What’s in your hands? 3D Reconstruction of Generic Objects in Hands

Yufei Ye\textsuperscript{12} \quad Abhinav Gupta\textsuperscript{12} \quad Shubham Tulsiani\textsuperscript{1}
\textsuperscript{1}Carnegie Mellon University \quad \textsuperscript{2}Meta AI
yufeiy2@cs.cmu.edu \quad gabhinav@fb.com \quad shubhtuls@cmu.edu

https://judyye.github.io/ihoi/

Figure 1. Given an RGB image depicting a hand holding an object, we infer the 3D shape of the hand-held object (rendered in the image frame and from a novel view).

Abstract

Our work aims to reconstruct hand-held objects given a single RGB image. In contrast to prior works that typically assume known 3D templates and reduce the problem to 3D pose estimation, our work reconstructs generic hand-held object without knowing their 3D templates. Our key insight is that hand articulation is highly predictive of the object shape, and we propose an approach that conditionally reconstructs the object based on the articulation and the visual input. Given an image depicting a hand-held object, we first use off-the-shelf systems to estimate the underlying hand pose and then infer the object shape in a normalized hand-centric coordinate frame. We parameterized the object by signed distance which are inferred by an implicit network which leverages the information from both visual feature and articulation-aware coordinates to process a query point. We perform experiments across three datasets and show that our method consistently outperforms baselines and is able to reconstruct a diverse set of objects. We analyze the benefits and robustness of explicit articulation conditioning and also show that this allows the hand pose estimation to further improve in test-time optimization.

1. Introduction

Humans interact with their surrounding world with their hands. Not only do we understand interaction as abstract concepts such as touching screens, squeezing balls, holding pens, etc., we also perceive their underlying 3D shape. Holding a pen means a stick lying on the purlicue and gripped by thumb, index, and middle fingers; holding a bowl is placing it on top of an up-facing palm. We aim to build a recognition system that can perceive and reason about the geometric information of hand-object interactions (HOI) for generic objects.
Over the past decade, we have made significant advances in inferring the 3D shape of both hands and objects in isolation. Hand poses can be recovered in the form of 2D keypoints, 3D skeletons [34, 45–47, 74], or even full 3D meshes [2, 56] via either fitting templates or direct regression. On the other hand, recent works have also pursued scaling object reconstruction from estimating the 6D pose of one specific known object [51] to more generic objects, such as various instances within one category [23, 35, 39], or even pursuing a joint model for cross-category reconstruction [11, 22]. But one area where the progress has been quite limited is understanding human-object interactions (HOI) specifically for manipulable objects [14, 58].

Reconstruction of objects in hand in the wild is highly challenging and ill-posed due to lack of data, presence of mutual and self-occlusion. Current works [4, 17, 30, 41, 62] typically focus on reconstructing a handful of known object (3D model is given) which reduces reconstruction to a 6D pose estimation problem. We argue that knowing the 3D template of the object as a priori during inference is a strong assumption and prevents these systems from reconstructing unknown objects. Furthermore, they struggle to handle various object shapes in the wild as these templates are rigid and instance-specific. In contrast, our work studies hand-object reconstruction without object templates and instead focuses on reconstructing HOI for novel objects from images.

Our key observation is that hand articulation is driven by the local geometry of the object. Thus, hand articulation provides strong cues for the object in interaction. Fingers curled like fists indicate thin handles in between while open palms are likely to interact with flat surfaces. Instead of treating the hand occlusion as noise to marginalize over, we explicitly consider hand pose as informative cues for the object it interacts with. We operationalize this idea by conditionally predicting the object shape based on hand articulation and the input image. Instead of estimating both hand pose and object shape jointly, we leverage advances in hand pose reconstruction to estimate hand pose first. Given the inferred articulated hand along with the input image, our approach then reconstructs the object in a normalized hand-centric coordinate frame.

We evaluate our method across three datasets including synthetic and real-world benchmarks and compare ours with prior explicit and implicit HOI reconstruction methods that infer the shape of unknown objects independent of hand pose. Our articulation-conditioned object shape prediction consistently outperforms prior works by large margins and can reconstruct various objects in a wide range of shapes. We also analyze how our model benefits from articulation-aware coordinates. Lastly, we show that the initial hand pose estimation could be further improved by encouraging interaction between the predicted hand and the object.

2. Related Work

Hand pose estimation. Approaches tackling hand pose estimation from RGB(-D) images can be broadly categorized as being model-free and model-based. Model-free methods [10, 34, 45–47, 52, 53, 74] typically detect 2D keypoints and lift them to 3D joints position or hand skeletons. Some works [10, 18, 49] then directly predict 3D meshes vertices from the 3D skeleton by coarse-to-fine generation. Model-based methods [2, 56, 59, 72, 73] leverage statistical models like MANO [55] whose low-dimensional pose and shape parameters can be directly regressed [2, 56] or optimized [59, 72, 73]. These model-based methods are generally robust to occlusion, domain gap etc., and we build on these in our work. In particular, we use a state-of-the-art model-based reconstruction system [56] to first estimate hand pose and condition the inference of the hand-held object shape on this prediction. While our work primarily focuses on inferring the object shape given an off-the-shelf hand pose estimate, we also show that jointly reasoning about the geometric interaction between the predicted 3D object and the inferred hand pose can help improve the initial hand pose estimate.

Single-view Object Reconstruction. Reconstructing objects from images is a long-standing problem, dating back to the seminal thesis from Larry Roberts [51], where 3D models were known and the problem was reduced to 6-DoF pose estimation. In the following decades, several works have since aimed to reconstruct more generic objects from images [1, 26, 33, 38]. Recent learning-based methods can learn category-specific networks for 3D prediction [6, 12, 36] across a broad class of objects, and can even do so without direct 3D supervision [23, 35, 39, 69]. Using stronger 3D supervision, other approaches have shown that it is possible to learn a common model across multiple categories, with output representations such as voxels [11, 20, 68], meshes [22, 25, 67], point clouds [16, 40], or primitives [15, 19, 63]. Closer to our work, neural implicit representations [9, 43, 48] have shown the impressive capacity for accurate reconstruction for different topology, and our work extends these to be articulation-conditioned for inferring 3D of hand-held objects. While the field of object reconstruction has witnessed remarkable progress, the state-of-the-art methods typically assume isolated and unoccluded objects in images – and cannot be directly leveraged for reconstructing hand-held objects. Even approaches that are robust to occlusion consider it as noisy context to marginalize over, instead of a source of signal for the shape of the underlying object. In contrast, our work shows that explicitly taking hand pose into account helps infer the 3D structure of objects more accurately.

Reconstructing hand-held objects. Understanding HOI in 3D is a very challenging problem due to mutual occlusion.
Figure 2. Given an image of a hand-held object, we first use an off-the-shelf system to estimate hand articulation $\theta_A$ and the camera pose $\pi_w$. With the predicted articulated hand along with the image, the object shape is reconstructed by an implicit network. For each query point $x$ in canonical hand wrist frame, it is transformed to image space $x_p$ to get visual feature $\phi = g(x_p; I)$. In parallel, we also encode its articulation-aware representation $\psi = h(x; \theta)$. Then we use an implicit decoder to predict signed distance value $s = f(x, \phi, \psi)$.

3. Method

Given an image depicting a hand holding an object, we aim to reconstruct the 3D shape of the underlying object. Our key insight is that the hand articulation is predictive of the object shape within it, for example, fingers pinching together indicate a thin stick-like structure between them. We operationalize this intuition by explicitly conditioning the inference of the object shape on (predicted) hand articulation.

As shown in Fig 2, we first use an off-the-shelf system to estimate hand articulation and predict the camera transformation that projects the canonical articulated hand to the image coordinates. Given the predicted hand along with the input image, we then infer the object shape via an articulation-conditioned reconstruction network. This network is implemented as a point-wise implicit function [48] that maps a query 3D point to a signed distance from the object surface, and the zero-level set of this function can be extracted as the object surface [42]. Instead of predicting this 3D shape in the image coordinate frame, our implicit reconstruction network infers it in a normalized frame around the hand wrist. This allows the network to learn relations between the hand articulation and object shape that are invariant to global transformations.

More formally, given an input image $I$, we first infer the underlying the hand pose $\theta$ and the camera pose $\pi$. Then, for any point $x$ in the normalized wrist frame, the object inference model takes in the query point with the image and predicts its signed distance function $s$. More specifically, the projection of the point to image coordinates is used to obtain corresponding visual features $\phi = g(\pi(x); I)$. In parallel, we also encode its position relative to each hand joint to extract an articulation-aware representation $\psi = h(x; \theta)$. The point-wise visual feature and articulation embedding are then used by an implicit decoder to predict signed distance value $s = f(\phi, \psi)$ at the query point $x$.

3.1. Preliminary: Hand Reconstruction

We use an off-the-shelf system [56] to estimate hand articulation and associated camera pose from an input image.
The reconstruction method is model-based which directly regresses a 45-dimensional articulation parameter \( \theta_A \) and a 6-dim global rotation and translation \( \theta_w \) along with a weak perspective camera. We rig the parametric MANO model by the predicted articulation pose \( \theta_A \) to obtain an articulated hand mesh in a canonical frame around the wrist. To relate a point in the wrist frame to the image space, we first transform the hand by the predicted global transformation and then project it by the camera matrix. As an implementation detail, we convert the predicted weak perspective camera to a full perspective one as it helps to account for large perspective effects. In summary, we project a query point in the canonical wrist frame to the image by

\[
x_p = \pi_{\theta_w}(x) = KT_{\theta_w}x
\]

where \( K \) is the camera intrinsic and \( T_{\theta_w} \) is the global rigid transformation of the hand.

### 3.2. Articulation-conditioned SDF

Given the predicted hand articulation \( \theta_A \), and camera matrix \( \pi_{\theta_w} \), our articulation-conditioned object inference network takes an additional input image \( I \) to output the articulation-aware encoding \( \psi = h(X; \theta_A) \). The encoding is a concatenation of the coordinates relative to every joint. Given the articulation parameter \( \theta_A \), we run forward kinematics to derive transformation \( T(\theta_A): \mathbb{R}^3 \rightarrow \mathbb{R}^{45} \) that maps a point in wrist frame to each joint coordinate. The 15 joint coordinates are position encoded \([66]\) and concatenated together as the final representation \( h(x; \theta_A) = \gamma(T(\theta_A)x) \) where \( \gamma \) is the positional encoder. For more details please refer to appendix.

**Implicit decoder.** The decoder maps the query points with visual feature and articulation embedding to a signed distance value \( f(x, \phi, \psi) = s \). These two representations \( \phi, \psi \) are concatenated together and passed along with the query point to the decoder. The decoder simply follows the design in DeepSDF \([48]\) which consists of 8 layers of MLP with a skip layer.

### 3.3. Training

To learn our articulation conditioned neural implicit field, we rely on a training dataset where we assume known hand pose and 3D shape of the object in the image frame. We preprocess the data by sampling points inside and outside of the object around the hand to calculate the ground-truth SDF. 95% of the points are sampled around the surface of the objects and others are sampled uniformly in the space. During training, the network optimizes to match the predicted SDF to the ground-truth at the sampled points with the eikonal term as a regularizer.

\[
\mathcal{L} = \|s - \hat{s}\| + \lambda(\|\nabla s\| - 1)^2
\]

After the network is learned, at inference time, we do not require knowledge of the object 3D shape in the canonical frame a priori which is a major limitation of most most prior works.

### 3.4. Refining hand pose

While our work primarily focuses on object reconstruction conditioned on a predicted hand pose, this initial pose
prediction, while reliable, is not perfect. As object reconstruction also leverages visual cues, our insight is that it can provide complementary information to further refine the predicted hand pose. During inference, we show that the predicted hand pose and object shape can therefore be further (jointly) optimized by enforcing physical plausibility – by encouraging contact while discouraging intersection.

We optimize the articulation pose parameters with respect to these two interaction terms, which can be naturally incorporated with an SDF representation. To discourage intersection between the hand and object, we penalize if the points on the hand surface are predicted to have negative SDF values by the object reconstruction model. Following prior work [31], we encourage hand-object contact for specifically defined regions – if the surface points in these contact regions are near the object surface, they are encouraged to come even closer.

\[
\min_{\theta} \sum_{x \in H} \| \max(-f(x), 0) \| + \\
\sum_{x \in C} \max(\| \min(f(x) - \tau, 0) \| - \epsilon, 0)
\]

Note that the SDF \( f \) is conditioned on articulation thus it is also a function of the hand pose \( \theta \). As we refine the hand pose, the SDF of the object also changes accordingly. One could continuously update the SDF every time the pose is updated but we use a simpler solution that fixes the SDF during hand pose optimization and only update it once using the final optimized pose \( \theta \).

4. Experiments

We compare our method with two model-free baselines [31, 37] on three datasets – one synthetic, and two real-world. We show that our approach outperforms baselines across these datasets, both in terms of object reconstruction and modeling hand-object interaction. We further analyze the benefit from explicitly considering hand pose and the benefit from our particular form of articulation-aware positional encoding. Lastly, we show that our reconstructed hand-held object could further refine the initial hand pose estimation and improve hand-object interaction.

Datasets and Setup. We evaluate our method across three datasets.

- ObMan [31] is a synthetic dataset that consists of 2772 objects from 3D warehouse [7]. The grasps are automatically generated by GraspIt [44], resulting in a total of 21K grasps. The grasped objects are rendered over random backgrounds using Blender. We follow the standard splits where there is no overlap between the objects used in training and testing.

- HO-3D [29] is a real-world video dataset consisting of 103k annotated images capturing 10 subjects interacting with 10 common YCB objects [75]. The ground-truth is annotated using multi-camera reconstruction pipelines. To test on more shapes, we create a custom split by holding out one video sequence per object as test set. Please refer to the appendix for more details.

- MOW [5] dataset consists of a curated set of 442 images, spanning 121 object templates, collected from in-the-wild hand-object interaction datasets [14, 58]. It is more diverse in terms of both appearance and object shape compared to the HO3D dataset, but only provides approximate ground-truth. These object shape and hand pose annotations are obtained via a single-frame optimization-based method [5]. We use 350 randomly selected examples for training and the remaining 92 for evaluation.

For the ObMan dataset, we use the hand pose predictor from Hasson et al. [31] as this system is specifically trained on this synthetic dataset. For HOI and MOW, we use the FrankMocap [56] system that is trained on multiple real-world datasets. Since HOI data with ground truth in the real world is scarce and lacks diversity in terms of object shape and appearance, we initialize our method and baselines with models pretrained on ObMan and finetune them on HO3D and MOW datasets.

Evaluation metrics. We evaluate the quality of both, object reconstruction and the relation between object and hand. To evaluate the reconstruction quality, we first extract a mesh from the predicted SDF. Following prior works, we then evaluate the object reconstruction by reporting Chamfer distance (CD), but also report the F-score at 5mm and 10mm thresholds as Chamfer distance is more vulnerable to outliers [61]. Another desirable property for HOI reconstruction is that the interpenetration between the hand and
the object should be minimal, we also report the intersection volume between two meshes (in $cm^3$) as a measure of understanding the relations between the hand and object.

**Baselines.** While most prior approaches require a known object template, recent work by Hasson et al. (HO) [31] and Karunratanakul et al. (GF) [37] can tackle the same task as ours – inferring the shape of a generic object from a single interaction image. HO jointly regresses MANO parameters to estimate hand pose and reconstructs the object in the camera frame. It is based on Atlas-Net [25], and deforms a sphere to infer the object mesh. Closer to our approach, GF is also based on a point-wise implicit network. It takes an image as input and outputs an implicit field that maps a point in the camera frame to a signed distance to both, the hand and object, while also predicting hand part labels.

Our approach differs from these baselines in three main aspects. First, both prior approaches infer object shape independent of the predicted hand pose, while we formulate hand-held object reconstruction as conditional inference. Second, these baselines encode the visual information only via a global feature while we additionally use pixel-aligned local features. Third, while both baselines reconstruct objects in the camera frame, we predict them in a normalized wrist frame with articulation-aware positional encoding. Note that while our approach predicts 3D in a hand-centric frame, the evaluations are all performed in the image coordinate frame for fair comparison (using the predicted hand pose to transform our prediction).

### 4.1. Results on ObMan

We visualize the reconstructed objects and the corresponding hand poses in Figure 3. While baselines can predict the coarse shape of the object, they typically lose sharp details such as the corner of the phone and sometimes miss part of the surface of the object. This may be because they only use a global feature of the image that loses spatial resolution. In contrast, our method reconstructs shape that better aligns with the visual inputs from the original view and hallucinates the invisible part of the objects occluded by hands.

This is also empirically reflected in quantitative results reported in Table 1. We outperform baselines by a large margin on F-score. Our improvement over baselines is particularly significant on the smaller threshold, indicating that our method is better to reconstruct local shape. In terms of Chamfer distance, ours is better than GF that is also based on implicit fields. Ours is not as good as HO in Chamfer distance probably because HO explicitly trains to minimize Chamfer distance with a regularizer on edge length which discourages large displacements from a sphere thus producing fewer outliers.

### 4.2. Evaluation on real-world datasets

We visualize the reconstruction in the image frame and a novel view in Figure 4 and Figure 5. GF can predict blobby cylinders but the reconstructed objects lack details in shape such as around the neck of the mustard bottle, and sometimes reconstructs a different object shape such as predicting boxes instead of scissors. In contrast, our method is able to generate diverse object shape more accurately including boxes, power drills, bottles, pens, cup, spray bottles etc.

### 4.3. Ablations

**Importance of articulation conditioning.** We analyze how hand articulation conditioning helps hand-held object reconstruction by constructing a variant of our method that only conditions on pixel-aligned image features. This approach is analogous to the one proposed by Saito et al. [57] where human 3D shape is inferred by a pixel-aligned implicit network. Table 3 reports results of this variant that do not explicitly consider hand articulation and we observe that the object reconstruction degrades by a large margin.

---

<table>
<thead>
<tr>
<th>ObMan</th>
<th>F-5↑</th>
<th>F-10↑</th>
<th>CD↓</th>
<th>Vol ↓</th>
<th>HO [31]</th>
<th>F-5↑</th>
<th>F-10↑</th>
<th>CD↓</th>
<th>Vol ↓</th>
<th>GF [37]</th>
<th>F-5↑</th>
<th>F-10↑</th>
<th>CD↓</th>
<th>Vol ↓</th>
<th>Ours</th>
<th>F-5↑</th>
<th>F-10↑</th>
<th>CD↓</th>
<th>Vol ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-5↑</td>
<td>0.23</td>
<td>0.56</td>
<td><strong>0.64</strong></td>
<td>8.64</td>
<td>0.11</td>
<td>0.22</td>
<td>4.19</td>
<td>9.44</td>
<td>0.03</td>
<td>0.06</td>
<td>49.8</td>
<td>25.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-10↑</td>
<td>0.30</td>
<td>0.51</td>
<td>1.39</td>
<td>1.84</td>
<td>0.12</td>
<td>0.24</td>
<td>4.96</td>
<td>6.31</td>
<td>0.06</td>
<td>0.13</td>
<td>40.1</td>
<td><strong>8.82</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD↓</td>
<td>0.42</td>
<td>0.63</td>
<td>1.02</td>
<td><strong>1.74</strong></td>
<td><strong>0.28</strong></td>
<td><strong>0.50</strong></td>
<td>1.53</td>
<td><strong>4.77</strong></td>
<td><strong>0.13</strong></td>
<td><strong>0.24</strong></td>
<td>23.1</td>
<td>19.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vol ↓</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td>5900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Quantitative results for object reconstruction error using F-score ($5mm, 10mm$), Chamfer distance ($mm$) and intersection volume ($cm^3$). We compare our method with prior works [31, 37] on ObMan, HO3D, MOW datasets.

<table>
<thead>
<tr>
<th>train set</th>
<th>F-5↑</th>
<th>F-10↑</th>
<th>CD↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ObMan</td>
<td>0.14</td>
<td>0.27</td>
<td>4.36</td>
</tr>
<tr>
<td>MOW</td>
<td>0.15</td>
<td>0.30</td>
<td>4.09</td>
</tr>
</tbody>
</table>

Table 2. Cross-dataset generalization: we report quantitative results on HO3D for models pretrained on ObMan and MOW.

**Zero-shot transfer to HO3D.** We also directly evaluate models that are only trained on ObMan and MOW datasets and report their reconstruction results on HO3D dataset. Both models without finetuning still outperform baselines trained on HO3D dataset. Interestingly, even though the MOW dataset only consists of 350 training images, which is significantly less compared to 21K images from the synthetic dataset, learning from MOW still helps cross-dataset generalization. It indicates the importance of diversity for in-the-wild training. Please see the qualitative result in the appendix.
Figure 4. Visualizing reconstruction of our method and two baselines [31, 37] on HO3D dataset in the image frame and from a novel view.

Figure 5. Visualizing reconstruction of our method and two baselines [31, 37] on MOW dataset in the image frame and from a novel view.

Table 3. Analysis of articulation-conditioning: we report quantitative results of object error in F-score, Chamfer distance (CD), intersection volume on 3 datasets and compare ours with the ablation that only consider visual feature.

<table>
<thead>
<tr>
<th></th>
<th>Ours</th>
<th>Ours w/o Art.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ObMan</td>
<td>0.42</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>1.02</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>1.74</td>
<td>3.93</td>
</tr>
<tr>
<td>HO3D</td>
<td>0.33</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>0.58</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>0.93</td>
<td>1.18</td>
</tr>
<tr>
<td></td>
<td>4.77</td>
<td>6.30</td>
</tr>
<tr>
<td>MOW</td>
<td>0.13</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>23.1</td>
<td>29.0</td>
</tr>
<tr>
<td></td>
<td>19.4</td>
<td>17.3</td>
</tr>
</tbody>
</table>

Figure 6. Visualizing reconstruction of hand-held object with or without explicitly considering hand pose.

while the intersection volume also doubles. This suggests that hand information provides a strong cue that is complementary to visual inputs. Figure 6 visualizes comparison between ours and the variant where our method can better respect hand-object physical relations such as objects do not penetrate the hands and the area around fingertips are likely to be in contact with objects.

Representation of hand articulation matters for generalization. To represent articulation information, a natural alternative to our proposed articulation-aware positional encoding is to simply concatenate the query point with the MANO pose parameter \( \theta_A \), i.e., \( \tilde{h}(x; \theta_A) = [x, \theta_A] \). The result in Table 4 shows that although it performs comparably when provided with ground-truth hand pose parameters, it degrades significantly when with predicted hand pose despite that we perform jitter augmentation on hand articulation for both methods. More interestingly, it performs even worse than the variant without articulation-aware encoding. This indicates that the object shape overfits to pose parameters which are constant within one example. In contrast, our articulation-aware positional encoding generalizes better.

Robustness against hand prediction quality. We use
Table 4. Analysis of articulation-aware encoding: We compare different ways to incorporate hand articulation: articulation-aware positional encoding and pose parameters. Star indicates reconstruct object shape given ground truth hand articulation.

<table>
<thead>
<tr>
<th>Method</th>
<th>F5 ↑</th>
<th>F10 ↑</th>
<th>CD ↓</th>
<th>Vol↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art.-aware PE*</td>
<td>0.49</td>
<td>0.70</td>
<td>0.92</td>
<td>1.73</td>
</tr>
<tr>
<td>Pose param.*</td>
<td>0.46</td>
<td>0.66</td>
<td>1.25</td>
<td>1.44</td>
</tr>
<tr>
<td>Art.-aware PE</td>
<td>0.42</td>
<td>0.63</td>
<td>1.02</td>
<td>1.74</td>
</tr>
<tr>
<td>Pose param.</td>
<td>0.23</td>
<td>0.42</td>
<td>1.82</td>
<td>2.57</td>
</tr>
<tr>
<td>W/o art.</td>
<td>0.37</td>
<td>0.56</td>
<td>1.89</td>
<td>3.93</td>
</tr>
</tbody>
</table>

Table 5. Error analysis against hand pose noise. $\sigma$ is the average prediction error. * marks our unabladed method.

| Noise Level | Gaussian | 50% $\sigma$ | 0.40 | 0.63 | 1.01 | 0.28 | 0.50 | 1.51 |
|            | 100% $\sigma$ | 0.31 | 0.53 | 1.40 | 0.25 | 0.46 | 1.68 |
|            | 150% $\sigma$ | 0.24 | 0.42 | 1.94 | 0.22 | 0.42 | 1.93 |
| Prediction  | 50%   | 0.46 | 0.67 | 0.96 | 0.29 | 0.52 | 1.48 |
|            | 100% * | 0.42 | 0.63 | 1.02 | 0.28 | 0.50 | 1.53 |
|            | 200%  | 0.35 | 0.56 | 1.28 | 0.26 | 0.47 | 1.67 |
| Baselines   | HO    | 0.23 | 0.56 | 0.64 | 0.11 | 0.22 | 4.19 |
|            | GF    | 0.30 | 0.51 | 1.39 | 0.12 | 0.24 | 4.96 |
|            | GT%   | 0.49 | 0.70 | 0.92 | 0.30 | 0.53 | 1.46 |

Table 6. Test-time refinement. We report object error, intersection volume, simulation displacement and hand error before and after test-time refinement.

<table>
<thead>
<tr>
<th></th>
<th>F5 ↑</th>
<th>CD ↓</th>
<th>Vol↓</th>
<th>Sim↓</th>
<th>EPE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>ours w/o rf</td>
<td>0.17</td>
<td>1.02</td>
<td>1.74</td>
<td>3.32</td>
<td>8.9</td>
</tr>
<tr>
<td>ours w rf</td>
<td>0.17</td>
<td>1.00</td>
<td>1.28</td>
<td>3.00</td>
<td>8.7</td>
</tr>
<tr>
<td>ours w GT pose</td>
<td>0.20</td>
<td>0.92</td>
<td>1.73</td>
<td>2.44</td>
<td>–</td>
</tr>
</tbody>
</table>

Figure 7. Top: Object reconstruction given hand pose corrupted by Gaussian on Obman dataset. Bottom: Object reconstruction given hand pose corrupted by prediction error on HO3D dataset.

Figure 8. Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.
References

[18] Luhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 2
[34] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *ECCV*, 2018. 2


[38] Abhijit Kundu, Yin Li, and James M Rehg. 3d-rccn: Instance-level 3d object reconstruction via render-and-compare. In CVPR, 2018. 2


[40] Chen-Hsuan Lin, Chen Kong, and Simon Lucey. Learning efficient point cloud generation for dense 3d object reconstruction. In AAAI, 2018. 2


[50] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. TPAMI, 2017. 3


[52] Grégory Roquez, Maryum Khademi, JS Supančič III, Jose Maria Martinez Montiel, and Deva Ramanan. 3d hand pose detection in egocentric rgb-d images. In ECCV, 2014. 2


[64] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. IJCV, 2016. 3, 8

[65] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In ICCV, 2015. 3


[70] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. CVPR, 2021. 4

[72] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, 2019. 2

