3D Reconstruction of Objects in Hands without Real World 3D Supervision

Aditya Prakash, Matthew Chang, Matthew Jin, Ruisen Tu, and Saurabh Gupta

University of Illinois Urbana-Champaign
{adityap9,mc48,mjin11,ruisent2,saurabhg}@illinois.edu
https://bit.ly/WildHOI

Abstract. Prior works for reconstructing hand-held objects from a single image train models on images paired with 3D shapes. Such data is challenging to gather in the real world at scale. Consequently, these approaches do not generalize well when presented with novel objects in in-the-wild settings. While 3D supervision is a major bottleneck, there is an abundance of a) in-the-wild raw video data showing hand-object interactions and b) synthetic 3D shape collections. In this paper, we propose modules to leverage 3D supervision from these sources to scale up the learning of models for reconstructing hand-held objects. Specifically, we extract multiview 2D mask supervision from videos and 3D shape priors from shape collections. We use these indirect 3D cues to train occupancy networks that predict the 3D shape of objects from a single RGB image. Our experiments in the challenging object generalization setting on in-the-wild MOW dataset show 11.6% relative improvement over models trained with 3D supervision on existing datasets.

Keywords: hand-held objects · shape priors · multiview supervision

1 Introduction

While 3D reconstruction of hand-held objects is important for AR/VR [4,20] and robot learning applications [39,40,47,48,68,71], *lack of 3D supervision outside of lab settings* has made it challenging to produce models that work in the wild. This paper develops techniques to improve the generalization capabilities of *single image* hand-held object reconstruction methods by extracting supervision from in-the-wild videos & synthetic shape collections showing hand-object interactions.

Collecting image datasets with ground truth 3D shapes for hand-held objects is hard. Any visual scanning setups (via multiple RGB/RGB-D cameras or motion capture) require full visibility of the object which is not available. Synthesizing realistic hand-object interaction is an open problem in itself [28,31,49,65]. Manual alignment of template shapes [5] is expensive, yet only approximate. Thus, there is very little in-the-wild real-world data with ground truth 3D shapes for hand-held objects. And while many past works have designed expressive models to predict shapes of hand-held objects [22,31,73], they are all held back due to the limited amount of real-world 3D data available for training and suffer from unsatisfactory performance on novel objects encountered in the wild.

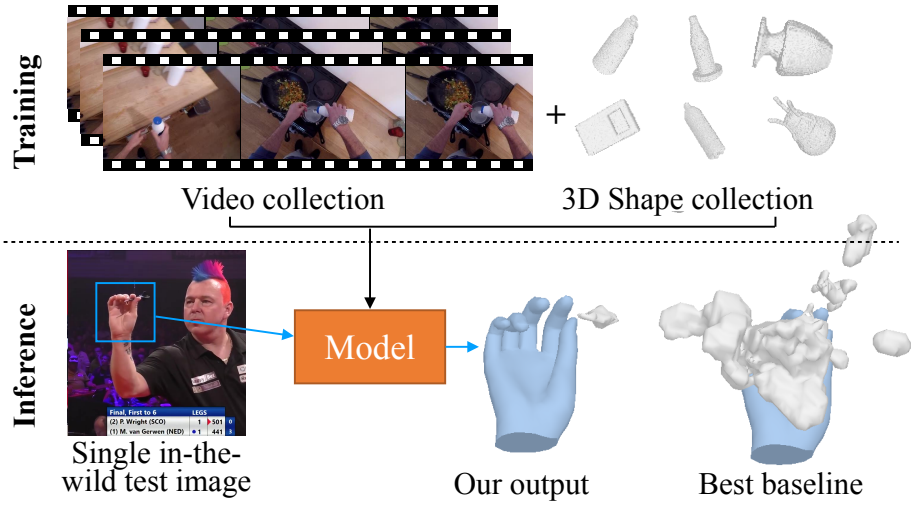

Fig. 1: We propose modules to extract supervision from in-the-wild videos (Sec. 3.2) & learn shape priors from 3D object collections (Sec. 3.3), to train occupancy networks which predict the 3D shapes of hand-held objects from a single image. This circumvents the need for paired real world 3D shape supervision used in existing works [22, 73].

While in-the-wild images with paired 3D shapes are rare, there are a) plenty of in-the-wild videos containing multiple views of hand-held objects [12, 17] (Fig. 1), b) large catalogues of 3D object shapes [6] (Fig. 1). Shape collections provide 3D supervision but lack realistic hand grasps, videos showcase realistic hand-object interaction but don't provide direct 3D supervision. Either by itself seems insufficient, but can we combine supervision from these diverse sources to improve generalization of single-image hand-held object reconstruction methods?

Let's consider each cue one at a time. While videos show multiple views of the object, we unfortunately don't know the relative object pose in the different views. Automatically extracting the object pose using structure from motion techniques, *e.g.* COLMAP [56] doesn't work due to insufficient number of feature matches on the object of interaction. We sidestep this problem by using *hand pose as a proxy for object pose* (Fig. 2). This is based on the observation that humans rarely conduct in-hand manipulation in pick & place tasks involving rigid objects. Thus, if we assume that the hand and the object are rigidly moving together, then the relative 6 DoF pose of the hand between pairs of frames reveals the relative 6 DoF pose of the object. This reduces the SfM problem to an easier setting where the motion is known. Specifically, we use off-the-shelf FrankMocap system [54] to obtain 6 DoF pose for the hand and consequently the object's. We then use our proposed 2D mask guided 3D sampling module (Sec. 3.2) to generate 3D supervision for the object shape using object segmentation masks (Fig. 2). This lets us train on objects from 144 different categories, where as most methods currently train on only a handful of categories (< 20).

While this works well for unoccluded parts of the object, this does not generate reliable supervision for parts of the object that are occluded by the hand (Fig. 1). This brings us to the 3D shape catalogues, which we use to extract shape priors. This enables the model to learn to output contiguous shapes even when the object is interrupted by the hand in the image, *e.g.* it can hallucinate a handle for a jug even when it is covered by the hand, because jugs typically have one. We adopt an adversarial training framework [16] to train a discriminator to differentiate between real shapes (from ObMan [22]) and shapes predicted from the model (Fig. 3). Unlike prior works [67] which train the discriminator on 3D inputs, we instead propose a 2D slice-based 3D discriminator (Sec. 3.3), which is computationally efficient and learns better fine-grained shape information.

Our overall framework consists of an occupancy network [43] that predicts the 3D shape of hand-held objects from a single image. We train this model on sequences curated from the VISOR dataset [13] and use the Obman dataset [22] to build the shape prior. Training on diverse real world data outside of lab settings, enabled by our innovations, leads our model (HORSE) to good generalization performance. HORSE outperforms previous state-of-the-art models by 11.6% in the challenging object generalization setting on MOW [5].

2 Related Work

Reconstructing objects in hands: Several works [9,10,22,31,73,77] have trained expressive architectures for predicting 3D shape from a single image using paired real world 3D supervision. Fitting object templates [5,21] or learned 3D shapes [14,25,72,74] to videos using appearance cues [5,14,21,25] or geometric priors [72,74] have also been explored. The most relevant work to ours is [73], which uses paired 3D supervision from synthetic [22] and small-scale real-world datasets to predict 3D shape from a single image. However, it does not generalize to novel object categories in the wild due to limited 3D supervision. Instead, we train our model on diverse object categories from in-the-wild videos by extracting multiview 2D supervision and learning shape priors from existing datasets, without any real-world 3D supervision. Note that our setting involves a single image input at test time and we use in-the-wild videos for training only.

Hand-Object datasets with 3D object models: Existing real-world hand-object datasets with 3D annotations are captured in lab settings and contain limited variation in objects, *e.g.* HO3D [18]:10, H2O [32]:8, FPHA [15]:4, FreiHAND [81]:35, ContactDB [2]:50, ContactPose [3]:25, DexYCB [8]:20, GRAB [58]: 51, HOI4D [34]: 16 object categories. Collecting datasets with ground truth 3D shapes is difficult to scale since it often requires visual scanning setups (multiple cameras or motion capture). Synthesising realistic hand-object interaction is an open problem in itself [28,31,49,65]. In this work, we curate sequences from in-the-wild VISOR dataset containing 144 object categories and design modules to extract supervision for training occupancy networks. The closest to ours is MOW with 120 objects that we only use to test models to assess generalization.

Hand-Object Interactions in the wild: There is a growing interest in understanding hands and how they interact with objects around them. Researchers have collected datasets [8,18,19,22,32,34,58] and trained models for detecting & segmenting hands and associated objects of interaction [13,57,62,63]. Recognizing what hands are doing in images [7,46,79] is also relevant: through grasp classification [31], 2D pose estimation [51,80], and more recently 3D shape and pose estimation [21,22,53,54,61,73] for both hands and objects in contact.

3D from single image without direct 3D supervision. Several works relax the need for direct 3D supervision by incorporating auxiliary shape cues during training, *e.g.* multi-view consistency in masks [64], depth from single image [26,37,78] or stereo [24], appearance [11,27,60,76]. These have been applied to reconstruction of category specific [27,29,30,37] as well as generic objects [11,75,76]. However, directly applying these approaches to hand-held objects in the wild poses several challenges, *e.g.* unknown camera, novel object categories, heavy occlusion, inaccurate depth estimates. In this work, we propose modules to extract supervision from in-the-wild videos using object masks [13] & hand pose [54] and learn priors from synthetic collections of hand-held objects [22].

3 Approach

We propose a novel framework for training 3D shape predictors from a single image without using any real world 3D supervision. Following prior work [73], we use implicit shape representation [43,45] for 3D objects.

3.1 Preliminaries

Consider the recent AC-SDF model for this task from Ye *et al.* [73]. Given an input RGB image, AC-SDF uses a neural network to predict the SDF of 3D points. The prediction is done in the hand coordinate frame obtained using FrankMocap [54], which outputs (a) hand articulation parameters θ^a (45 dimensional MANO hand pose [52]), (b) global rotation θ^w of the wrist joint w.r.t. camera, (c) weak perspective camera θ^c, with scale factor s & 2D translation (t_x, t_y), which is converted into a full perspective camera K. These can be used to project a 3D point $\mathbf{x}$ into the image (f is the focal length) as $\mathbf{x}_p = K[T_{\theta^w}\mathbf{x} + (t_x, t_y, f/s)]$

Given a 3D point $\mathbf{x}$ & image I, AC-SDF conditions the SDF prediction on: (a) global image features from a ResNet-50 [23], (b) pixel-aligned features [55] from intermediate layers of ResNet-50 at the projection $\mathbf{x}_p$ of $\mathbf{x}$ in the image, (c) hand articulation features obtained by representing $\mathbf{x}$ in the coordinate frame of 15 hand joints. This is realized as, $\mathbf{s} = \mathcal{F}(\mathbf{x}; I, \theta, K)$. Training $\mathcal{F}$ requires sampling 3D points x around the object and corresponding SDF values s, $\theta = (\theta^a, \theta^w, \theta^c, K)$ are estimated from FrankMocap.

3.2 2D Mask Guided 3D Sampling

Training models with implicit shape representation require supervision in the form of occupancy [43] or SDF [45] for 3D points sampled inside and outside

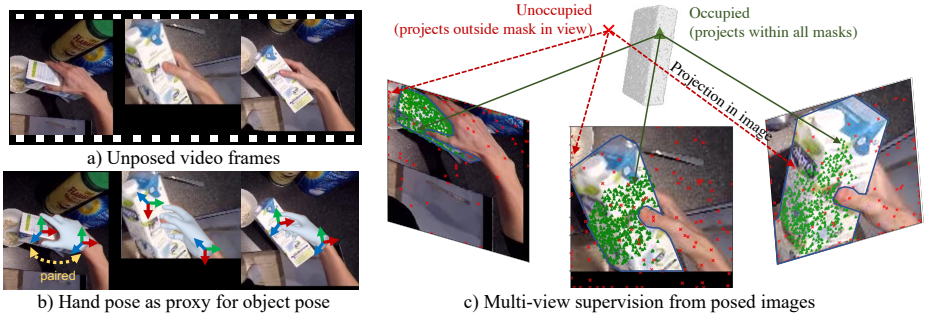

Fig. 2: Registering objects via hand pose and 2D Mask guided 3D sampling. (a) Consider unposed frames from in-the-wild videos. (b) We use hand pose from FrankMocap [54] as a proxy for object pose, thereby registering the different views. (c) We then use 2D object masks for labeling 3D points with occupancy (Sec. 3.2). 3D points that project into the object mask in all views are considered as occupied (green triangles), all other points are considered unoccupied (red crosses). (3D object in the figure is for visualization only, not used for sampling.)

the object. Note that the balanced sampling of points inside and outside the object is an important consideration for training good predictors. While existing approaches [22,31,73] on this task use datasets with paired 3D supervision (3D object shape corresponding to 2D image), we operate in in-the-wild settings which do not contain 3D supervision. Instead, we propose a 2D mask guided 3D sampling strategy to obtain occupancy labels for training.

Consider multiple views $\{I_1, \ldots, I_n\}$ of a hand-held object (Fig. 2), along with their masks $\{M_1, \ldots, M_n\}$. We can sample points $\mathbf{x}$ in 3D space and project them into different views. Any point x which projects into the object mask in all views is consider as occupied whereas if it projects outside the mask in even one of the views, it is considered as unoccupied. Thus, we get occupancy labels for a point $\mathbf{x}$ as $\mathbf{s}^{gt} = \cap_{i=1}^{n} M_i^{\mathbf{x}_{p_i}}$. Here, $M_i^{\mathbf{x}_{p_i}}$=1 if x_{p_i} lies inside the mask M_i & 0 otherwise. Note that it is not possible to obtain SDF values in this manner, since distance to the object surface cannot be estimated in the absence of 3D objects models. While we can obtain 3D occupancy labels using this strategy, there are two important considerations: camera poses are unknown (required for projection) & how to balance the sampling of points inside & outside the object.

Camera pose: We assume that the hand is rigidly moving with the object. This is not an unreasonable assumption, as humans rarely do in-hand manipulation in pick & place tasks involving small rigid objects. Thus, the relative pose of hand between different views reveals the relative pose of the object. This lets use the hand pose predicted by FrankMocap $\{\theta_1, \ldots, \theta_n\}$ to register the different views.

Balanced sampling: In the absence of 3D object models, a natural choice is to sample points uniformly in 3D space. However, this leads to most points lying outside the object because the object location is unknown. Instead, we sample points in the hand coordinate frame. Consider the total number of points to be q.

We adopt several strategies for balanced sampling for points inside ($s^{gt} = 1$) and outside the object ($s^{gt} = 0$). We uniformly sample $q/2$ 3D points $\mathbf{x} \in \mathbb{R}^3$ in the normalized hand coordinate frame and project these into all the available views. Since all these $q/2$ points may not be occupied, we use rejection sampling to repeat the procedure, for maximum of $t = 50$ times or until we get $q/2$ occupied points. Also, all points projecting into the hand mask in all views and vertices of the MANO [53] hand are labeled as unoccupied.

Formally, for images $\{I_1, \ldots, I_n\}$ with object masks $\{M_1, \ldots, M_n\}$, hand masks $\{H_1, \ldots, H_n\}$ and MANO vertices $\{V_1, \ldots, V_n\}$, $\mathbf{s}^{gt}$ for $\mathbf{x}$ is:

$$\mathbf{s}^{gt} = \begin{cases} 1 & \text{if } \cap_{i=1}^{n} M_i^{\mathbf{x}_{p_i}} \text{ and } \cap_{i=1}^{n} \neg H_i^{\mathbf{x}_{p_i}} \text{ and } \cup_{i=1}^{n} \neg V_i^{\mathbf{x}} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{x}_{p_i}$ is the projection of $\mathbf{x}$, $M_i^{\mathbf{x}_{p_i}}$=1 if x_{p_i} lies inside M_i, $H_i^{\mathbf{x}_{p_i}}$=1 if x_{p_i} lies inside H_i, $V_i^{\mathbf{x}}$= 1 if $\mathbf{x}$ belongs to V_i and $\neg$ is the logical negation operator.

Note that, due to hand occlusions and errors in FrankMocap predictions, it is possible that some 3D points belonging to the object are not projected into the object masks but we do not want to label these points as unoccupied. So we disregard points which project onto the object mask in some views and hand mask in other views as these points could belong to object due to hand occlusion.

This is reminiscent of the visual hull algorithm [33,42], which generates 3D reconstruction by carving out space that projects outside the segmentation in any view. Visual hull algorithms need multiple views at test time to generate any output. In contrast, we are doing this at training time to obtain supervision for $\mathcal{F}(\mathbf{x}; I_1, \theta_1, K_1)$, which makes predictions from a single view.

Training: We use cross-entropy loss (CE) to train $\mathcal{F}$ using ground truth $\mathbf{s}^{gt}$:

$$\mathcal{L}_{\text{visual-hull}} = \text{CE}(\mathcal{F}(\mathbf{x}), \mathbf{s}^{gt}) \tag{2}$$

To further regularize training, we also encourage the occupancy prediction from different views to be consistent with each other. Since our predictions are already in the hand coordinate frame, which is common across all views, this can be done by minimizing $\mathcal{L}_{\text{consistency}}$ for different views i & j of the same object.

$$\mathcal{L}_{\text{consistency}} = \sum_{\mathbf{x} \in \mathbb{R}^3, i \neq j} \text{CE}\left(\mathcal{F}(\mathbf{x}; I_i, \theta_i, K_i), \mathcal{F}(\mathbf{x}; I_j, \theta_j, K_j)\right) \tag{3}$$

3.3 2D Slice based 3D Discriminator as Shape Prior

We adopt an adversarial training framework [16] to build a prior on shapes of hand-held objects and use it to supervise the training of the occupancy prediction function $\mathcal{F}(\mathbf{x}; I_1, \theta_1^a, \theta_1^w, K_1)$. As such a prior can be challenging to hand-craft, we build it in a data-driven way. We use 3D shape repository from synthetic datasets [22], which contain more than 2.5K hand-held objects, to learn the prior. Specifically, we train a discriminator $\mathcal{D}$ to differentiate between 3D shapes from

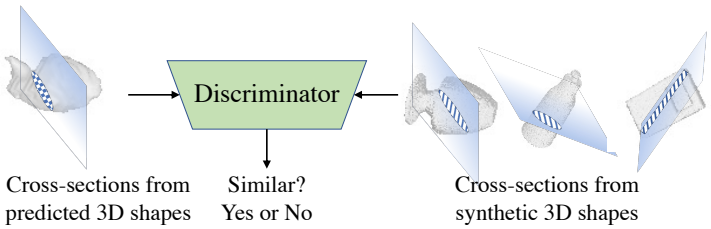

Fig. 3: 2D slice based 3D discriminator. We learn data-driven 3D shape priors using hand-held objects from ObMan dataset. We sample planes through the object (shown above in blue), resulting in a 2D cross-section map. We pass occupancy predictions on points from these cross-sections through a discriminator which tries to distinguish cross-sections of predicted 3D shapes from cross-sections of ObMan objects (Sec. 3.3).

ObMan [22] and generated shapes as predicted by $\mathcal{F}$. We derive supervision for $\mathcal{F}$ by encouraging it to predict shapes that are real as per $\mathcal{D}$.

A natural choice is to train the discriminator with 3D input, *e.g.* $N \times N \times N$ cube in 3D voxel space [67]. One way to do this is to sample N^3 3D points in the hand coordinate frame and run a forward pass through $\mathcal{F}$ to get the occupancy for each of these points. However this is computationally expensive and often leads to large imbalance as most points lie outside the object (we ablate this in Sec. 4.3). Instead, we propose a novel 2D slice based 3D discriminator which operates on arbitrary 2D slices. There are computed by taking the cross-section of 2D planes with 3D shapes and sampling 3D points that lie on these 2D cross-sections. The key intuition here is that the discriminator sees *different randomly sampled* 2D slides during the course of training, which helps it to learn fine-grained shape information. *E.g.* for a sphere, all cross-sections are circular but for a cylinder, most are oval. This helps distinguish between different 3D shapes.

Sampling 2D slices: There are several important considerations in sampling 2D slices. First, uniformly sampling 2D planes often leads to most points lying outside the object, which is not useful for training the discriminator. Instead, we sample 2D planes that pass through the origin in the hand coordinate system. Since the objects are in contact with the hand, the sampled points are more likely to encompass the object. Then, we rotate the sampled 2D planes by arbitrary angles so that they are not axis aligned to better capture fine-grained shape information. We ablate all these design choices in Sec. 4.3. This sampling function $\mathcal{Z}$ results in a set of 2D planes on which 3D points are uniformly sampled.

Training: We pass the sampled points from 2D slices of the generated 3D shape through $\mathcal{F}$ to get the corresponding occupancy values S^{gen}. This represents the generated 3D shape. We adopt the same strategy for representing 3D shapes from ObMan (used as real shapes) but use the predictions S^{real} of the occupancy network overfitted on ObMan. As they come from a overfitted model, they generally match the ground truth slices well but at the same time are soft and prevent the discriminator from cheating.

Fig. 4: VISOR visualizations. Using existing hand pose estimation techniques [54], we are able to track the objects in relation to hands through time in in-the-wild videos. We visualize these tracks along with object masks from the VISOR dataset [13]. This form of data, where objects move rigidly relative to hands, is used to train our model to learn 3D shape of hand-held objects.

We train the discriminator $\mathcal{D}$ to differentiate between S^{gen} & S^{real} using the least squares formulation [41] for discriminator loss. We derive supervision for $\mathcal{F}$ by computing gradients through $\mathcal{D}$ on the occupancy values at the sampled points to maximize the realism of the generated shapes.

$$\begin{aligned} \mathcal{L}_{\text{adv}}^{\mathcal{D}} &= [\mathcal{D}(S^{\text{real}}) - 1]^2 + [\mathcal{D}(S^{\text{gen}})]^2 \\ \mathcal{L}_{\text{adv}}^{\mathcal{F}} &= [\mathcal{D}(S^{\text{gen}}) - 1]^2 \\ \mathcal{L}_{\text{shape-prior}} &= \lambda_f \mathcal{L}_{\text{adv}}(\mathcal{F}) + \lambda_d \mathcal{L}_{\text{adv}}(\mathcal{D}) \end{aligned} \tag{4}$$

3.4 Training Details

We train $\mathcal{F}$ & $\mathcal{D}$ in an alternating manner with 2 iterations of F for every iteration of D. The total loss for training our framework is:

$$\begin{aligned} \mathcal{L}_{\mathcal{F}} &= \lambda_v \mathcal{L}_{\text{visual-hull}} + \lambda_c \mathcal{L}_{\text{consistency}} + \lambda_f \mathcal{L}_{\text{adv}}^{\mathcal{F}} \\ \mathcal{L}_{\mathcal{D}} &= \lambda_d \mathcal{L}_{\text{adv}}^{\mathcal{D}} \end{aligned} \tag{5}$$

Following standard practice [73], we pretrain on synthetic ObMan. We train our model jointly on ObMan (3D supervision, shape priors) & VISOR (2D supervision) with a dataset ratio of ObMan:VISOR as 1:2. We use batch size of 64, learning rate of 1e-5 across 4 NVIDIA A40 GPUs & loss weights as $\lambda_v = 1, \lambda_c = 1$, $\lambda_f = 0.25$, $\lambda_d = 0.25$. Please refer to supplementary for more details.

3.5 Constructing Wild Objects in Hands Dataset

Our framework requires dataset containing multi-view images of rigid hand-object interactions in the wild, with 3D hand pose and 2D object masks. To construct such a dataset, we consider VISOR [13] which provides 2D tracks for hands, objects they are interacting with and their segmentation masks. It contains a rich set of hand-object interactions, *e.g.* taking out milk from the fridge, pouring oil from bottles, kneading dough, cutting vegetables, and stirring noodles in a wok. Our interest is in the 3D reconstruction of *rigid* objects which are *in-contact* with a hand, but there are no 3D object annotations in VISOR. Hence, we process it to prepare a dataset for training our model.

Table 1: Generalization to novel objects in the wild. We report F-score at 5mm & 10 mm, Chamfer distance (CD, mm) for object generalization splits on MOW. We compare with AC-OCC & AC-SDF trained on different combinations of datasets with full 3D supervision. Our approach outperforms baselines across all metrics without using real-world 3D supervision (Relative % improvement w.r.t. best baseline in green).

Method	Dataset and supervision used	F@5 ↑	F@10 ↑	CD ↓
AC-OCC	ObMan (Synthetic 3D)	0.095	0.179	8.69
AC-SDF [73]	ObMan (Synthetic 3D)	0.108	0.199	7.82
AC-SDF [73]	ObMan (Synthetic 3D) + HO3D (Lab 3D)	0.082	0.159	7.52
AC-SDF [73]	ObMan (Synthetic 3D) + HO3D (Lab 3D) + HOI4D (3D)	0.095	0.193	7.43
HORSE (Ours)	ObMan (Synthetic 3D) + VISOR (2D Masks) + Shape priors	**0.121** +10.7%	**0.220** +10.6%	**6.76** +13.5%

We first sample a subset of VISOR involving hand-object contact, using available contact annotations. We select object tracks where only one hand is in consistent contact with the object. This leaves us with 14768 object tracks from the original VISOR dataset. We then manually filter this subset to select a subset that showcases manipulation of rigid objects with a single hand. This leaves us with 604 video snippets showing hands interacting with different objects.

Processing hands on VISOR: We rely on the 3D hand poses to set up the output coordinate frame, compute hand articulation features, and more importantly to register the different frames together [38, 66]. These hand poses are estimated using FrankMocap, which may not always be accurate. To remove erroneous poses, we employ automated filtering using the uncertainty estimate technique from Bahat & Shakhnarovich [1] following 3D human pose literature [50]. Specifically, we obtain 3D hand pose predictions on five different versions of the image, augmented by different fixed translations . The uncertainty estimate for a given image is computed as the standard deviation of reprojection locations of MANO vertices across these 5 image versions. This sidesteps the need to hand-specify the trade-off between translation, rotation, and articulation parameters that are part of the 3D hand pose output. This leaves us with 473 video snippets consisting of 144 object categories. This object diversity is 4× larger than existing datasets [18, 19, 32, 34, 69] used for our task, typically containing 10 to 32 object categories. We refer to this dataset as Wild Objects in Hands, some example object sequences are shown in Fig. 4. Note the *incidental* multiple views and relative consistency in hand and object pose over the course of interaction.

4 Experiments

4.1 Protocols

We use 4 datasets for training (ObMan [22], VISOR [13], HO3D [18], HOI4D [34]) and 2 datasets (MOW [5], HO3D) for evaluation. Different methods are trained on different datasets, depending on the specific evaluation setting.

Training datasets: ObMan is a large scale synthetic hand-object dataset with 2.5K objects and 3D supervision. HO3D & HOI4D are real world datasets collected

Table 2: HO3D Object generalization. We outperform AC-OCC & AC-SDF trained on different datasets with 3D supervision.

Method	Supervision (ObMan +)	F@5	F@10	CD
AC-OCC	-	0.18	0.33	4.39
AC-SDF	-	0.17	0.33	3.72
AC-SDF	MOW (3D)	0.17	0.33	3.84
AC-SDF	MOW (3D) + HOI4D (3D)	0.17	0.33	3.63
Ours	VISOR (Multi-view 2D)	**0.20**	**0.35**	**3.39**

Table 3: HO3D View generalization. We outperform HO [22] & GF [31], trained on HO3D with full 3D supervision.

Method	Supervision (ObMan +)	F@5	F@10	CD
AC-SDF	-	0.17	0.32	3.72
HO [22]	HO3D (3D)	0.11	0.22	4.19
GF [31]	HO3D (3D)	0.12	0.24	4.96
Ours	HO3D (Multi-view 2D)	**0.23**	**0.43**	**1.41**

in lab settings with 3D annotations. HO3D contains 10 YCB [82] objects whereas HOI4D contains 16 object categories, out of which 7 are rigid. VISOR does not contain any 3D supervision. Instead, we use the process described in Sec. 3.5, to extract supervision from VISOR, resulting in 144 object categories.

The baselines are trained with different combinations of HO3D & HOI4D [34]. As our method does not require 3D ground truth, we do not use these datasets for training. Instead, we use auxiliary supervision from Wild Objects in Hands (Sec. 3.5) & learn shape priors using ObMan. VISOR does not have 3D annotations and can not be used to train the baselines. Note that all models are initialized from the model pretrained on ObMan for fair comparisons, following protocol [73].

Evaluation datasets: We focus on the challenging zero-shot generalization to novel objects in-the-wild setting. We use MOW [5] dataset which contains images from YouTube, spanning 120 object templates. Note that these types of images have not been seen during training. To be consistent with prior work [73], we also use HO3D for evaluation, consisting of 1221 testing images across 10 objects. While [73] operate in view generalization setting, *i.e.*, making predictions on novel views of training objects, we also consider the more challenging object generalization setting. Almost all of our experiments are conducted in the *object generalization setting* where we assess predictions on novel objects across datasets.

Metrics: Following [59,73], we report Chamfer distance (CD) and F-score at 5mm & 10mm thresholds. F-score evaluates the distance between object surfaces as the harmonic mean between precision & recall. Precision measures accuracy of the reconstruction as % of reconstructed points that lie within a certain distance to ground truth. Recall measures completeness of the reconstruction as % of points, on the ground truth, that lie within a certain distance to the reconstruction. CD computes sum of distances for each pair of nearest neighbors in the two point clouds. We report mean CD & F-score over all test objects.

Baselines: We compare our model with AC-SDF trained in supervised manner using 3D ground truth on different combination of datasets in different settings: (1) For object generalization on MOW in the wild, AC-SDF is trained on ObMan, ObMan + HO3D, ObMan + HO3D + HOI4D, (2) For object generalization on HO3D, AC-SDF is trained on ObMan, ObMan + MOW, ObMan + MOW + HOI4D, (3) For view generalization on HO3D, AC-SDF is trained on ObMan + HO3D. We also compare with an occupancy variant of AC-SDF (AC-OCC) and recent published methods with different forms of SDF representation, *e.g.*

Table 4: Comparison with relevant methods. Our approach also outperforms gSDF, AlignSDF & DDFHO (trained in the same setting as ours) in zero-shot generalization to MOW across most metrics.

Method	F@5 ↑	F@10 ↑	CD ↓
AC-SDF [73]	0.108	0.199	7.82
AlignSDF [10]	0.099	0.182	8.30
gSDF [9]	0.107	0.197	7.50
DDFHO [77]	0.094	0.166	3.06
HORSE (Ours)	0.121	0.220	6.76

Table 5: 3D *vs*. 2D input to discriminator. Training with 3D inputs (at different resolutions) perform worse, likely due to coarse sampling resulting in very few points inside the object.

Disc. input	F@5 ↑	F@10 ↑	CD ↓
No disc.	0.117	0.216	6.93
$10 \times 10 \times 10$	0.120	0.218	7.29
$16 \times 16 \times 16$	0.115	0.209	7.79
$32 \times 32 \times 32$	0.104	0.191	7.83
2D slices	**0.121**	**0.220**	**6.76**

AlignSDF [10], gSDF [9], DDFHO [77]. Note that the VISOR dataset cannot be used for training since it does not have 3D supervision. For the view generalization setting on HO3D, we also compare with HO [22] & GF [31] trained with 3D ground truth on ObMan + HO3D. Recent works [44,70] on unsupervised reconstruction of objects require several views or depth, which are not available in our setting.

4.2 Results

Object generalization in the wild:. We first examine if the auxiliary supervision from visual hull and shape prior is useful for generalization to novel objects in the wild. We evaluate on MOW in Tab. 1 and compare with AC-OCC & AC-SDF trained on different combinations of ObMan, HO3D, HOI4D datasets with 3D supervision. Our approach provides gains of 24.3% compared to AC-OCC (trained on ObMan) and 11.6% on AC-SDF (trained on ObMan). This shows the benefits of our supervision cues in the wild over training on just large scale synthetic data with 3D supervision. We also outperform AC-SDF trained on ObMan + HO3D + HOI4D with full 3D supervision by 16.8% across all metrics. This indicates that our supervision cues from in-the-wild VISOR are better than using 3D supervision on lab datasets with limited diversity in objects. We also outperform relevant methods that use different forms of SDF representations, *e.g*. AlignSDF, gSDF & DDFHO across most metrics (Tab. 4). Note that our contributions are orthogonal and could be combined with these works.

Adding 3D supervision to AC-SDF. In Tab. 1, we observe that adding more data from HO3D & HOI4D to AC-SDF training did not help in zero-shot generalization to MOW. Instead, the performance drops compared to AC-SDF trained on ObMan. This is likely due to limited diversity in HO3D: 10 YCB objects, HOI4D: 7 rigid object categories & the model overfitting to these categories.

Object generalization on HO3D: Our approach is better than AC-OCC & AC-SDF trained on different datasets with 3D supervision (Tab. 2). This further shows the benefits of auxiliary supervision from VISOR for object generalization. Also, AC-SDF does not benefit from MOW & HOI4D. This could because HO3D evaluates on 10 objects only and they may not be present in MOW or HOI4D.

Table 6: Supervision quality on HO3D. Automated filtering to remove incorrect hand poses improves results & using ground truth hand pose differs little compared to predicted pose.[1]

	F@5 ↑	F@10 ↑	CD ↓
HORSE (base setting)	0.234	0.434	1.41
no training on HO3D	0.175	0.329	3.72
w/o filtering	0.213	0.405	1.42
w/ ground truth pose[1]	0.243	0.444	1.39

Table 7: Role of different loss functions. We report F-score at 5mm & 10mm, Chamfer distance (CD, mm) for different variants of our model on MOW. All losses are effective & multiview supervision leads to largest gain.

$\mathcal{L}_{\text{ObMan}}$	$\mathcal{L}_{\text{visual-hull}}$	$\mathcal{L}_{\text{consistency}}$	$\mathcal{L}_{\text{shape-prior}}$	F@5 ↑	F@10 ↑	CD ↓
✓				0.095	0.181	8.69
✓	✓			0.111	0.205	7.26
✓		✓		0.073	0.132	12.75
✓			✓	0.097	0.175	10.29
✓	✓	✓		0.117	0.216	6.93
✓	✓	✓	✓	**0.121**	**0.220**	**6.76**

Occupancy vs SDF. We see that SDF formulation is better than occupancy when trained with full 3D supervision (AC-OCC *vs*. AC-SDF). In contrast, we find SDF training to be unstable (does not give meaningful predictions) with auxiliary supervision. This could be because regressing continuous SDF values with weak supervision is harder than binary classification for occupancy values.
View generalization results on HO3D. In Tab. 3, we see gains with using supervision cues over just training on synthetic data, consistent with trends in the object generalization setting. We also outperform HO [22] & GF [31], both trained on HO3D using full 3D supervision. We outperform these methods even without any images from HO3D (last row in Tab. 1 *vs*. GF & HO in Table 3), likely due to use of more expressive pixel-aligned & hand articulation features.

4.3 Ablation Study

Analysis of supervision quality. We also observe in Tab. 3 that our method is able to bridge more than 40% of the gap between no training on HO3D to training with full 3D supervision. We further use the view generalization setting to assess the quality of 2D object mask supervision used in our method in Tab. 6. Our automated filtering of frames with inaccurate hand poses (as described in Sec. 3.5) is crucial for good performance. Also, little is lost from using hand pose as a proxy for object pose on the HO3D dataset. [1]
Role of different loss terms: We experiment with multiple variants of our model to assess the importance of different loss terms. We start with the AC-OCC model trained on ObMan and gradually add $\mathcal{L}_{\text{visual-hull}}$, $\mathcal{L}_{\text{consistency}}$, and $\mathcal{L}_{\text{shape-prior}}$. From the results in Tab. 7, we observe that $\mathcal{L}_{\text{visual-hull}}$ is more effective than $\mathcal{L}_{\text{consistency}}$ and using them together provides further benefits. Moreover, $\mathcal{L}_{\text{shape-prior}}$ improves performance on top of $\mathcal{L}_{\text{consistency}}$ and $\mathcal{L}_{\text{visual-hull}}$.
3D vs 2D input to discriminator: We also consider 3D volumes as input to the discriminator (instead of 2D cross-sections). For this, we need to sample 64x64x64

[1] While [73] uses similar contrast between predicted *vs*. ground truth hands to make claims, we note that those claims & this result should be taken with a grain of salt. FrankMocap is trained on HO3D, so its predictions on HO3D are better than they would be on unseen data. As most of our models are trained on VISOR (not used for training FrankMocap), our other experiments do not suffer from this issue.

Table 8: Design choices for mask guided sampling. Uniformly sampling points is much worse than the rejection sampling used in our method. Using negative points from hand masks is useful.

Sampling method	F@5 ↑	F@10 ↑	CD ↓
Uniform	0.093	0.166	10.29
Ours (no hand points)	0.113	0.207	7.69
Ours	**0.117**	**0.216**	**6.93**

Table 9: Sampling method for 2D planes. Sampling planes through origin of hand coordinate system & rotated randomly performs the best compared to sampling axis-aligned planes either uniformly or through origin.

Sampling method	F@5 ↑	F@10 ↑	CD ↓
Uniform (axis-aligned)	0.115	0.208	7.01
Origin (axis-aligned)	0.098	0.183	8.52
Origin (random rotation)	**0.121**	**0.220**	**6.76**

(=262144) points & run several forward passes of our model to get occupancies Since this is computationally expensive, we sample points at coarser resolutions: 32x32x32, 16x16x16, 10x10x10. We use 32x32 size 2D slices, so 10x10x10 3D volume has no. of points & takes similar compute. We see that 2D slices perform better than 3D volumes (Tab. 5). Also, the performance gets worse with increase in the sampled 3D volume, likely due to 3D sampling being so coarse that very few points lie inside the object, thus unable to capture fine-grained shape.

Sampling 2D slices for discriminator: We ablate different design choices (Sec. 3.3) in Tab. 9. We observe that sampling 2D planes through origin of the hand coordinate system and rotated randomly performs the best compared to sampling axis-aligned frames either uniformly or through origin.

Design choices for mask guided sampling: We run rejection sampling (with hand & object masks) to sample points in the hand coordinate frame (Sec. 3.2). We compare with 2 variants: uniformly sampling in the hand frame & removing negative points from hand masks. We find our strategy to work the best (Tab. 8).

4.4 Visualizations

We compare the mesh generated by our model and AC-SDF (trained on ObMan-best baseline) on zero-shot generalization to MOW (Fig. 5) and Core50 [35] (Fig. 6). For this, we sample points uniformly in a $64 \times 64 \times 64$ volume, predict their occupancies or SDF from the network and run marching cubes [36]. We project the mesh into the input image & render it in different views. Our model captures the visual hull of the object, as evidenced by the projection of the mesh onto the image, and generates more coherent shapes than AC-SDF, which often reconstructs disconnected and scattered shapes. More visualizations are in supplementary.

4.5 Limitations

Inaccurate hand pose. We use predictions from FrankMocap for hand pose & camera parameters. Note that the sampled points do not cover the entire object if the hand pose is not accurate, due to mis-projection into the image plane. This leads to exclusion of points in certain parts of the object (Fig. 7).

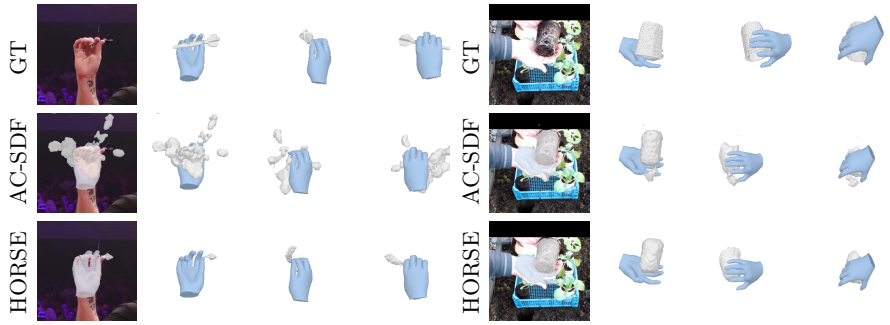

Fig. 5: Visualizations on MOW object generalization split. We show the object mesh projected onto the image and rendered in different views for our HORSE model and compare with the AC-SDF model trained on ObMan dataset with 3D supervision (best baseline model). We also show the ground truth (GT) object model. We observe that our model is able to predict the object shape more accurately than AC-SDF which often reconstructs smaller and disconnected shapes.

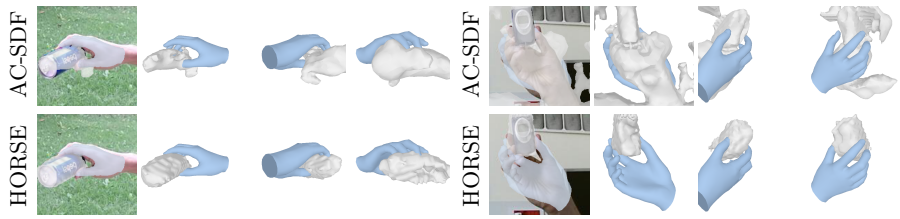

Fig. 6: Visualizations on zero-shot generalization to Core50 [35]. We show the object mesh projected onto the image and rendered in different views on Core50. HORSE predicts better shapes than AC-SDF (best baseline, often leads to artifacts).

Limited object views. Videos in the wild often do not capture 360° view of the object, *e.g.* kettle in Fig. 7. This is different than lab settings where the interactions are often constrained & multi-camera setup is used to capture all sides of the object.

Fig. 7: Sampled points do not cover the entire object if hand pose is inaccurate.

5 Conclusion

We present an approach for reconstructing hand-held objects in 3D from a single image. We propose modules to extract supervision from in-the-wild videos & learn data-driven 3D shape priors from synthetic ObMan to circumvent the need for direct 3D supervision. Experiments show that our approach generalizes better to novel objects in the wild than baselines trained using 3D supervision. Future directions include jointly optimizing the hand pose with the object shape to deal with inaccurate hand poses or incorporating additional cues, *e.g.* contact priors.

Acknowledgements: We thank Ashish Kumar, Erin Zhang, Arjun Gupta, Shaowei Liu, Anand Bhattad, Pranay Thangeda & Kashyap Chitta for feedback on the draft. This material is based upon work supported by NSF (IIS2007035), NASA (80NSSC21K1030), DARPA (Machine Common Sense program), an Amazon Research Award, an NVIDIA Academic Hardware Grant, and the NCSA Delta System (supported by NSF OCI 2005572 and the State of Illinois).

References

1. Bahat, Y., Shakhnarovich, G.: Confidence from invariance to image transformations. arXiv (2018)
2. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
3. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
4. Buckingham, G.: Hand tracking for immersive virtual reality: Opportunities and challenges. Frontiers in Virtual Reality (2021)
5. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
6. Chang, A.X., Funkhouser, T.A., Guibas, L.J., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: Shapenet: An information-rich 3D model repository. ArXiv (2015)
7. Chang, M., Prakash, A., Gupta, S.: Look ma, no hands! agent-environment factorization of egocentric videos. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
8. Chao, Y., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Wyk, K.V., Iqbal, U., Birchfield, S., Kautz, J., Fox, D.: Dexycb: A benchmark for capturing hand grasping of objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
9. Chen, Z., Chen, S., Schmid, C., Laptev, I.: gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
10. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
11. Choi, H., Chavan-Dafle, N., Yuan, J., Isler, V., Park, H.: Handnerf: Learning to reconstruct hand-object interaction scene from a single rgb image. In: International Conference on Robotics and Automation (2024)
12. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. Proceedings of the European Conference on Computer Vision (ECCV) (2018)
13. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. In: NeurIPS Track on Datasets and Benchmarks (2022)

14. Fan, Z., Parelli, M., Kadoglou, M.E., Kocabas, M., Chen, X., Black, M.J., Hilliges, O.: Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. arXiv preprint arXiv:2311.18448 (2023)
15. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.: First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
17. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
18. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
19. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
20. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T., Tai, C., Akbay, M., Wang, Z., Nitzan, A., Dong, G., Ye, Y., Tao, L., Wan, C., Wang, R.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Transactions on Graphics (TOG) (2020)
21. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
22. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Heppert, N., Irshad, M.Z., Zakharov, S., Liu, K., Ambrus, R.A., Bohg, J., Valada, A., Kollar, T.: CARTO: category and joint agnostic reconstruction of articulated objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
25. Huang, D., Ji, X., He, X., Sun, J., He, T., Shuai, Q., Ouyang, W., Zhou, X.: Reconstructing hand-held objects from monocular video. In: ACM Transactions on Graphics (2022)
26. Irshad, M.Z., Zakharov, S., Ambrus, R., Kollar, T., Kira, Z., Gaidon, A.: Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
27. Irshad, M.Z., Zakharov, S., Liu, K., Guizilini, V., Kollar, T., Gaidon, A., Kira, Z., Ambrus, R.: Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)

28. Jiang, H., Liu, S., Wang, J., Wang, X.: Hand-object contact consistency reasoning for human grasps generation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
29. Kanazawa, A., Tulsiani, S., Efros, A.A., Malik, J.: Learning category-specific mesh reconstruction from image collections. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
30. Kar, A., Tulsiani, S., Carreira, J., Malik, J.: Category-specific object reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
31. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: Proceedings of the International Conference on 3D Vision (3DV) (2020)
32. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2O: two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2021)
33. Laurentini, A.: The visual hull concept for silhouette-based image understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **16**, 150–162 (1994)
34. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: HOI4D: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
35. Lomonaco, V., Maltoni, D.: Core50: a new dataset and benchmark for continuous object recognition. In: Proceedings of the Conference on Robot Learning (CoRL) (2017)
36. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3D surface construction algorithm. ACM Transactions on Graphics (1987)
37. Lunayach, M., Zakharov, S., Chen, D., Ambrus, R., Kira, Z., Irshad, M.Z.: FSD: fast self-supervised single RGB-D to categorical 3d objects. arXiv **abs/2310.12974** (2023)
38. Ma, W.C., Yang, A.J., Wang, S., Urtasun, R., Torralba, A.: Virtual correspondence: Humans as a cue for extreme-view geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
39. Mandikal, P., Grauman, K.: Dexvip: Learning dexterous grasping with human hand pose priors from video. In: Proceedings of the Conference on Robot Learning (CoRL) (2021)
40. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2021)
41. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
42. Matusik, W., Buehler, C., Raskar, R., Gortler, S.J., McMillan, L.: Image-based visual hulls. In: ACM Transactions on Graphics (2000)
43. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
44. Niemeyer, M., Mescheder, L.M., Oechsle, M., Geiger, A.: Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

45. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
46. Prakash, A., Tu, R., Chang, M., Gupta, S.: 3d hand pose estimation in everyday egocentric images. In: Proceedings of the European Conference on Computer Vision (ECCV) (2024)
47. Qin, Y., Su, H., Wang, X.: From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. Proceedings of the International Conference on Intelligent Robots and Systems (IROS) (2022)
48. Qin, Y., Wu, Y., Liu, S., Jiang, H., Yang, R., Fu, Y., Wang, X.: Dexmv: Imitation learning for dexterous manipulation from human videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
49. Rijpkema, H., Girard, M.: Computer animation of knowledge-based human grasping. In: Thomas, J.J. (ed.) ACM Transactions on Graphics (1991)
50. Rockwell, C., Fouhey, D.F.: Full-body awareness from partial observations. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
51. Rogez, G., Khademi, M., Supančič III, J., Montiel, J.M.M., Ramanan, D.: 3d hand pose detection in egocentric rgb-d images. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014)
52. Romero, J., Kjellström, H., Kragic, D.: Hands in action: real-time 3D reconstruction of hands in interaction with objects. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2010)
53. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) (2017)
54. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3D hand and body motion capture by regression and integration. Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops) (2021)
55. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
56. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
57. Shan, D., Geng, J., Shu, M., Fouhey, D.F.: Understanding human hands in contact at internet scale. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
58. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
59. Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T.: What do single-view 3d reconstruction networks learn? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
60. Truong, P., Rakotosaona, M., Manhardt, F., Tombari, F.: SPARF: neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
61. Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Laina, I., Larlus, D., Damen, D., Vedaldi, A.: EPIC fields: Marrying 3d geometry and video understanding. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
62. Tschernezki, V., Laina, I., Larlus, D., Vedaldi, A.: Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In: Proceedings of the International Conference on 3D Vision (3DV) (2022)

63. Tschernezki, V., Larlus, D., Vedaldi, A.: Neuraldiff: Segmenting 3d objects that move in egocentric videos. In: Proceedings of the International Conference on 3D Vision (3DV) (2021)
64. Tulsiani, S., Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
65. Turpin, D., Wang, L., Heiden, E., Chen, Y., Macklin, M., Tsogkas, S., Dickinson, S.J., Garg, A.: Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. In: Proceedings of the European Conference on Computer Vision (ECCV) (2022)
66. Tzionas, D., Gall, J.: 3d object reconstruction from hand-object interactions. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
67. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)
68. Wu, Y., Wang, J., Wang, X.: Learning generalizable dexterous manipulation from human grasp affordance. In: Proceedings of the Conference on Robot Learning (CoRL) (2022)
69. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
70. Yariv, L., Kasten, Y., Moran, D., Galun, M., Atzmon, M., Basri, R., Lipman, Y.: Multiview neural surface reconstruction by disentangling geometry and appearance. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
71. Ye, J., Wang, J., Huang, B., Qin, Y., Wang, X.: Learning continuous grasping function with a dexterous hand from human demonstrations. arXiv (2022)
72. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-HOP: generative hand-object prior for interaction reconstruction and grasp synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
73. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3D reconstruction of generic objects in hands. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
74. Ye, Y., Hebbar, P., Gupta, A., Tulsiani, S.: Diffusion-guided reconstruction of everyday hand-object interaction clips. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2023)
75. Ye, Y., Tulsiani, S., Gupta, A.: Shelf-supervised mesh prediction in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
76. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
77. Zhang, C., Di, Y., Zhang, R., Zhai, G., Manhardt, F., Tombari, F., Ji, X.: DDF-HO: hand-held object reconstruction via conditional directed distance field. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
78. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
79. Zhu, Z., Damen, D.: Get a grip: Reconstructing hand-object stable grasps in egocentric videos. arXiv preprint arXiv:2312.15719 (2023)

80. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
81. Zimmermann, C., Ceylan, D., Yang, J., Russell, B.C., Argus, M.J., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2019)
82. Çalli, B., Singh, A., Walsman, A., Srinivasa, S.S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: Proceedings of the International Conference on Advanced Robotics (ICAR) (2015)