Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips

Yufei Ye    Poorvi Hebbar    Abhinav Gupta    Shubham Tulsiani
Carnegie Mellon University
{yufeiy2, phebbar, gabhinav, shubhtuls}@andrew.cmu.edu
https://judyye.github.io/diffhoi

Abstract

We tackle the task of reconstructing hand-object interactions from short video clips. Given an input video, our approach casts 3D inference as a per-video optimization and recovers a neural 3D representation of the object shape, as well as the time-varying motion and hand articulation. While the input video naturally provides some multi-view cues to guide 3D inference, these are insufficient on their own due to occlusions and limited viewpoint variations. To obtain accurate 3D, we augment the multi-view signals with generic data-driven priors to guide reconstruction. Specifically, we learn a diffusion network to model the conditional distribution of (geometric) renderings of objects conditioned on hand configuration and category label, and leverage it as a prior to guide the novel-view renderings of the reconstructed scene. We empirically evaluate our approach on egocentric videos across 6 object categories, and observe significant improvements over prior single-view and multi-view methods. Finally, we demonstrate our system’s ability to reconstruct arbitrary clips from YouTube, showing both 1st and 3rd person interactions.

1. Introduction

Our hands allow us to affect the world around us. From pouring the morning coffee to clearing the dinner table, we continually use our hands to interact with surrounding objects. In this work, we pursue the task of understanding such everyday interactions in 3D. Specifically, given a short clip of a human interacting with a rigid object, our approach can infer the shape of the underlying object as well as its (time-varying) relative transformation w.r.t. an articulated hand (see Fig. 1 for sample results).

This task of recovering 3D representations of hand-object interactions (HOI) has received growing interest. While initial approaches [4, 16, 22, 40, 69] framed it as 6-DoF pose task estimation for known 3D objects/templates, subsequent methods have tackled the reconstruction of a priori unknown objects [24, 33, 83]. Although single-view 3D reconstruction approaches can leverage data-driven techniques to reconstruct HOI images [24, 83, 90], these approaches cannot obtain precise reconstructions given the fundamentally limited nature of the single-view input. On the other hand, current video-based HOI reconstruction methods primarily exploit multi-view cues and rely on purely geometry-driven optimization for reconstruction. As
a result, these methods are suited for in-hand scanning where a user carefully presents exhaustive views of the object of interest, but they are not applicable to our setting as aspects of the object may typically be unobserved.

Towards enabling accurate reconstruction given short everyday interaction clips, our approach (DiffHOI) unifies the data-driven and the geometry-driven techniques. Akin to the prior video-based reconstruction methods, we frame the reconstruction task as that of optimizing a video-specific temporal scene representation. However, instead of purely relying on geometric reprojection errors, we also incorporate data-driven priors to guide the optimization. In particular, we learn a 2D diffusion network which models the distribution over plausible (geometric) object renderings conditioned on estimated hand configurations. Inspired by recent applications in text-based 3D generation [37, 58], we use this diffusion model as a generic data-driven regularizer for the video-specific 3D optimization.

We empirically evaluate our system across several first-person hand-object interaction clips from the HOI4D dataset [42], and show that it significantly improves over both prior single-view and multi-view methods. To demonstrate its applicability in more general settings, we also show qualitative results on arbitrary interaction clips from YouTube, including both first-person and third-person clips.

2. Related Works

Reconstructing Hand-Object Interactions. Hands and objects inherently undergo mutual occlusions which makes 3D reconstruction extremely ill-posed during interactions. Hence, many works [1, 5, 14, 22, 23, 51, 56, 69, 71, 86] reduce the problem to 6D pose estimation by assuming access to instance-specific templates. Their frameworks of 6D pose optimization can be applied to both videos and images. Meanwhile, template-free methods follow two paradigms for videos and images. On one hand, video-based methods take in synchronized RGB(D) videos [18, 31, 64, 65, 85] or monocular videos [19, 29, 77] and fuse observation to a canonical 3D representation [48]. This paradigm does not use any prior knowledge and requires all regions being observed in some frames, which is often not true in everyday video clips. On the other hand, methods that reconstruct HOI from single images [7, 8, 24, 33, 83] leverage learning-based prior to reconstruct more general objects. While they are able to generate reasonable per-frame predictions, it is not trivial to aggregate information from multiple views in one sequence and generate a time-consistent 3D shape. Our work is template-free and unifies both geometry-driven and data-driven methods.

Generating Hand-Object Interactions. Besides reconstructing the ongoing HOIs, many works have explored generating plausible HOIs in different formulations. Some works model their joint distributions [6, 28, 33, 66]. Works that are usually called affordance or grasp prediction study the conditional distribution that generates hands/humans given an object/scenes, in 3D representation [2, 11, 30, 35, 67] or 2D images [10, 34, 74, 84]. Recently, some other works explore the reverse problem that generates plausible scenes for a given human/hand pose [3, 50, 57, 80, 82, 88]. In our work, we follow the latter formulation to learn an image-based generative model of hand-held objects given hands, since hand pose estimation [62] is good enough to bootstrap the system.

Neural Implicit Fields for Dynamic Scenes. Neural implicit fields [45, 46, 52] are flexible representation that allows capturing diverse shape with various topology and can be optimized with 2D supervision via differentiable rendering [73, 81]. To extend them to dynamic scenes, a line of work optimizes per-scene representation with additional general motion priors [36, 43, 53, 54, 59]. Though these methods can synthesize highly realistic novel views from nearby angles, they struggle to extrapolate viewpoints [15]. Another line of work incorporates category-specific priors [47, 70, 76, 78, 79] to model articulations. They typically work on single deformable objects of the same category such as quadrupeds or drawers. In contrast, we focus on hand-object interactions across six rigid categories where dynamics are mostly due to changing spatial relations and hand articulations.

Distilling diffusion models. Diffusion models [25, 60] have made significant strides in text-to-image synthesis. They can also be quickly adapted to take additional inputs or generate outputs in other domains [27, 87]. Recent works have shown that these image-based diffusion models can be distilled [58, 72] into 3D scene representations for generation or reconstruction [13, 32, 39, 44, 58, 75, 89]. Inspired by them, we also adopt a conditional diffusion model for guiding 3D inference of HOIs. However, instead of learning prior over appearance, we use a diffusion model to learn a prior over a-modal geometric renderings.

3. Method

Given a monocular video of a hand interacting with a rigid object, we aim to reconstruct the underlying hand-object interaction, i.e., the 3D shape of the object, its pose in every frame, along with per-frame hand meshes and camera poses. We frame the inference as per-video optimization of an underlying 3D representations. While the multiple frames allow leveraging multi-view cues, they are not sufficient as the object of interests is often partially visible in everyday video clips, due to limited viewpoints and mutual occlusion. Our key insight is to incorporate both view consistency across multiple frames and a data-driven prior.
of the HOIs geometry. The learned interaction prior captures both category priors, e.g., mugs are generally cylindrical, and hand priors, e.g., pinched fingers are likely to hold thin handles. We train a conditional diffusion model for the prior that guides the HOI to be reconstructed during per-video optimization.

More specifically, given a monocular video $I^t$ with corresponding hand and object masks $M^t = (M^t_h, M^t_o)$, we aim to optimize a HOI representation (Sec. 3.1) that consists of a time-persistent implicit field $\phi$ for the rigid object, a time-varying morphable mesh for the hand $H^t$, the relative transformation between hand and object $T^t_{h\rightarrow o}$, and time-varying camera poses $T^t_{c\rightarrow h}$. The optimization objective consists of two terms (Sec. 3.3): a reprojection error from the estimated original viewpoint and data-driven prior term that encourages the object geometry to appear more plausible given category and hand information when looking from another viewpoint. The prior is implemented as a diffusion model conditioned on a text prompt $C$ about the category and renderings of the hand $\pi(O)$ with geometry cues (Sec. 3.2). It denominates the rendering of the object $\pi(O)$ and backpropagates the gradient to the 3D HOI representation by score distillation sampling (SDS) [58].

3.1. HOI Scene Representation

Implicit field for the object. The rigid object is represented by a time-persistent implicit field $\phi$ that can handle unknown topology and has shown promising results when optimizing for challenging shapes [73, 79, 81]. For every point in the object frame, we use multi-layer perceptrons to predict the signed distance function (SDF) to the object surface, $s = \phi(X)$.

Time-varying hand meshes. We use a pre-defined parametric mesh model MANO [61] to represent hands across frames. The mesh can be animated by low-dimensional parameters and thus can better capture more structured motions, i.e., hand articulation. We obtain hand meshes $H^t$ in a canonical hand wrist frame by rigging MANO with a 45-dim pose parameters $\theta^t_A$ and 10-dim shape parameters $\beta$, i.e., $H^t = \text{MANO}(\theta^t_A, \beta)$. The canonical wrist frame is invariant to wrist orientation and only captures finger articulations.

Composing to a scene. Given the time-persistent object representation $\phi$ and a time-varying hand mesh $H^t$, we then compose them into a scene at time $t$ such that they can be reprojected back to the image space from the cameras. Prior works [21, 23, 56] typically track 6D object pose directly in the camera frame $T^t_{c\rightarrow o}$, which requires an object template to define the object pose. In our case, since we do not have access to object templates, the object pose in the camera frame is hard to estimate directly. Instead, we track object pose with respect to hand wrist $T^t_{h\rightarrow o}$ and initialize them to identity. It is based on the observation that the object of interest usually moves together with the hand and undergoes “common fate” [63]. A point in the rigid object frame can be related to the predicted camera frame by composing the two transformations, camera-to-hand $T^t_{c\rightarrow h}$ and hand-to-object $T^t_{h\rightarrow o}$. For notation convention, we denote the implicit field transformed to the hand frame at time $t$ as $\phi^t(\cdot) = \phi(T^t_{h\rightarrow o}(\cdot))$. Besides modeling camera extrinsics, we also optimize for per-frame camera intrinsics $K^t$ to account for zoom-in effect, cropping operation, and inaccurate intrinsic estimation.

In summary, given a monocular video with corresponding masks, the parameters to be optimized are

$$\phi, \beta, \theta^t_A, T^t_{h\rightarrow o}, T^t_{c\rightarrow h}, K^t$$

(1)

Differential Rendering. To render the HOI scene into an image, we separately render the object (using volumetric rendering [81]) and the hand (using mesh render-
interacting with it common object geometry given its category and the hand
reconstruct the 3D shape of the full object. To do so, we
directly observe all aspects of the object because of occlusions

3.2. Data-Driven Prior for Geometry

Given an arbitrary viewpoint \( v \), both differentiable render-
cans render geometry images including mask, depth,
and normal images, i.e. \( G_h \equiv (M_h, D_h, N_h), G_o \equiv
(M_o, D_o, N_o) \) To compose them into a semantic mask
\( M_{HOI} \) that is later used to calculate the reprojection loss,
we softly blend the individual masks by their predicted

Figure 3. Geometry-informed Diffusion Model: The diffusion
model takes in a noisy geometry rendering of the object, the
geometry rendering of the hand, and a text prompt, to output the
denoised geometry rendering of objects.

By de-noising corrupted images. During training, they take
in corrupted images with a certain amount of noise \( \sigma_t \) along
with conditions and learn to reconstruct the signals [26]:

\[
\mathcal{L}_{DDPM}[x; c] = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, I), i} \| x - D_{\psi}(x_i, \sigma_i, \epsilon) \|^2 \quad (3)
\]

where \( x_i \) is a linear combination of signal \( x \) and noise \( \epsilon \) while \( D_{\psi} \) is the denoiser.

In our case, as shown in Fig. 3, the diffusion model de-
oines the a-modal geometry rendering of an object given
text prompt and hand. Additionally, the diffusion model
is also conditioned on the rendering of uv-coordinate of
MANO hand \( U_h \) because it can better disambiguate if the
hand palm faces front or back. More specifically, the train-
ing objective is \( \mathcal{L}_{\text{diff}} = \mathcal{L}_{DDPM}[G_o; C, G_h, U_h] \). The text
prompt comes from a text template: “an image of a hand
holding \{category\}”.

Implementation Details. When we train the diffusion
model with the rendering of ground truth HOI, we draw
viewpoints with rotation from the uniform distribution in
\( \text{SO}(3) \). We use the backbone of a text-to-image model [49]
with cross attention and modify it to diffuse 5-channel ge-
ometry images (3 for normal, 1 for mask and 1 for depth).
We initialize the weights from the image-conditioned dif-
fusion model [49] pretrained with large-scale text-image
pairs. The additional channels in the first layer are loaded
from the average of the pretrained weights.

3.3. Reconstructing Interaction Clips in 3D

After learning the above interactions prior, at inference
time when given a short monocular clip with semantic
masks of hand and object, we optimize a per-sequence HOI
representation to recover the underlying hand-object inter-
actions. We do so by differentiable rendering of the 3D
scene representation from the original views and from ran-
don novel views. The optimization objectives consist of the
following terms.

Reprojection error. First, the HOI representation is opti-
mized to explain the input video. We render the seman-
tic mask of the scene from the estimated cameras for each
frame and compare the rendering of the semantic masks
(considering hand-object occlusion ) with the ground truth
masks: \( \mathcal{L}_{\text{reproj}} = \sum_t \| M^t - \hat{M}^t \|_1 \)

Learned prior guidance. In the meantime, the scene is
guided by the learned interactino prior to appear more likely
from a novel viewpoint following Scored Distillation Sam-
pling (SDS) [58]. SDS treats the output of a diffusion model
as a critic to approximate the gradient step towards more
likely images without back-propagating through the diffusion
model for compute efficiency:

\[
\mathcal{L}_{SDS} = \mathbb{E}_{v, t, i} [w_t \| \pi(\phi^t) - \hat{C}^*_t \|_2^2] \quad (4)
\]
where $\tilde{G}_h$ is the reconstructed signal from the pre-trained diffusion model. Please refer to relevant works [44, 58] or supplementary for full details.

Other regularization. We also include two regularization terms: one Eikonal loss [17] that encourages the implicit field $\phi$ to be a valid distance function $L_{eik} = \| \nabla \phi \|^2 - 1$, and another temporal loss that encourages the hand to move smoothly with respect to the object $L_{\text{smooth}} = \sum_t \| T_{h \rightarrow o}(t) - T_{h \rightarrow o}(t-1) \|^2$

**Initialization and training details.** While the camera and object poses are learned jointly with object shape, it is crucial to initialize them to a coarse position [38]. We use FrankMocap [62], an off-the-shelf hand reconstruction system, to initialize the hand parameters, camera-to-hand transformations, and camera intrinsic. More specifically, FrankMocap predicts finger articulation $\theta_{A}$, wrist orientation $\theta_{w}$, and a weak perspective camera. The last two are used to compute camera-to-hand transformation and intrinsics of a full perspective camera. See appendix for derivation. We initialize the object implicit field to a coarse sphere [81] and the object poses $T_{h \rightarrow o}$ to identity such that the initial object is roughly round hand palm.

The per-frame hand pose estimation sometimes fails miserably in some challenging frames due to occlusion and motion blur. We run a lightweight trajectory optimization on wrist orientation to correct the catastrophic failure. The optimization objective encourages smooth joint motion across frames while penalizing the difference to the per-frame prediction, i.e., $L = \| H(x^t) - H(\hat{x}) \| + \lambda \| H(x^{t+1}) - H(x^t) \|$ where $\lambda$ is 0.01. Please see appendix for full details.

**4. Experiment**

We first train the diffusion model on the egocentric HOI4D [42] dataset and visualize its generations in Section 4.1. Then, we evaluate the reconstruction of hand-object interactions quantitatively and qualitatively on the held-out sequences and compare DiffHOI with two model-free baselines (Section 4.2). We then analyze the effects of both category-prior and hand-prior respectively, ablate the contribution from each geometry modality, and analyze its robustness to initial prediction errors (Section 4.3). In Section 4.4, we discuss how DiffHOI compares with other template-based methods. Lastly, in Section 4.5, we show that our method is able to reconstruct HOI from in-the-wild video clips both in first-person and from third-person view.

**Dataset and Setup.** HOI4D is an egocentric dataset consisting of short video clips of hand interacting with objects. It is collected under controlled environments and recorded by head-wear RGBD cameras. Ground truth is provided by fitting 6D pose of scanned objects to the RGBD videos. We use all of the 6 rigid object categories in portable size (mug, bottle, kettle, knife, toy car, bowl). To train the diffusion model, we render one random novel viewpoint for each frame resulting in 35k training points. We test the object reconstruction on held-out instances, two sequences per category. All of baselines and our method use the segmentation masks from ground truth annotations and the hand poses from the off-the-shelf prediction system [62] if required.

For in-the-wild dataset, we test on clips from EPIC-KITCHENS [12] videos and casual YouTube videos downloaded from the Internet. The segmentation masks are obtained using an off-the-shelf video object segmentation system [9].

**4.1. Visualizing Data-Driven Priors**

We show conditional generations by the pre-trained diffusion model in Fig. ?? given the geometry rendering of hand (only visualizing surface normal), as well as a text prompt, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the objects and bottom row visualizes the generated object masks overlayed on the given hand masks. Note the left and middle column share the same text condition while middle and right column share the same hand condition.

Figure 4. Generations from conditional diffusion model: Given the geometry rendering of hand $G_h$ (only showing surface normals) and a text prompt $C$, we visualize 4 different generations from the diffusion model. Middle row shows the generated surface normal of the objects and bottom row visualizes the generated object masks overlayed on the given hand masks. Note the left and middle column share the same text condition while middle and right column share the same hand condition.
Table 1. **Comparison with baselines:** We compare our method along with prior works HHOR [29] and iHOI [83] on the HOI4D dataset and report object reconstruction error in $F@5$mm and $F@10$mm scores and Chamfer Distance ($CD$).

<table>
<thead>
<tr>
<th></th>
<th>Mug</th>
<th>Bottle</th>
<th>Kettle</th>
<th>Bowl</th>
<th>Knife</th>
<th>ToyCar</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F@5$</td>
<td>$F@10$</td>
<td>$CD$</td>
<td>$F@5$</td>
<td>$F@10$</td>
<td>$CD$</td>
<td>$F@5$</td>
</tr>
<tr>
<td>HHOR</td>
<td>0.18</td>
<td>0.37</td>
<td>7.0</td>
<td>0.26</td>
<td>0.56</td>
<td>3.1</td>
<td>0.31</td>
</tr>
<tr>
<td>iHOI</td>
<td>0.44</td>
<td>0.71</td>
<td>2.1</td>
<td>0.48</td>
<td>0.77</td>
<td>1.5</td>
<td>0.38</td>
</tr>
<tr>
<td>Ours (DiffHOI)</td>
<td><strong>0.64</strong></td>
<td><strong>0.86</strong></td>
<td><strong>1.0</strong></td>
<td><strong>0.54</strong></td>
<td><strong>0.92</strong></td>
<td><strong>0.7</strong></td>
<td><strong>0.79</strong></td>
</tr>
</tbody>
</table>

Figure 5. **Qualitative evaluation on HOI4D:** We show reconstruction by our method (DiffHOI) along with two baselines [29, 83] in the image frame (left) and another novel view with (top right) or without (bottom right) hand. Please see project website for reconstruction videos.

The generated object match the category information in the prompt while the generations are diverse in position, orientation, and size. Yet, all of the hand-object interactions are plausible, *e.g.* different generated handles all appear at the tip of the hand. Comparing middle and right examples, different category prompts lead to different generations given the same hand rendering. With the same prompt but different hands (left and middle), the generated objects flip the orientation accordingly. In summary, Fig. ?? indicates that the learned prior is aware of both the hand prior and the category-level prior hence being informative to guide the 3D reconstruction from clips.

4.2. **Comparing Reconstructions of HOI4D**

**Evaluation Metric.** We evaluate the object reconstruction errors. Following prior works [20,29], we first align the reconstructed object shape with the ground truth by Iterative Closest Point (ICP), allowing scaling. Then we compute Chamfer distance (CD), F-score [68] at $5mm$ and $10mm$ and report mean over 2 sequences for each category. Chamfer distance focuses on the global shapes more and is affected by outliers while F-score focuses on local shape details at different thresholds [68].

**Baselines.** While few prior works tackle our challenging setting – 3D HOI reconstruction from casual monocular...
Table 2. Analysis of the effect of data-driven priors: Quantitative results on HOI4D for object reconstruction error in the object-centric frame ($F_{@5}$, $F_{@10}$, $CD$) and for hand-object alignment in the hand frame ($CD_h$). We compare our method with ablations that do not use prior, or use other variants of diffusion models that only conditions on hand or category.

<table>
<thead>
<tr>
<th></th>
<th>$F_{@5}$</th>
<th>$F_{@10}$</th>
<th>$CD$</th>
<th>$CD_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No prior</td>
<td>0.47</td>
<td>0.73</td>
<td>2.7</td>
<td>37.0</td>
</tr>
<tr>
<td>Hand prior</td>
<td>0.39</td>
<td>0.65</td>
<td>2.8</td>
<td>55.0</td>
</tr>
<tr>
<td>Category prior</td>
<td>0.56</td>
<td>0.87</td>
<td>1.6</td>
<td>85.2</td>
</tr>
<tr>
<td>Ours</td>
<td>0.62</td>
<td>0.91</td>
<td>0.8</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Figure 6. Ablation Study: We show reconstruction in the image frame (top) and from a novel viewpoint (bottom) by our method along with ablations using other variants of diffusion models that only conditions on category or hand, and one that does not use prior.

clips without knowing the templates, the closest works are two template-free methods from Huang et al. [29] (HHOR) and Ye et al. [83] (iHOI).

HHOR is proposed for in-hand scanning. It optimizes a deformable semantic implicit field to jointly model hand and object. HHOR captures the dynamics by a per-frame warping field while no prior is used during optimization. iHOI is a feed-forward method and reconstructs 3D objects from single-view images by learning the hand prior between hand poses and object shapes. The method does not leverage category-level prior and do not consider time-consistency of shapes. We finetune their pretrained model to take in segmentation masks. We evaluate their result by aligning their predictions with ground truth for each frame and report the average number across all frames.

Results. We visualize the reconstructed HOI and object shapes from the image frame and a novel viewpoint in

Table 3. Ablation without surface normal, mask and depth in distillation: Quantitative results on HOI4D for object reconstruction error in the object-centric frame ($F_{@5}$, $F_{@10}$, $CD$) and for hand-object alignment in the hand frame ($CD_h$). We compare our method with other ablations that do not distill normals, masks, and depths respectively.

<table>
<thead>
<tr>
<th></th>
<th>$F_{@5}$</th>
<th>$F_{@10}$</th>
<th>$CD$</th>
<th>$CD_h$</th>
</tr>
</thead>
<tbody>
<tr>
<td>− normal</td>
<td>0.36</td>
<td>0.58</td>
<td>4.3</td>
<td>220.2</td>
</tr>
<tr>
<td>− mask</td>
<td>0.56</td>
<td>0.82</td>
<td>1.3</td>
<td>128.1</td>
</tr>
<tr>
<td>− depth</td>
<td>0.66</td>
<td>0.90</td>
<td>0.9</td>
<td>88.0</td>
</tr>
<tr>
<td>Ours</td>
<td>0.62</td>
<td>0.91</td>
<td>0.8</td>
<td>48.7</td>
</tr>
</tbody>
</table>

Figure 7. Ablation Study: We show reconstruction in the image frame (top) and from a novel view (bottom) by our method along with other variants that do not distill on depth, mask, and normals.

Fig. 5. HHOR generates good-looking results from the original view but actually degenerates to a flat surface since it does not incorporate any prior knowledge besides the visual observation. It also cannot decompose the hand and the object on the unobserved side of the scene because HHOR distinguishes them by per-point classification predicted from the neural field, which does not get gradient from the observations. iHOI reconstructs better object shapes and interactions but it is not very accurate as it cannot aggregate information across different frames. Its prediction is not time consistent either (better visualized as videos). In contrast, we are able to reconstruct time-persistent object shapes with time changing hand poses. The reconstructed object is more accurate, e.g. knife blade is thinner and the kettle body is more cylindrical.

This is consistent with quantitative results in Tab. 1. HHOR generally performs unfavorably except for knife category. While iHOI performs better, its quality is limited by only relying on information from a single frame. DiffHOI outperforms the baseline methods by large margins in most sequences and performs the best on all three metrics for mean values.
Table 4. Error analysis against hand pose noise: * marks our unablated method. Numbers in parentheses are per-frame prediction errors before optimization.

<table>
<thead>
<tr>
<th></th>
<th>Object Reconstruction</th>
<th>Hand Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$F@5$ ↑ $F@10$ ↑ $CD$ ↓</td>
<td>MPJPE ↓ AUC ↑</td>
</tr>
<tr>
<td>GT</td>
<td>0.68 0.91 0.75 - -</td>
<td>- -</td>
</tr>
<tr>
<td>Prediction*</td>
<td>0.62 0.91 0.77 26.9(28.4) 0.49(0.47)</td>
<td>- -</td>
</tr>
<tr>
<td>Pred. Error × 2</td>
<td>0.63 0.87 1.01 40.7(44.6) 0.31(0.27)</td>
<td>- -</td>
</tr>
</tbody>
</table>

4.3. Ablation Studies

We ablate our system carefully to analyze the contribution of each component. Besides the object reconstruction errors in the aligned object-centric frame, we further evaluate the hand-object arrangement by reporting the Chamfer distance of objects in hand frame, i.e. $CD_h \equiv CD(T_{h\rightarrow}\hat{O}, T_{h\rightarrow}\hat{O})$. We only report mean value in the main paper. Please refer to supplementary for category-wise results.

How does each learned prior help? We analyze how the category and hand priors affect reconstruction by training two more diffusion models conditioned only on text-prompt or hand renderings respectively. We also compare with the variant without optimizing $\mathcal{L}_{\text{SDE}}$ (no prior). As reported quantitatively, we find that category prior helps object reconstructions while hand prior helps hand-object relation (Tab. 2). And combining them both results in best performance.

We highlight an interesting qualitative result of reconstructing the bowl in Fig. 6. Neither prior can reconstruct the concave shape on its own – the hand pose alone is not predictive enough of the object shape while only knowing the object to be a bowl cannot make the SDS converge to a consensus direction that the bowl faces. Only knowing both can the concave shapes be recovered. This example further highlights the importance of both priors.

Which geometry modality matters more for distillation? Next, we investigate how much each geometry modality (mask, normal, depth) contributes when distilling them into 3D shapes. Given the same pretrained diffusion model, we disable one of the three input modalities in optimization by setting its weight on $\mathcal{L}_{\text{SDE}}$ to 0.

As visualized in Fig. 7, the surface normal is the most important modality. Interestingly, the model collapses if not distilling surface normals and even performs worse than the no-prior variant. Without distillation on masks, the object shape becomes less accurate probably because binary masks predict more discriminative signals on shapes. Relative depth does not help much with global object shape but it helps in aligning detailed local geometry ($F@5$) and aligning the object to hand ($F@10$).

How robust is the system to hand pose prediction errors? We report the object reconstruction performance when using GT vs predicted hand pose in Tab. 4, and find that our system is robust to some prediction error. Moreover, even if we artificially degrade the prediction by doubling the error, our performance remains better than the baselines (Tab. 1). We also report the hand pose estimation metrics and find that our optimization improves the initial predictions (in parentheses).

4.4. Comparing with Template-Based Methods

We compare with HOMAN [23], a representative template-based method that optimizes object 6D poses and hand articulations with respect to reprojection error and multiple interaction objectives including contact, intersection, distance, relative depth, temporal smoothness, etc.

We show quantitative and qualitative results in Tab. 5 and 8. Note that evaluating HOMAN in terms of object...
reconstruction is equivalent to evaluating templates since the objects are aligned in the object-centric frame. We first report the average object reconstruction errors when optimizing with different templates from training sets. While the gap indicates potential room to improve object shapes for template-free methods, DiffHOI is favorable over some templates in the training set. Nevertheless, when evaluating the objects in the hand frame, DiffHOI outperforms HOMAN by a large margin. The numbers along with visualizations in Fig. 8 indicate that template-based methods, even when optimizing with multiple objectives to encourage interactions, still struggle to place objects in the context of hands, especially for subtle parts like handles. Furthermore, optimizing with random templates degrades CDh significantly, highlighting the inherent drawbacks of template-based methods to demand the accurate templates.

4.5. Reconstructing In-the-Wild Video Clips

Lastly, we show that our method can be directly applied to more challenging video clips. In Fig. 9 top, we compare between our method and iHOI [83]. iHOI predicts reasonable shapes from the front view but fails on transparent objects like the plastic bottle since it is never trained on such appearance. In contrast, we transfer better to in-the-wild sequences as the learned prior only take on geometry cues. In Fig. 9 bottom, we visualize more results from our method. By incorporating learned priors, our method is robust to mask prediction inaccuracy, occlusion from irrelevant objects (the onion occludes knife blade), truncation of the HOI scene (bowl at the bottom left), etc. Our method can also work across ego-centric and third-person views since the learned prior is trained with uniformly sampled viewpoints. The reconstructed shapes vary from thin objects like knives to larger objects like kettles.

5. Conclusion

In this work, we propose a method to reconstruct hand-object interactions without any object templates from daily video clips. Our method is the first to tackle this challenging setting. We represent the HOI scene by a model-free implicit field for the object and a model-based mesh for the hand. The scene is optimized with respect to re-projection error and a data-driven geometry prior that captures the object shape given category information and hand poses. Both of these modules are shown as critical for successful reconstruction. Despite the encouraging results, there are several limitations: the current method can only handle small hand-object motions in short video clips up to a few (∼5) seconds; the reconstructed objects still miss details of shape. Despite the challenges, we believe that our work takes an encouraging step towards a holistic understanding of human-object interactions in everyday videos.

Acknowledgements. The authors would thank Di Huang for HHOR comparison. We thank Dandan Shan, Sudeep Dasari for helping with EPIC-KITCHENS datasets. We also thank Sudeep Dasari, Hanzhe Hu, Helen Jiang for detailed feedback on the manuscript. Yufei was partially supported by the NVIDIA fellowship.
References


[29] Di Huang, Xiaopeng Ji, Xingyi He, Jianming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-object interactions from video in the wild. In SIGGRAPH Asia, 2022.


[71] 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. 2


[77] 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 CVPR, 2023. 2


[80] 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 CVPR, 2022. 2