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

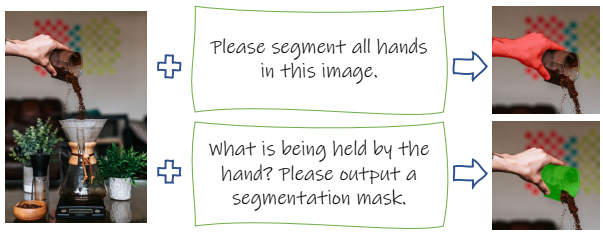


Figure 8. Given an input image, we use predefined prompt to reason the segmentation of hand and object.

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.

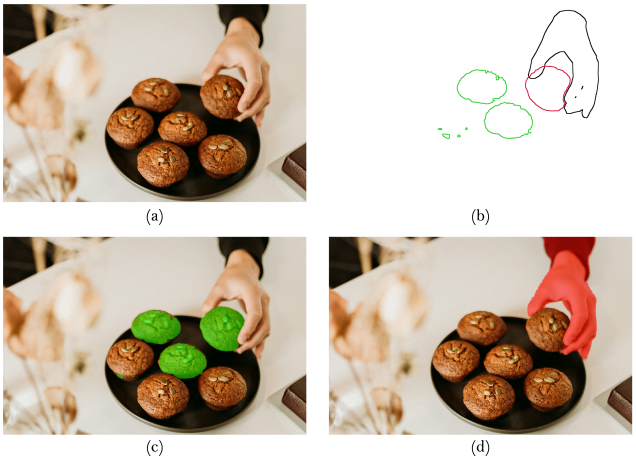


Figure 9. The figure illustrates the segmentation and contour extraction for hand-object interaction analysis. Image (a) is the input image. Image (b) displays the corresponding contours extracted from the object and hand masks. Black contours represent the hand, while red contours highlight the target 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.

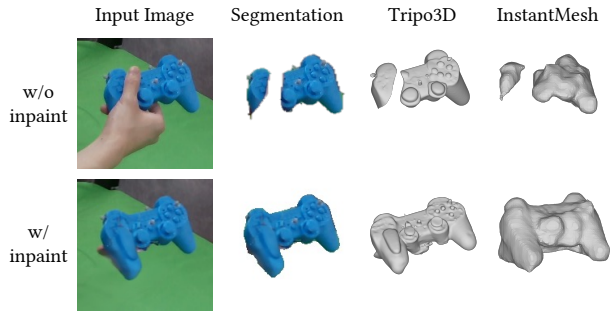
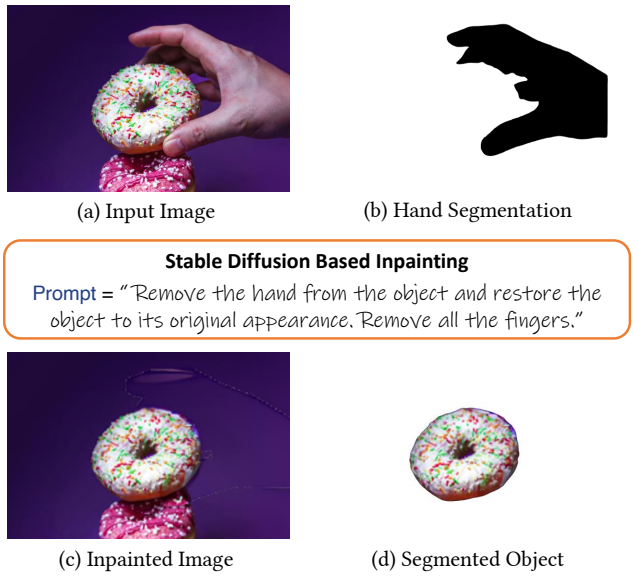


Figure 10. We conducted a comparative analysis of reconstruction results between original images and those subjected to inpainting. The top row displays results from the original image, while the bottom row presents results obtained from images after applying the inpainting process.



(a) Input Image (b) Hand Segmentation

(c) Inpainted Image (d) Segmented Object

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.

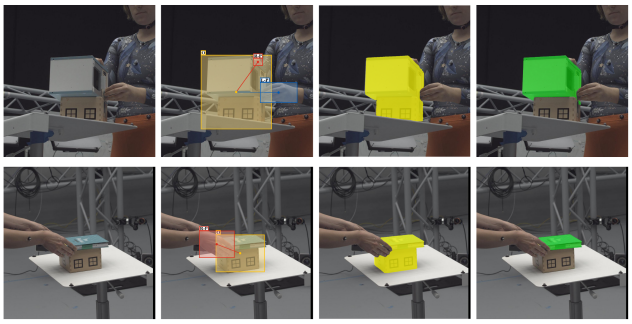


Figure 12. Visualization of the ablation on HOI reasoning large model.

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

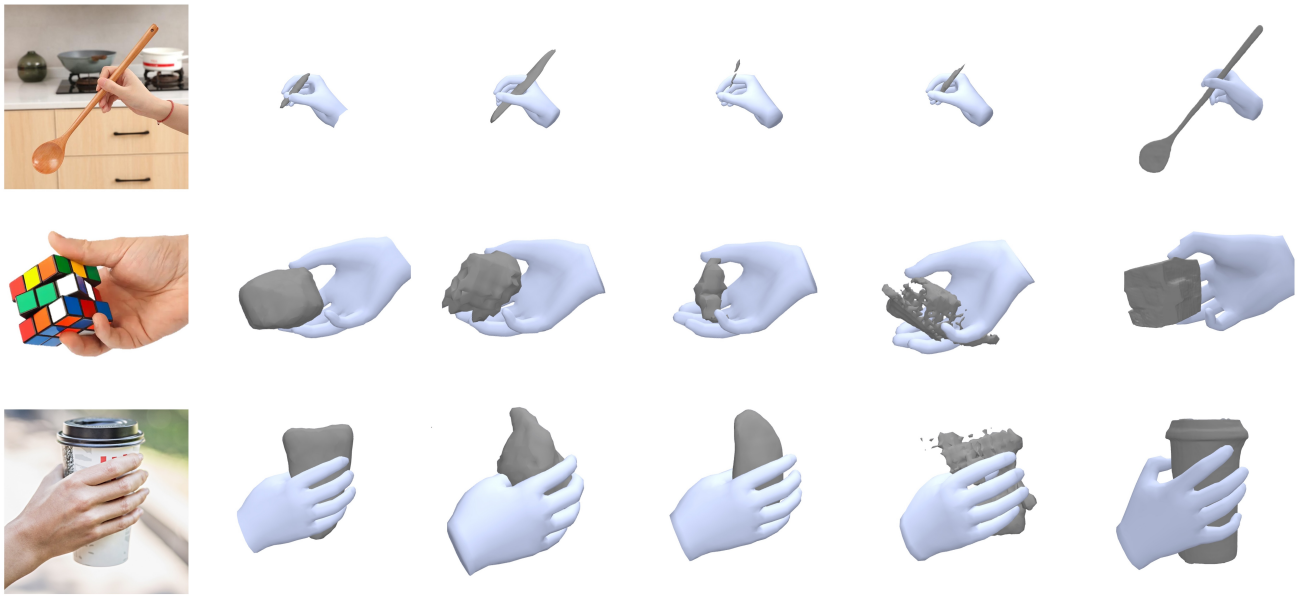


Figure 13. Qualitative Comparison of IHOI, AlignSDF, gSDF, MOHO, and Our Method on in-the-wild Images

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.

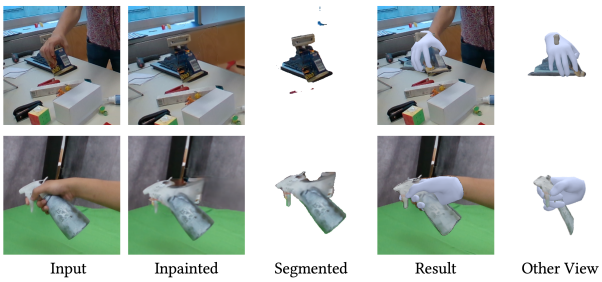


Figure 14. Examples of failure Cases in the OakInk dataset. Failures primarily result from artifacts introduced by the inpainting model.

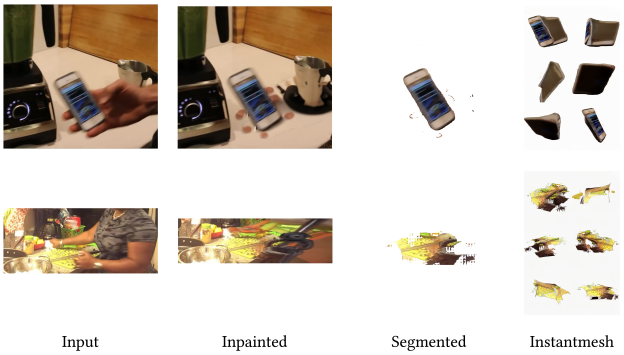


Figure 15. Examples of failure cases on the MOW dataset. The last column is the synthesized 6 views from InstantMesh, it shows that motion blur and ambiguous grasping present significant challenges for object reconstruction.

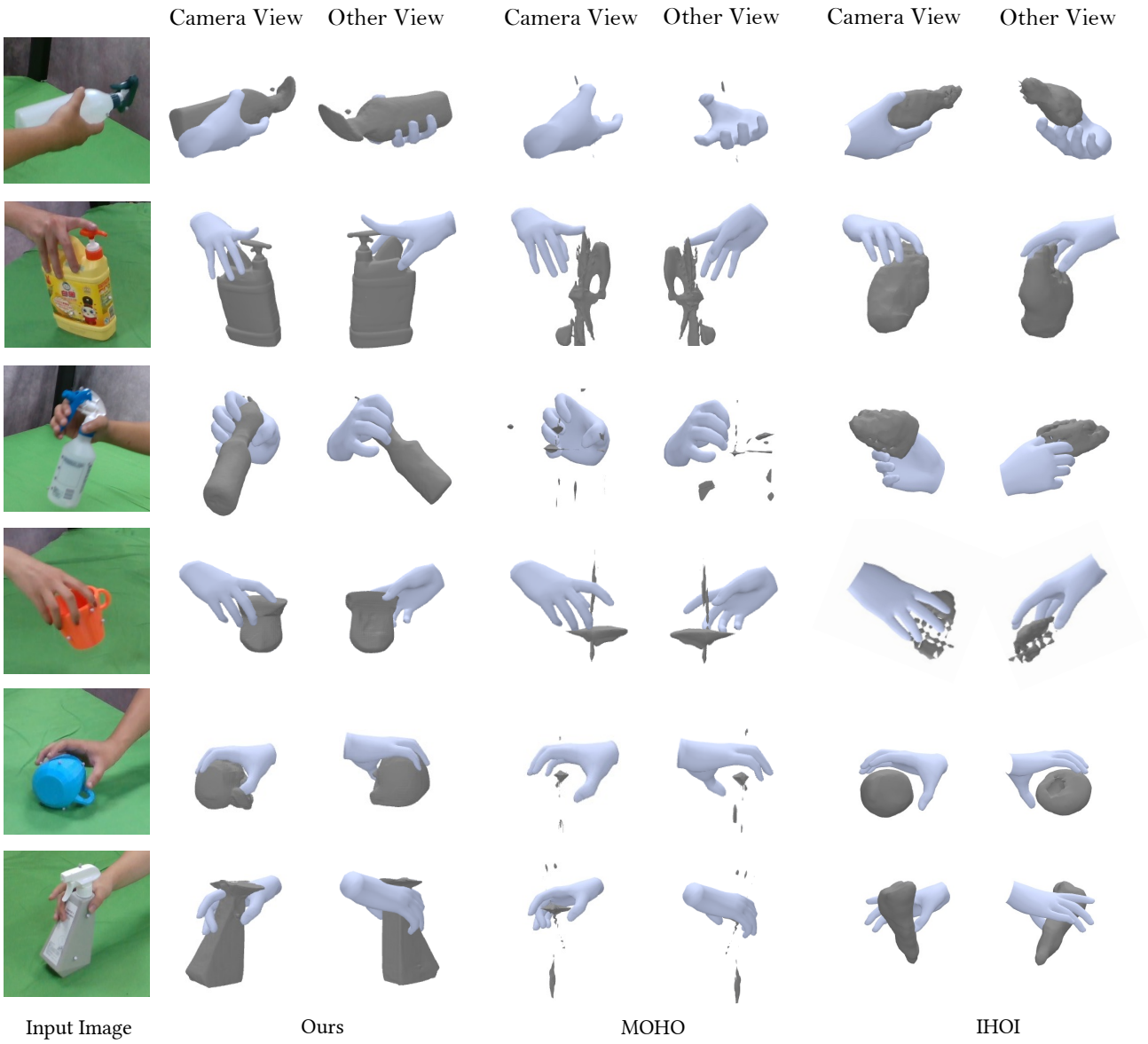


Figure 16. This gallery showcases the outcomes of our hand-object reconstruction on the dataset OakInk. The first column is the input image, we present the camera view and another view to display the reconstructed HOI meshes.

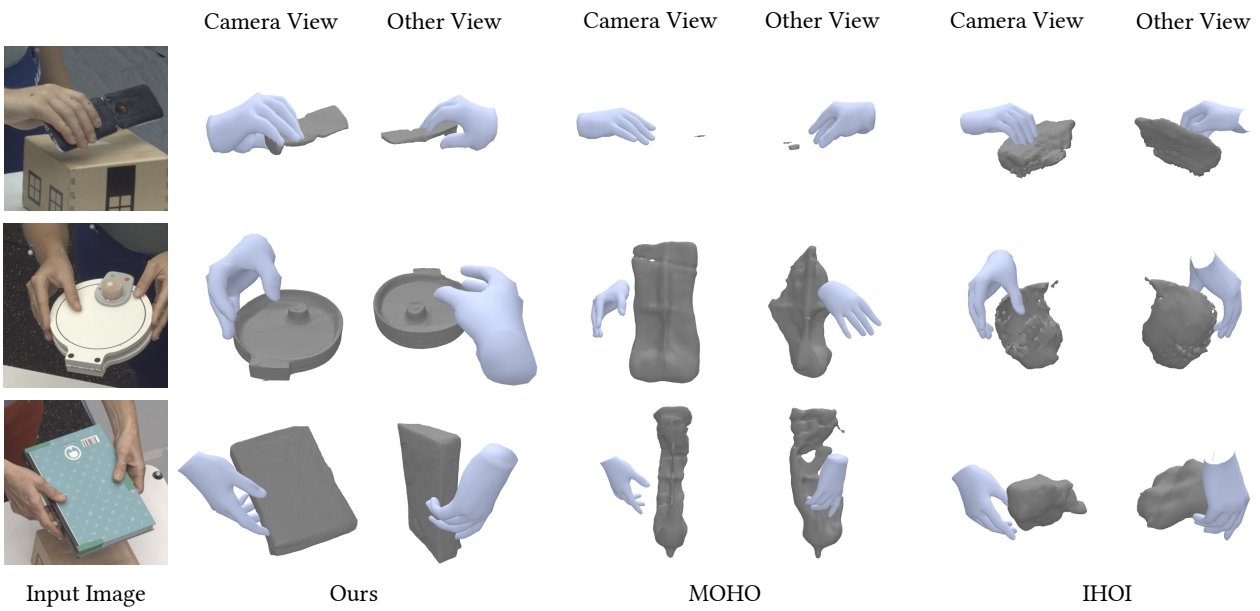


Figure 17. This gallery showcases the outcomes of our hand-object reconstruction on the dataset Arctic.

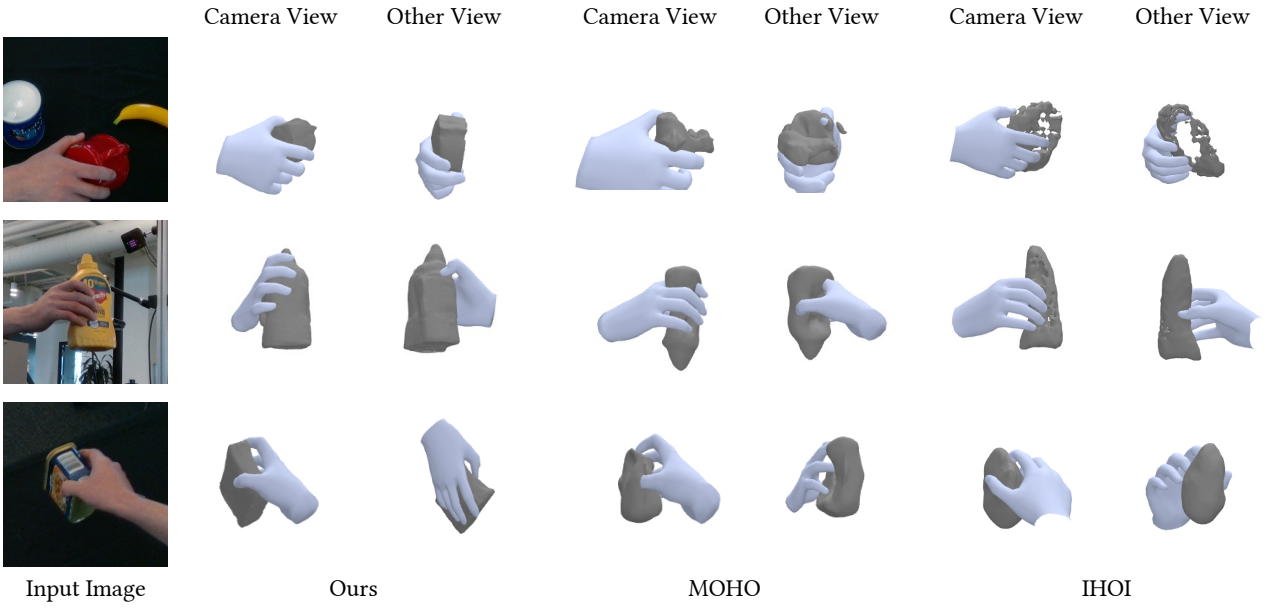


Figure 18. This gallery showcases the outcomes of our hand-object reconstruction on the dataset DexYCB.

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.