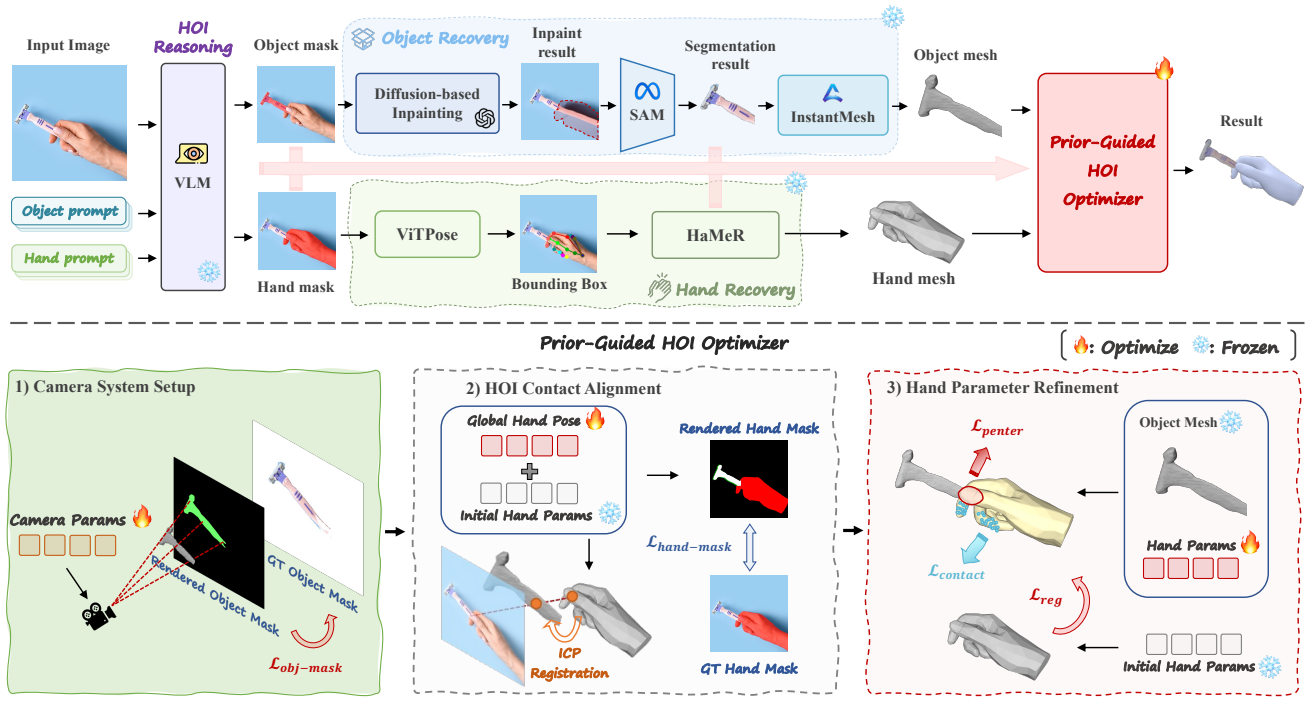
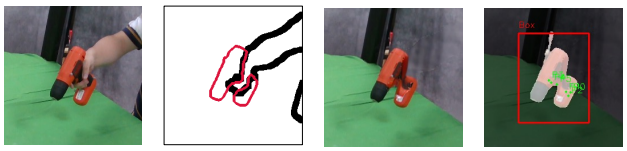


Figure 2. The illustration of our pipeline. We first extract hand and object masks through HOI reasoning for object reconstruction and recovering hand mesh from the input image. With these initial reconstructions, we employ a three-stage prior-guided optimizer to establish a camera system for object, align hand and object by analyzing contact points, and finally refines hand parameters to ensure physical plausibility.



(a) Input (b) Contours (c) Inpaint (d) Segmentation

Figure 3. Illustration of the segmentation process after inpainting. (a) Original input image. (b) Hand and object contours, showing the object split into two disconnected regions due to hand occlusion. (c) Inpainted image with the hand region removed. (d) Object segmentation results, consisting of sampled points and a bounding box, used as prompts to segment the inpainted image.

two systems of hand and object can be simplified by using the reconstructed object's coordinate frame as the global reference, and then estimating the camera parameters of the input image defined in the object coordinate system. Using a differentiable renderer Ψ, we would like to obtain the optimal camera parameters (K, $[R|t]$) by minimizing the soft IoU loss between rendered object mask $M_o^r = \Psi(\Omega_o, K, R, t)$ and ground truth mask $\hat{M}_o$ (the segmentation mask of the inpainted object with complete appearance), as shown in Eq. (1).

$$\mathcal{L}_{\text{obj-mask}} = \text{IOU}(M_o^r, \hat{M}_o) = \frac{M_o^r \cdot \hat{M}_o}{|M_o^r| + |\hat{M}_o| - M_o^r \cdot \hat{M}_o}, \tag{1}$$

where $|.|$ denotes the sum of all elements in the mask.

However when initial and ground truth masks don't overlap, the IoU loss becomes ineffective since its gradient is zero. We address this by incorporating Sinkhorn distance [13], a regularized Wasserstein distance that measures the minimal cost of transforming the rendered mask distribution to match the ground truth. Let M_α and M_β represent two normalized mask images, treated as discrete distributions supported on a finite grid. The Wasserstein distance between them is defined by:

$$W(M_\alpha, M_\beta) = \left(\inf_{\gamma \in \Pi(M_\alpha, M_\beta)} \sum_{i,j} \sum_{k,l} ||(i,j) - (k,l)||^2 \gamma_{ijkl} \right)^{1/2}, \tag{2}$$

where γ_{ijkl} represents the transport flow from position (i, j) in M_α to position (k, l) in M_β, and $\Pi(M_\alpha, M_\beta)$ denotes the set of all possible transport flows. We employ the Sinkhorn-Knopp algorithm [13] to compute the optimal transport loss:

$$\mathcal{L}_{\text{OT}} = W\left(\frac{M_o}{|M_o|}, \frac{\hat{M}_o}{|\hat{M}_o|}\right). \tag{3}$$

Combining this with IoU loss yields optimal camera parameters $\bar{K}$ and $[\bar{R}|\bar{t}]$, ensuring robust alignment even for non-overlapping masks.

2) HOI Contact Alignment. Since we adopt the object's coordinate system as the global reference, the object is po-

sitioned at the system origin, allowing us to focus solely on optimizing the hand parameters for hand-object interaction (HOI) reconstruction. In this stage, we retain the articulated hand pose θ_h as estimated by HaMeR, and optimize the hand's global 6D pose ϕ_h to bring the hand and object into approximate contact. The optimization alternates between two steps: ***Mask-Constrained Optimization*** and ***Contact-Based Registration***, iteratively improving contact alignment.

(a) Mask-constrained Hand Pose Optimization. We first need an initial 2D alignment between the estimated hand configuration and the input image, leveraging hand segmentation as supervision. The differentiable renderer Ψ generates a hand mask $M_h^r = \Psi(\theta_h, \phi_h, \bar{K}, \bar{R}, \bar{t})$. We then compute the soft IoU loss between this rendered mask M_h^r and the ground truth hand mask M_h obtained from input image segmentation.

$$\mathcal{L}_{\text{hand-mask}} = \text{IOU}(M_h^r, M_h) = \frac{M_h^r \cdot M_h}{|M_h^r| + |M_h| - M_h^r \cdot M_h}. \tag{4}$$

The hand mask loss ensures hand alignment with the input image in the 2D plane. However, differentiable rendering only provides gradients parallel to the image plane, not in the camera viewing direction. This fundamental limitation makes it challenging to achieve precise 3D alignment between the hand and the object.

(b) ICP-based Hand-Object Registration. To achieve accurate hand-object alignment, analyzing their contact relationships is essential. Mutual occlusions provide inherent contact cues, which we use to detect 2D contact regions (detailed methodology in the supplementary material). These 2D contact regions are then converted to 3D contact points through ray-casting. As shown in Fig. 4, ray-casting from image pixels creates multiple intersection points due to monocular ambiguity. For objects, we select the nearest and farthest intersection points along each ray. For hands, we focus only on the palm side as the functional grasping area, excluding the dorsal side (the back of the hand) using pre-marked regions on the MANO template. We then identify valid hand contact points by first keeping only palmar intersections, then selecting the nearest and farthest points along each ray.

Once all potential contact points on both the hand and the object are identified, we apply the Iterative Closest Point (ICP) method to compute the optimal hand translation, aligning the contact points and providing a rough estimation of the hand's pose.

3) Hand Parameter Refinement. While initial global pose optimization provides coarse alignment, finger configurations remain unadapted to the object's geometry. We then jointly optimize the hand's global pose θ_h (6 DoF) and articulation parameters ϕ_h (45 DoF) to achieve physically plausible, penetration-free configurations. The objective function is defined as:

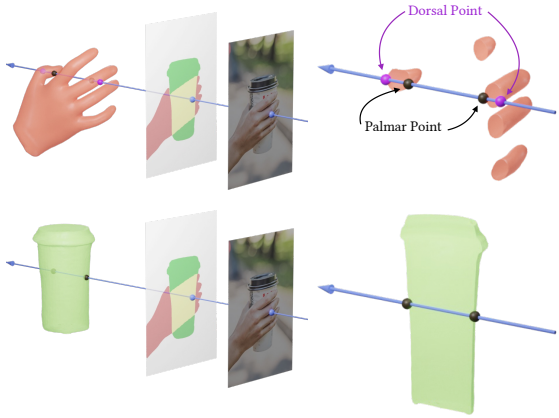


Figure 4. Converting 2D contact regions to 3D contact points. Rays emitted from contact mask pixels intersect object and hand geometries. Contact point candidates are constrained to the extremal ray intersections: nearest or farthest points relative to the camera for the object, and palmar-side extremal points for the hand.

$$\mathcal{L}_{\text{hand}} = \lambda_1 \mathcal{L}_{\text{hand-mask}} + \lambda_2 \mathcal{L}_{\text{penetr}} + \lambda_3 \mathcal{L}_{\text{contact}} + \lambda_4 \mathcal{L}_{\text{reg}}, \tag{5}$$

where $\mathcal{L}_{\text{hand-mask}}$ represents the IoU loss of the hand mask, as shown in Eq. (4), $\mathcal{L}_{\text{penetr}}$ is the penetration loss to prevent hand-object penetration, $\mathcal{L}_{\text{contact}}$ encourages a reasonable hand-object contact relationship, and $\mathcal{L}_{\text{reg}}$ is a regularization loss that ensures the articulated hand pose remains close to the result from HaMeR. The parameters $\lambda_1, \lambda_2, \lambda_3$ and λ_4 balance the different losses.

Penetration Loss. We introduce a penetration loss in this stage, defined as the mean distance from hand vertices inside the object to its boundary, to discourage intersections.

$$\mathcal{L}_{\text{penetr}} = \frac{1}{N} \sum_{v \in \mathcal{H}} \max(0, -d(v)), \tag{6}$$

where $d(\cdot)$ is the signed distance function defined around the object mesh, $\mathcal{H}$ represents the hand mesh and v is arbitrary vertex of the hand mesh.

Contact Loss. To ensure sufficient contact and stability of grasps, we incorporated a contact loss. We use the contact zones $\mathcal{H}_C$ introduced by ObMan [24], which are the regions of hand mesh that frequently make contact during grasping. We then calculate the total distance from the exterior points sampled in the regions to the object surfaces.

$$\mathcal{L}_{\text{contact}} = \sum_{v \in \mathcal{H}_C} \max(0, d(v)). \tag{7}$$

Regularization Loss. To ensure that the optimized hand pose does not deviate significantly from the initial estimates, we employ a regularization loss. Specifically, we utilize an L1 loss to compare the optimized pose with the initial pose θ_h^0 estimated by the HaMeR system.

$$\mathcal{L}_{\text{reg}} = \|\theta_h - \theta_h^0\|_1. \tag{8}$$

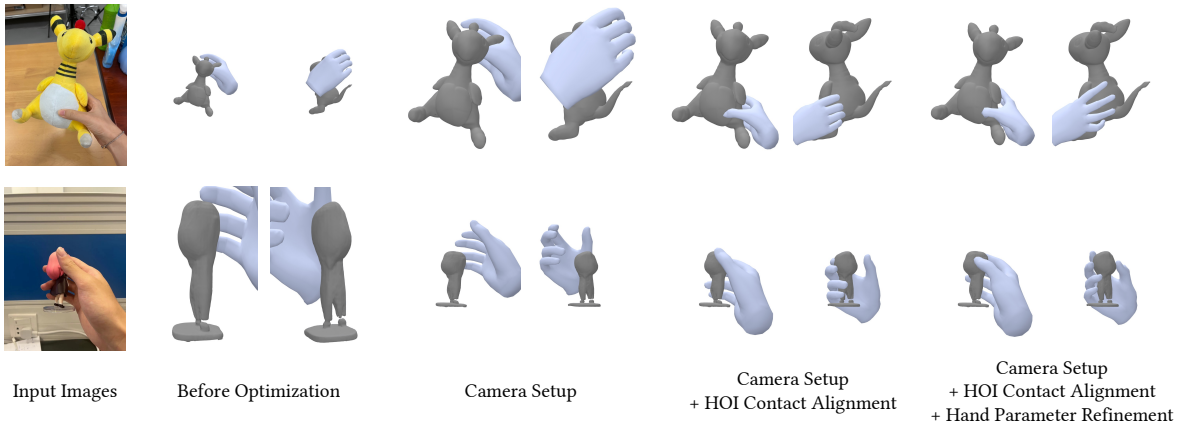


Figure 5. We perform ablation studies by examining the outputs at each processing stage. Starting with reconstruction results from foundational models, we present the progressive improvements through Camera Setup, HOI Contact Alignment, and Hand Parameter Refinement.

This regularization maintains a balance between accuracy and realism, preventing excessive deviations from the HaMeR predictions in the pose adjustments.

4. Experiments

4.1. Experimental Setup

Our method is implemented on an Ubuntu server equipped with an NVIDIA A40 GPU. For the camera system setup, we use a learning rate of 10^{-2} for optimizing camera parameters, terminating when $\mathcal{L}_{\text{obj-mask}} < 0.1$ or after 1000 iterations. In human-object interaction (HOI) contact alignment, global hand parameters are optimized with a learning rate of 10^{-2}. Hand-object registration are performed only at the 100th and 200th iterations, mask-constrained hand pose optimization continue until $\mathcal{L}_{\text{hand-mask}} < 0.1$ or 1000 iterations. During hand parameter refinement, we set learning rates of 10^{-4} for the global hand pose and 10^{-2} for articulated pose. The loss weights in Eq. (5) are set to $\lambda_1 = 5$, $\lambda_2 = 10$, $\lambda_3 = 5$, and $\lambda_4 = 0.1$.

4.2. Datasets and Baselines

To evaluate our method's generalization capability, we conduct a zero-shot comparison with IHOI [69], AlignSDF [10], gSDF [11], and MOHO [72]. Specifically, we test on three public datasets: Arctic [15], OakInk [68], and DexYCB [8]. These datasets contain videos with 3D hand-object pose and shape annotations: Arctic includes 11 articulated objects, OakInk features 100 diverse objects, and DexYCB presents 20 distinct YCB-video objects. To ensure valid grasping in the images, we use a force closure tester [38] to filter out instances where the hand and object are not in contact. We then randomly select 500 images from each dataset for evaluation. All datasets represent unseen domains for IHOI and MOHO. For AlignSDF and gSDF, DexYCB is part of their training data, so we exclude it from testing these methods to ensure a fair comparison.

4.3. Evaluation Criteria

To evaluate object reconstruction quality, we use the **Chamfer Distance** to measure the discrepancy between predicted outputs and ground truth. To minimize the impact of outliers, we also report the **F-score** at 5mm and 10mm thresholds. For assessing hand-object interaction quality, we calculate the **Intersection Volume** between the hand and object models. Additionally, we perform a simulation with a fixed hand pose and gravity applied to the environment. By measuring the distance the object falls over 1,000 simulation steps, termed **Simulation Displacement**, we effectively evaluate grasp stability.

4.4. Comparison Results

Quantitative Results. We evaluate our method's performance in terms of object reconstruction accuracy, grasping quality and hand accuracy. To ensure fair comparison, we employ HaMeR for hand pose detection when evaluating all baseline methods. As shown in Tab. 1, our method achieves the lowest Chamfer distance and highest F-scores. These metrics demonstrate our method's capability to accurately reconstruct objects from single-view images. In terms of hand-object interaction quality, Tab. 1 shows that our method outperforms competing approaches, consistently achieving the lowest simulation displacement and smallest intersection volume across all datasets. These results underscore the effectiveness of our hand-object interaction modeling. In Tab. 2, we report the reconstruction errors for hand vertices and joints. Since the object geometry recovered by LRM is not entirely accurate, optimizing for physically plausible HOI results with such imperfect object geometry inevitably compromises hand accuracy. Consequently, our final hand results are slightly worse than HaMER's. However, with ground truth object geometry, our proposed HOI optimization can further refine the initial hand results produced by HaMER, demonstrating the

	Arctic					OakInk					DexYCB				
	F5 ↑	F10 ↑	$C.D.$ ↓	S.D.↓	I.V.↓	F5 ↑	F10 ↑	$C.D.$ ↓	S.D.↓	I.V.↓	F5 ↑	F10 ↑	$C.D.$ ↓	S.D.↓	I.V.↓
IHOI [69]	0.083	0.164	1.375	2.79	4.94	0.097	0.152	1.742	3.14	4.66	0.084	0.143	1.897	2.59	4.93
AlignSDF [10]	0.102	0.196	1.289	2.46	4.73	0.095	0.148	1.813	3.27	4.61	-	-	-	-	-
gSDF [11]	0.115	0.247	1.247	2.31	4.89	0.106	0.173	1.992	3.15	4.16	-	-	-	-	-
MOHO [72]	0.072	0.136	12.878	3.94	4.72	0.175	0.323	3.883	3.55	4.86	0.119	0.249	1.695	2.62	4.61
Ours	**0.155**	**0.272**	**1.089**	**2.25**	**4.67**	**0.247**	**0.445**	**1.035**	**3.08**	**4.11**	**0.134**	**0.253**	**1.628**	**2.43**	**4.52**

Table 1. Quantitative evaluation for object quality in HOI reconstruction. Since AlignSDF and gSDF were trained on DexYCB, we exclude their DexYCB results from our zero-shot comparisons. The metrics F5 and F10 measure the F score of points from reconstructed object within 5mm and 10mm of the GT object, respectively. The metric $C.D.$ denotes the Chamfer Distance between reconstructed object and GT object, $S.D.$ denotes Simulation Displacement(in cm) and $I.V.$ represents Intersection Volume(in cm^3).

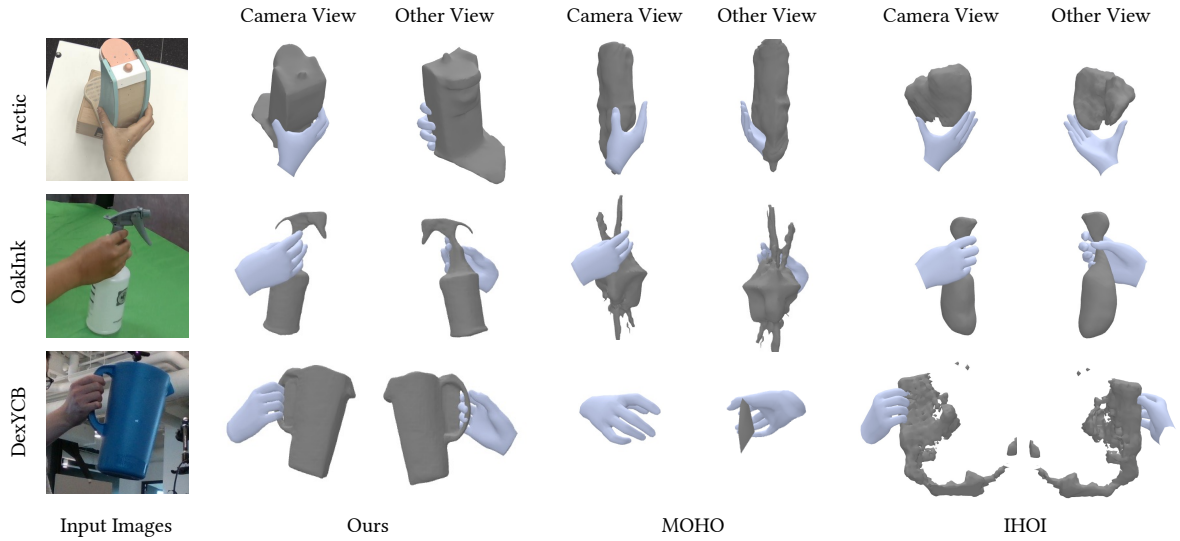


Figure 6. This gallery showcases the outcomes of our hand-object reconstruction on three public dataset Arctic, OakInk and DexYCB (more results in the supp.). The first column is the input image, we present the camera view and another view to display the reconstructed HOI meshes.

	Arctic		OakInk	
	MPVPE↓	MPJPE↓	MPVPE↓	MPJPE↓
IHOI	1.36	1.57	1.36	1.32
AlignSDF	1.38	1.36	1.25	1.26
gSDF	1.34	1.37	1.27	1.14
MOHO	1.38	1.35	1.27	1.33
HaMeR	1.14	1.05	1.06	1.03
Ours (Pred-object)	1.48	1.36	1.19	1.22
Ours (GT-object)	**1.12**	**0.95**	**0.86**	**0.77**

Table 2. Hand pose accuracy in HOI reconstruction is evaluated using MPVPE (Mean Per Vertex Position Error) and MPJPE (Mean Per Joint Position Error), both measured in centimeters (cm).

effectiveness of our method.

Qualitative Results. We evaluate our method through visual comparisons with two baselines, IHOI and MOHO. Fig. 6 showcases representative results on Arctic, OakInk, and DexYCB datasets, demonstrating our method's effectiveness across varied hand-object interactions. Our approach successfully reconstructs hand-object interactions in challenging scenarios with significant occlusions and complex hand poses, highlighting its robustness for single-view hand-object reconstruction in real-world settings.

The supplementary material contains additional examples from the benchmark datasets that demonstrate consistent performance across diverse scenarios. We also present results on our collected "in-the-wild" images, which showcase our method's robustness in various real-world environments. Furthermore, we provide a qualitative comparison on the selection of large-scale reconstruction models, which shows that integrating our method with Tripo3D [61] yields enhanced object reconstruction results and brings notable improvements in hand-object interaction results. This successful integration demonstrates our framework's extensibility and suggests that our approach can effectively leverage future advances in large-scale reconstruction models to achieve even better performance.

4.5. Ablation Study and Discussions

To analyze each component of our proposed method, we conducted a series of ablation experiments to quantify the impact of individual elements on the overall performance of our method.

Ablation of HOI Optimization Stages. To analyze the three stages of HOI optimization scheme, we quantitatively and qualitatively evaluate the intermediate results on datasets of Arctic, OakInk and DexYCB. As illustrated in Fig. 5, each stage of our optimization pipeline yields incre-

Camera Setup	HOI Contact Alginment	Hand Parameter Refinement	Arctic S.D.↓ Mean	Arctic S.D.↓ Std	Arctic I.V.↓ Mean	Arctic I.V.↓ Std	OakInk S.D.↓ Mean	OakInk S.D.↓ Std	OakInk I.V.↓ Mean	OakInk I.V.↓ Std	DexYCB S.D.↓ Mean	DexYCB S.D.↓ Std	DexYCB I.V.↓ Mean	DexYCB I.V.↓ Std
✘	✘	✘	5.79	6.35	**0.49**	**1.63**	5.81	5.94	**0.47**	**1.82**	6.32	6.29	**0.44**	**1.76**
✔	✘	✘	5.79	6.35	0.49	1.63	5.81	5.94	0.47	1.82	6.32	6.29	0.44	1.76
✔	✔	✘	2.87	2.84	8.36	9.06	3.57	3.66	10.13	11.74	2.93	2.86	9.64	9.75
✔	✔	✔	**2.25**	**2.09**	4.67	4.54	**3.08**	**2.92**	4.11	4.77	**2.43**	**2.41**	4.52	4.64

Table 3. Ablation study for the HOI prior-guided optimization scheme.

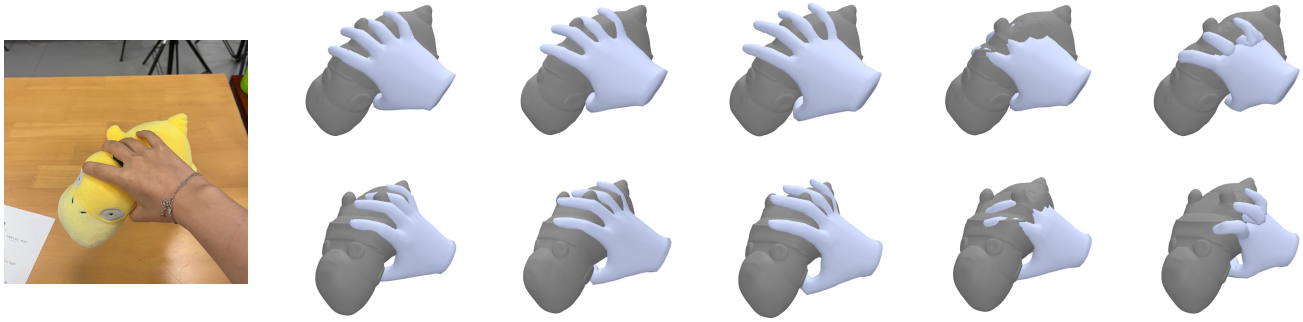


Figure 7. A visualization of the ablation study on loss terms in hand parameter refinement. Each loss term was individually removed from the total loss function, and hand parameter refinement was performed to observe the resulting effects. The top row shows the input viewpoint, while the bottom row provides an alternative viewpoint to more clearly illustrate the differences.

	S.D.↓ Mean	S.D.↓ Std	I.V.↓ Mean	I.V.↓ Std	MPVPE↓	MPJPE↓
Full losses	**3.08**	**2.92**	**4.11**	4.77	**1.19**	**1.22**
w/o $\mathcal{L}_{\text{hand-mask}}$	3.21	3.69	4.28	**4.69**	1.73	1.68
w/o $\mathcal{L}_{\text{penetr}}$	2.95	3.16	9.62	10.13	1.35	1.33
w/o $\mathcal{L}_{\text{reg}}$	3.24	3.73	4.47	4.82	4.57	4.41
w/o $\mathcal{L}_{\text{contact}}$	3.94	3.81	4.26	4.79	1.30	1.27

Table 4. Ablation study of each loss term in hand parameter refinement on OakInk dataset.

mental improvements over its predecessor. *1) Camera System Setup:* Before optimization, imprecise camera parameters lead to misaligned object and hand positions, with grasp contacts deviating from the image. The first stage aligns the rendered object with the input image, however it leaves hand inconsistent with the image. *2) HOI Contact Alignment:* The second stage optimizes hand to achieve both precise image alignment and appropriate contacts, but introduces significant hand-object intersections. *3) Hand Parameter Refinement:* The final stage ensures physical plausibility and stability of the grasping by eliminating penetrations while preserving fingertip contacts. The quantitative results of Tab. 3 further validates the conclusion that the whole pipeline yields the best results.

Ablation of Loss Terms in Hand Parameter Refinement. To further examine Hand Parameter Refinement, we conducted an ablation study on the loss terms, with results shown in Fig. 7 and Tab. 4. Both qualitative and quantitative evaluations offer insights into the significance of each loss term. Although $\mathcal{L}_{\text{hand-mask}}$ was optimized in a prior step, its removal still increases simulation displacement and introduces deviations of hand shape from the input image. Additionally, removing $\mathcal{L}_{\text{penetr}}$ results in more intersections, highlighting its essential role in maintaining physical realism. Without $\mathcal{L}_{\text{reg}}$, we observe twisted fingers, higher simulation displacement, and larger intersection volumes, which reduce grasp fidelity. Lastly, omitting $\mathcal{L}_{\text{contact}}$ positions the fingertips farther from the object surface and increases simulation displacement, underscoring its critical role in preserving grasp stability. These findings confirm that each loss term is crucial for achieving physically plausible and stable hand-object interactions.

5. Conclusion

In this paper, we explore the use of large-scale models to reconstruct HOI from single-view RGB images. We start by initial reconstruction of hand and object and propose a prior-guided optimization scheme that ensures physical plausibility while preserving the visual fidelity of the input image. Extensive experiments on both public datasets and our own collected data validate the effectiveness of our approach, demonstrating superior generalization capabilities and underscoring the potential of large-scale models in HOI reconstruction. As our method relies heavily on the adopted large models, it is susceptible to unsatisfactory results when these models fail. In future work, we aim to enhance efficiency and robustness by exploring how to improve large models to specifically address the unique requirements of hand-object interaction tasks.

6. Acknowledgements

This work was supported by NSFC (No.62206173), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (KLIP-HuMaCo). The authors thank Qi Jiang and Yaxun Yang (ShanghaiTech University) for their helpful discussions on this project.

References

[1] Marco Attene. A lightweight approach to repairing digitized polygon meshes. *The visual computer*, 26:1393–1406, 2010. 2

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 3

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 3

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 361–378. Springer, 2020. 2

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 2

[6] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *CVPR*, 2022. 2

[7] etc. Chao, Yu-Wei. Dexycb: A benchmark for capturing hand grasping of objects. In *Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[9] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30: 4008–4021, 2021. 2

[10] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 2, 6, 7

[11] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-Driven signed distance functions for 3D hand-object reconstruction. In *CVPR*, 2023. 2, 6, 7

[12] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 4

[14] George Fahim, Khalid Amin, and Sameh Zarif. Single-view 3d reconstruction: A survey of deep learning methods. *Computers & Graphics*, 94:164–190, 2021. 2

[15] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 2

[17] Kui Fu, Jiansheng Peng, Qiwen He, and Hanxiao Zhang. Single image 3d object reconstruction based on deep learning: A review. *Multimedia Tools and Applications*, 80:463–498, 2021. 2

[18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2

[19] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018.

[20] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 2

[21] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. In *ICLR*, 2021. 2

[22] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010. 2

[23] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 2

[24] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 5

[25] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 2

[26] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3

[27] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022. 2

[28] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European conference on computer vision (ECCV)*, pages 118–134, 2018. 3

[29] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[30] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, 2019. 2

[31] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[32] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[33] Sixu Li, Chaojian Li, Wenbo Zhu, Boyang Yu, Yang Zhao, Cheng Wan, Haoran You, Huihong Shi, and Yingyan Lin. Instant-3d: Instant neural radiance field training towards on-device ar/vr 3d reconstruction. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–13, 2023. 2

[34] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020. 2

[35] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023. 3

[36] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 3

[37] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 2

[38] Tengyu Liu, Zeyu Liu, Ziyuan Jiao, Yixin Zhu, and Song-Chun Zhu. Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator. *IEEE Robotics and Automation Letters*, 7(1): 470–477, 2021. 6

[39] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3

[40] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. *CVPR*, 2024. 2, 3

[41] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–59, 2018. 3

[42] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4):1–13, 2019. 3

[43] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3, 2

[44] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[45] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018. 3

[46] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 2, 3

[47] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017. 2

[48] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3

[49] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *2010 IEEE International Conference on Robotics and Automation*, pages 458–463. IEEE, 2010. 2

[50] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 3

[51] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 2

[52] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2

[53] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 3

[54] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv*, 2023. 3

[55] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 2

[56] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2015. 3

[57] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 294–310. Springer, 2016. 2

[58] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv preprint arXiv:2402.05054*, 2024. 3

[59] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019. 2

[60] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 2

[61] Tripo3D. Tripo3d: Generate 3d design powered by ai in one click, within seconds, 2024. 2, 7, 3

[62] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–737, 2015. 2

[63] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016. 2

[64] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3

[65] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. *arXiv preprint arXiv:2403.05034*, 2024. 3

[66] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3

[67] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 2, 3

[68] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[69] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 2, 6, 7

[70] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 2

[71] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023. 3, 2

[72] Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2024. 2, 6, 7

[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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 34–51. Springer, 2020. 2

[74] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 3

[75] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 3

EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild

Supplementary Material

This supplementary material provides additional details on our method and results that complement the main paper. Sec. 7 details the segmentation and reconstruction of the object. Sec. 8 elaborates on the HOI contact alignment stage of the HOI optimization process. Finally, Sec. 9 presents further experiments and analyses, demonstrating the robustness and versatility of our method.

7. Initial Reconstruction of Hand and Object

7.1. Hand-Object Interaction Reasoning

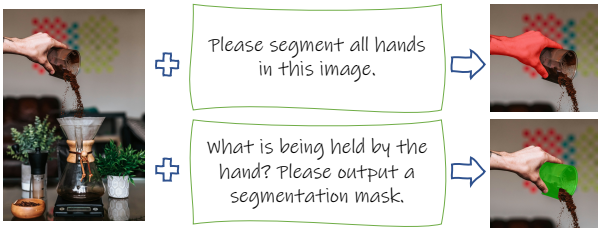


Figure 8. Given an input image, we use predefined prompt to reason the segmentation of hand and object.

Before reconstructing the Hand-Object interaction, we first need to identify the region of interest, specifically, the area in the input image where the object is in interaction with the hand. This is a challenging task, as many in-the-wild images contain multiple objects, but only one is being actively interacted with.

Reasoning with Vision-Language Model. Inspired by the recent success of vision-language models in image understanding, we employ LISA, a context-aware segmentation model, to analyze and segment hand-object interactions. As illustrated in Fig. 8, given a single input image, we prompt the LISA model with two queries to obtain segmentation masks for the hand and the object: 1) *"Please segment all hands in this image."*; 2) *"What is being held by the hand? Please provide its segmentation mask."* LISA's visual-language capabilities enable precise segmentation masks for both the hand and its interacting object.

Contour-guided Filtering. Although the LISA model can successfully reason about and segment the hand and the object it interacts with in most cases, we observed there still exist imperfections in the segmentation masks that hinder further processing. As shown in Fig. 9, LISA incorrectly segments redundant masks of objects that are not interacted with hands. This error may arise because the cookies share the same language description and similar visual appearance.

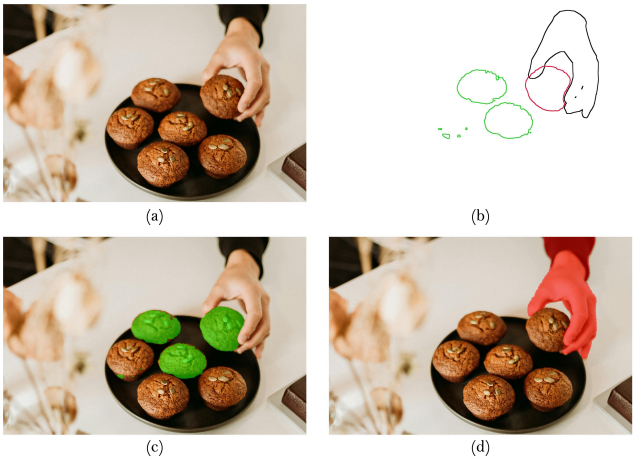


Figure 9. The figure illustrates the segmentation and contour extraction for hand-object interaction analysis. Image (a) is the input image. Image (b) displays the corresponding contours extracted from the object and hand masks. Black contours represent the hand, while red contours highlight the target object parts crucial for HOI understanding. Green contours indicate redundant masks identified for removal, as they do not contribute to the hand-object interaction being analyzed. Image (c) and (d) depict the segmented object and hand masks.

To address the issue mentioned above, we propose a contour-guided filtering strategy. Specifically, we first extract the contours of the hand and all segmented objects. If an object is being interacted with, its contour should be adjacent to the hand's. Based on this assumption, we discard objects whose contours are not neighboring the hand's. This approach enables us to accurately obtain the segmentation mask for the objects that hands interacts with.

7.2. Object Reconstruction

Here we present details on how to reconstruct the object from input image. First we remove occlusions from the image, then re-segment the complete object image, and finally generate the corresponding object mesh using this object image.

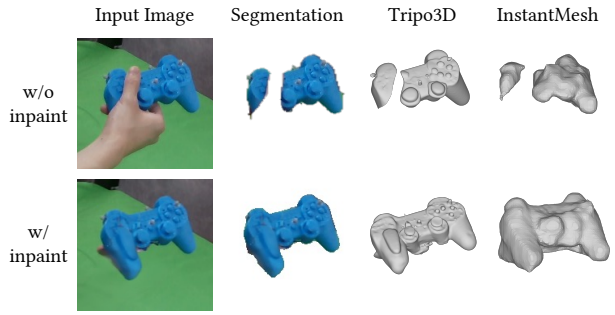


Figure 10. We conducted a comparative analysis of reconstruction results between original images and those subjected to inpainting. The top row displays results from the original image, while the bottom row presents results obtained from images after applying the inpainting process.

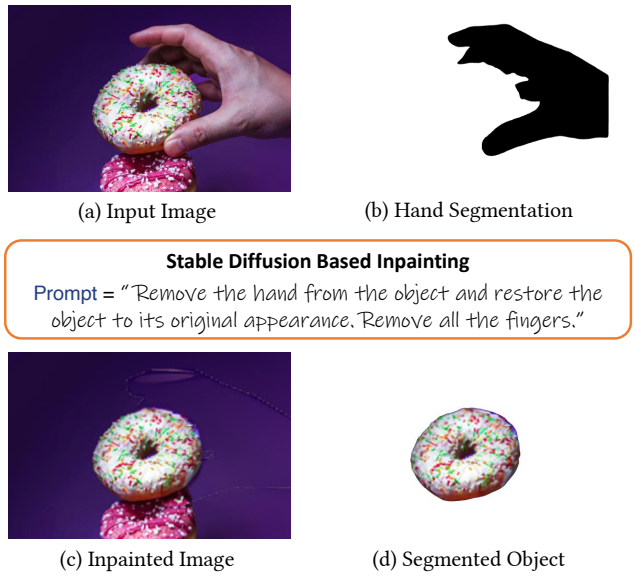


Figure 11. Illustration of the inpainting process. Given an input image containing a hand and a corresponding hand mask, a text-guided diffusion model effectively removes the hand from the image and inpaints the masked region.

Object occlusion removal via image inpainting. Since objects interacting with hands are often partially occluded in the image, directly using the original input to reconstruct the object's 3D geometry can result in distorted and incomplete shapes. To obtain a more accurate 3D geometry, we first use a diffusion model [43, 71] to recover the complete appearance of the object in 2D image.

As illustrated in Fig. 11, we employ a stable diffusion model for object inpainting, using the input image and hand mask alongside a tailored text prompt. The hand mask identifies regions requiring inpainting, while the text prompt guides the reconstruction of the object's original appearance. Thanks to the robust generalization capabilities of stable diffusion, this inpainting approach successfully synthesizes the occluded object regions across diverse scenarios, producing photorealistic results.

Re-segment from Inpainted Image. With the inpainted image, we utilize a large reconstruction model, InstantMesh, to reconstruct the object's geometry. Since InstantMesh requires a background-free input, we must first obtain the segmentation mask of the inpainted object. To generate this mask, we use the occluded object mask as an indicator. As shown in Fig. 10, the occluded mask typically consists of multiple sub-masks due to the hand separating the object. We randomly sample points within each sub-mask and compute a bounding box that loosely covers the occluded mask. These sampled points and the bounding box serve as prompts for the SAM model, which extract the object from the inpainted image. Finally, InstantMesh takes the completed object as input and reconstructs its geometry.

Watertight Post-processing. In hand-object interactions, mutual occlusions naturally occur and are intrinsically linked to contact relationships. The reconstructed meshes from the LRM are sometimes non-watertight, which can hinder robust and accurate hand-pose optimization. To address this, we convert the non-watertight meshes into watertight ones when needed. For a non-watertight mesh, we first render depth maps from multiple viewpoints that cover the entire object. These depth maps are then fused into a unified point cloud, which helps eliminate isolated and occluded parts. Next, we apply the Poisson reconstruction method to generate a mesh from the point cloud. Finally, a hole-filling algorithm [1] is used to ensure the mesh meets the watertight requirement.

8. Hand-Object Interaction Optimization

HOI Contact Alignment. We identify potential contact regions by analyzing two types of hand-object overlaps in the input image. For front-side contacts, where the object is occluded by the hand, we compute the contact mask $M_{\text{front}} = \hat{M}_o \setminus M_o$ as the difference between the inpainted object mask $\hat{M}_o$ and the original object mask M_o. For back-side contacts, where the hand is occluded by the object, we derive the contact mask $M_{\text{back}} = \hat{M}_h \setminus M_h$ as the difference between complete hand mask $\hat{M}_h$ and the segmented hand mask M_h. The complete hand mask is obtained by rendering on the pose and camera parameters estimated by HaMeR.

From the contact masks M_{front} and M_{back}, we recover 3D contact points via ray-casting to hand and object geometries seperately. As shown in the Fig. 4 in the main paper, we can emit a ray from each pixel on contact masks to hit the

reconstructed object and hand. Through the application of rasterization and depth peeling techniques, we extract multiple depth values from different layers of the 3D models. In our implementation, we utilize four depth layers, which we have empirically found to be sufficient for all test cases in our experiments.

For ray-object intersections, we select the minimum depth values within M_{front} and maximum depth values within M_{back}, corresponding to the nearest and farthest points from the camera respectively.

Regarding the ray-hand intersection, it is important to note that the functional area for grasping is limited to the palmar surface. The dorsal side of the hand, comprising the back of the hand and fingers, is not involved in object manipulation. We manually select and label faces corresponding to the palmar and dorsal regions on the MANO template model as a preprocessing step. This anatomical annotation serves as prior knowledge, allowing us to efficiently exclude 3D points located on the dorsal side. Therefore, valid hand contact points are determined for each pixel in M_{front} and M_{back} by filtering ray intersections based on face indices to retain only palmar-side points, then selecting the nearest and farthest intersections based on depth value.

Once all potential contact points on both the hand and the object are identified, we apply the Iterative Closest Point (ICP) method to compute the optimal hand translation, aligning the contact points and providing a rough estimation of the hand's pose.

9. Experiments

9.1. Running Time Analysis

We measured the processing time for 65 images at each stage and calculated the average runtime. The initial reconstruction stage required an average of 58.97 seconds, with the breakdown as follows: HOI Reasoning (11.68 seconds), Hand Reconstruction (5.54 seconds), Object Inpainting (15.55 seconds), and Object Reconstruction (26.20 seconds). For the optimization stage, the average runtime was 57.03 seconds, consisting of Camera Setup (4.41 seconds), Contact Alignment (30.51 seconds), and Hand Refinement (22.11 seconds).

9.2. Ablation on Large Models

The Selection of HOI Reasoning Models. In addition to the vision-language model, we examine the Hand-Object detector (HODet)[53] for segmenting hands and objects in input images. To compare HOI reasoning performance, we evaluate LISA against HODet on the Arctic dataset. As HODet predicts only object bounding boxes, we employ SAM to generate object segmentations based on these predictions. LISA outperforms HODet, achieving an average IoU of 0.74 compared to 0.61. A visual comparison in Figure 12 shows that HODet frequently misidentifies objects and detects extraneous elements.

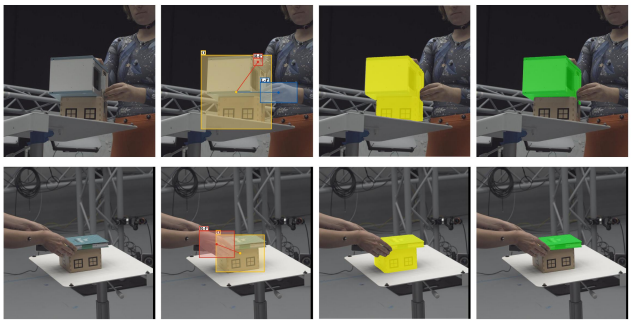


Figure 12. Visualization of the ablation on HOI reasoning large model.

The Selection of Large Reconstruction Models. While our pipeline incorporates the open-source model InstantMesh for object reconstruction, it could significantly benefit from a more advanced model. For comparison, we employ the state-of-the-art commercial model Tripo3D [61]. Fig. 19 displays the reconstructed meshes produced by both approaches on a range of challenging in-the-wild images. This comparison highlights the potential of our approach to combine the strengths of multiple large-scale models to achieve highly accurate object reconstruction across diverse scenarios.

To assess the impact of reconstruction quality, we evaluate the optimization results using InstantMesh reconstruction, Tripo3D reconstruction, and ground truth (GT) meshes on Oakink dataset. As shown in the table below, lower-quality reconstructions (indicated by higher Chamfer Distance (C.D.) values) result in poorer HOI performance.

	Object quality			HOI results		final hand results	
	F5 ↑	F10 ↑	$C.D.$ ↓	S.D.↓	I.V.↓	MPVPE↓	MPJPE↓
GT	**1.00**	**1.00**	**0.00**	**1.92**	**2.44**	**0.86**	**0.77**
Tripo3D	0.280	0.503	0.842	2.76	4.05	1.15	1.21
InstantMesh	0.247	0.445	1.035	3.08	4.11	1.19	1.22

Table 5. The impact of object reconstruction quality on HOI optimization performance.

9.3. Qualitative Comparison Results

Comparative results on In-th-wild Images As depicted in Figure 13, we showcase qualitative comparisons with IHOI, AlignSDF, gSDF, MOHO, and our method on in-the-wild images. Our approach excels in accurately reconstructing intricate object geometries and details.

Additional comparative results on public datasets. Here we provide additional comparative results on public

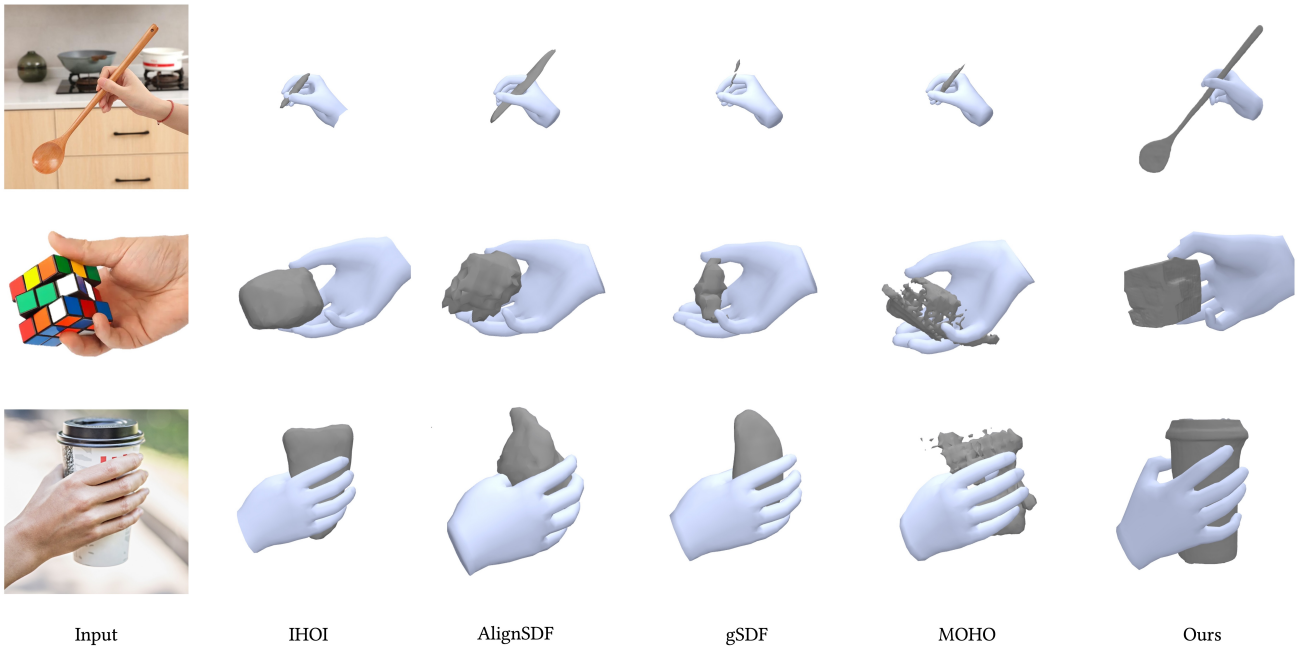


Figure 13. Qualitative Comparison of IHOI, AlignSDF, gSDF, MOHO, and Our Method on in-the-wild Images

datasets. Fig. 16 demonstrates comparisons with IHOI and MOHO on the OakInk dataset, while Fig. 17 and Fig. 18 show our method's performance on the Arctic and DexYCB datasets, respectively. These additional examples demonstrate our method's performance across diverse scenarios.

9.4. Failure Cases Analysis

Most failures originate from the initial reconstruction stage rather than the optimization stage. As shown in Figure 14, the inpainting model occasionally introduces artifacts, causing the object being held to blend with the background and making segmentation challenging. Consequently, the input images for the large reconstruction model become incoherent and deviate significantly from real-world objects. We believe that incorporating more advanced inpainting models and leveraging hand contours in the segmentation process are promising directions for future exploration.

We also evaluated our pipeline on the MOW dataset. The images in this dataset are of relatively low quality, frequently displaying ambiguous grasping poses and motion blur. These factors present significant challenges for both inpainting and object reconstruction. Figure 15 illustrates typical examples of failure cases observed on the MOW dataset.

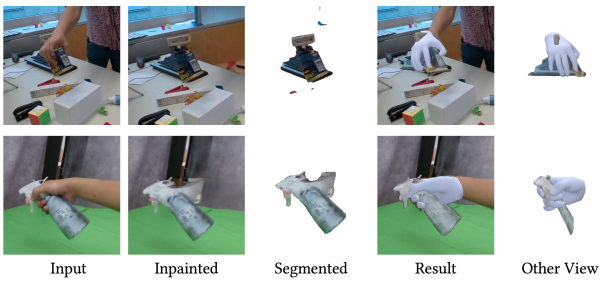


Figure 14. Examples of failure Cases in the OakInk dataset. Failures primarily result from artifacts introduced by the inpainting model.

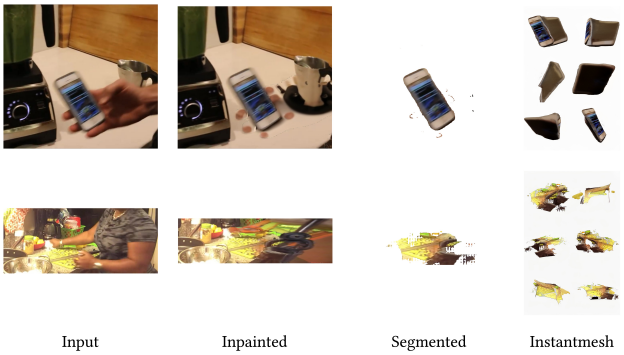


Figure 15. Examples of failure cases on the MOW dataset. The last column is the synthesized 6 views from InstantMesh, it shows that motion blur and ambiguous grasping present significant challenges for object reconstruction.

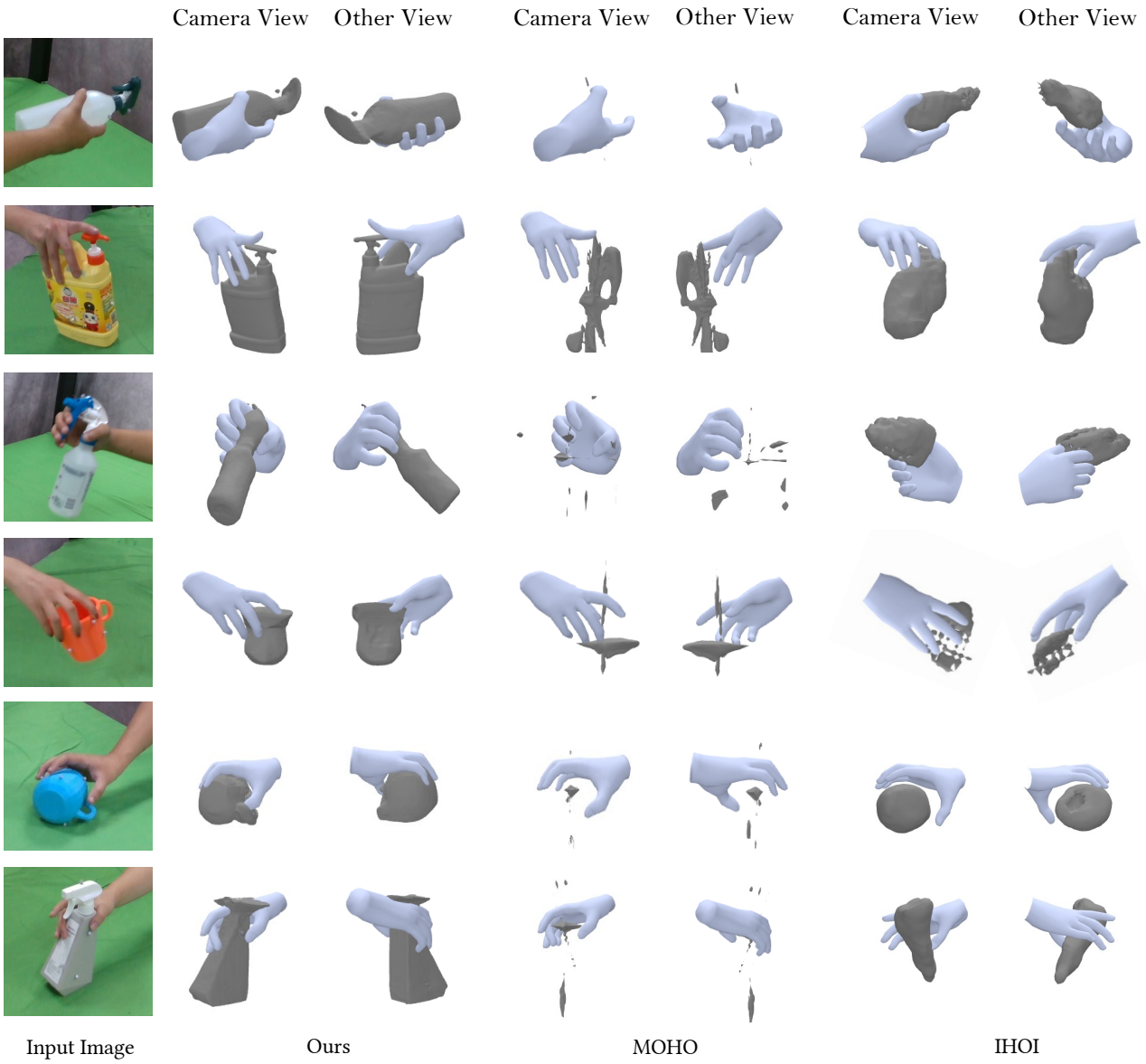


Figure 16. This gallery showcases the outcomes of our hand-object reconstruction on the dataset OakInk. The first column is the input image, we present the camera view and another view to display the reconstructed HOI meshes.

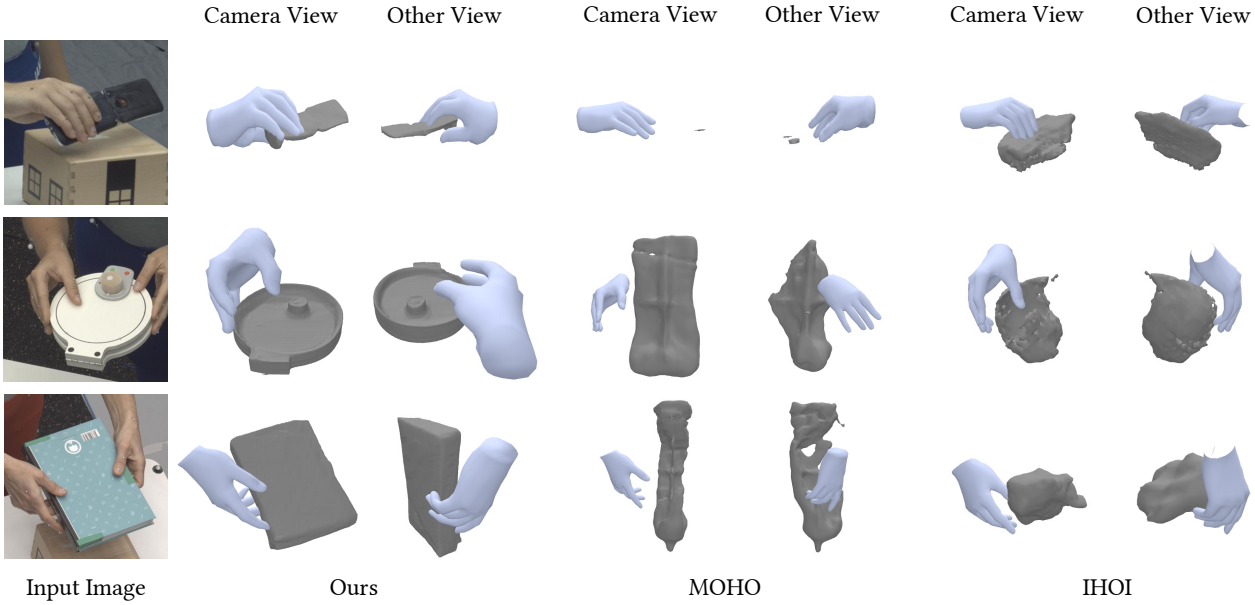


Figure 17. This gallery showcases the outcomes of our hand-object reconstruction on the dataset Arctic.

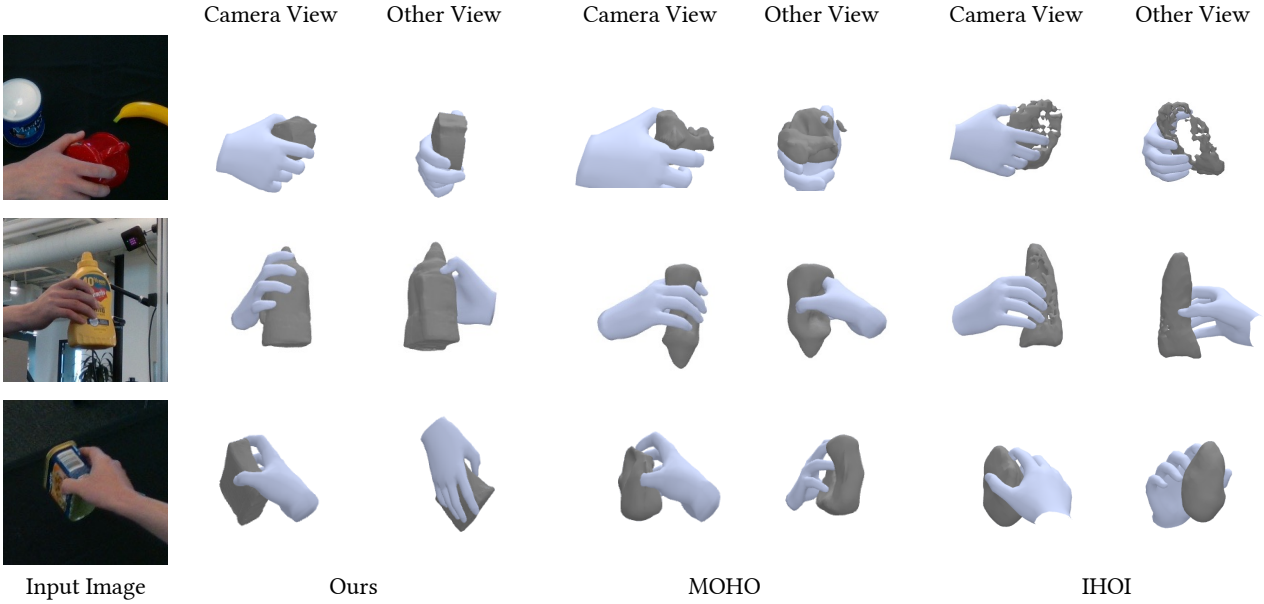


Figure 18. This gallery showcases the outcomes of our hand-object reconstruction on the dataset DexYCB.

Figure 19. This gallery showcases the outcomes of our hand-object reconstruction results on in-the-wild images, we test the reconstruction result on two LRM, instantmesh and tripo3d.