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

Yufei Ye[1,2] Abhinav Gupta[1,2] Shubham Tulsiani[1]
[1]Carnegie Mellon University [2]Meta AI
yufeiy2@cs.cmu.edu gabhinav@fb.com shubhtuls@cmu.edu

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

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

Abstract

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

1. Introduction

Humans interact with their surrounding world with their hands. Not only do we understand interaction as abstract concepts such as touching screens, squeezing balls, holding pens, *etc*., we also perceive their underlying 3D shape. Holding a pen means a stick lying on the purlicue and gripped by thumb, index, and middle fingers; holding a bowl is placing it on top of an up-facing palm. We aim to build a recognition system that can perceive and reason about the geometric information of hand-object interactions (HOI) for generic objects.

Over the past decade, we have made significant advances in inferring the 3D shape of both hands and objects in isolation. Hand poses can be recovered in the form of 2D keypoints, 3D skeletons [34, 45–47, 74], or even full 3D meshes [2, 56] via either fitting templates or direct regression. On the other hand, recent works have also pursued scaling object reconstruction from estimating the 6D pose of one specific known object [51] to more generic objects, such as various instances within one category [23,35,39], or even pursuing a joint model for cross-category reconstruction [11,22]. But one area where the progress has been quite limited is understanding human-object interactions (HOI) specifically for manipulable objects [14,58].

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

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

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

2. Related Work

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

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

Reconstructing hand-held objects. Understanding HOI in 3D is a very challenging problem due to mutual occlu-

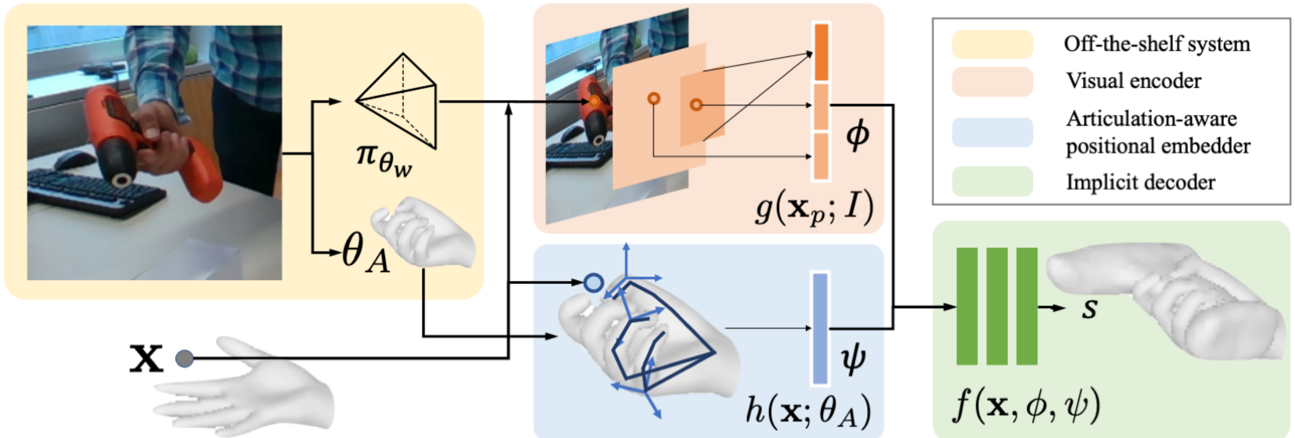

Figure 2. Given an image of a hand-held object, we first use an off-the-shelf system to estimate hand articulation θ_A and the camera pose π_w. With the predicted articulated hand along with the image, the object shape is reconstructed by an implicit network. For each query point $\mathbf{x}$ in canonical hand wrist frame, it is transformed to image space $\mathbf{x}_p$ to get visual feature $\phi = g(\mathbf{x}_p, I)$. In parallel, we also encode its articulation-aware representation $\psi = h(\mathbf{x}; \theta)$. Then we use an implicit decoder to predict signed distance value $s = f(\mathbf{x}, \phi, \psi)$.

sion and lack of data with annotation. To make inference tractable, prior works typically make the simplifying assumption of having access to known instance-specific templates (typically around 10 objects) [17, 27, 27, 28, 54, 60, 62, 64, 65] and infer these known objects under lab controlled environments. With access to instance-specific templates, they reduce the object reconstruction to 6D pose estimation and jointly predict both hand and object pose by reasoning about their interactions. The joint reasoning could be implicit feature fusion [8, 21, 41, 58, 62], explicitly leveraging geometric constraints like contact or collision [3, 4, 13, 24, 27, 71], or encouraging physical realism [50, 64]. In contrast, we focus on inferring hand-held object shape without knowing a corresponding template. Sharing this goal, work by Hasson *et al.* [31] predicts explicit genus-0 meshes and Karunratanakul *et al.* [37] predict a joint implicit field for reconstructing the hand and object. While these methods also perform model-free reconstruction, unlike our approach they *independently* infer the object shape and hand pose in a feed-forward manner. Instead, we formulate object reconstruction as conditional inference of 3D shape given the hand-articulation, and make our predictions in a normalized wrist frame, and show that this significantly improves performance.

3. Method

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

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

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

3.1. Preliminary: Hand Reconstruction

We use an off-the-shelf system [56] to estimate hand articulation and associated camera pose from an input image.

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

$$\mathbf{x}_p = \pi_{\theta_w}(\mathbf{x}) = KT_{\theta_w}\mathbf{x}$$

where K is the camera intrinsic and T_{θ_w} is the global rigid transformation of the hand.

3.2. Articulation-conditioned SDF

Given the predicted hand articulation θ_A, and camera matrix π_{θ_w}, our articulation-conditioned object shape inference network takes an additional input image I to output a signed distance field of the object. For a query point in the wrist coordinate frame, the point-wise network takes into account the query's corresponding visual feature from a visual encoder and its relative position to each joint from an articulation-aware positional embedder. The visual feature and the embedding are then passed to an implicit decoder along with the query to predict the signed distance.

Visual encoder. The visual encoder first extracts the image feature pyramid at different resolutions. For a query 3D point in the wrist frame, the visual encoder projects it to the image coordinate and compute global and local feature from the pyramid. The global part allows us to reason about global context and generate more coherent object shapes. For example, realizing the object is a bottle helps the network to generate a cylinder shape. The local feature allows the prediction more consistent with the visual observation [57, 70].

The backbone of the visual encoder is implemented as ResNet [32]. The global feature is a linear combination of the averaged conv5 feature. The local feature for each point is an interpolated feature at image coordinate from where it is projected by the predicted camera $\phi[\pi_{\theta_w}(\mathbf{x})]$, where ϕ denotes resnet feature and $\phi[x]$ represents a bilinear sample of the feature at a 2D location x. The local feature sampling is implemented for every layer of the feature pyramid $\phi_{1,\ldots,5}[x]$. It allows the model to draw visual cues with various resolutions and receptive fields.

Articulation positional embedder. Our key idea is that the hand pose is predictive of the object shape it interacts with. This is especially informative and complements the visual cues for reconstructing hand-held objects that are often occluded. We explicitly encode hand pose information for the query point via the articulation embedder. To do this, one naive way is to simply use identity mapping $\psi = \theta_A$. However, this representation is not robust to hand prediction error as we show in ablation. Furthermore, it is not trivial for the network to relate the reconstruction metric space with the hand pose joint space. For example, if a point is within 2mm from both index and thumb, it is very likely to have some object passing through. To better capture the structure of the problem, we encode the hand pose information by the position of the query points relative to the articulated joints.

More specifically, the articulation embedder takes as input an articulation parameter θ_A and a point position in wrist frame X to output the articulation-aware encoding $\psi = h(X; \theta_A)$. The encoding is a concatenation of the coordinates relative to every joint. Given the articulation parameter θ_A, we run forward kinematics to derive transformation $T(\theta_A) : \mathbb{R}^3 \to \mathbb{R}^{45}$ that maps a point in wrist frame to each joint coordinate. The 15 joint coordinates are position encoded [66] and concatenated together as the final representation $h(\mathbf{x}; \theta_A) = \gamma(T(\theta_A)\mathbf{x})$ where γ is the positional encoder. For more details please refer to appendix.

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

3.3. Training

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

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

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

3.4. Refining hand pose

While our work primarily focuses on object reconstruction conditioned on a predicted hand pose, this initial pose

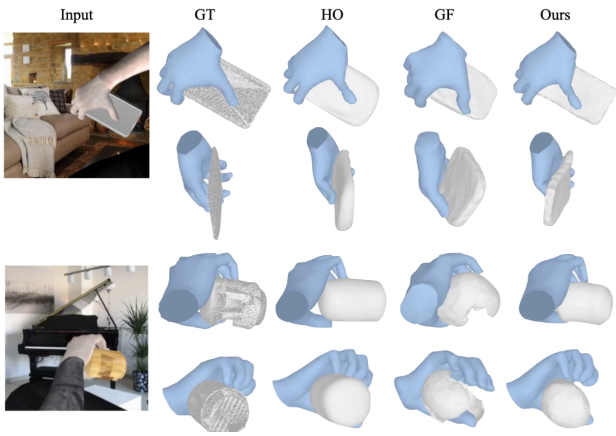

Figure 3. Visualizing reconstruction from our method and two baselines [31, 37] on ObMan dataset from the image frame and a novel view.

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

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

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

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

4. Experiments

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

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

- ObMan [31] is a synthetic dataset that consists as 8 categories of 2772 objects from 3D warehouse [7]. The grasps are automatically generated by GraspIt [44], resulting in a total of 21K grasps. The grasped objects are rendered over random backgrounds using Blender. We follow the standard splits where there is no overlap between the objects used in training and testing.
- HO-3D [29] is a real-world video dataset consisting of 103k annotated images capturing 10 subjects interacting with 10 common YCB objects [75]. The ground-truth is annotated using multi-camera reconstruction pipelines. To test on more shapes, we create a custom split by holding out one video sequence per object as test set. Please refer to the appendix for more details.
- MOW [5] dataset consists of a curated set of 442 images, spanning 121 object templates, collected from in-the-wild hand-object interaction datasets [14, 58]. It is more diverse in terms of both appearance and object shape compared to the HO3D dataset, but only provides approximate ground-truth. These object shape and hand pose annotations are obtained via a single-frame optimization-based method [5]. We use 350 randomly selected examples for training and the remaining 92 for evaluation.

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

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

	ObMan				HO3D				MOW			
	F-5 ↑	F-10 ↑	CD ↓	Vol ↓	F-5 ↑	F-10 ↑	CD ↓	Vol ↓	F-5 ↑	F-10 ↑	CD ↓	Vol ↓
HO [31]	0.23	0.56	**0.64**	8.64	0.11	0.22	4.19	9.44	0.03	0.06	49.8	25.6
GF [37]	0.30	0.51	1.39	1.84	0.12	0.24	4.96	6.31	0.06	0.13	40.1	**8.82**
Ours	**0.42**	**0.63**	1.02	**1.74**	**0.28**	**0.50**	**1.53**	**4.77**	**0.13**	**0.24**	**23.1**	19.4

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

the object should be minimal, we also report the intersection volume between two meshes (in cm^3) as a measure of understanding the relations between the hand and object.

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

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

4.1. Results on ObMan

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

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

4.2. Evaluation on real-world datasets

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

train set	F-5 ↑	F-10 ↑	CD ↓
ObMan	0.14	0.27	4.36
MOW	**0.15**	**0.30**	**4.09**

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

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

4.3. Ablations

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

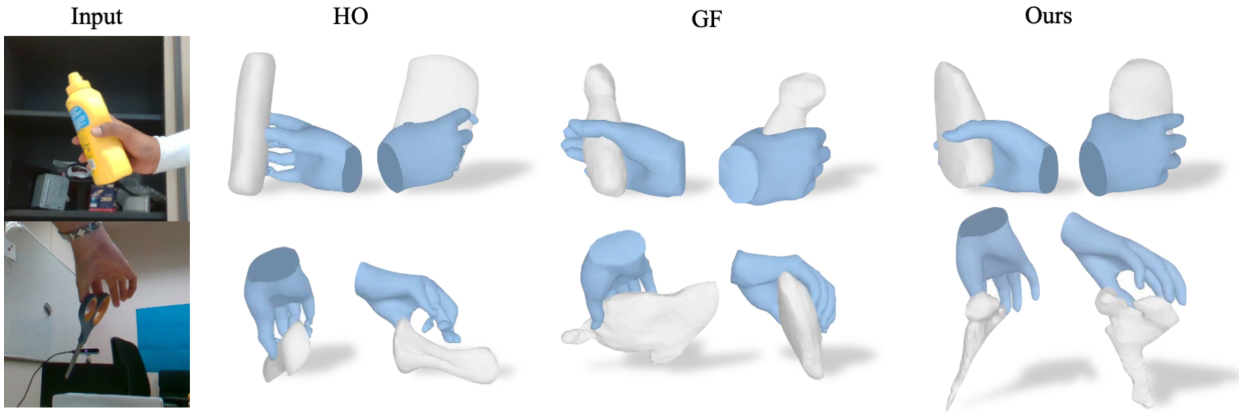

Figure 4. Visualizing reconstruction of our method and two baselines [31, 37] on HO3D dataset in the image frame and from a novel view.

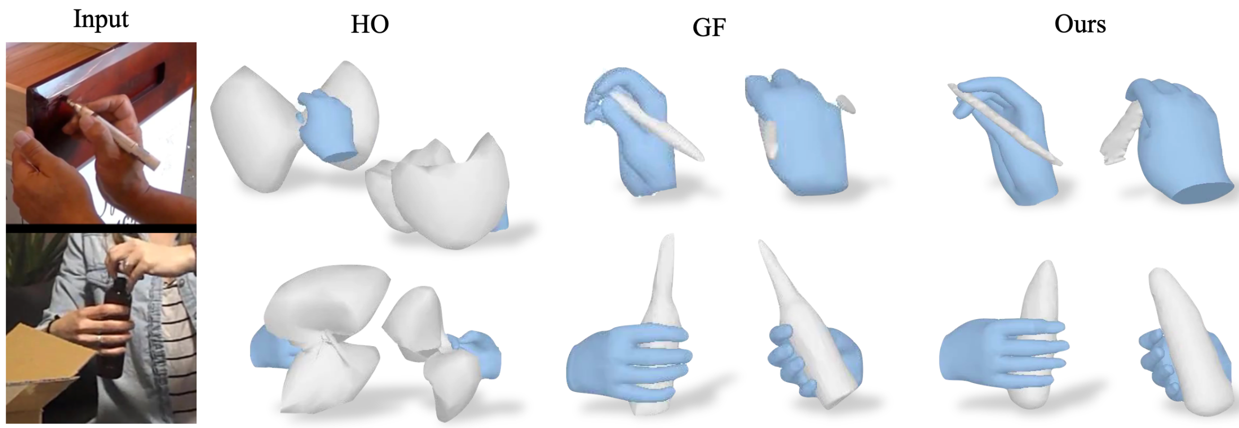

Figure 5. Visualizing reconstruction of our method and two baselines [31, 37] on MOW dataset in the image frame and from a novel view.

		F-5 ↑	F-10 ↑	CD ↓	Vol ↓
ObMan	Ours	**0.42**	**0.63**	**1.02**	**1.74**
	Ours w/o Art.	0.37	0.56	1.89	3.93
HO3D	Ours	**0.33**	**0.58**	**0.93**	**4.77**
	Ours w/o Art.	0.27	0.48	1.18	6.30
MOW	Ours	**0.13**	**0.24**	**23.1**	19.4
	Ours w/o Art.	0.10	0.19	29.0	**17.3**

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

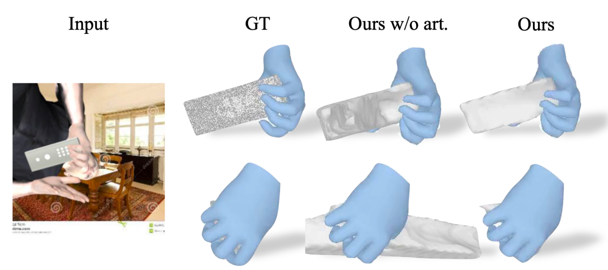

Figure 6. Visualizing reconstruction of hand-held object with or without explicitly considering hand pose.

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

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

Robustness against hand prediction quality. We use

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

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

	Noise Level	ObMan F5 ↑	ObMan F10 ↑	ObMan CD ↓	HO3D F5 ↑	HO3D F10 ↑	HO3D CD ↓
Gaussian	50% σ	0.40	0.63	1.01	0.28	0.50	1.51
	100% σ	0.31	0.53	1.40	0.25	0.46	1.68
	150% σ	0.24	0.42	1.94	0.22	0.42	1.93
Prediction	50%	0.46	0.67	0.96	0.29	0.52	1.48
	100% *	0.42	0.63	1.02	0.28	0.50	1.53
	200%	0.35	0.56	1.28	0.26	0.47	1.67
Baselines	HO	0.23	0.56	0.64	0.11	0.22	4.19
	GF	0.30	0.51	1.39	0.12	0.24	4.96
GT	0%	0.49	0.70	0.92	0.30	0.53	1.46

Table 5. Error analysis against hand pose noise. σ is the average prediction error. * marks our unablated method.

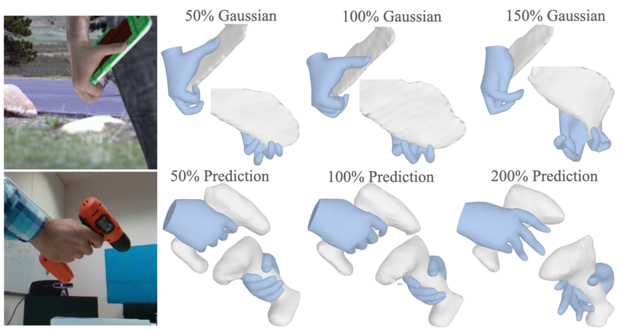

Figure 7. Top: Object reconstruction given hand pose corrupted by Gaussian on Obman dataset. Bottom: Object reconstruction given hand pose corrupted by prediction error on HO3D dataset.

	F-5↑	CD↓	Vol↓	Sim↓	EPE↓
ours w/o rf	0.17	1.02	1.74	3.32	8.9
ours w rf	0.17	**1.00**	**1.28**	**3.00**	**8.7**
ours w GT pose	0.20	0.92	1.73	2.44	–

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

hand poses corrupted by different levels of noise, either from Gaussian or more structured prediction noise. For the latter, we linearly interpolate (and even extrapolate) the true poses and off-the-shelf predictions. Our method still outperform baselines even when the predicted hand pose is *with twice more error* (Tab 5 and Fig 7).

Test-time refinement improves hand pose. The object reconstruction above is obtained by direct feed-forward prediction. We then show that our articulation-conditioned object shape can in turn refine the initial hand pose estimation and improve the HOI quality. We report end point error (EPE in mm) for each joint on Obman dataset in Table 6. To evaluate HOI quality, we report intersection volume along with simulation displacement of the object. We follow prior works [31, 37, 64] to pass the HOI reconstruction to a simulator and report how much the object slips from hand after running simulation for a fixed amount of time.

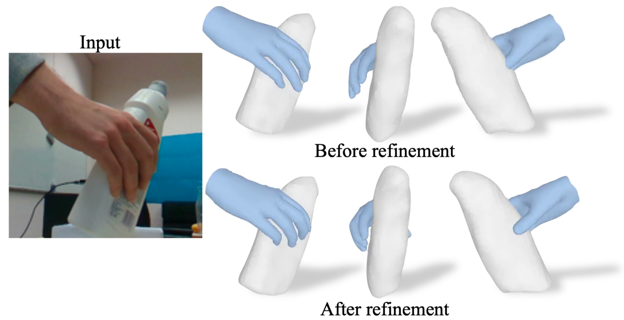

Figure 8. Visualizing hand-object reconstruction before and after test-time refinement in the image frame and from two novel views.

As shown in Table 6, both object and hand reconstruction improve after test-time refinement. The simulation displacement of the object drops with less intersection region. When ground truth hand articulation is provided, the object error and simulation displacement continue to improve. Figure 8 visualizes one example before and after refinement. Four finger tips are attracted to object surface while the thumb is pushed out of the object.

5. Conclusion

In this paper, we propose a method to infer implicit 3D shape of generic objects in hand. We explicitly treat predicted hand pose as a cue for object inference via an articulation-aware positional encoding. We have shown that this complements visual cues, especially when the hand-held object is occluded. While the results are encouraging, there are several limitations. For example, our work cannot be directly adapted to reconstructing dynamic grasps from videos where object consistency given varying articulation is required. Additionally, we require 3D ground-truth for training and it would be interesting to extend it with differentiable rendering techniques. Despite these challenges, we believe that our work on reconstructing hand-held generic objects takes an encouraging step towards understanding HOI for in-the-wild videos.

Acknowledgements. The authors would like to thank Zhe Cao and Ilija Radosavovic for helping set up MOW dataset. We would also like to thank Aljosa Osep, Pedro Morgado, and Jason Zhang for their detailed feedback on manuscript.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *ICCV*, 2009. 2

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[3] 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. 3

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021. 2, 3

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 5

[6] Thomas J Cashman and Andrew W Fitzgibbon. What shape are dolphins? building 3d morphable models from 2d images. *TPAMI*, 2012. 2

[7] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, 2015. 5

[8] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *TIP*, 2021. 3

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2

[11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *ECCV*, 2016. 2

[12] Timothy F Cootes and Christopher J Taylor. Active shape models—'smart snakes'. In *BMVC*. 1992. 2

[13] Enric Corona, Albert Pumarola, G. Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *CVPR*, 2020. 3

[14] 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. *ECCV*, 2018. 2, 5

[15] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *CVPR*, 2020. 2

[16] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 2

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2, 3

[18] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 2

[19] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, 2020. 2

[20] Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *ECCV*, 2016. 2

[21] Georgia Gkioxari, Ross B. Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. *CVPR*, 2018. 3

[22] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *ICCV*, 2019. 2

[23] Shubham Goel, Angjoo Kanazawa, and Jitendra Malik. Shape and viewpoint without keypoints. In *ECCV*, 2020. 2

[24] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. Contactopt: Optimizing contact to improve grasps. *CVPR*, 2021. 3

[25] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *CVPR*, 2018. 2, 6

[26] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010. 2

[27] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 3

[28] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and objects poses. 3

[29] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 5

[30] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. *CVPR*, 2020. 2

[31] 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. 3, 5, 6, 7, 8

[32] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016. 4

[33] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *SIGGRAPH*. 2005. 2

[34] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 2

[35] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 2

[36] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Category-specific object reconstruction from a single image. In *CVPR*, 2015. 2

[37] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*. 3, 5, 6, 7, 8

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

[39] 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. 2

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

[41] 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. 2, 3

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

[43] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[44] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 2004. 5

[45] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 2

[46] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (TOG)*, 2019. 2

[47] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, 2018. 2

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2, 3, 4

[49] Akila Pemasiri, Kien Nguyen Thanh, Sridha Sridharan, and Clinton Fookes. Im2mesh gan: Accurate 3d hand mesh recovery from a single rgb image. *arXiv*, 2021. 2

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

[51] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963. 2

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

[53] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 2

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

[55] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ToG*, 2017. 2

[56] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *ICCV Workshop*, 2021. 2, 3, 5

[57] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, 2019. 4, 6

[58] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 2, 3, 5

[59] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, 2015. 2

[60] 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. 3

[61] 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. 5

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

[63] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *CVPR*, 2017. 2

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

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

[66] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 4

[67] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 2

[68] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NeurIPS*, 2016. 2

[69] Yufei Ye, Shubham Tulsiani, and Abhinav Gupta. Shelf-supervised mesh prediction in the wild. In *CVPR*, 2021. 2

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

[71] 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. *ECCV*, 2020. 3

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

[73] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 2

[74] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 2

[75] 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. 5