CHORD: Category-level Hand-held Object Reconstruction via Shape Deformation

Kailin Li[1]* Lixin Yang[1,2]* Haoyu Zhen[1] Zenan Lin[3]
Xinyu Zhan[1] Licheng Zhong[1] Jian Xu[4] Kejian Wu[4] Cewu Lu[1,2]†
[1]Shanghai Jiao Tong University [2]Shanghai Qi Zhi Institute [3]South China University of Technology [4]XREAL
[1]{kailinli, siriusyang, anye_zhen, kelvin34501, zlicheng, lucewu}@sjtu.edu.cn
[3]auzenanlin@mail.scut.edu.cn [4]{jianxu, kejian}@nreal.ai

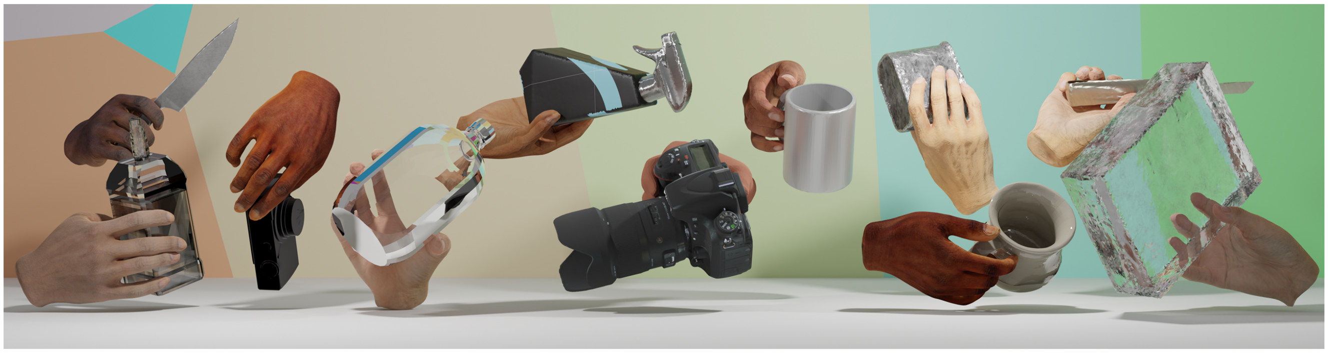


Figure 1: **Examples of Hand-held Objects at Category Level.** We proposed a new method CHORD which exploits the categorical shape prior for reconstructing the shape of intra-class objects. In addition, we constructed a new dataset, COMIC, of category-level hand-object interaction. COMIC encompasses a diverse collection of object instances, materials, hand interactions, and viewing directions, as illustrated.

Abstract

In daily life, humans utilize hands to manipulate objects. Modeling the shape of objects that are manipulated by the hand is essential for AI to comprehend daily tasks and to learn manipulation skills. However, previous approaches have encountered difficulties in reconstructing the precise shapes of hand-held objects, primarily owing to a deficiency in prior shape knowledge and inadequate data for training. As illustrated, given a particular type of tool, such as a mug, despite its infinite variations in shape and appearance, humans have a limited number of 'effective' modes and poses for its manipulation. This can be attributed to the fact that humans have mastered the shape prior of the 'mug' category, and can quickly establish the corresponding relations between different mug instances and the prior, such as where the rim and handle are located. In light of this, we propose a new method, CHORD, *for* ***C****ategory-level* ***H****and-held* ***O****bject* ***R****econstruction via shape* ***D****eformation.* CHORD *deforms a categorical shape prior for reconstructing the intra-class objects. To ensure accurate reconstruction, we empower* CHORD *with three types of awareness: appearance, shape, and interacting pose. In addition, we have constructed a new dataset,* COMIC, *of category-level hand-object interaction.* COMIC *contains a rich array of object instances, materials, hand interactions, and viewing directions. Extensive evaluation shows that* CHORD *outperforms state-of-the-art approaches in both quantitative and qualitative measures. Code, model, and datasets are available at* https://kailinli.github.io/CHORD

1. Introduction

In daily life, we perform complex tasks by continually manipulating a limited number of simple objects with our hands. Understanding the physical nature of hand-object interaction is crucial for AI to comprehend human activities. This requires us to pursue a geometric representation (reconstruction) of the hand-held object, especially for manipulable tools. Many significant efforts [26, 31, 56, 54, 57, 11, 5] have been made for reconstructing

*These authors contributed equally.

†Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China, and Shanghai Qi Zhi institute.

both the hand and hand-held objects. These works, where the hand is represented by a kinematic prior, such as MANO [43], have achieved high qualities on hand reconstruction via keypoints estimation. However, producing the object shape in high quality is more challenging, primarily due to the mutual occlusion, lack of geometrical prior, and insufficient data of variant shapes and appearances.

Most previous works thus resorted to reconstructing objects with known template shape [47, 36, 25, 54, 24]. This setting is commonly referred to as "object pose estimation". However, it would fail on the unseen object instances. To address this limitation, some works [26, 31, 57, 11] have attempted to directly regress object-agnostic (and also pose-agnostic) surfaces from images using large-scale synthetic data. However, their results are not robust to varied object shapes and appearances, and often produce "bubble-like" shapes and broken geometries. These issues can be attributed to either a lack of shape basis [31, 57, 11] or the use of an underlying shape basis with fixed topology and resolution [26, 20, 29].

In this paper, we aim to leverage the best of both worlds: utilizing shape information from a known template while also generalizing well to unseen instances. With this in mind, we propose to first estimate the object pose at category-level, and then resort to a categorical object shape prior (later we call it *object-prior*) to reconstruct the surface of unseen objects. We design **CHORD**, stand for **C**ategory-level **H**and-held **O**bject **R**econstruction via shape **D**eformation. CHORD is a deep-learning model that learns to "deform" an implicit surface from an explicit object-prior. Given the estimated pose of the object-prior, CHORD takes two steps. It first deforms the 2D surface-aware feature maps (*i.e.* normal and depth map) of the object-prior to those of the actual object instance. From these feature maps, CHORD then deforms the 3D shape of the object-prior to the actual object instance.

Inspired by [45, 53, 39], the first step is achieved by leveraging the design of an image-to-image translation network ($\mathcal{G}_{\text{N}}$, Sec. 3.2), Specifically, we rendered the normal and depth map of the object-prior in the estimated poses, along with the image, as inputs for $\mathcal{G}_{\text{N}}$. Additionally, we use the rendered depth and normal maps of the estimated MANO hand mesh as extra inputs to help with decoupling the surfaces of the hand and object.

To perform the second step, we use a point-wise implicit function (IF, denoted as $\mathcal{G}_{\text{S}}$) [40] to regress the signed distance of query points to the surface of object instance. The final reconstruction is then obtained as the zero-level set of these query points. The use of IF enables us to reconstruct objects with fine-grained geometry and arbitrary topology. We argue that for robustly and accurately reconstructing the hand-held object, it is necessary for CHORD to possess three different types of awareness: (1) the **appearance** awareness, (2) the **shape** awareness of the categorical object-prior, and (3) the **pose** awareness of the interacting hand. To integrate these awareness, we develop three types of local features (Sec. 3.3-A,B,C). Consequently, $\mathcal{G}_{\text{S}}$'s prediction of signed distance is conditioned on these features.

The final issue is that the current datasets for hand-object interaction (HOI) are not suitable for reconstruction in category-level. These datasets commonly lack diverse samples within the same category [23, 3, 17, 2] or lack real-world human interacting behaviors [26]. To address this limitation, we propose a new dataset, named COMIC, which is built from high-fidelity rendering and targets on category-level hand-object reconstruction (Sec. 3.5). This dataset contains a large number of images that depict the interaction between the hand and categorical objects with diverse shapes and appearances (see Figs. 1 and 5). Unlike the simulated grasping poses as in GraspIt [38], which do not reflect the intention of human behaviors, the interacting poses in COMIC are based on real-world human demonstrations captured in [55].

We conduct an extensive evaluation of CHORD utilizing the COMIC dataset, including qualitative outcomes on several additional HOI datasets (not incorporated for training) as well as unseen, in-the-wild images. Our quantitative comparisons verify that CHORD exceeds the state-of-the-art (SOTA) in performance. Furthermore, our qualitative results illustrate that CHORD generalizes more effectively to unseen, in-the-wild instances. We summarize our contribution in three-folds:

- We propose the reconstruction of hand-held objects at category-level via our novel model, CHORD. The model explicitly encapsulates shape from an object-prior, facilitating the deformation of the implicit surface to actual object instances.
- Within CHORD, we incorporate three types of awareness - appearance, shape, and interacting pose - to ensure the accuracy of reconstruction. An extensive ablation study and gain analysis substantiate the theory that performance enhancement corresponds with increased awareness.
- We introduce a new dataset for category-level hand-object reconstruction, known as COMIC. This dataset comprises a wealth of high-fidelity images featuring a vast variety of objects and interacting hand poses.

2. Related Work

Hand-held Object Reconstruction. Previous studies have utilized RGB or RGB-D data for reconstructing hand-held objects with optimization [42, 49]. However, these methods generally only reconstruct a limited number of known objects for which a 3D model is available [17, 2, 47, 23, 50], using approaches such as implicit feature fusion [10, 36, 47], explicit geometric constraints like contact or collision [1, 2, 13, 18, 22, 58], or physical realism to aid in hand joint reasoning and object reconstruction [42, 50]. These methods commonly assume the availability of the 3D object template during inference, which is a limiting assumption. Reconstructing hand-held objects without a known template is significantly more challenging as it requires the algorithm to handle various object appearances and shapes while only partially observing them. Hasson *et al*. [26] utilize a view-centric variant of AtlasNet [21] to handle generic object categories without the need for instance-specific knowledge. Karunratanakul *et al*. [31] characterize each point in 3D space using signed distances to the surface of the hand and object, and combine the hand, object, and contact area in a shared space represented by implicit surfaces. Additionally, works like [57, 11] also focus on reconstructing hand-held objects from images without knowledge of their 3D templates. They use an implicit network to infer the signed distance to parameterize objects and infer the object shape in a normalized hand-centric coordinate space. In this paper, we focus on what the above approaches miss: reconstructing hand-held objects by using a priori knowledge of shape at the category-level.

Category-level Object Pose Estimation. Category-level object pose estimation involves localizing and estimating the 3D pose of an object within a specific category (e.g., chair, table, or car) without knowledge of the particular instance. Sahin *et al*. [44] was the first work to address the problem of 6DoF object pose estimation at the category level. However, its generalization capability across unseen object instances is limited. One of the significant challenges for category-level pose estimation is the intra-class variation, including appearance and shape. To address this challenge, Wang *et al*. [52] proposed a shared canonical representation, called Normalized Object Coordinate Space (NOCS), that uses the dense image-shape correspondence for estimating target instances under the same category. Tian *et al*. [48] used the NOCS and also explicitly modeled the deformation from a pre-learned categorical shape prior. Chen *et al*. [4] modeled canonical shape space (CASS) as a latent space with normalized poses for estimating the 6DoF poses of objects. Furthermore, Chen *et al*. [7] proposed a Decoupled Rotation Representation that directly learns the two perpendicular vectors of a categorical-specified object, just name a few. Mainstream category-level object pose estimation follows the NOCS formulation, [52, 4, 48, 16], which requires the extra depth map as input. As an exception, OLD-Net [15] eliminates the need for depth by additionally performing depth reconstruction. In our task, where the interacting hand is predictive of the object shape and scale, the requirement for a depth map can be replaced by incorporating the hand's pose estimation. For category-level object pose estimation, we employ decoupled rotation axes for the sake of simplicity. This representation eliminates the need for dense image-to-shape correspondences such as NOCS, which can be particularly challenging to establish in the presence of severe occlusions between the hand and object.

3. Method

Inferring the detailed shape of both the hand and object of known category is a challenging and ill-posed problem: the hand commonly occludes the object, and vice versa, making it necessary for the deep network to decouple the conjunct parts and reconstruct the surfaces separately, based on the observable parts. CHORD jointly considers a MANO hand model and the object-prior to reduce the ambiguities (Sec. 3.1). Specifically, CHORD takes an RGB image cropped around the hand, along with the estimated hand (MANO) parameters and the pose of a category-level object-prior, and outputs the shape reconstruction of the object instance aligned with the input image. CHORD uses two consecutive submodules to this end: (1) pixel-level 2D deformation (Sec. 3.2) and (2) point-level 3D deformation (Sec. 3.3). After completing these two steps, CHORD outputs the shape of an unseen object instance in the form of zero-level set of signed distance field (SDF).

3.1. Preceding Task Model

To estimate the MANO hand mesh, we leverage the 3D keypoints prediction $\mathbf{J} \in \mathbb{R}^{21\times 3}$ as proxy following [60], since keypoints are more robust to occlusion. The 3D keypoints are then transferred to MANO parameter via an inverse kinematics network (IKNet) [60]. The MANO parameter consist of pose $\boldsymbol{\theta} \in \mathbb{R}^{K\times 3}$, and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$, which together are mapped to the MANO hand mesh $\mathbf{V}_H = \mathcal{M}(\boldsymbol{\theta}, \boldsymbol{\beta}) \in \mathbb{R}^{N_V\times 3}$ via a skinning function $\mathcal{M}$ [43, Eq.(1)]. The $K = 16$ and $N_V = 778$ are the numbers of joint rotations and vertices of MANO, respectively. For estimating the rotation of the object-prior from the camera, we use the category-aware Decoupled Rotation representation following [6], where the two (or one, depending on the object's symmetrical property) shape-aligned rotation axes $\mathbf{R}_1, \mathbf{R}_2 \in \mathbb{R}^3$ are predicted separately. These rotation axes are then used to construct the rotation matrix $\mathbf{R}_{\mathbb{O}} \in SO(3)$ of the **object-prior (denoted as $\mathbb{O}$)**. For estimating the translation of the object-prior relative to hand $\mathbf{t}_{\mathbb{O}} \in \mathbb{R}^3$, we employ 1-channel volumetric likelihood heatmap following AlignSDF [11]. Importantly, CHORD is also compatible with other MANO-based mesh recovery models [8, 35, 33], as well as other category-level object pose representation, such as NOCS [52]. The simple preceding task models used in CHORD allow us to isolate the benefits of the shape prior guided 2D and 3D deformations. Network details are provided in Supp. Mat.

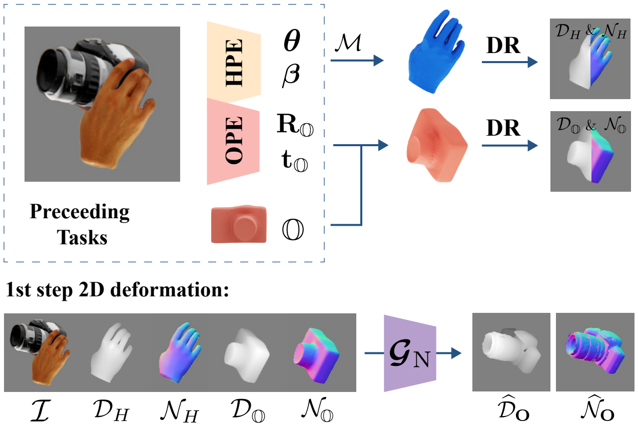

HPE
OPE
$\boldsymbol{\theta}$
$\boldsymbol{\beta}$
$\mathbf{R}_{\mathbb{O}}$
$\mathbf{t}_{\mathbb{O}}$
$\mathcal{M}$
DR
DR
$\mathcal{D}_H$ & $\mathcal{N}_H$
$\mathcal{D}_{\mathbb{O}}$ & $\mathcal{N}_{\mathbb{O}}$
Preceeding Tasks
$\mathbb{O}$
1st step 2D deformation:
$\mathcal{G}_{\mathrm{N}}$
$\mathcal{I}$
$\mathcal{D}_H$
$\mathcal{N}_H$
$\mathcal{D}_{\mathbb{O}}$
$\mathcal{N}_{\mathbb{O}}$
$\widehat{\mathcal{D}}_{\mathbf{O}}$
$\widehat{\mathcal{N}}_{\mathbf{O}}$

Figure 2: **Illustration of the first step 2D deformation $\mathcal{G}_{\mathrm{N}}$. HPE**: hand pose estimation, **OPE**: category-level object-prior pose estimation, **DR**: differentiable rendering.

3.2. Object-prior Guided 2D Deformation

The success of 3D clothed human digitization [45, 53] has demonstrated two key findings: (1) common convolution neural networks are more adept at inferring 2D feature maps aligned with images than directly estimating detailed 3D surfaces, and (2) the 2D normal map (a common 2D feature map) can serve as a useful guide for 3D surface reconstruction. Building upon these observations, in the first step, we utilize an image-to-image translation network $\mathcal{G}_{\mathrm{N}}$ to predict the surface-aware 2D features (*i.e*. normal and depth map) of the hand and object from the input image. These image-aligned 2D feature maps are lately used to aid the 3D surface reconstruction, which we introduce in Sec. 3.3-A.

To reduce the ambiguity of the conjunct 3D surfaces, we incorporate the poses of the MANO and object-prior estimated from the preceding task model (Sec. 3.1) as additional inputs to $\mathcal{G}_{\mathrm{N}}$. Given the estimated mesh of hand and object-prior, we use a differentiable renderer in PyTorch3D to generate four 2D feature maps: $\mathcal{D}_H$ and $\mathcal{N}_H$ for the depth and normal maps of hand, and $\mathcal{D}_{\mathbb{O}}$ and $\mathcal{N}_{\mathbb{O}}$ for those of the object-prior. The separated hand and object depths and normals provide initial guidance for decoupling the 3D surfaces. Accordingly, the network learns to deform the initial feature maps of the **object-prior (denoted as $\mathbb{O}$)** so as to synchronize them with the **object instance (denoted as O)** observed in the input image.

Based on these initial 2D feature maps $\{\mathcal{D}_\star, \mathcal{N}_\star\}$, where $\star \in \{H, \mathbb{O}\}$, $\mathcal{G}_{\mathrm{N}}$ predicts the 2D feature maps of the actual object instance post-deformation, denoted as $\widehat{\mathcal{D}}_{\mathbf{O}}$ (for the depth map) and $\widehat{\mathcal{N}}_{\mathbf{O}}$ (for the normal map). Formally, the $\mathcal{G}_{\mathrm{N}}$ represents the following mapping

$$\mathcal{G}_{\mathrm{N}} : (\mathcal{I}, \mathcal{D}_H, \mathcal{N}_H, \mathcal{D}_{\mathbb{O}}, \mathcal{N}_{\mathbb{O}}) \mapsto (\widehat{\mathcal{D}}_{\mathbf{O}}, \widehat{\mathcal{N}}_{\mathbf{O}}). \tag{1}$$

The input of $\mathcal{G}_{\mathrm{N}}$ comprises the original RGB image $\mathcal{I}$, along with the four initial feature maps. The loss function for training network $\mathcal{G}_{\mathrm{N}}$ consists of two terms: (1) a pixel-wise $L1$ discrepancy loss of the two feature maps: $\mathcal{L}_{\mathrm{pix}} = \sum(|\widetilde{\mathcal{N}}_{\mathbf{O}} - \widehat{\mathcal{N}}_{\mathbf{O}}| + |\widetilde{\mathcal{D}}_{\mathbf{O}} - \widehat{\mathcal{D}}_{\mathbf{O}}|)$, where $\widetilde{\mathcal{N}}_{\mathbf{O}}$, $\widetilde{\mathcal{D}}_{\mathbf{O}}$ are the ground-truth normal and depth map of actual object instance, and (2) a perceptual loss $\mathcal{L}_{\mathrm{VGG}}$ [28] weighted by λ_{VGG}.

3.3. Object-prior Guided 3D Deformation

Given the estimated pose of the object-prior, the MANO hand parameters, and the image-aligned 2D feature maps, we propose an implicit network $\mathcal{G}_{\mathrm{S}}$ for regressing the 3D surface of the hand-held object. To achieve the three types of awareness, the network utilizes three types of local features aggregated from three different modalities. Specifically, for a given query point $\mathbf{x}$, we require its: (1) 2D pixel-aligned features in normals and depth maps, (2) 3D shape features interpolated within its nearest region on the object-prior, and (3) 3D articulated features represented in the local-registered MANO's pose space. Based on these inputs, the network $\mathcal{G}_{\mathrm{S}}$ deforms the object-prior by predicting the signed distance from $\mathbf{x}$ to its closest point on the object instance.

A. 2D Appearance-aware Feature Given the predicted 2D normal and depth maps of the actual hand and object instance, we extract the local 2D features of the query point $\mathbf{x}$ on these feature maps using bilinear interpolation. The local 2D features consist of four terms: $\mathcal{F}_H^{\mathcal{N}}(\mathbf{x})$, $\mathcal{F}_H^{\mathcal{D}}(\mathbf{x})$, $\mathcal{F}_{\mathbf{O}}^{\mathcal{N}}(\mathbf{x})$, and $\mathcal{F}_{\mathbf{O}}^{\mathcal{D}}(\mathbf{x})$. Specifically, $\mathcal{F}_H^{\mathcal{N}}(\mathbf{x}) = \mathcal{N}_H(\pi(\mathbf{x}))$ represents the normal value at the projection of query point $\pi(\mathbf{x})$ on the estimated hand normal map $\mathcal{N}_H$, and the same principle applies to the remaining terms. The final appearance-aware features $\mathcal{F}_{\mathrm{A}}$ are as:

$$\mathcal{F}_{\mathrm{A}}(\mathbf{x}) = [\mathcal{F}_H^{\mathcal{N}}(\mathbf{x}), \mathcal{F}_{\mathbf{O}}^{\mathcal{N}}(\mathbf{x}), \mathcal{F}_H^{\mathcal{D}}(\mathbf{x}), \mathcal{F}_{\mathbf{O}}^{\mathcal{D}}(\mathbf{x})]. \tag{2}$$

B. 3D Shape-aware Feature. To facilitate the capture of shape-aware features by CHORD, we incorporate a categorical shape prior for reconstruct different object instances. Accordingly, instead of learning the implicit surface of different objects separately, as is done in iHOI [57] and AlignSDF [11], CHORD learns to regress them from a shared set of latent codes on the object-prior. We draw inspiration from the Neural Body [41], which employs the structured latent codes anchored on the SMPL body vertices (inner surface) for reconstructing the clothed human body at the outer surface. Similarly, in CHORD, we anchor a set of latent codes $\mathcal{Z}$ to the vertices of a mesh-formed object-prior. These codes are learned alongside the CHORD model using the images of different object instances within the same category. To account for the shape variance between the object-prior and all individual instances, we employ the SparseConvNet [19] (denoted as $\mathcal{SP}$) to diffuse the structured latent codes on the object-prior to its nearby

3D volume space following [46]. Then, for a given query point $\mathbf{x}$, we extract its 3D shape feature by trilinear (tri-) interpolation in this diffused volume space: $\mathcal{SP}(\mathcal{Z})$. Notably, as learning on the structured latent code should be independent of the pose of the object-prior, we first transform $\mathbf{x}$ from world system to object-prior's canonical system using the inverse transformation of $\mathbf{R}_{\mathbb{O}}$, $\mathbf{t}_{\mathbb{O}}$ before the tri-interpolation. We denote the query point in the object canonical-space as $\mathbf{x}^{\ominus}$, where $\mathbf{x}^{\ominus} = \mathtt{inv}(\mathbf{R}_{\mathbb{O}}, \mathbf{t}_{\mathbb{O}}) \cdot \mathbf{x}$. Finally, the shape-aware feature is expressed as:

$$\mathcal{F}_{\mathrm{S}}(\mathbf{x}) = \mathcal{SP}(\mathcal{Z}, \mathbf{x}^{\ominus}), \tag{3}$$

where the $\mathcal{SP}(\mathcal{Z}, \cdot)$ denotes the tri-interpolation on the diffused latent code space.

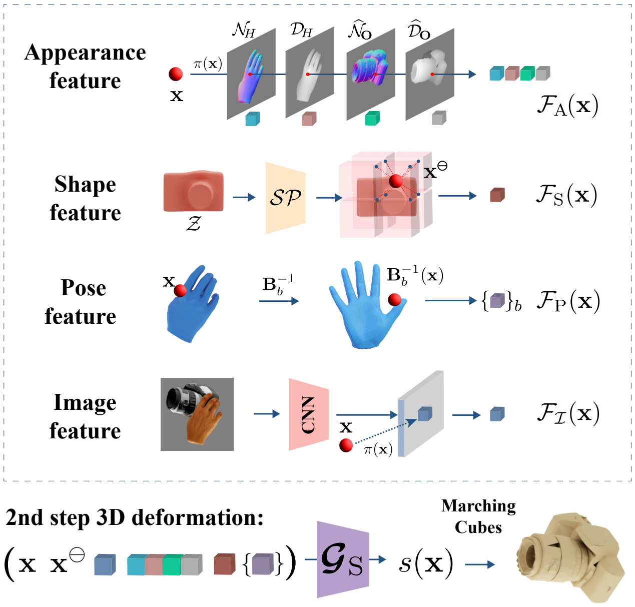
C. 3D Pose-aware Feature. To reconstruct an object instance held by hand, it is also essential for the network to also encapsulate the articulated hand poses. Therefore, we represent the pose-aware feature as the pose-conditioned value of the query point $\mathbf{x}$. The hand pose is described as a set of rigid transformation matrices $\{\mathbf{B}\}_{b=1}^{K}$, where each $\mathbf{B}_b$ represents the pose of the b-th joint' s local frame in the world system. $\mathbf{B}$ can be transformed from MANO pose $\boldsymbol{\theta}$ using Rodrigues' rotation formula. As demonstrated in the neural articulated shape approximation literature [14, 9], the query point $\mathbf{x}$ is more expressive when represented in the hand's rest-pose system (bone's local frame: $\mathbf{B}_b^{-1}$). In light of this, we design $\mathbf{x}$'s pose-conditioned feature $\mathcal{F}_{\mathrm{P}}$ as its coordinates in the total of $K = 16$ joints' local frames:

$$\mathcal{F}_{\mathrm{P}}(\mathbf{x}) = \{\mathbf{B}_b^{-1}(\mathbf{x})\}_{b=1}^{K}. \tag{4}$$

The Implicit Network. After collecting the three types of local features, we concatenate them channel-wise to form one input to $\mathcal{G}_{\mathrm{S}}$. In addition, we extract the image feature at the projected coordinate $\pi(\mathbf{x})$ on the four feature pyramid layers of the ResNet-50 [27] encoder. Inclusion of these features enables the $\mathcal{G}_{\mathrm{S}}$ to capture a broader range of visual cues at various receptive fields. The 4-layer extracted features are concatenated channel-wise and subsequently mapped to a 16-dimensional feature vector, denoted by $\mathcal{F}_{\mathcal{I}}(\mathbf{x})$. Beside, the query point's positoins in world $\mathbf{x}$ and object-canonical spaces $\mathbf{x}^{\ominus}$ are also used by $\mathcal{G}_{\mathrm{S}}$ as positional encoding, following [11]. The $\mathcal{G}_{\mathrm{S}}$ maps these query points with image feature, appearance-aware, shape-aware, and pose-aware features to the signed distance $s(\mathbf{x})$ of the object:

$$\mathcal{G}_{\mathrm{S}} : (\mathbf{x}, \mathbf{x}^{\ominus}, \mathcal{F}_{\mathcal{I}}(\mathbf{x}), \mathcal{F}_{\mathrm{A}}(\mathbf{x}), \mathcal{F}_{\mathrm{S}}(\mathbf{x}), \mathcal{F}_{\mathrm{P}}(\mathbf{x})) \mapsto s(\mathbf{x}). \tag{5}$$

We train the $\mathcal{G}_{\mathrm{S}}$ with the $L1$ loss between the predicted $s(\mathbf{x})$ and the ground-truth $\widetilde{s}(\mathbf{x})$. When performing inference, we follow the query points sampling strategy as DeepSDF [40]. The meshes surface are extracted from zero-level set by surface construction algorithm [37].



Appearance feature
$\mathcal{N}_H$
$\mathcal{D}_H$
$\widehat{\mathcal{N}}_{\mathrm{O}}$
$\widehat{\mathcal{D}}_{\mathrm{O}}$
$\pi(\mathbf{x})$
$\mathbf{x}$
$\mathcal{F}_{\mathrm{A}}(\mathbf{x})$
Shape feature
$\mathcal{Z}$
$\mathcal{SP}$
$\mathbf{x}^{\ominus}$
$\mathcal{F}_{\mathrm{S}}(\mathbf{x})$
Pose feature
$\mathbf{x}$
$\mathbf{B}_b^{-1}$
$\mathbf{B}_b^{-1}(\mathbf{x})$
$\{\ \}_b$
$\mathcal{F}_{\mathrm{P}}(\mathbf{x})$
Image feature
CNN
$\mathbf{x}$
$\pi(\mathbf{x})$
$\mathcal{F}_{\mathcal{I}}(\mathbf{x})$
2nd step 3D deformation:
$(\mathbf{x}\ \mathbf{x}^{\ominus}\ \{\ \})$
$\mathcal{G}_{\mathrm{S}}$
$s(\mathbf{x})$
Marching Cubes

Figure 3: **Illustration of the second step 3D deformation** $\mathcal{G}_{\mathrm{S}}$, where the three types of awareness ($\mathcal{F}_{\mathrm{A}}$, $\mathcal{F}_{\mathrm{S}}$, $\mathcal{F}_{\mathrm{P}}$) are integrated to predict the signed distance of $s(\mathbf{x})$.

3.4. Constructing Object Shape Prior

Given a set of objects of the same category, we explore three different approaches to find a representative object shape prior. **(1) Vanilla voxel mean.** We follow the apporach in [51] and explicitly construct the object-prior by averaging the voxel representations. **(2) Deep latent mean.** We employ the DeepSDF [40] network to jointly learn a latent code c of each object and a decoder that generates the implicit surface based on c. In this case, the object-prior is retrieved by forwarding the mean of the learned latent codes to the decoder. In DeepSDF, the latent codes are learned separately, without knowing the common structure amongs the objects. **(3) Deep implicit template (DIT).** To obtain a more representative object-prior, we leverage the advanced method: DIT [59], which jointly learns the latent code c, an instance-specified warping function $\mathcal{W}$, and an instance-irrelevant implicit template $\mathcal{T}$. During training, it wraps the implicit template according to different latent codes to model the sign distance of different objects. In this way, the shape of the object-prior is represented as $\mathcal{T}(\cdot)$ (see Fig. 4). Since the approach (3) establishes strong correspondences across shapes, it achieves the best reconstruction quality on unseen objects (see Sec. 4.2-C and Tab. 3).

3.5. COMIC Dataset

To address the limitations of existing datasets for category-level hand-held object reconstruction, we constructed a new dataset, named **COMIC**, which is built at **C**ategory-level and contains rich **O**bjects, **M**aterials, **I**nteractions and **C**amera-views.

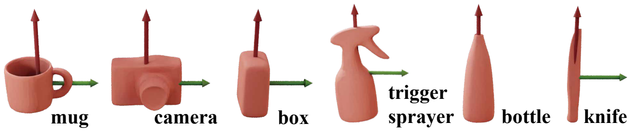


Figure 4: **Illustration of the object-prior.** These prior shapes are extracted using the DIT method, which learns individually across the six categories within the COMIC datset.

We argue that the human's interactions towards a given object category tend to have semantically similar poses, especially for those artificially designed objects. Thus, the main focus of COMIC is to increase the diversity of object instances. However, recording hand's interaction with real-world objects on a large scale is both time-consuming and expensive. To overcome this, we have synthesized images based on **real-world interactions** on **virtual objects**. We use the OakInk [55], a recent dataset of 3D hand-object interactions, as the source. OakInk features realistic hand interactions with multitude virtual objects, transferred from human demonstrations on a few real-world counterparts. We select the six categories from OakInk, namely **mug**, **camera**, **box**, **trigger sprayer**, **bottle**, and **knife**. The shape priors of these categories are shown in Fig. 4. To generate diverse appearances for the hand, we convert the current MANO hand poses to those of NIMBLE [34], and randomly sample an appearance for it. NIMBLE provides high surface resolution, natural muscle shape, and realistic skin tone. To add more variation to object appearance, we additionally applied a random material to the object models. We also randomly pose the camera orientation and set the intrinsics around the object for more diverse viewpoints and perspectives. Using the Blender [12] software with a ray tracing engine, we rendered the images in high fidelity. COMIC comprises 426 K images of 90 K hand-held objects from six frequently used categories. Several examples are shown in Fig. 5 and Supp. Mat.

4. Experiments and Results

4.1. Datasets and Metrics

When reporting the quantitative and ablation results, the CHORD is trained and tested **exclusively** on the COMIC dataset. For generalizing CHORD to in-the-wild images, we also incorporate several hand/hand-object datasets that contain real-world images with COMIC dataset for training. We list these datasets as follows: (1) FreiHand [61] and YouTubeHand [32] for training the hand pose estimation (HPE); (2) OakInk-Image [55] and DexYCB [3] for training HPE, category-level object pose estimation (C-OPE), and our CHORD; For qualitative evaluation of in-the-wild images, we additionally capture several 'out-of-domain' images with 'unseen' object instances.

To evaluate the quality of our method, we report Chamfer Distance (CD, in $\mathbf{1} \times \mathbf{10}\, mm^2$, CD is measured in terms of squared distance). We randomly sample 30,000 points on the surface of both the ground-truth model and our reconstructed mesh and calculate the average bi-directional point-to-point distances. In addition, we report two physical metrics, *i.e.* penetration depth (PD, in cm) and volume (PV, in cm^3), to verify the physical plausibility of our method in modeling hand-object interactions, following [57, 56]. All metrics are reported in camera coordinates system.

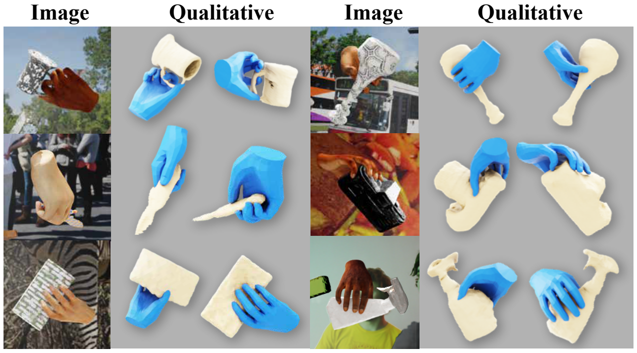


Figure 5: **Illustration of the images and CHORD's prediction on COMIC dataset.**

4.2. Evaluation

For benchmarking CHORD on COMIC dataset, we report the evaluation results on a total of six categories. To explore the effectiveness of different designs in CHORD, we primarily focus on the mug. The reasons are in two-fold. Firstly, the mug represents the only genus-1 object among the six categories, thus exhibiting the most complex geometric topology. Secondly, the mug's thin walls and deep non-convex interior pose challenges for reconstruction.

The quantitative evaluation of the six categories are reported in Tab. 1 (mug) and Tab. 6 (the remaining).

A. CHORD -vs- SOTA. We compare our method with two recent state-of-the-art hand-held object reconstruction methods: AlignSDF [11] and iHOI [57]. AlignSDF first regresses both the hand pose and object translation, and then performs implicit reconstruction in the aligned pose spaces. iHOI utilizes the estimated hand pose to guide object reconstruction, which corresponds to our **pose-awareness** in Sec. 3.3-C. However, neither of these methods is trained at category-level. For fairly comparing CHORD with them, we simulate two category-aware designs, namely, AlignSDF$_C$ and iHOI$_C$. In short, we incorporate an additional module for object-prior's pose estimation into these two models. These two models then predict the signed distance of $\mathbf{x}$, reliant on object-prior's pose. Both networks are trained using our COMIC dataset. Both the iHOI$_C$ and CHORD rely on the preceding hand pose estimation. For a fair comparison, we use the same estimated hand pose during the testing phase. Technical details are in Supp. Mat.

Methods	CD ↓		PD ↓		PV ↓		MPJPE ↓
	◐	●	◐	●	◐	●	◐
AlignSDF$_C$	8.28	3.89	0.54	0.48	3.71	3.16	6.18
iHOI$_C$	8.14	3.80	0.54	0.48	3.61	3.03	6.20
CHORD	**7.69**	**3.11**	**0.42**	**0.36**	**2.79**	**2.30**	**6.08**

Table 1: **Exp. A. Quantitative evaluations on CHORD *vs.* SOTAs.** ↓: lower is better; ◐: ***+Est.pose***; ●: ***+GT.pose***.

To be specific, when the input pose for the CHORD are derived from preceding tasks' estimations, we refer to this setting as '***+Est.pose***'. Moreover, we conduct an additional set of evaluations utilizing the ground-truth poses of hand and object-prior as inputs to CHORD. This approach is executed to reveal potential upper-bound performance. We refer to this setting as '***+GT.pose***'.

Although CHORD is not targeting on hand's reconstruction, we empirically find that incorporating the HPE module (recall Sec. 3.1) along with CHORD's second-step, where the two modules share the same ResNet backbone during training, can further improve the HPE results. With this in mind, in SOTA comparison, we also report the mean per joint position error (MPJPE, in mm) of hand. From Tab. 1, we can conclude that our method outperforms the two previous state-of-the-art methods in both reconstruction and physical-related qualities. We visualize CHORD's reconstruction results under the ***+Est.pose*** setting in Fig. 5.

B. Ablation on Three Awareness. To validate the effectiveness of the three proposed awareness for object reconstruction, we design eight experimental settings. We started with a baseline model that merely used image features as input, gradually incorporating **appearance**, **shape**, and **pose awareness**, and final reach CHORD's full design. Tab. 2 shows that the reconstruction performance improves with more awareness types incorporated. The three types of awareness improve the baseline under the predicted pose setting by 36%, 38%, and 37%, respectively. Our full model, CHORD, achieves the best performance and outperforms the baseline by 40% and 62% under the predicted and ground-truth poses settings, respectively.

$\mathcal{F}_\mathcal{I}$	$\mathcal{F}_\text{A}$	$\mathcal{F}_\text{S}$	$\mathcal{F}_\text{P}$	Chamfer Distance ↓	
				+Est.pose	***+GT.pose***
✓	✗	✗	✗	12.88	8.16
✓	✓	✗	✗	8.25	3.78
✓	✗	✓	✗	7.99	3.45
✓	✗	✗	✓	8.14	3.80
✓	✓	✓	✗	7.93	3.38
✓	✓	✗	✓	7.96	3.49
✓	✗	✓	✓	7.94	3.29
✓	✓	✓	✓	**7.69**	**3.11**

Table 2: **Exp. B.** Gain analysis of the three types of awareness. The ✓ indicates that the corresponding feature is incorporated by the variant of CHORD. The $\mathcal{F}_\mathcal{I}$ is always used by all variants.

Methods	CD ↓
Voxel Mean	10.27
DeepSDF Mean	9.41
DIT [59]	**7.99**

Table 3: **Exp. C.** Comparison of the different methods for generating object-prior.

Methods	CD ↓
$\mathbf{x}$ to $\mathbf{J}$ & $\mathbf{V}_H$ dist.	10.00
$\mathbf{x}$'s signed dist.	7.50
$\mathbf{B}_b^{-1}(\mathbf{x})$ w/o global	7.41
$\mathbf{B}_b^{-1}(\mathbf{x})$ w/ global	**7.27**

Table 4: **Exp. D.** Comparison of different methods for embedding hand pose into $\mathcal{G}_\text{S}$.

C. Different Object Shape Prior. In this study, we compare the performance of our CHORD with different methods of generating object-prior. To simplify the experimental setup, we exclusively utilize the $\mathcal{F}_\text{S}$ feature as input for $\mathcal{G}_\text{S}$, keeping all other variables constant, and report scores under the ***+Est.pose*** setting. From Tab. 3, we conclude that object-prior generated by implicit function outperforms that of explicit voxel mean. Among the tested methods, the object-prior from DIT achieves the best reconstruction performance. We attribute this improvement to that the learned implicit template has embedded the shape deformation within categories and thus enables the $\mathcal{SP}$ to extract spatial features more effectively.

D. Different Pose Feature. To explore the way of embedding hand pose into object reconstruction, we propose four experiments under the ***+Est.pose*** setting. (1) Inspired by the ContactPose [1], we calculate the distance from each query point $\mathbf{x}$ to 21 hand joints, as well as the vector to the nearest point on the hand mesh surface, and its dot product with the surface normal, resulting in $\mathcal{F}_\text{P} \in \mathbb{R}^{23}$. (2) We compute the signed distance of $\mathbf{x}$ to the hand mesh using the released neural occupancy model of hand: HALO [30], resulting in $\mathcal{F}_\text{P} \in \mathbb{R}^1$. (3) Following iHOI, we transfer the $\mathbf{x}$ the hand canonical space using the inversion of MANO transformation. Notably, iHOI ignores the $\mathbf{x}$ in the wrist (root) aligned system, resulting in $\mathcal{F}_\text{P} \in \mathbb{R}^{45}$. (4) Additionally, we consider the hand's root transformation. Therefore, our pose-aware feature: $\mathcal{F}_\text{P} \in \mathbb{R}^{48}$. As shown in Tab. 4 the incorporation of information regarding $\mathbf{x}$ expressed in the hand canonical space improves the object reconstruction quality.

E. Different Appearance Feature. As presented in Tab. 5, we conduct an ablation study on the **appearance awareness** and report score using ***+Est.pose*** setting. Two input forms are explored for normal and depth information: rendering the hand and object together, or rendering them separately. These are indicated by the green ✓ and blue ✓ checkmarks, respectively. To minimize the effect of noise and demonstrate the upper bound of CHORD, we used the rendered ground-truth normal and depth map as input to

$\widetilde{\mathcal{N}}_H, \widetilde{\mathcal{N}}_\mathbf{O}$	✗	✓	✗	✓	✓
$\widetilde{\mathcal{D}}_H, \widetilde{\mathcal{D}}_\mathbf{O}$	✗	✗	✓	✓	✓
CD ↓	7.88	8.03	7.58	7.45	**7.07**

Table 5: **Exp. E.** Ablation study on different appearance features.

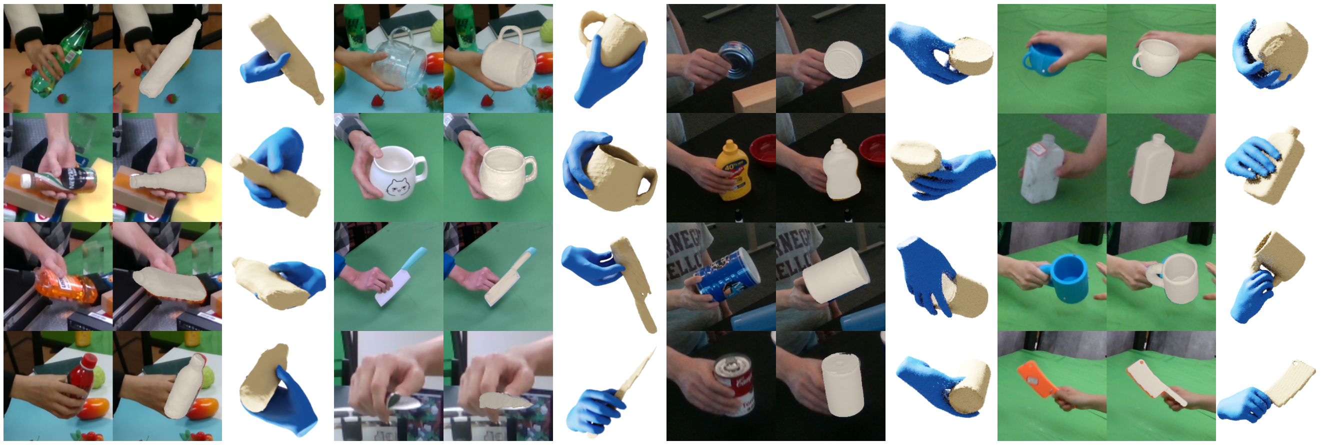


Figure 6: **Qualitative results of CHORD's prediction on in-the-wild images.** In the first six columns, the model is tested on images of unseen objects collected in the wild. The results show the reconstruction mesh in the camera view and another free view. The two sets of results on the right demonstrate that CHORD is evaluated on the unseen camera views of DexYCB and OakInk.

Catgory	Chamfer Distance ↓, under *+Est.pose*		
	COMIC	OakInk	DexYCB
bottle	19.86	34.37	10.24
knife	69.11	32.63	N/A
camera	37.85	74.97	N/A
trigger sprayer	43.07	38.81	N/A
box	15.50	N/A	19.16

Table 6: **Results of CHORD on category-level.** Evaluation on COMIC uses unseen objects, and on DexYCB and OakInk use unseen views. 'N/A': the dataset does not contain such category.

$\mathcal{F}_{\text{A}}$. The result in Tab. 5 shows that using separated normal and depth maps as input significantly improves the reconstruction accuracy of the network. This is because the HOI dataset contains a large amount of occlusion, and decoupling the 2D information benefits the subsequent network.

F. Generalization Ability. The generalization ability of CHORD is evaluated on two different scenarios. The first scenarios relates to existing datasets, whereas the second pertains to real-world 'in-the-wild' images. The CHORD model used in both experiments is trained using the mixture of COMIC, OakInk, and DexYCB datasets.

Existing Dataset. Following [57, 11], we assess CHORD's generalization ability under the 'unseen view' splits using the OakInk and DexYCB datasets. Specifically, the objects presented during testing are not new to the training set, but are observed from a previously unseen camera viewpoint. The corresponding scores are listed in Tab. 6. Qualitative results are shown in Fig. 6 right.

In-the-wild. Second, to unveil potential applications of our CHORD on real-world in-the-wild scenario, we capture images of hand-held objects in the real-world setting and evaluate CHORD utilizing these images. Following iHOI [57], we incorporate hand-object segmentation as an additional input to this process. We also estimate the global wrist translation via known camera intrinsic and learned wrist-relative joints' position using [60, Eq.5]. Our model's performance in a real-world context is demonstrated in Fig. 6 left, wherein CHORD exhibits adept reconstructions of previously unseen objects from multiple viewpoints. By incorporating prior information, our model achieves remarkable reconstruction results even on objects with complex topology like mug handle. Furthermore, by leveraging our COMIC dataset, the model can generate convincing object meshes, including those for transparent objects. For further details on the real-world settings and additional visualization results, please refer to Supp. Mat..

5. Conclusion

This study introduces a novel method, named CHORD, for category-level hand-held object reconstruction that overcomes the limitations of prior methods. CHORD utilizes a pre-trained categorical object-prior and incorporates three types of awareness, namely appearance, shape, and pose, to ensure accurate reconstruction. To address the lack of hand-object data at the category-level, we also introduce a large-scale synthetic dataset, named COMIC, which includes rich object models, realistic materials, diverse hand-object interactions, and camera views. Extensive evaluations demonstrate that CHORD outperforms SOTA methods in both quantitative and qualitative metrics. The proposed approach has the potential to advance AI's understanding of human activities by accurately modeling the shape of objects that are interacting with hands.

Acknowledgments. This work was supported by the National Key R&D Program of China (No. 2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, and Shanghai Science and Technology Commission (21511101200).

References

[1] 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 *European Conference on Computer Vision (ECCV)*, 2020. 3, 7

[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. 2, 3

[3] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6

[4] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[5] Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. In *AAAI Conference on Artificial Intelligence*, 2023. 1

[6] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. FS-Net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[7] Wei Chen, Xi Jia, Zhongqun Zhang, Hyung Jin Chang, Linlin Shen, and Ales Leonardis. Category-level 6d object pose estimation with flexible vector-based rotation representation. *arXiv preprint arXiv:2212.04632*, 2022. 3

[8] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *International Conference on Computer Vision (ICCV)*, 2021. 5

[10] 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*, 2021. 3

[11] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 5, 6, 8

[12] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[13] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[14] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA: neural articulated shape approximation. In *European Conference on Computer Vision (ECCV)*, 2020. 5

[15] Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[16] Yang Fu and X. Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *ArXiv*, abs/2206.15436, 2022. 3

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

[18] 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. 3

[19] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 4

[20] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[21] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[22] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *Computer Vision and Pattern Recognition (CVPR)*, 2010. 3

[23] 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. 2

[24] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint Transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

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

[26] Yana Hasson, Gul 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, 3

[27] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[29] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[30] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, 2021. 7

[31] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, 2020. 1, 2, 3

[32] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[33] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 3

[34] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. NIMBLE: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 2022. 6

[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] 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, 3

[37] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM SIGGRAPH computer graphics*, 1987. 5

[38] A.T. Miller and P.K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics Automation Magazine*, 2004. 2

[39] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-based clothed people. *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[40] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[41] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[42] 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 (TPAMI)*, 2017. 2, 3

[43] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 2, 3

[44] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images. In *European Conference on Computer Vision (ECCV)Workshops*, 2018. 3

[45] Shunsuke Saito, Tomas Simon, Jason M. Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 4

[46] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3d object detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[47] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2, 3

[48] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[49] Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[50] 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*, 2016. 2, 3

[51] Bram Wallace and Bharath Hariharan. Few-shot generalization for single-image 3d reconstruction via priors. In *International Conference on Computer Vision (ICCV)*, 2019. 5

[52] He Wang, Srinath Sridhar, Jingwei Huang, Julien P. C. 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

[53] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit clothed humans obtained from normals. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 4

[54] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[55] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[56] 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, 6

[57] Yufei Ye, Abhinav Kumar 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, 3, 4, 6, 8

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

[59] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 5, 7

[60] 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. 3, 8

[61] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHand: A dataset for markerless capture of hand pose and shape from single rgb images. In *International Conference on Computer Vision (ICCV)*, 2019. 6