HOLD: Category-agnostic 3D Reconstruction of Interacting Hands and Objects from Video

Zicong Fan[1,2] Maria Parelli[1] Maria Eleni Kadoglou[1]
Xu Chen[1,2,†] Muhammed Kocabas[1,2] Michael J. Black[2] Otmar Hilliges[1]
[1]ETH Zürich, Switzerland [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany

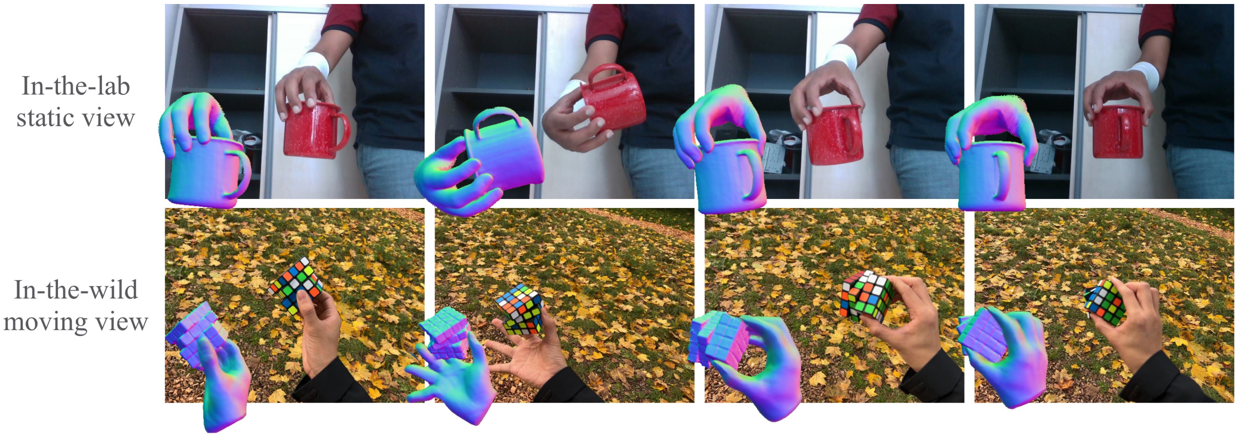

Figure 1. **HOLD:** Given a monocular video sequence of a hand interacting with an unknown object, our method, HOLD, reconstructs high-quality 3D hand and object surfaces in both in-the-lab videos from a static camera and in-the-wild egocentric-view videos. Here we show the input images and the reconstructed surface normals. Best viewed in color.

Abstract

Since humans interact with diverse objects every day, the holistic 3D capture of these interactions is important to understand and model human behaviour. However, most existing methods for hand-object reconstruction from RGB either assume pre-scanned object templates or heavily rely on limited 3D hand-object data, restricting their ability to scale and generalize to more unconstrained interaction settings. To address this, we introduce HOLD – the first category-agnostic method that reconstructs an articulated hand and an object jointly from a monocular interaction video. We develop a compositional articulated implicit model that can reconstruct disentangled 3D hands and objects from 2D images. We also further incorporate hand-object constraints to improve hand-object poses and consequently the reconstruction quality. Our method does not rely on any 3D hand-object annotations while significantly outperforming fully-supervised baselines in both in-the-lab and challenging in-the-wild settings. Moreover, we qualitatively show its robustness in reconstructing from in-the-wild videos. See here for code, data, models, and updates.

1. Introduction

We interact with a diverse set of objects in our everyday lives: We hold our morning cup of coffee; we hold a drill in making home renovations; and we pour cereal from a box. Studies show that on average, we interact with 140 objects per day [45]. To understand, model, and synthesize these interactions [7, 9, 59, 74, 75], it is critical to be able to reconstruct them in 3D. Towards this goal, we tackle the challenging problem of reconstructing diverse 3D objects and the articulated hands holding them from only monocular videos of the hand-object interaction, as shown in Fig. 1.

Most hand-object reconstruction methods assume a pre-scanned object template [12, 23, 24, 68], making it infeasible to scale to in-the-wild scenarios [2]. Other methods do not assume object templates [22, 70], but are trained using datasets with a limited number of objects, leading to poor generalization. Very recently, Ye *et al*. [71] introduced a data-driven prior that is trained on six object categories and they leverage this prior to reconstruct hand and object surfaces from segmentation mask observations. Although

† Work done prior joining Google

they can reconstruct novel objects and articulated hands, their method is limited to these training categories. Another emerging line of work focuses on in-hand object scanning [21, 26, 79] from monocular videos. They adapt multi-view reconstruction techniques to aggregate observations of hand-held objects in multiple rigid poses. While achieving promising reconstruction quality on novel objects, these methods do not consider hand articulation and hence cannot handle more dexterous hand-object interaction.

In this paper, we go beyond prior works to tackle the new task of *category-agnostic interaction reconstruction*. Given a monocular video as input, our method HOLD (Hand and Object reconstruction by Leveraging interaction constraints in three Dimensions) reconstructs hand and object 3D surfaces for every frame without assuming an object template. Our key insight is that hands and objects in interaction provide complementary cues to each other's shapes and poses. For example, when one holds a mug, the hand geometry constrains the possible shape of the mug via contact. Therefore, we jointly model the object and the articulated hand with a compositional neural implicit model.

To jointly reconstruct the hand and object surfaces from a video, HOLD performs initial hand pose estimation via an off-the-shelf hand regressor and object pose estimation with structure-from-motion (SfM). With the initial noisy hand and object poses, we train HOLD-Net, our compositional neural implicit model of an articulated hand, and an object. The model is volumetrically rendered and supervised with auxiliary losses to obtain the 3D hand and object surfaces. After initializing the hand and object shapes by training HOLD-Net, we optimize hand and object poses via interaction constraints. Finally, we use the refined poses to train HOLD-Net for better shape reconstruction.

We empirically show that by jointly modelling the hand and object in this category-agnostic reconstruction setting through interaction constraints, we achieve better reconstruction quality than methods that only consider objects. We quantitatively evaluate our method with an existing hand-object dataset and further show that our method can generalize to both in-the-lab and in-the-wild videos. We also demonstrate generalization to videos captured by a moving camera from both 3rd person and 1st person views with diverse lighting and background conditions.

To summarize our contributions: 1) We present a novel method that accurately reconstructs 3D hand and object surfaces from monocular 2D interaction videos without requiring a pre-scanned object template or pre-trained object categories; 2) We formulate a compositional implicit model that facilitates the disentanglement and the reconstruction of 3D hands and objects; 3) We show that by jointly optimizing hand-object constraints, we can obtain better reconstruction quality than treating the hand and object separately; 4) We evaluate our method both qualitatively and quantitatively for 3D reconstruction, and we demonstrate realistic reconstruction on challenging in-the-wild videos.

2. Related Work

3D hand pose and shape recovery: The field of monocular RGB 3D hand reconstruction has been evolving since the foundational work of Rehg and Kanade [49]. A significant portion of the existing literature is focused exclusively on reconstructing the hand [1, 4, 10, 11, 14, 22, 27, 34, 41, 43, 53–56, 64, 73, 78, 81, 82]. For instance, Zimmermann *et al.* [82] employ a deep convolutional network, implementing a multi-stage approach to achieve 3D hand pose estimation. Ziani *et al.* [81] adopt a self-supervised time-contrastive method to improve in-the-wild generalization. Recently, there are also methods that reconstruct 3D hand poses of strongly interacting hands [18, 31, 33, 34, 39–42, 44]. For example, Fan *et al.* [11] introduce a semantic feature fusion layer to address appearance ambiguities when two hands strongly interact. Tse *et al.* [63] introduce a spectral graph-based transformer for two-hand reconstruction. Unlike these, we focus on hand-object reconstruction.

Hand-object reconstruction: Reconstructing the hand and object in 3D from images and videos is also a well-established research area [8, 15, 22–24, 37, 61, 62, 68, 68, 80]. Most methods in the literature assume an object template and only estimate the hand and object poses [2, 8, 13, 37, 61, 68]. For example, Tekin *et al.* [61] infer 3D control points for both the hand and the object in videos, using a temporal model to propagate information across time. Fan *et al.* [12] estimate articulated object pose with hands in dexterous manipulation. Liu *et al.* [37] devise a semi-supervised learning approach by first constructing pseudo-groundtruth in hand-object interaction videos based on temporal heuristics and train the model with the new annotation. Yang *et al.* [68] introduce a contact potential field for better hand-object contact. Despite accurate object pose estimation quality, it is hard to generalize such work to novel objects and in-the-wild videos because it requires known object templates. There are methods that do not assume an object template by training on 3D hand-object data [5, 22, 70]. Unfortunately, these methods have poor generalization ability due to limited 3D hand-object data. Recently, there are more generalizable approaches [26, 47, 48, 58, 71] with differentiable rendering and data-driven priors. However, they require either the hand to be rigid when interacting with objects [26, 47], multi-view observations [48], or category-level hand-object supervision [71]. In contrast, ours allows articulated hands, only requires monocular videos, and is category-agnostic.

In-hand object scanning: There has been increasing interest in in-hand object scanning. The goal of this task is to reconstruct the canonical 3D object shape from a video of a human interacting with an object; the hand is often not

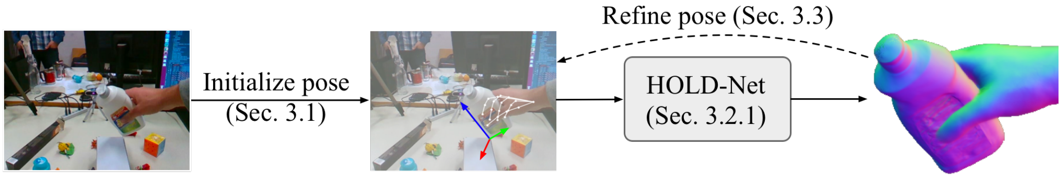

Figure 2. **Method overview**. For each image in a video, our method, HOLD, first initializes the hand and object poses using off-the-shelf estimators. Then we briefly pre-train HOLD-Net, a compositional implicit signed distance field to learn hand and object shapes. The learned shapes of HOLD-Net are then used to refine poses with hand-object interaction constraints. Finally, we use the refined poses to fully train HOLD-Net to learn accurate 3D geometries of hand and object.

reconstructed. For example, early work, such as Tzionas *et al*. [65], leverages hand motion as a prior for object scanning. Recently, BundleSDF [67] estimates the object pose with the help of sequential RGBD images and simultaneously reconstructs the implicit surface defined by a Signed Distance Field (SDF). HHOR [26] also employs SDFs for object surface representation but distinguishes itself by concurrently reconstructing both the object and the hand, assuming the object is securely gripped. Hampali *et al*. [21] propose a novel approach, incorporating a camera trajectory alignment technique and utilizing volumetric rendering for enhanced object surface reconstruction. Very recently, Zhong *et al*. [79] introduce a global coloring and relighting network that significantly improves texture extraction during the object scanning process. In contrast to our work, the methods above do not reconstruct hands with articulation and mainly focus on capturing the object's canonical shape.

3. Method: HOLD

Figure 2 summarizes our method, HOLD, for reconstructing hand-object surfaces from a monocular RGB video. To achieve this, HOLD first initializes hand and object poses (Sec. 3.1) for each frame in a video. Then we use the poses to train HOLD-Net (Sec. 3.2), a compositional implicit signed distance field for hand and object shapes. With the learned shapes, we refine hand-object poses using interaction constraints (Sec. 3.3). Finally, with the refined poses we fully train HOLD-Net (Sec. 3.4), resulting in accurate 3D hand-object geometry.

3.1. Pose initialization

For each frame, to obtain hand poses $\theta \in \mathbb{R}^{48}$ (including global orientation), shape β, and translation $\mathbf{t}_h \in \mathbb{R}^3$, we use an off-the-shelf hand pose estimator [35]. Estimating object pose is more challenging because our approach is category-agnostic and existing category-level object pose estimators are unsuitable for out-of-category objects [3, 66]. Consequently, we first create object-only images for each video using object pixels with an off-the-shelf segmentation network [29]. We then use HLoc [51, 52] to perform structure-from-motion (SfM) to obtain a point cloud defining the object and its rotation $\mathbf{R}_o \in SO(3)$ and translation $\mathbf{t}_o \in \mathbb{R}^3$ for each frame. Since SfM only reconstructs point clouds up to a scale, to align the hand and object in the same space and to estimate the object scale $s \in \mathbb{R}$, we perform a simple optimization procedure that encourages hand-object contact while enforcing the 2D reprojection of hand joints and the object point cloud to match with the original 2D projection. This optimization updates the hand and object translation $\{\mathbf{t}_h, \mathbf{t}_o\}$ for each frame, the hand shape β, and an object scale s. For a detailed explanation, see SupMat.

3.2. HOLD-Net training

3.2.1 HOLD-Net

Figure 3 outlines HOLD-Net, our compositional neural implicit model. In detail, we represent the hand and object surfaces as two neural representations that can be volumetrically rendered into an RGB image. Following [17], we use a time-dependent NeRF++ [77] to model the dynamic background. HOLD shares the time-independent canonical 3D geometries for the hand and the object across frames. Thus, if an object region is occluded in one frame, the region can be observed from another non-occluded frame.

Hand model: We model the hand as an implicit network, driven by MANO pose θ, and translation $\mathbf{t}_h$ [50]. To model the hand shape and appearance in canonical space, we use a signed distance and texture field parameterized by a multi-layer perception (MLP):

$$f_h : \mathbb{R}^3 \to \mathbb{R} \times \mathbb{R}^3 \quad (1)$$

$$\mathbf{x} \mapsto d, \mathbf{c}, \quad (2)$$

where the MLP f_h, with learnable parameters ψ_h, takes in a canonical point $\mathbf{x}$, and predicts its signed distance values to the hand surface d and color $\mathbf{c}$.

To determine the signed distance and color in the deformed observation space, we map points in the observation space $\mathbf{x}'$ back to the canonical space using inverse Linear Blend Skinning (LBS):

$$\mathbf{x} = (\textstyle\sum_{i=1}^{n_b} w_i(\mathbf{x}') \cdot \mathbf{B}_i)^{-1}\mathbf{x}', \quad (3)$$

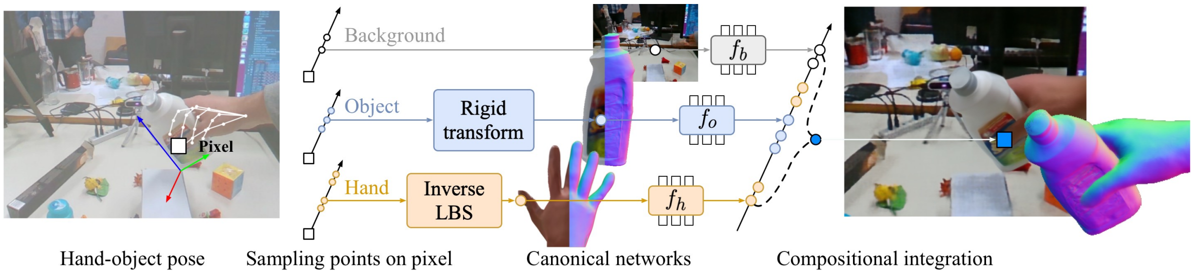

Figure 3. **HOLD-Net**. Given as input hand and object poses and a query pixel, HOLD-Net determines the pixel color in the following steps: 1) HOLD-Net first samples points along the ray independently for object, hand, and background using error-bounded sampling; 2) These sampled points in the observation space are then mapped to the canonical space via rigid transformation for the object and inverse linear blend skinning for the hand; 3) The SDF and color values for the sampled points are queried from the canonical hand, object, and background networks; 4) All object, hand, and background points are merged by sorting via their z-values, and their color and density values are integrated to determine the pixel color. Images on the right are the rendered RGB and normal images.

where $\{\mathbf{B}_i\}_{i=1,\ldots,n_b}$ are the bone transformations derived from θ with forward kinematics, and $\{w_i(\mathbf{x}')\}_{i=1,\ldots,n_b}$ are the skinning weights of each deformed point determined by averaging the skinning weights of the K-nearest vertices of the MANO model [50] weighted by the distance.

Object model: Similar to the hand model, our object model is driven by the relative object scale s, rotation $\mathbf{R}_o$ and translation $\mathbf{t}_o$ between the canonical and deformed space respectively. The object canonical shape and texture are modelled via a neural signed distance and texture field f_o, with learnable parameters ψ_o:

$$f_o : \mathbb{R}^3 \times \mathbb{R}^{n_{z,o}} \to \mathbb{R} \times \mathbb{R}^3 \tag{4}$$

$$\mathbf{x}, \mathbf{z_o} \mapsto d, \mathbf{c}, \tag{5}$$

where $\mathbf{z_o} \in \mathbb{R}^{n_{z,o}}$ of dimension $n_{z,o} = 32$ is an optimizable time-dependent latent code to model the changing object appearance due to varying pose, occlusion and shadows.

To determine the signed distance and color of the object in the deformed observation space, we map points in the observation space $\mathbf{x}'$ back to the canonical space using a simple rigid transformation:

$$\mathbf{x} = (s\mathbf{R}_o)^{-1} \cdot (\mathbf{x}' - \mathbf{t}_o). \tag{6}$$

Background: Following [17, 69], we define a bounding sphere of the foreground scene, in our case the hand and the object. For a given sample $\mathbf{x}'$ outside the bounding sphere, the signed distance and color are predicted by a background network with learnable parameters ψ_b:

$$f_b : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^{n_{z,b}} \to \mathbb{R} \times \mathbb{R}^3 \tag{7}$$

$$\mathbf{x}, \mathbf{v}, \mathbf{z_b} \mapsto d, \mathbf{c}, \tag{8}$$

where $\mathbf{v} \in \mathbb{R}^3$ is the viewing direction and $\mathbf{z} \in \mathbb{R}^{n_{z,b}}$ of dimension $n_{z,b} = 32$ is an optimizable latent code with distinct value for each frame to model dynamic backgrounds. Since we are only interested in modelling hands and objects, and images of human interaction often include other parts of the body, this model is also used to explain partial observation of the human body as part of the changing background. Following NeRF++ [77], we use their inverted sphere parametrization in our background model. For more details, we refer readers to SupMat.

Compositional volumetric rendering: Following [69], to convert hand and object Signed Distance Function (SDF) values to density σ for volume rendering, we use the cumulative distribution function of the scaled Laplace distribution, denoted as $\Gamma_{\alpha_1,\alpha_2}(s)$, where $\alpha_1, \alpha_2 > 0$ are optimizable. More details can be found in [69].

For each frame, to render the foreground, *i.e.*, the hand and the object, we first sample points along the corresponding ray $\mathbf{r}$ parameterized by a camera center $\mathbf{o}$ and a viewing direction $\mathbf{v}$ using error-bounded sampling [69]. We sample n points for the hand $\{\mathbf{x}'\}^h_{i=1,\ldots,n}$, transform them to canonical space using inverse LBS, and query their opacity and color values $\{(\sigma_i, \mathbf{c}_i)\}^h_{i=1,\ldots,n}$ from the canonical hand model f_h. Similarly for the object, we sample n points $\{\mathbf{x}_i\}_{i=1,\cdots,n}$ along the same ray, and obtain their density and color $\{(\sigma_i, \mathbf{c}_i)\}^o_{i=1,\ldots,n}$ by transforming them rigidly back to the canonical object model. We then sort and merge the two sets of samples via their depth values to obtain $\{(\sigma_i, \mathbf{c}_i)\}_{i=1,\ldots,2n}$ and perform volumetric rendering:

$$C_F(\mathbf{r}) = \textstyle\sum_{i=1}^{2n} \tau_i \mathbf{c}_i \tag{9}$$

$$\text{where } \tau_i = \exp\left(-\textstyle\sum_{j<i} \sigma_j \delta_j\right) \left(1 - \exp\left(-\sigma_i \delta_i\right)\right)$$

and δ_i is the distance between two consecutive samples. Similarly, we determine the background color $C_B(\mathbf{r})$ by querying the density and color of sampled points from the background network. To composite the background and foreground, we render the foreground mask probability of a ray $\mathbf{r}$, which can be derived as $M_F(\mathbf{r}) = \sum_{i=1}^{2n} \tau_i \in \mathbb{R}$. To render with the dynamic background, the final color value of the ray is defined as

$$C(\mathbf{r}) = C_F(\mathbf{r}) + (1 - M_F(\mathbf{r}))C_B(\mathbf{r}) \tag{10}$$

where $C_B(\mathbf{r})$ is the background color value. Similar to determining the foreground probability, our model also determines the amodal mask [32] probability of a pixel belonging to hand $M_h(\mathbf{r}) \in \mathbb{R}$ or object $M_o(\mathbf{r}) \in \mathbb{R}$ by accumulating the transmittance of hand or object samples independently. In addition, our model renders the class probability $S(\mathbf{r}) \in \mathbb{R}^3$ between hand, object, and background of each pixel by following the rendering procedure in Equation 9 and Equation 10, while replacing the color $\mathbf{c}$ of each sample point with a one-hot three-vector for each class.

3.2.2 Training losses

Since reconstructing 3D shapes from a monocular video is highly under-constrained, we devise a loss $\mathcal{L}$ consisting of several terms to optimize for the texture and shape network parameters $\{\psi_h, \psi_o, \psi_b\}$, the per-frame parameters $\{\theta, \mathbf{t}_h, \mathbf{R}_o, \mathbf{t}_o, \mathbf{z_o}, \mathbf{z_b}\}$, and global parameters $\{\beta, s\}$.

In particular, we first encourage RGB values to be consistent with the input image via

$$\mathcal{L}_{\text{rgb}} = \sum_{\mathbf{r}} \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\| \tag{11}$$

where $\mathbf{r}$ is a ray casted from a sampled pixel on an image, and $C(\mathbf{r})$ and $\hat{C}(\mathbf{r})$ are the rendered and ground-truth color.

To encourage the disentanglement between the hand, the object, and the background, we supervise the networks with a multi-class segmentation loss

$$\mathcal{L}_{\text{segm}} = \sum_{\mathbf{r}} \left\| S(\mathbf{r}) - \hat{S}(\mathbf{r}) \right\|, \tag{12}$$

where $\hat{S}(\mathbf{r}) \in \mathbb{R}^3$ is a one-hot vector representing the predicted class of a pixel, obtained with an off-the-shelf segmentation network [29]. To regularize the hand and object shapes, we sample points uniformly at random as well as around the surface of the hand and object in their canonical space. We then enforce the eikonal loss $\mathcal{L}_{\text{eikonal}}$ [16] to regularize the canonical hand and object shapes. To provide a shape prior for the hand, using the same set of samples, we enforce the SDF predicted by our canonical hand model to be similar to the one from MANO using the following loss:

$$\mathcal{L}_{\text{sdf}} = \sum_{\mathbf{x} \in \mathcal{X}} \| f_h(\mathbf{x}) - SDF(\mathbf{x}) \| \tag{13}$$

where $\mathcal{X}$ is a set of randomly sampled points in canonical space and $SDF(\mathbf{x})$ is the signed distance from the MANO mesh. To obtain a smooth SDF from the MANO mesh, we sub-divide MANO using Loop subdivision [38].

Finally, to enforce sparsity of the hand density outside of its surface, for a ray $\mathbf{r}$ that is far from the MANO hand mesh, we enforce its amodal mask probability $M_h(\mathbf{r})$ to be zero. A ray $\mathbf{r}$ is far away from a mesh if its closest distance to the mesh exceeds a threshold. Similarly, we periodically construct an object mesh via marching cubes and use it to enforce the object sparsity loss when the ray of a pixel is far away from the object. Formally,

$$\mathcal{L}_{\text{sparse}} = \sum_{\mathbf{r} \in \mathcal{F}_h} \| M_h(\mathbf{r}) \| + \sum_{\mathbf{r} \in \mathcal{F}_o} \| M_o(\mathbf{r}) \| \tag{14}$$

where $\mathcal{F}_h$ and $\mathcal{F}_o$ are the set of rays far from the hand and object meshes respectively. The total loss $\mathcal{L}$ is defined as

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{rgb}} &+ \lambda_{\text{segm}} \mathcal{L}_{\text{segm}} + \lambda_{\text{sdf}} \mathcal{L}_{\text{sdf}} \\ &+ \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}} + \lambda_{\text{eikonal}} \mathcal{L}_{\text{eikonal}} \end{aligned} \tag{15}$$

where λ_* are the weights for the loss terms. Note that since predicted segmentation masks are often noisy, we gradually decrease λ_{segm} over time and gradually increase the prior weights λ_{sdf} and λ_{sparse} over time.

3.3. Pose refinement

The poses from Sec. 3.1 are imperfect because object point clouds from SfM are noisy, and the hand shape parameters are not optimized. After the training from Sec. 3.2, HOLD-Net learns a custom object template, which is more precise than a SfM point cloud for refining hand and object poses with contact constraints. While jointly training HOLD-Net and optimizing the poses could theoretically resolve noisy poses, we empirically find that this strategy is inefficient as the pose of each training frame gets only sparse training signals, i.e. only when the the corresponding frame is sampled. To obtain accurate poses efficiently, we first train HOLD-Net for a small number of epochs to obtain a coarse estimate of the object shape. Then we follow [72] and refine the hand and object pose parameters $\{\mathbf{t}_h, \mathbf{R}_o, \mathbf{t}_o, \beta, s\}$ with mesh-based interaction constraints, using the object mesh extracted from HOLD, and MANO.

In particular, we encourage contact between frequently contacted hand vertices $\mathbf{V}_{tips}$ (vertex IDs from [22]) and the object vertices by encouraging each such hand vertex to be close to an object vertex. Formally, the loss is defined as:

$$\mathcal{L}_{\text{contact}} = \sum_i \min_j \left\| \mathbf{V}^i_{\text{tips}} - \mathbf{V}^j_o \right\|. \tag{16}$$

To provide better pixel-alignment for the hand and the object, we use Soft Rasterizer [36] to render the hand amodal masks $\mathcal{M}_h$ and object amodal masks $\mathcal{M}_o$ and encourage them to match the masks from off-the-shelf semantic segmentation using an occlusion-aware term $\mathcal{L}_{\text{mask}}$ similar to [76]. These simple terms work well in practice. See SupMat for more details and discussion.

3.4. Final training

Using the refined hand pose parameters $\{\theta\}$ from Sec. 3.2 and $\{\mathbf{t}_h, \mathbf{R}_o, \mathbf{t}_o, \beta, s\}$ from Sec. 3.3, we fully train HOLD-Net with the loss $\mathcal{L}$ following the formulation in Sec. 3.2 to reconstruct the 3D hand and object geometries for every frame of an input video. To avoid artifacts that f_h, f_o, f_b could have learnt during pre-training due to inaccurate poses, we train $\{\psi_h, \psi_o, \psi_b, \mathbf{z}_o, \mathbf{z}_b\}$ from scratch. For brevity, we ignore the timestamp for frame-specific parameters. HOLD-Net is pre-trained with half the number of epochs compared to this full-training stage for computational efficiency as we observe that the hand and object shape stabilizes in the early training. The two training procedures are identical except the poses used and the epochs.

4. Experiments

In this section, we compare our method with existing baselines for our new *category-agnostic interaction reconstruction* task. The goal is to reconstruct accurate 3D surfaces for the hand and object from a monocular video, where we assume neither an object template nor an object category.

In-the-lab dataset: We use HO3D-v3 [20] for quantitative and qualitative evaluation. The dataset consists of RGB videos of a hand manipulating a rigid object. The hand is articulated and it provides 3D annotations for MANO parameters and 6D object poses. Since HO3D does not release ground-truth annotations on the test set, we use two sequences for each object with 3D annotations in their training set for evaluation: one matching Hampali *et al.* [21] for consistency, and a second random sequence where hands and objects remain within the frame throughout, simplifying preprocessing. We omit the banana and the scissors as SfM does not converge. These two objects either lack texture or have thin structures and are also failure cases for [19, 21]. For completeness, we report results for the two objects in SupMat using random poses.

HOLD dataset: To evaluate if our method can generalize to diverse in-the-wild settings, we capture sequences of household items in both in-door and out-door scenes. We capture in both 1st-person moving views and 3rd-person static views using an iPhone 14 main camera under different lighting conditions. For each video, we downsample it every 10 frames for our experiments.

Metrics: We use the root-relative mean-per-joint error (MPJPE) in millimeters to measure hand pose error, and Chamfer distance in squared centimeters to evaluate object reconstruction quality [5]. Since Chamfer distance is sensitive to outliers, we also use F-score in percentage to measure local shape details [60, 71]. In particular, to evaluate object template quality independent from object pose, following [71], we perform ICP alignment to the ground-truth mesh of the HO3D meshes allowing scale, rotation and translation and compute the Chamfer distance (CD) and F-score at 5mm (F5) and 10mm (F10). To measure object pose and shape relative to the hand in 3D, we subtract each object mesh by the predicted hand root and compute the hand-relative Chamfer distance for the object (CD$_h$).

	MPJPE [mm] ↓	CD [cm^2] ↓	F10 [%] ↑	CD$_h$ [cm^2] ↓
HOMan† [24]	32.0	N/A	N/A	78.2
iHOI‡ [70]	38.4	3.8	75.8	41.7
DiffHOI [71]	32.3	4.3	68.8	43.8
Ours	**24.2**	**0.4**	**96.5**	**11.3**

Table 1. **Comparison with SOTA hand-object reconstruction methods**. We evaluate our method and the baselines on the HO3D dataset. †*HOMan assumes a ground-truth object template.* ‡*During training, iHOI uses 3D annotation of the test objects, while DiffHOI and ours do not use such information.*

Implementation details: We train each sequence using Adam. In each iteration we optimize 10 randomly sampled images from the sequence. For stability, we perform gradient clipping, which is crucial for convergence of the hand model. We perform the initial training for 100 epochs, which requires around 10 hours using an A100 GPU. The final training takes 200 epochs. We use SAM-track [6] to derive the hand and object segmentation masks by using point-prompting for the first frame of each video. See SupMat for details. We use AITViewer [28] for visualization.

4.1. State-of-the-art comparison

Hand-object reconstruction: Table 1 compares HOLD with existing hand-object reconstruction methods. We observe that HOLD significantly outperforms existing methods in terms of hand pose (MPJPE) and object shape (CD, F10) quality. Our method also infers the relative spatial alignment of the hand and object more accurately as shown by the superior hand-relative Chamfer distance (CD$_h$).

This improvement is also reflected in the qualitative comparison in Fig. 4. Our method consistently produces reconstructions that are closer to the ground-truth than those of iHOI and DiffHOI, with notable improvements in capturing the fine structures, such as the mug handle and the car frame, as well as the dynamic postures of the hand and object. In contrast, the reconstructions from the two baseline methods lack details and suffer from erroneous hand and object poses, even on the easier in-the-lab dataset. Notably, both baseline methods use 3D supervision - iHOI is trained on the HO3D dataset sequences with ground-truth 3D shape and DiffHOI uses 3D shapes of diverse bottles and mugs as training supervision. In contrast, our method only uses the input 2D monocular video without requiring any 3D annotation, while still achieving superior quality. Our method can also reconstruct hands and objects reliably under different backgrounds, and lighting conditions in both 3rd-person view and moving egocentric views (see Fig. 5).

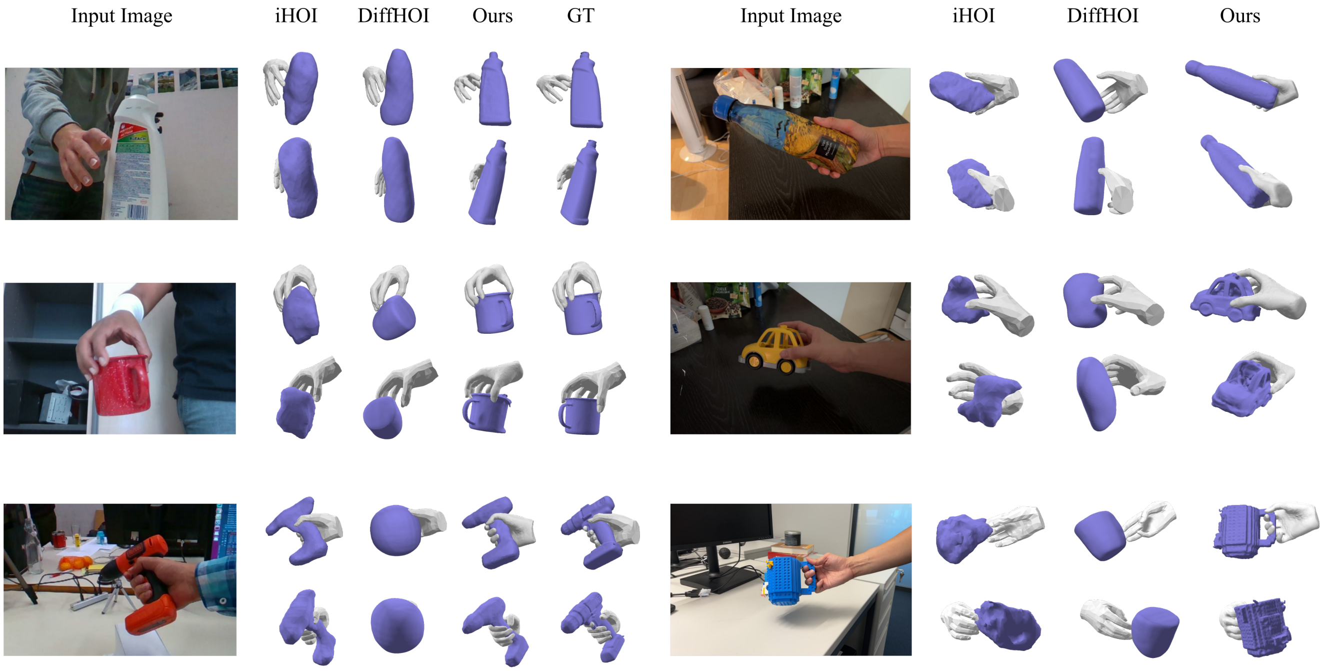

Figure 4. **Qualitative comparison with SOTA**. We show hands and objects reconstructed by our method and SOTA baselines from in-the-lab (*left*) and in-the-wild (*right*) videos. Our reconstruction demonstrates more accurate shapes, richer details, and more accurate poses. In addition, our method works consistently well on various objects, even those with unique shapes (e.g. the Lego mug at the bottom right).

	Object categories	MPJPE [mm] ↓	CD [cm^2] ↓	F10 [%] ↑	CD$_h$ [cm^2] ↓
DiffHOI	DiffHOI training	34.2	1.3	83.5	42.5
Ours		**22.5**	**0.4**	**95.9**	**10.4**
DiffHOI	DiffHOI unseen	30.9	6.5	57.8	44.8
Ours		**25.5**	**0.3**	**96.9**	**12.0**

Table 2. **Generalization comparison**. We compare the generalization ability of HOLD and DiffHOI. We report results on objects within and beyond DiffHOI's training categories. DiffHOI's performance degrades significantly on unseen object categories while ours produces more accurate reconstruction consistently.

Generalization: To quantify our method's ability to generalize compared to DiffHOI, in Table 2 we split the HO3D sequences according to whether they belong to the training categories of DiffHOI. We see that while DiffHOI's performance significantly drops across all metrics for unseen categories, our method has consistent performance on all categories. This is also reflected in Fig. 4: our method can accurately reconstruct objects such as the drill, while the baseline methods do not generalize to instances that are outside their training distributions.

Notably, our method significantly outperforms DiffHOI even for its training categories. We can gain insight into this from the water bottle example at the top-right of Fig. 4: DiffHOI tries to reconstruct bottles seen in their data-driven prior training set, which leads to a generic bottle. In comparison, our reconstructed bottle realistically captures the shape details of the one in the image.

	CD [cm^2] ↓	F5 [%] ↑	F10 [%] ↑
Hampali [21]	1.4	57.4	79.9
Ours	**0.5**	**84.3**	**94.4**

Table 3. **Comparison with a SOTA in-hand scanning method**.

In-hand object scanning: Table 3 compares with Hampali *et al*. [21], the SOTA method for in-hand object scanning. Since no code is released, we compare our canonical shapes with their released object point clouds. We observe better object canonical shapes (CD) and local details (F5 and F10) in HOLD. SupMat contains a qualitative comparison.

4.2. Ablation

Joint hand-object reconstruction: To verify that hand reconstruction is complementary to object reconstruction, we implement an ablative baseline without modeling the hand. To be specific, we mask out the hand from all video frames and train the object network on these processed frames. As demonstrated in Table 4, removing the hand from our model leads to degraded reconstruction accuracy (CD and F10). A qualitative example is shown in Fig. 6 (a). Without hand modeling, the reconstructed object has a hole at the hand-grasping region since the object model needs to explain the

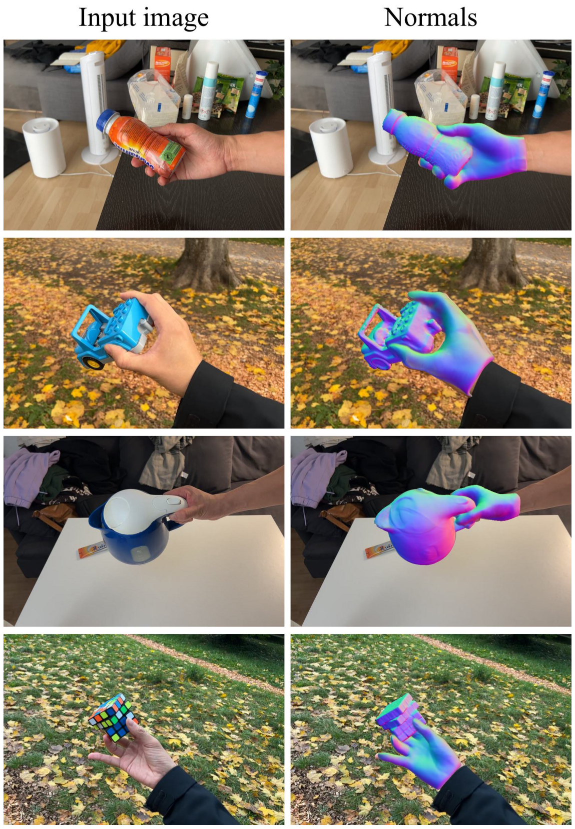

Figure 5. **More qualitative results**. We render the normals of hands and objects reconstructed by HOLD. Our method can reliably reconstruct in both static views and moving egocentric views.

images with the hand masked out. By jointly modeling the hand, the object, and their occlusion, our method can faithfully reconstruct the object despite hand-object occlusion.

Contact-based hand-object pose refinement: To assess the impact of pose refinement as described in Section 3.3, we compare our model to a baseline that omits this process. Figure 6 (b) provides a rotated-view illustration that highlights the disparity between the baseline model and our full approach. Without pose refinement, there is an unrealistic separation between the hand and object, a common issue in monocular reconstructions due to significant depth ambiguity leading to spatial misalignments.

Our refinement strategy mitigates this by encouraging hand-object contact, thereby diminishing the relative depth uncertainty. The improvements in pose accuracy for both the hand and the object, as well as their spatial arrangement, are quantitatively evidenced in Table 4. Our method outperforms the baseline by achieving superior hand pose accuracy, indicated by lower MPJPE, and improved hand-relative Chamfer distance (see CD_h). These improvements in pose accuracy also translate into more accurate object template reconstructions (see CD and F10).

	MPJPE [mm] ↓	CD [cm^2] ↓	F10 [%] ↑	CD_h [cm^2] ↓
w/o hand	-	0.41	95.9	-
w/o pose ref.	24.6	0.55	94.2	122.1
Ours	**24.2**	**0.38**	**96.5**	**11.3**

Table 4. **Ablation study**. Modelling the hand and object jointly improves object reconstruction accuracy. Pose refinement improves hand-object poses and consequently object reconstruction.

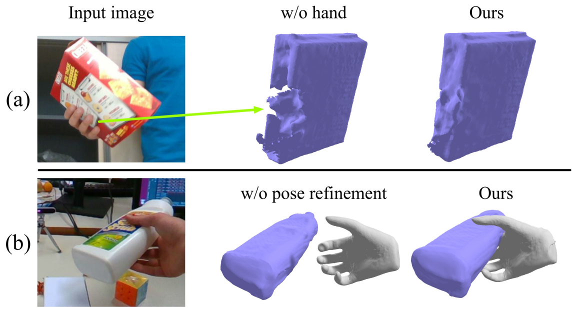

Figure 6. **Ablation study**. *(a)* Jointly reconstructing the hand and object effectively reduces artifacts. *(b)* Without contact-based pose refinement, the hand and object can have an erroneous spatial arrangement due to depth ambiguity.

5. Conclusion

In this paper, we present HOLD – the first category-agnostic method that reconstructs an articulated hand and object jointly from a monocular interaction video. We introduce a novel compositional implicit model of the object and articulated hand that disentangles and reconstructs 3D hands and objects from 2D observations. We further show that jointly optimizing the hand and object via interaction constraints leads to better reconstruction of object surfaces than reconstructing objects in isolation. Our method significantly outperforms fully-supervised SOTA baselines in both in-the-lab and in-the-wild settings while not relying on 3D hand-object annotation data. We qualitatively demonstrate our method's robustness on challenging in-the-wild videos.

Limitations and discussion: While HOLD successfully reconstructs category-agnostic interactions, it does face challenges. The reconstruction of thin/textureless objects is limited by detector-based SfM for pose initialization. Advances in detector-free SfM [25, 57] could potentially address this. Further, our reliance on RGB supervision may hinder the reconstruction of rarely observed object regions. This could be regularized with priors [46]. Lastly, training time can be reduced with faster representations [30].

Acknowledgements: This work was partially supported by the Swiss SERI Consolidation Grant "AI-PERCEIVE". **Disclosure**: https://files.is.tue.mpg.de/black/CoI_CVPR_2024.txt

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3D hand shape and pose from images in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[2] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[3] Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 3

[4] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[5] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-driven signed distance functions for 3D hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 6

[6] Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023. 6

[7] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[8] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[9] Markos Diomataris, Nikos Athanasiou, Omid Taheri, Xi Wang, Otmar Hilliges, and Michael J. Black. WANDR: Intention-guided human motion generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[10] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J. Black. HMP: Hand motion priors for pose and shape estimation from video. In *Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2

[11] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J. Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *International Conference on 3D Vision (3DV)*, 2021. 2

[12] 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. 1, 2

[13] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, Fei Li, Liu Zheng, Feng Lu, Karim Abou Zeid, Bastian Leibe, Jeongwan On, Seungryul Baek, Aditya Prakash, Saurabh Gupta, Kun He, Yoichi Sato, Otmar Hilliges, Hyung Jin Chang, and Angela Yao. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. *arXiv preprint arXiv: 2403.16428*, 2024. 2

[14] Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M. Kitani. Deformer: Dynamic fusion transformer for robust hand pose estimation. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[15] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[16] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, 2020. 5

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

[18] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. HandNeRF: Neural radiance fields for animatable interacting hands. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[19] Shreyas Hampali. *3D Pose and Shape Estimation of Objects and Hands in Challenging Scenarios*. PhD thesis, TU Graz, 2023. 6

[20] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[21] Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3D object scanning from an RGB sequence. *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 6, 7

[22] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[23] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[24] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *International Conference on 3D Vision (3DV)*. IEEE, 2021. 1, 2, 6

[25] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *arxiv*, 2023. 8

[26] 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 2022 Conference Papers*, 2022. 2, 3

[27] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D

heatmap regression. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[28] Manuel Kaufmann, Velko Vechev, and Dario Mylonopoulos. aitviewer, 2022. 6

[29] 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 *International Conference on Computer Vision (ICCV)*, 2023. 3, 5

[30] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splats. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 8

[31] Jihyun Lee, Minhyuk Sung, Honggyu Choi, and Tae-Kyun Kim. Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[32] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision (ECCV)*. Springer, 2016. 5

[33] Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, and Chen Chen. RenderIH: A large-scale synthetic dataset for 3D interacting hand pose estimation. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[34] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[35] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[36] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3D reasoning. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[37] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3D hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[38] Charles Loop. Smooth subdivision surfaces based on triangles. 1987. 5

[39] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In *European Conference on Computer Vision (ECCV)*. Springer, 2022. 2

[40] Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2023.

[41] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[42] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3D interacting hands. *NeurIPS*, 2023. 2

[43] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[44] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[45] Valeria Perasso. What have you touched today?, 2015. 1

[46] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *International Conference on Learning Representations (ICLR)*, 2022. 8

[47] Aditya Prakash, Matthew Chang, Matthew Jin, and Saurabh Gupta. Learning hand-held object reconstruction from in-the-wild videos. *arXiv*, 2305.03036, 2023. 2

[48] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[49] James M. Rehg and Takeo Kanade. Visual tracking of high DOF articulated structures: An application to human hand tracking. In *European Conference on Computer Vision (ECCV)*, 1994. 2

[50] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *Transactions on Graphics (TOG)*, 36(6), 2017. 3, 4

[51] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[52] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[53] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[54] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[55] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3D hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision (ECCV)*, 2020.

[56] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning. In *International Conference on Computer Vision (ICCV)*, 2021. 2

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

[58] Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. SHOWMe: Benchmarking object-agnostic hand-object 3D reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[59] Omid Taheri, Yi Zhou, Dimitrios Tzionas, Yang Zhou, Duygu Ceylan, Soren Pirk, and Michael J. Black. GRIP: Generating interaction poses using latent consistency and spatial cues. In *International Conference on 3D Vision (3DV)*, 2024. 1

[60] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3D reconstruction networks learn? In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[61] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[62] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[63] Tze Ho Elden Tse, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, et al. Spectral graphormer: Spectral graph-based transformer for egocentric two-hand reconstruction using multi-view color images. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[64] Dimitrios Tzionas and Juergen Gall. A comparison of directional distances for hand pose estimation. In *German Conference on Pattern Recognition (GCPR)*, 2013. 2

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

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

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

[68] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[69] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 4

[70] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 6

[71] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 6

[72] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4D: 4D instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[73] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3D pose and shape reconstruction from single color image. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[74] Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. GraspXL: Generating grasping motions for diverse objects at scale. *arXiv preprint*, 2024. 1

[75] Hui Zhang, Sammy Christen, Zicong Fan, Luocheng Zheng, Jemin Hwangbo, Jie Song, and Otmar Hilliges. ArtiGrasp: Physically plausible synthesis of bi-manual dexterous grasping and articulation. In *International Conference on 3D Vision (3DV)*, 2024. 1

[76] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3D human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*. Springer, 2020. 5

[77] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3, 4

[78] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[79] Licheng Zhong, Lixin Yang, Kailin Li, Haoyu Zhen, Mei Han, and Cewu Lu. Color-NeuS: Reconstructing neural implicit surfaces with color. In *International Conference on 3D Vision (3DV)*, 2024. 2, 3

[80] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[81] Andrea Ziani, Zicong Fan, Muhammed Kocabas, Sammy Christen, and Otmar Hilliges. TempCLR: Reconstructing hands via time-coherent contrastive learning. In *International Conference on 3D Vision (3DV)*, 2022. 2

[82] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *International Conference on Computer Vision (ICCV)*, 2017. 2