Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips

Anonymous ICCV submission

Paper ID 7185

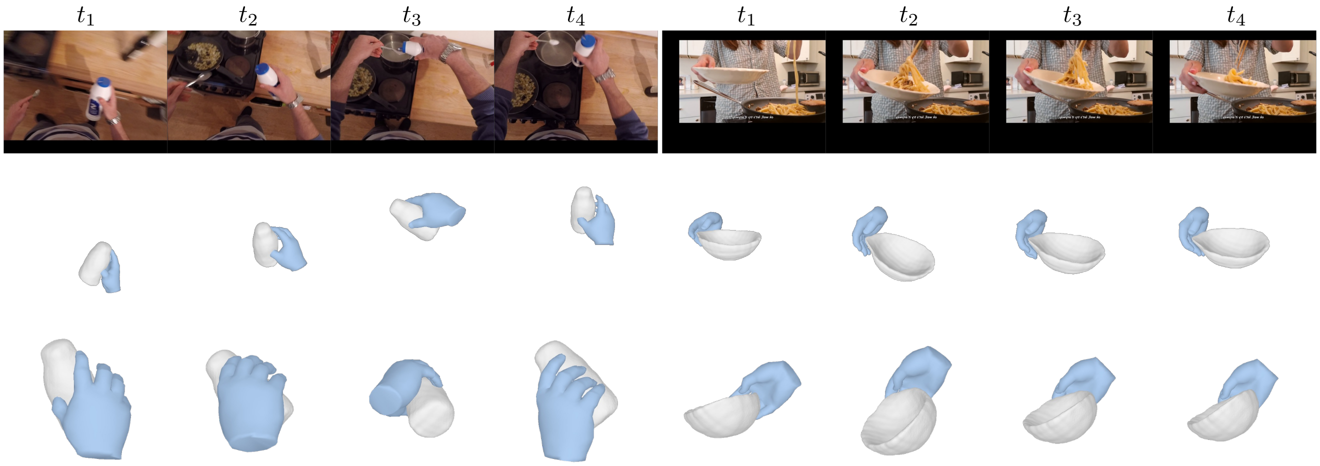

Figure 1: Given a video clip depicting a hand-object interaction, we infer the underlying 3D shape of both the hand and the object. **Top:** sampled input frames; **Middle:** reconstruction in the image frame; **Bottom:** reconstruction from a novel view. Please see the supplementary for reconstruction videos.

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 1^{st} and 3^{rd} 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 [14, 50, 26, 9, 2] framed it as 6-DoF pose task estimation for known 3D objects/templates, subsequent methods have tackled the reconstruction of apriori unknown objects [61, 21, 17]. Although single-view 3D reconstruction approaches can leverage data-driven techniques to reconstruct HOI images, their successes have been limited to simple objects [17, 64]. While incorporating pixel-aligned prediction architectures can further improve performance [61], 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 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, 24], 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 [28], 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

Hand-object interactions from Images Single-view reconstruction of individual generic objects has achieved great progress in recent years[39, 62, 22, 4, 10] by incorporating learned prior. Reconstructing hand-objects interactions is even more challenging due to heavy mutual occlusions. Most of the prior works make the simplifying assumption of knowing the instance-specific object template and then reduce this ill-posed problem to 6DoF pose estimation [3, 40, 11, 42, 48, 53, 63]. Recent works [21, 14, 61] explore a template-free approach to reconstruct more general objects by learning data-driven priors of interaction from large-scale datasets. 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.

Hand-object interactions from Videos There have been many efforts in capturing hand-object interactions with multiple cameras or monocular RGBD cameras. Known (scanned) templates of either rigid or articulated objects are fitted to multiple sequences and can achieve very accurate reconstructions to even serve as pseudo ground truth of datasets [46, 13, 50, 52, 1, 8]. Another line of works recover the 6D object pose from monocular RGB videos [15, 16, 36]. While all previous works assume the reconstructed object to be known, a few very recent works focus on template-free in-hand scanning from monocular videos [20, 12]. However, the scanning setup requires every region of the objects to be fully observed, which is often not true for everyday video clips. In contrast to all prior works, we tackle template-free 3D HOI reconstruction from everyday video clips.

Neural Implicit Fields for Dynamic Scene Neural radiance field and neural implicit fields [60, 30, 33, 31] have shown great potential in novel view synthesis and representing 3D static scenes. The following works incorporate dynamics scenes with different formulations including warping field [38, 34, 57, 23] that models dynamics in general scenes, to hierarchy deformable fields [59] for articulated objects. Other works incorporate the dynamics in the scene by explicitly capturing surfaces [51, 56, 58]. We focus on a specific class of dynamics with special interests – hand-object interactions and represent the object by a neural implicit field [60] which can be supervised with 2D images.

Diffusion models Diffusion models [18, 41] have made significant strides in text-image synthesis. Recently, view-conditioned diffusion models like DreamFusion[37], and Magic3D[24] have demonstrated the potential of diffusion models in optimizing 3D scenes using conditioned text prompts. On the other hand, approaches like NeRDi[7] and RealFusion[29] focus on 3D reconstruction from images. These methods rely heavily on RGB information to obtain novel views of the object. Different from previous methods, we leverage geometry-based information to reconstruct 3D models, which can be beneficial in terms of generalizing to novel scenes under distinct RGB appearances.

3. Method

Given a monocular video of a hand interacting with a rigid object, we aim to reconstruct the underlying hand-object interactions, *i.e.*, the 3D shape of the object, its pose in every frame, along with per-frame hand meshes and camera poses. Our key insight is to incorporate both view consistency across multiple frames and a data-driven prior of the geometry of hand-object interactions as the object of interest is often partially observed from everyday clips. The learned interaction prior captures both category cues, *e.g.* mugs are generally cylindrical, and hand cues, *e.g.* pinched

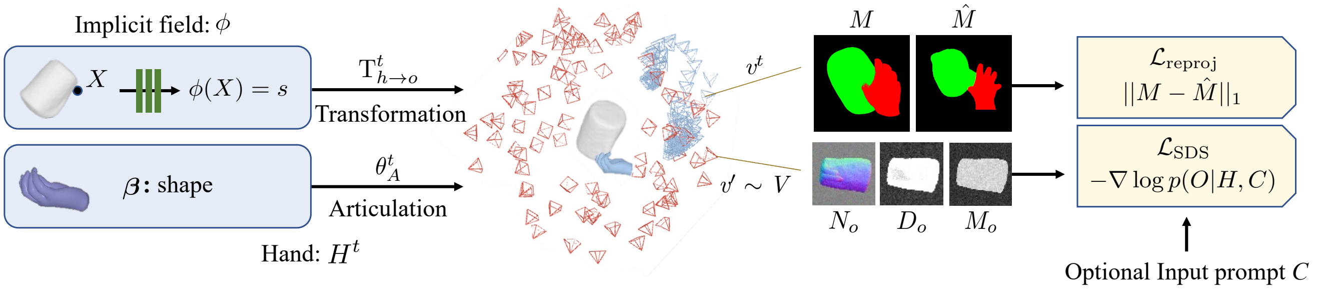

Figure 2: **Method Overview :** We model the HOI scene (middle) by a time-persistent implicit field ϕ for the object, hand meshes H^t parameterized by hand shape β, hand articulation θ^t_A, along a time-varying rigid transformation $T^t_{h\to o}$ for object pose. We register the cameras in the hand frame. We optimize a video-specific scene representation using reprojection loss from the original view and diffusion distillation loss from a novel view v'.

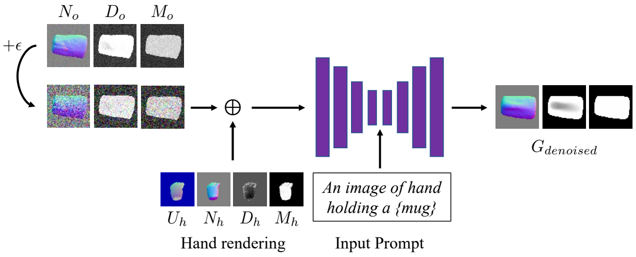

Figure 3: **Geometry-informed Diffusion Method:** Our diffusion model takes in a noisy geometry rendering of the object, the geometry rendering of the hand, and a text prompt, to output the denoised geometry rendering of objects.

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 $\hat{I}^t$ with corresponding hand and object masks $\hat{M}^t \equiv (\hat{M}^t_h, \hat{M}^t_o)$, we aim to optimize a HOI representation (Sec. 3.1) that consists of a time-persistent implicit field ϕ for the rigid object, a time-varying morphable mesh for the hand H^t, the relative transformation between hand and object $T^t_{h\to o}$, and time-varying camera poses $T^t_{c\to h}$. 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 $\pi(H)$ with geometry cues (Sec. 3.2). It denoises the rendering of the object $\pi(O)$ and backpropagates the gradient to the 3D HOI representation by score distillation sampling (SDS) [37].

3.1. HOI Scene Representation

Implicit field for object The rigid object is represented by a time-persistent implicit field ϕ that can handle unknown topology and has shown promising results when optimizing for challenging shapes [60, 54, 59]. For every point in the object frame, we use multi-layer perceptrons to predict signed distance function (SDF) to the object surface, $s = \phi(X)$.

Time-varying hand meshes We use a pre-defined parametric mesh model MANO [43] to represent hands dynamics across frames. The mesh can be animated by low-dimensional parameters and thus can handle more structured hand dynamics like articulation better. The MANO hand meshes are parameterized by 10-dim hand shapes β, 45-dim articulation parameter θ^t_A, and wrist orentation θ^t_w. We obtain hand meshes H^t in a canonical hand wrist frame by rigging MANO with only articulations and shape while the wrist orientation is used to register per-frame cameras, *i.e.* $H^t = \text{MANO}(\theta^t_A, \beta)$. Please see supplementary for details.

Composing to a scene Given the time-persistent object representation ϕ 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 [16, 36, 13] typically track 6D object pose directly in the camera frame $T_{c\to o}$ 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 $T^t_{h\to o}$ 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" [47]. A point in the rigid object frame can be related to the predicted camera frame by composing the two transformations, camera-to-hand $T^t_{c\to h}$ and hand-to-object $T^t_{h\to o}$. For notation convention, we denote the implicit field transformed to the hand frame at time t as $\phi^t(\cdot) \equiv \phi(T_{h\to o}(\cdot))$. As an implementation detail, we also optimize for per-frame camera intrinsics to account for zoom-in effect and inaccurate intrinsic estimation.

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

$$\phi, \beta, \theta^t_A, T^t_{h\to o}, T^t_{c\to h}, K^t \tag{1}$$

Differentialble Rendering To render the HOI scene into an image, we separately render the object (using volumetric rendering[60]) and the hand (using mesh rendering[27, 35]) to obtain geometry cues. We then blend their renderings into HOI images by their predicted rendered depth.

Given an arbitrary viewpoint v, both differentiable renders can render geometry images including mask, depth, and normal images, $G_h \equiv (M_h, D_h, N_h), G_o \equiv (M_o, D_o, N_o)$. 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 depth. Similar to blending two-layer surfaces of in mesh rendering, the final semantic masks can be computed as expected light transported to the cameras: $M = B(M_h, M_o, D_h, D_o)$. Please refer to supplementary material for the full derivation of the blending function B.

3.2. Data-Driven Prior for Geometry

When observing everyday interactions, we do not directly observe all aspects of the object because of occlusions and limited viewpoint variability. Despite this, we aim to reconstruct the 3D shape of the full object. To do so, we rely on data-driven prior that captures the likelihood of a common object geometry given its category and the hand interacting with it $p(\phi^t|H, C)$. More specifically, we use a diffusion model which learns a data-driven distribution over geometry rendering of objects given that of hands and category.

$$\log p(\phi^t|H, C) \approx \mathbb{E}_{v\sim V} \log p(\pi(\phi^t; v)|\pi(H; v), C) \tag{2}$$

where $v \sim V$ is a viewpoint drawn from a prior distribution, C as category label and π as rendering function. Since this learned prior only uses geometry cues, there is no domain gap to transfer the prior across daily videos with complicated appearances. We first pretrain this diffusion model with large-scale ground truth HOIs and then use the learned prior to guide per-sequence optimization (Sec. 3.3).

Learning a-modal HOI geometry Diffusion models are a class of probabilistic generative models that gradually transform a noise from a tractable distribution (Gaussian) to a complex (e.g. real image) data distribution. The diffusion model is supervised to capture the likelihood by de-noising corrupted images. During training, it takes in a corrupted image with a certain amount of noise σ_i along with conditions and learns to reconstruct the signal [19]

$$\mathcal{L}_{\text{DDPM}}[x; c] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),i}\|x - D\psi(x_i, \sigma_i, c)\|^2_2 \tag{3}$$

where x_i is a linear combination of signal x and noise ϵ while $D\psi$ is the denoiser.

In our case, the diffusion model denoises 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. Mor specifically, the training objective is $\mathcal{L}_{\text{diff}} = \mathcal{L}_{\text{DDPM}}[G_o; C, G_h, U_h]$. The text prompt comes from a text template: "an image of a hand holding $\{$*category*$\}$" and it is set to *null* when the category is unknown.

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 SO(3) . We use the backbone of a text-to-image model [32] with cross attention and modify it to diffuse 5-channel geometry images (3 for normal, 1 for mask and 1 for depth) and initialize the weights from the image-conditioned diffusion model [32] pretrained with large-scale text-image pairs.

3.3. Reconstructing Interaction Clips in 3D

Now, 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 interactions by differentiable rendering of geometry cues from the predicted views and from distilling the learned interaction prior.

Reprojection error First, the HOI representation is optimized to explain the input video. We render the semantic 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 distillation In the meantime, we optimize the scene to appear more likely from a novel viewpoint following Scored Distillation Sampling (SDS) [37]. 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 effi-

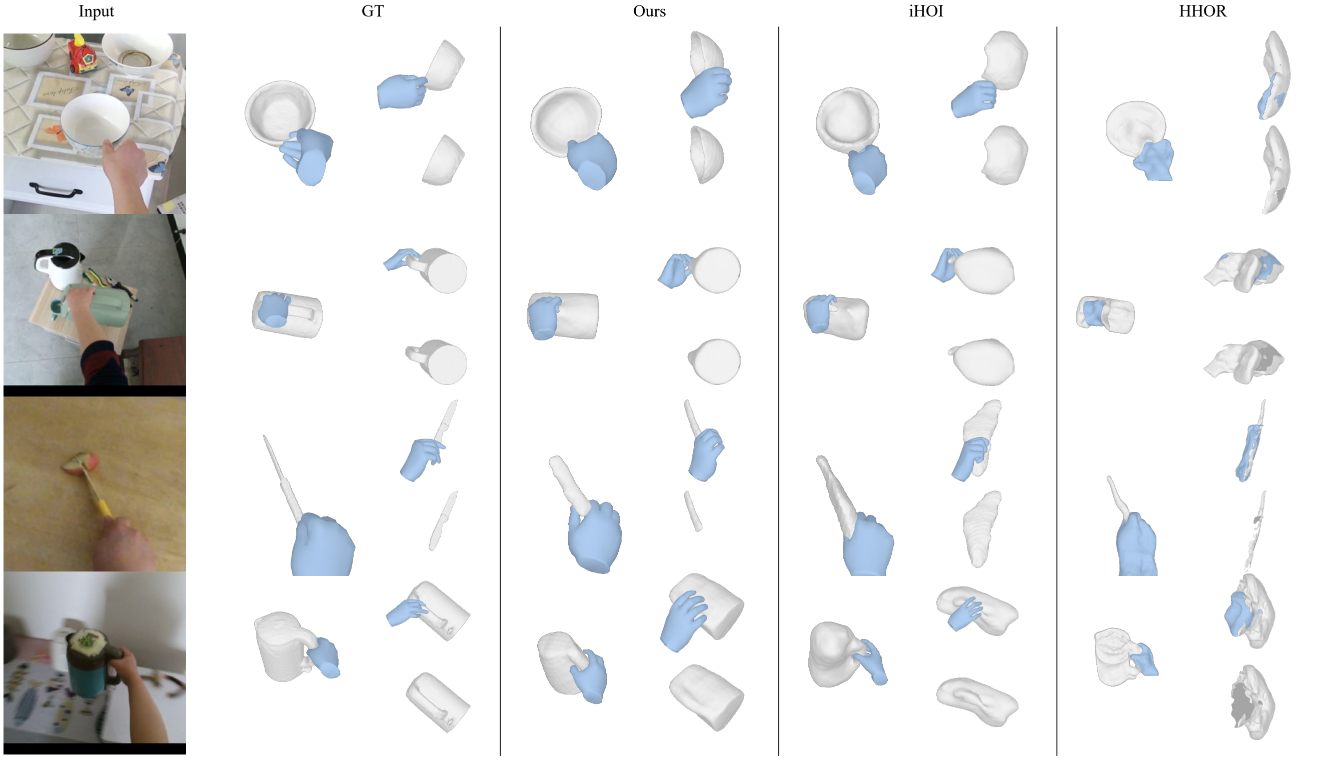

Figure 4: Comparing reconstruction visualizations of our method and 2 baselines [20, 61] on the HOI4D dataset. (Left: image frame, top right: novel view HOI, and bottom right: novel view object-only). Please see the supplementary for reconstruction videos.

ciency:

$$\mathcal{L}_{SDS} = \mathbb{E}_{v,\epsilon,i}[w_i \| \pi(\phi^t) - \hat{G}_o^i \|_2^2] \quad (4)$$

where $\hat{G}_o^i$ is the reconstructed signal from the pre-trained diffusion model. Please refer to relevant works [29, 37] or supplmentary for full details.

Other regularization We also include two regularization terms: one Eikonal loss that encourages the implicit field ϕ to be a valid distance function $\mathcal{L}_{\text{eik}} = \| |\nabla_X \phi|^2 - 1 \|^2$, and another temporal loss that encourages the hand to move smoothly with respect to the object $\mathcal{L}_{\text{smooth}} = \sum_t \| T^t_{h\to o} H^t - T^{t-1}_{h\to o} H^{t-1} \|_2^2$

Initialization and training details While the cameras pose and object pose are learned jointly with object shape, it is crucial to initialize them to a coarse position [25, 55]. We use off-the-shelf hand reconstruction system[45] to initialize the hand parameters, camera-to-hand transformations, and camera intrinsic. We initialize the object implicit field to a coarse sphere [60] and the object poses $T^t_{h\to o}$ 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.* $\mathcal{L} = \|H(x^t) - H(\hat{x}^t)\| + \lambda \|H(x^{t+1}) - H(x^t)\|$ where λ is 0.01. Please see supp for full details.

4. Experiment

We first train the diffusion model on the egocentric HOI4D [28] dataset. We evaluate the reconstruction of hand-object interactions quantitatively and qualitatively on the held-out sequences and compare our method with two model-free baselines. We then analyze the effects of both category-prior and hand-prior respectively and ablate the contribution from each geometry modality. Lastly, 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.

Table 1: **Comparison with baselines:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with prior works HHOR [20] and iHOI [61] on the HOI4D dataset.

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD
HHOR[20]	0.18	0.37	6.9	0.26	0.56	3.1	0.12	0.30	11.3	0.31	0.54	4.2	**0.71**	0.93	0.6	0.26	0.59	1.9	0.31	0.55	4.68
iHOI[61]	0.44	0.71	2.1	0.47	0.77	1.5	0.21	0.45	6.3	0.38	0.64	3.1	0.33	0.68	2.8	0.66	0.95	0.5	0.42	0.70	2.73
Ours	**0.67**	**0.86**	**1.0**	**0.62**	**0.92**	**0.7**	**0.47**	**0.73**	**1.6**	**0.68**	**0.93**	**0.6**	0.66	**0.96**	**0.6**	**0.81**	**0.98**	**0.3**	**0.65**	**0.90**	**0.79**

Table 2: **Analysis of category and HOI prior effects on object reconstruction:** Quantitative results for object reconstruction error using F1@5mm and F1@10mm scores and Chamfer Distance (mm). We compare our method with ablations that consider no prior, category prior only, and hand prior only on the HOI4D dataset.

	Mug			Bottle			Kettle			Bowl			Knife			ToyCar			Mean		
	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD	$F@5$	$F@10$	CD
no prior	0.39	0.66	2.2	0.39	0.63	2.3	0.15	0.35	11.6	0.41	0.70	2.1	0.70	0.93	0.5	0.66	0.92	0.6	0.45	0.70	3.22
hand prior	0.52	0.78	1.4	0.36	0.63	1.8	0.13	0.29	6.3	0.35	0.57	5.5	0.57	0.96	0.6	0.73	0.98	**0.4**	0.44	0.70	2.65
category prior	0.62	0.87	0.9	0.43	0.71	2.9	0.65	**0.89**	**1.0**	0.26	0.43	9.2	0.46	0.95	0.8	0.70	**0.98**	0.4	0.52	0.81	2.53
Ours	**0.66**	**0.89**	**0.8**	**0.84**	**0.99**	**0.3**	**0.71**	0.88	1.1	**0.50**	**0.86**	**0.9**	**0.74**	**0.98**	**0.4**	**0.75**	0.97	0.4	**0.70**	**0.93**	**0.66**

Table 3: **Analysis of category and HOI prior effects on hand-object relation:** Quantitative results for hand-object alignment using Chamfer distance (mm) in hand frame (CD_h). We compare our method with ablations that only incorporate the category prior, the hand prior, and no prior on the HOI4D dataset. The highlighted results are only compared among with-prior variants as the no-prior variant does not generate realistic shapes shown in Fig. 5 thus not comparable.

	Mug	Bottle	Kettle	Bowl	Knife	ToyCar	Mean
no prior	44.2	8.5	51.4	39.7	35.3	5.9	30.83
hand prior	27.3	15.0	52.0	**78.3**	**45.5**	51.5	44.94
category prior	26.1	31.8	53.6	129.2	86.4	43.8	61.82
Ours	**20.4**	**12.2**	**45.2**	92.8	53.8	**27.0**	**41.92**

Table 4: **Ablation without surface normal, mask and depth:** Quantitative results on the HOI4D dataset for object reconstruction error using mean F1 scores (5mm, 10mm), CD in object frame and for hand-object alignment using CD in hand frame (CD_h). We compare our method with other ablations that do not distill normals, masks, and depths respectively.

	$F@5$	$F@10$	CD	CD_h
− normal	0.37	0.57	4.5	282.6
− mask	0.57	0.84	1.2	106.7
− depth	0.66	0.93	0.7	49.6
Ours	**0.70**	**0.93**	**0.7**	**41.9**

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 [44] if required.

For in-the-wild dataset, we test on clips from EPIC-KITCHEN [6] videos and casual YouTube videos downloaded from the Internet. The segmentation masks are obtained using an off-the-shelf video object segmentation system [5].

Baselines While few prior works tackle our challenging setting – 3D HOI reconstruction from casual monocular clips without knowing the templates, the closest works are two template-free methods from Huang *et al.* [20] (HHOR) and Ye *et al.* [61] (iHOI).

HHOR is proposed for in-hand scanning. It optimizes a deformable semantic implicit field to jointly model hand and object. They capture 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. They do 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

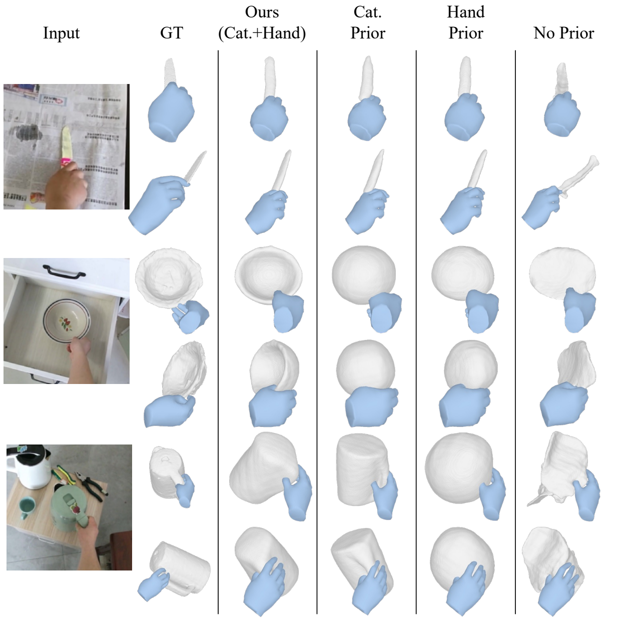

Figure 5: Visualizing HOI reconstruction comparisons of our method with other variants of diffusion models that only incorporate category prior, hand prior, and no prior. (Top: image frame, bottom: novel view)

with ground truth for each frames and report the average number across all frames.

4.1. Comparison on HOI4D

Evaluation Metric We evaluate the object reconstruction errors. Following prior works [20, 12], 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 [49] at $5mm$ and $10mm$ and reports 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 [49].

Results We visualize the reconstructed HOI and object shapes from the image frame and a novel viewpoint in Fig. 4. 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 because the opposite side of the scene is barely observed thus getting no gradient on the semantic field. The iHOI reconstructs more realistic object shapes and interactions but it is not very accurate due to inherent depth ambiguity as it cannot aggregate information across different frames. As shown in the supplementary video, its prediction is not time consistent too. In contrast, we are able to reconstruct time-consistent 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.

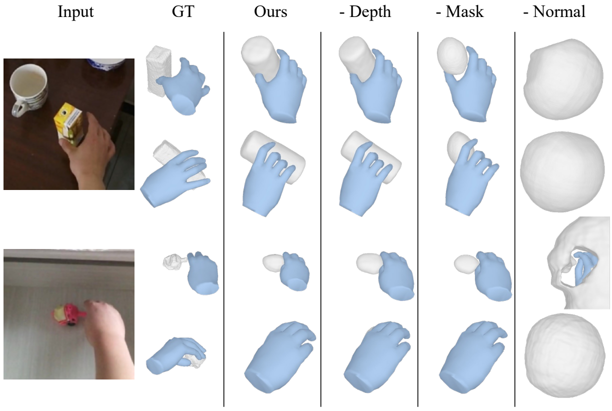

Figure 6: **Ablation Study:** Visualizing HOI reconstruction comparison of our method and variants that do not distill on depth, mask, and normals. (Top: image frame, bottom: novel view)

This is consistent with quantitative results in Tab. 1. HHOR generally performs unfavorably except for knife category. iHOI performs better. We outperform the baseline methods by large margins in most sequences and performs the best on all three metrics for mean values.

4.2. Ablation Studies

todo: *motivation, and one visual figure showing* CD_o, CD_h We ablate our system carefully to analyze the contribution of each component. We report the quantitative result on half of the test sequences. Besides the object reconstruction errors after 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*. $CD_h \equiv CD(T^t_{o\rightarrow h}O, \hat{T}^t_{o\rightarrow h}\hat{O})$.

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 $\mathcal{L}_{\text{SDS}}$ (no prior). As reported quantitatively, we find that *category prior helps object reconstructions (Tab. 2) while hand prior helps hand-object relation (Tab. 3)*. And combining them both results in best performance.

We highlight an interesting qualitative result of reconstructing the bowl in Fig. 5. 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

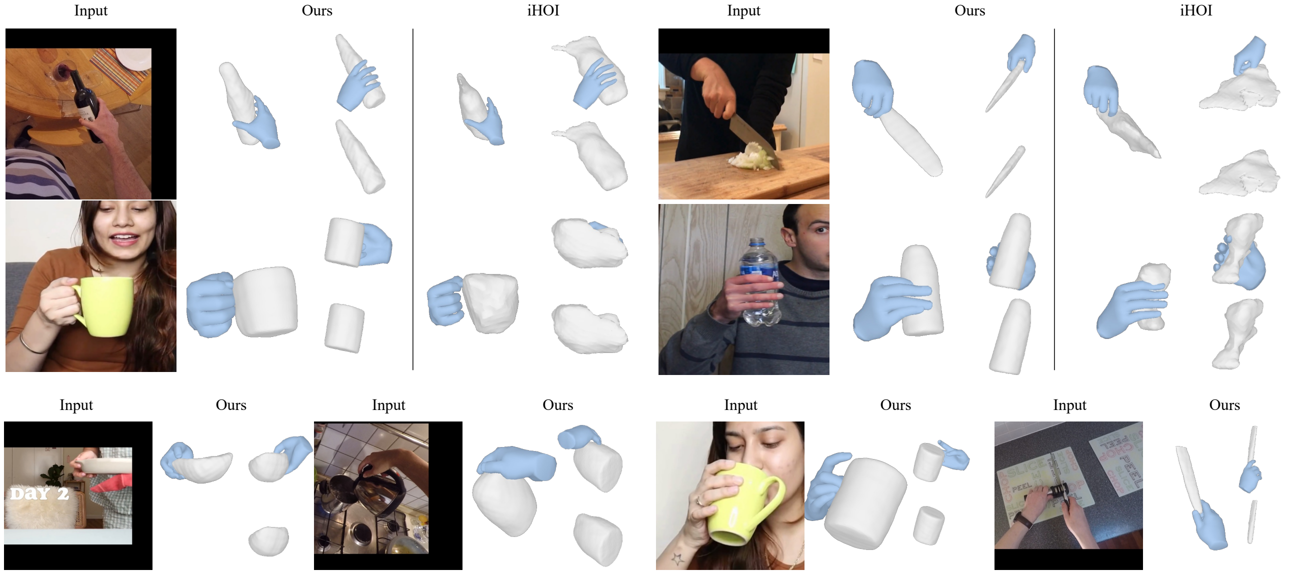

Figure 7: Comparing reconstructions of our method and the iHOI baseline [61] on 8 in-the-wild video clips taken from the Internet (Left: image frame, top right: novel view HOI, and bottom right: novel view object-only). Please see the supplementary for reconstruction videos.

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 hightlights the importance of both priors.

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

As visualized in Fig. 6, 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).

4.3. Reconstructing In-the-Wild Video Clips

Lastly, we show that our method together with the learned prior can be directly transferred to more challenging video clips.

In Fig. 7 top, we show comparisons between our method and iHOI [61]. iHOI predicts reasonable shapes from the front view but sometimes 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. 7 bottom, we visualize more results from our method. By incorporating learned priors, our method is robust to mask prediction inaccuracy, occlusion from irrelevant objects (knife blade is occluded by the onion), truncation of the HOI scene (bowl at the bottom left), *etc*. Our method can 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 ($\sim$5) seconds. Despite the challenges, we believe that our work takes an encouraging step towards a holistic understanding of human-object interactions in everyday videos.

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 *ECCV*, 2020.

[2] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021.

[3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021.

[4] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.

[5] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[7] Congyue Deng, Chiyu Max Jiang, C. Qi, Xinchen Yan, Yin Zhou, Leonidas J. Guibas, and Drago Anguelov. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *ArXiv*, abs/2212.03267, 2022.

[8] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *ArXiv*, abs/2204.13662, 2022.

[9] 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 *CVPR*, 2018.

[10] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021.

[11] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010.

[12] Shreyas Hampali, Tomás Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-hand 3d object scanning from an rgb sequence. *ArXiv*, abs/2211.16193, 2022.

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020.

[14] 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 *CVPR*, 2020.

[15] Yana Hasson, Gül Varol, Ivan Laptev, and Cordelia Schmid. Towards unconstrained joint hand-object reconstruction from rgb videos. In *ArXiv*, 2021.

[16] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021.

[17] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019.

[18] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[20] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Proceedings*, 2022.

[21] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020.

[22] Xueting Li, Sifei Liu, Kihwan Kim, Shalini De Mello, Varun Jampani, Ming-Hsuan Yang, and Jan Kautz. Self-supervised single-view 3d reconstruction via semantic consistency. In *ECCV*, 2020.

[23] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[24] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *arXiv preprint arXiv:2211.10440*, 2022.

[25] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

[26] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021.

[27] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019.

[28] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, June 2022.

[29] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360° reconstruction of any object from a single image. In *Arxiv*, 2023.

[30] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[33] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019.

[34] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[36] Austin Patel, Andrew Wang, Ilija Radosavovic, and Jitendra Malik. Learning to imitate object interactions from internet videos. *arXiv:2211.13225*, 2022.

[37] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020.

[39] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.

[40] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[41] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[42] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010.

[43] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017.

[44] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

[45] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCV Workshop*, 2021.

[46] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. *CVPR 2022*, 2022.

[47] Dandan Shan, Richard Higgins, and David Fouhey. Cohesiv: Contrastive object and hand embedding segmentation in video. *Advances in Neural Information Processing Systems*, 34:5898–5909, 2021.

[48] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016.

[49] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019.

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

[51] Shubham Tulsiani, Nilesh Kulkarni, and Abhinav Gupta. Implicit mesh reconstruction from unannotated image collections. *arXiv preprint arXiv:2007.08504*, 2020.

[52] 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.

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

[54] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[55] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[56] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. *arXiv preprint arXiv:2211.12497*, 2022.

[57] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021.

[58] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021.

[59] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022.

[60] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.

[61] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022.

[62] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021.

[63] 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.

[64] Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, P. Abbeel, and Aaron M. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. *ICAR*, 2015.