ContactArt: Learning 3D Interaction Priors for Category-level Articulated Object and Hand Poses Estimation

Zehao Zhu[1*], Jiashun Wang[2*], Deqing Sun[3], Varun Jampani[3], Xiaolong Wang[4]
[1] Zhejiang University [2]Carnegie Mellon University [3]Google Research [4]UC San Diego
https://zehaozhu.github.io/ContactArt/

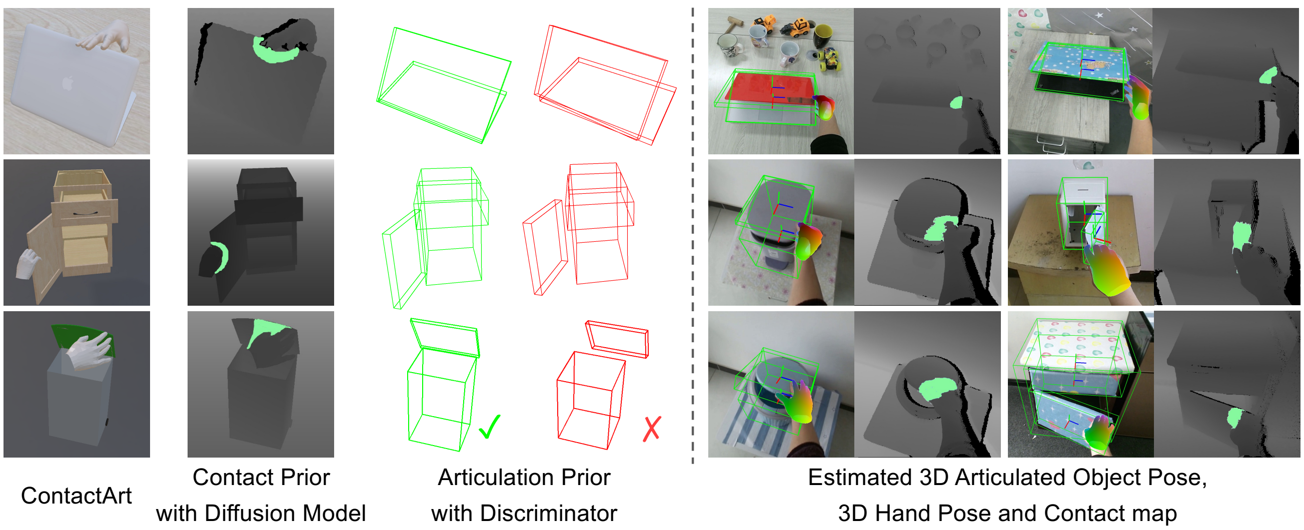


Figure 1. **Overview.** We collect a dataset named ContactArt, which is created by human interacting with the articulated objects in a simulator, using teleoperation. Two interaction priors are learned from ContactArt: (i) a contact prior predicted by a diffusion model to improve 3D hand pose estimation; (ii) an articulation prior with a discriminator to improve category-level articulated object pose estimation. We visualize the pose estimation results in real-world data, leveraging the learned priors.

Abstract

We propose a new dataset and a novel approach to learn hand-object interaction priors for hand and articulated object poses estimation. We first collect a dataset using visual teleoperation, where the human operator can directly play within a physical simulator to manipulate the articulated objects. We record the data and obtain the free and accurate annotations on object poses and contact information from the simulator. Our system only requires an iPhone to record human hand motion, which can be easily scaled up and largely lower the costs on data and annotation collection. With this data, we learn 3D interaction priors including a discriminator (in a GAN) capturing the distribution of how object parts are arranged, and a diffusion model which generates the contact regions on an articulated objects, guiding the hand pose estimation. Such structural and contact priors can easily transfer to the real-world data with barely any domain gap. By using our data and learned priors, our method significantly improves the performance on joint hand and articulated object poses estimation over existing state-of-the-arts.

*: equal contribution

1. Introduction

The understanding of the 3D articulated structure has caught a lot of attention in computer vision recently: Active studies have been conducted on estimating the articulated object poses [45, 46, 78]. Beyond studying the single object in isolation, understanding the interactions between human hands and articulated objects play an important role in wide applications such as robotics and Augmented Reality. However, there remains several challenges for hand and category-level articulated object pose estimation given the high Degree of Freedom on poses and mutual occlusions.

Most current research focusing on articulated object pose estimation has been limited by the high cost of annotations on real-world objects [48]. To alleviate this issue, approaches on using synthetic data with cheaper annotations have been proposed [34, 45, 46, 78]. However, this inevitably introduces sim2real gap when transferring pose estimation to images in the wild. The joint estimation of human hand and articulated object poses makes the problem even more challenging given their mutual occlusions. Recent efforts on collecting the real-world category-level human-object 3D poses annotations [47] have largely advanced this field. However, the expensive labeling process still makes it hard to scale and it is very difficult to obtain the accurate contact labels between hand and objects from observing the images. Is there a cheaper and a more scalable way to obtain the hand-object interaction annotations?

Our answer is affirmative and our key insight is that, while there is a large sim2real appearance gap, the geometric contacts between hand and objects are actually consistent across simulation and real world. In this paper, we collect the hand-object interaction data and accurate annotations by asking humans to directly play within a physical simulator using visual teleoperation (Fig. 1 1st column). We name this dataset (**ContactArt**): **Contact** with **Art**iculation. Specifically, we design a visual teleoperation system that only requires a single camera from an iPhone to record the human hand, which makes it scalable. The user will use their hand, which is mapped to a MANO hand [60], to operate and manipulate the articulated objects in a physical simulator. Within each object category, we collect interaction data across diverse articulated object instances. We can obtain the accurate hand-object poses and their contact points for free by reading from the simulator. This largely reduces the labeling cost from previous approaches [12,47].

The ContactArt dataset enables us to train real-world pose estimators with the free annotations. To minimize sim2real gap, we learn 3D interaction priors from ContactArt and use them to improve the real-world hand and object poses estimation. We train a generalizable model for each object category, and evaluate the model on unseen instances. We propose to learn two types of 3D hand-object interaction priors, which capture how object parts are generally articulated and where human generally touch the object for manipulation. The first prior is to learn the discriminator network, modeling the joint distribution of object part arrangement inside each object category (Fig. 1 2nd column). Following a GAN framework [24], we consider the pose estimators as the generators, and we train the discriminator by using the estimated hand and object poses as the fake data inputs, and the ground-truth CaptureArt poses as the real data. The discriminator then learns how should object parts "naturally" connect together, and we use this discriminator via back-prop to optimize the estimated object pose. The second prior is to learn a contact map diffusion generator [67] for modeling where the hand can touch the object (Fig. 1 3rd column). Given the input articulated object, this model predicts the plausible regions that the human hand operates (object affordance regions), using a diffusion process. With an initial hand pose estimation, we optimize the hand pose to match the estimated hand-object contact information. The two priors are complimentary to each other and are used jointly to optimize both estimated hand and articulated object poses.

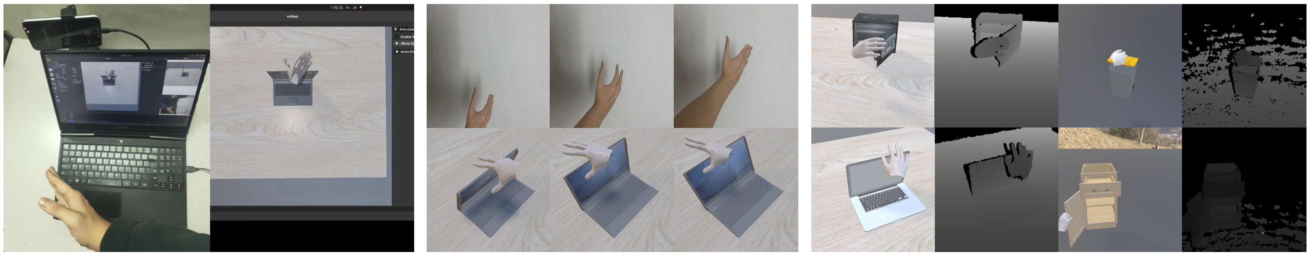
We perform our experiments on three in-the-wild articulated object datasets, HOI4D [47], BMVC [49] and RBO [48] including five categories in total. We find that with our ContactArt dataset and the proposed articulation and the contact prior, we can not only achieve large improvements over previous state-of-the-art methods on estimating articulated object poses, but also observe significant improvements on the hand pose estimation. Further, we find training on ContactArt first as a warm start then finetuning on HOI4D can bring better performance while requiring less data compared with training from scratch on HOI4D.

Our contributions include: (i) A new dataset with contact-rich hand articulated object interaction; (ii) A contact diffusion model used to estimate the contact map of interaction; (iii) An articulation discriminator which learns articulation prior and boosts articulated object pose estimation; (iv) Substantial performance improvement on articulated object and hand pose estimation.

2. Related Work

Articulated Object Pose Estimation. Beyond understanding single rigid objects, more attentions have been put on articulated object modeling and pose estimation recently [45, 46, 49, 51, 75, 77, 78, 82]. For example, Li *et al*. [45] propose to perform category-level articulated object pose estimation, and evaluate their approach on unseen instances during training. Weng *et al*. [78] adopt the Ancsh [75] to handle category-level pose tracking for both rigid and articulated object by leveraging the RotationNet and CoordinateNet. Liu *et al*. [46] reform the articulated object pose estimation setting for real-world environments and build an articulated object dataset ReArt-48. However, these approaches mainly focus on modeling the articulated object itself without considering how hand and objects contact and interact. In this paper, we study the joint pose estimation problem with hand and object together using different priors learned from our ContactArt dataset.

Conditional Diffusion Probabilistic Models. Recent progresses on diffusion probabilistic models [35, 68–70] have shown to be very effective on generating high-quality images. Inspired by these results, the conditional diffusion model has been widely applied in text-to-image genera-



(a) Hardware setup and interface (b) Sequence collection (c) RGB-D images of the collected data

Figure 2. To **collect ContactArt**, the hardware requirement is an iPhone and a laptop. The system allows us to easily scale up the dataset without human annotation effort. We can collect manipulation sequences and render images from different camera views.

tion [17,52,59,64,67,71], image super resolution [39,59,65] and image-to-image translation tasks [6, 15, 63, 87]. Different from the above diffusion models which are conditioned on input image or prompt, our proposed contact diffusion model is conditioned on the point-wise feature from the point cloud. There have been several work [3, 6, 79] performing semantic segmentation with a conditional diffusion model. For example, Wolleb *et al*. [79] uses the stochastic sampling process to implicitly ensembles the segmentation masks of medical images.

Inspired by these works, our contact diffusion model adopts the diffusion process to predict the contact map indicating where the hand should touch the articulated object.

Hand Object Interaction. Estimating hand object interaction has been a long standing problem in computer vision [5, 30, 54, 55, 72]. More recently, a line of studies [11,13,18,31,33,34,42,53,74,83,85] use data-driven and deep-learning based methods to jointly estimate or constructe the hand and object. For example, Hasson et al. [34] propose to use synthetic data to learn two separate deep neural networks to regress the hand and object mesh. Another line of research studies synthesizing plausible hand-object interactions [9,14,16,25,36,38,84,89]. For example, Jiang et al. [36] propose to generate the hand grasp pose and contact map at the same time and optimize the consistency between the hand and object during test time.

The success of these recent works is inseparable from the hand-object interaction datasets [8, 10, 12, 21, 27, 31, 34, 47, 58, 73, 80, 81], which are playing crucial roles in both estimation, synthesis and robot manipulation tasks. For example, DexYCB [12] and HOI4D [47] are two recent hand-object pose datasets annotated by humans, which is much more expensive compared to 2D labels. ContactDB [8] is proposed to capture the hand-object contact map with a thermal camera. But this is difficult to scale given the equipment requirement. To remove the constraints from annotation cost and hardware setup, we propose to collect the human and articulated object interaction using visual teleoperation in a physical simulator. We provide a scalable solution with free annotations from the simulator. Such geometric priors and contacts are transferable to the real world.

Vision-based manipulation teleoperation [4,19,20,32,41, 44, 57, 66] is a commonly applied technique in robotics. To reduce the device cost for scalable collection, we build our system upon [57], using only a single-camera to record the human hand to manipulate the articulated objects inside the Sapien [80] simulator. Different from the robotics application [57], our goal is to record the hand-object poses as well as the contact points for learning 3D interaction priors.

Adversarial Learning for Priors While adversarial learning is initially proposed for image generation [23], the discriminator trained with adversarial learning is also utilized in multiple tasks such as 3D Human pose estimation [1, 7, 28, 37, 40, 76] and 2D human trajectory prediction [2, 29, 43, 62]. Our articulation prior is inspired by [37,40], which are focusing on a specific articulation category: human. These approaches try to jointly learn a prior for what is a natural pose for human, how each articulated part is combined with the others. Similarly, in articulated objects, we have the upper drawer and the lower drawer are always parallel, and the keyboard and screen share a common side. Thus we propose to utilize the discriminator from adversarial learning to capture the articulation priors.

3. ContactArt Dataset

To the best of our knowledge, there is only one large-scale dataset [47] including 3D hand-articulated object interaction. However it still holds the following limitations. (i) HOI4D dataset is only captured in egocentric view and can not generalize to the third view. (ii) The annotation of HOI4D dataset is not accurate enough to provide contact information. Therefore we design a teleoperation system to build a large hand articulated object interaction dataset with no annotation effort and more accurate pose and contact information.

We design a single-camera human teleoperation system to manipulate articulated objects in the Sapien [80] simulation. This system allows us to get accurate pose annotation and contact information using only an iPhone and a laptop (Fig. 2 (a)). Since the teleoperation is in the simulation, the

Dataset	HOI	Hand GT	Multi Views	Contact Label	Frames
BMVC [49]	×	×	×	×	8K
RBO [48]	✓	×	×	×	12K
ReArtMix [46]	×	×	✓	×	100K
ReArtVal [46]	×	×	×	×	6K
HOI4D [47]	✓	✓	×	×	1.44M
ContactArt	✓	✓	✓	✓	332K

Table 1. Comparison with other articulated object dataset. HOI refers to hand object interaction and Hand GT refers to annotation of ground truth hand pose. ContactArt allows rendering in multi-view and has accurate contact information. Statistics is performed on articulated object.

annotations can be automatically recorded, which will make it easy to scale up the size of the dataset. It is also beneficial for us because we can render each frame with different camera view. The system greatly increases the ease of use. We train our models with this collected dataset.

Dataset collection. We use the front camera of an iPhone to stream the RGB-D video at 15 fps. The set up and collection interface is shown in Fig. 2 (a). We provide a sequence of ContactArt collection process in Fig. 2 (b). The teleportation system allows one to control the customized robot hand with his/her own hand motion as control signal in the simulation. One can easily manipulate the articulation object, such as opening the drawer. We render the RGB image and depth image respectively. We give examples in Fig. 2 (c). We record the object pose, bounding boxes and hand poses and the hand-object contact regions. Note that for rendering the depth image, we apply the active stereovision depth sensor simulation proposed in [86], which renders realistic depth images close to the depth camera captured in real world.

Dataset statistics. We select five common articulated object categories in our daily life including laptop, drawer, safe, microwave and trashcan, 80 instances in total to collect. All the object models are from Partnet dataset [50]. And it is convenient to scale up. Tab. 1 summarizes the statistics of ContactArt comparing previous datasets. ContactArt can provide accurate annotation, rich hand object interaction and contact information. One can also easily render more frames by using different camera views. Please see the supplementary materials for more details about dataset statistics. **We will release our dataset and the code for collection and one can easily incorporate more data.**

4. Method

In this paper, we target at the problem of hand and articulated objects estimation from known categories with interaction. Compared to [45,46,46,78] which only focus on the pose estimation of articulated object, our method pays attention to both hand and articulated object since they influence a lot each other during interaction. Our method takes an RGB-D image as input and output part-level 6D object pose (rotation and translation) and the hand pose parameterized by the MANO model [60].

We propose a NOCS-based [75] category-level pose estimator for articulated objects, together with two 3D interaction priors. In our framework, we train the articulated object pose estimator using both the reconstruction loss, and adversarial training with a discriminator. This **Articulation Discriminator** will serve as a prior of how object parts should be arranged together within a category. During test time, given an initial estimation from the pose estimator, we can use the discriminator to provide the gradients through back-propagation to optimize the pose of each object part. Meanwhile, to model the hand object interaction, we also propose a diffusion-based contact map generator, which estimates the regions where the hand will touch on the object, namely **Contact Diffusion Model**. We will use it as an optimization constraint to encourage the hand to reach the generated contact region. The architecture of our model is shown in Fig. 3 and the test time adaption framework is shown in Fig. 4.

4.1. Object Pose Estimator

We design a multi-branch pose estimator $\mathcal{E}$ to predict the articulated object pose (Fig. 3 dotted blue box). We first detect the hand and object and get the 2D bounding box of them with off-the-shelf method [22] and backproject the patch to point cloud $v \in \mathbb{R}^{N\times 3}$. Then we utilize PointNet++ [56] to extract the points feature. Building on the object points features, we use three separate MLPs to predict (i) the part segmentation, (ii) part-level NOCS map [45,75] and (iii) rotation of each parts. We adopt the 6D continuous rotation representation [88] for rotation. The part-level NOCS map is defined as a 3D space contained within a unit cube of each articulated parts and consistently aligns to a category-level canonical orientation.

In a forward pass, given the predictions of per-part rotation and dense correspondence between NOCS map and point cloud, we can analytically compute translation and scale via the Umeyama algorithm. Then we leverage the computed pose (rotation, translation, scale) to transform the canonical 3D bounding box to the camera space. Different from [45,46], by predicting the rotation of each parts using a neural network, we make the prediction of pose and bounding box fully differentiable, allowing end-to-end training. This also provides the opportunities for optimization using the discriminator from adversarial learning.

Specifically, we use the cross-entropy loss (CE) for part segmentation. For rotation loss, we calculate the L2 distance between our prediction and ground truth in this 6D space in the form of continuous rotation representation. We use L2 distance for NOCS map loss. The object pose esti-

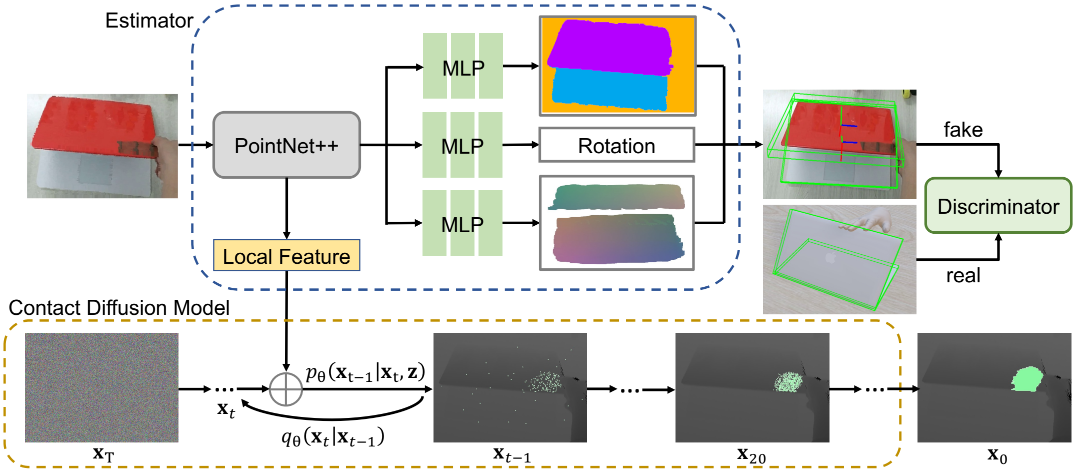


Figure 3. **Training framework.** We first adopt Pointnet++ [56] to extract a point-wise local feature and pass it into three individual branches to regress the part segmentation, NOCS map and part-level rotation. We then compute the 3D bounding box of each parts and feed it to a discriminator. We utilize a contact diffusion model conditioned on the Pointnet++ feature to estimate the contact map, which serves as the contact prior to further optimize hand pose. We visualize the contact points in green. $\bigoplus$ denotes concatenation.

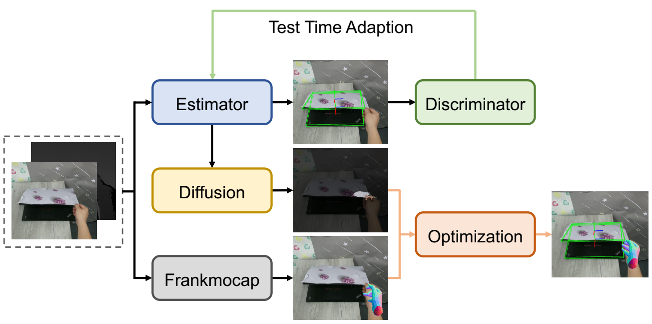


Figure 4. **Test time adaption framework.** We utilize the discriminator with fixed paramaters to calculate adversarial loss and back propagate the gradients to update the estimator. Then we optimize hand pose by minimizing the distance between hand vertices and the contact points at the predicted contact map.

mation loss can be written as,

$$L_{pose} =\lambda_{seg} \sum_{i}^{N} CE(s_i, s_i^*) + \lambda_{rot} \|r - r^*\|_2 + \lambda_{nocs} \sum_{i}^{N} \mathbb{1}(s_i^* > 0) \|n_i - n_i^*\|_2 \qquad (1)$$

where N is the number of sampled points, s^*, s are the ground truth and predicted part segmentation, n^*, n are the ground truth and predicted part-level NOCS maps and r^*, r are the ground truth and predicted rotation. λ_{nocs}, λ_{seg} and λ_{rot} are hyperparameters balancing the weights. In addition to L_{pose}, we introduce an adversarial loss with the Articulation Discriminator as described in the following section.

4.2. Articulation Discriminator

Our model jointly learns an Articulation Discriminator $\mathcal{D}$ (Fig. 3 green box) as the articulation structure prior during training the estimator $\mathcal{E}$. This discriminator will improve the naturalness on how parts are arranged together. The discriminator takes inputs as the 3D bounding boxes, which fully reflect the part placement rules. Furthermore, there is only a very small sim2real domain gap on the 3D bounding box space. We can calculate each parts' bounding box $\hat{b}$ with the outputs from the estimator. During training, we use the estimated boxes $\hat{b} \sim p_{\mathcal{E}}$ as negative samples. We use the accurate bounding boxes b from simulation data p_S as positive samples. We define the loss function for the discriminator as,

$$L_{\mathcal{D}} = \mathbb{E}_{b \sim p_S}[(\mathcal{D}(b) - 1)^2] + \mathbb{E}_{\hat{b} \sim p_{\mathcal{E}}}[(\mathcal{D}(\hat{b}))^2]. \qquad (2)$$

And the adversial loss term for the estimator is,

$$L_{adv} = \mathbb{E}_{\hat{b} \sim p_{\mathcal{E}}}[(\mathcal{D}(\hat{b}) - 1)^2]. \qquad (3)$$

4.3. Contact Diffusion Model

Diffusion model have shown state-of-the-art performance in generation tasks. In our work, we extend it to generate realistic 3D contact map between the object and hand point cloud (Fig. 3 bottom). Once we get such contact information, we use it to guide the optimization of 3D hand.

We first define the definition of contact map. For an input point cloud set $v \in \mathbb{R}^{N \times 3}$, the contact map $x \in \mathbb{R}^{N \times 1}$ is defined as a binary vector indicates whether each point belonging to the contact region or not. We calculate the L2 distance between the points from the object and its nearest points from the hand. If this distance is smaller than a threshold, we take this point as contacted.

We formulate the details of our contact diffusion model as following. Let $X_0 = (x_0, z)$ denote the input, where $x_0 \in \mathbb{R}^{N \times 1}$ is the contact map and $z \in \mathbb{R}^{N \times 3}$ is the PointNet++ local feature. T is the number of steps in the diffusion model and the intermediate results can be denoted as $X_t = (x_t, z)$, where $0 \leq t \leq T$. Diffusion models are composed of forward and backward processes. The forward process gradually injects random noise to the distribution, while the generative process learns to remove noise to obtain realistic samples by mimicking the reverse process. The forward process converts the original contact map distribution into a noise distribution, which can be described by the formulation,

$$q(x_{1:T}|x_0) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1}), \tag{4}$$

The reverse process p_θ is learned by the model parameters θ. Different from the forward process which simply adds noise to the contact map, the reverse process recovers the desired contact map from the input noise, encoded by the Pointnet++ feature z. The reverse diffusion process is,

$$p_\theta(x_{0:T}|z) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, z). \tag{5}$$

A parameterization trick [35] is used to simplify the training objective. The simplified training objective becomes,

$$E_{X_0 \sim q(x_0), x_{1:T} \sim q(x_{1:T}|x_0, z_0)}[\sum_{t=1}^{T} \log p_\theta(x_{t-1}|x_t, z)]. \tag{6}$$

Since posterior $q(x_{t-1}|x_t, x_0, z)$ is known and its derivation is similar to the unconditional generative model, we define the L_{diff} as,

$$L_{diff} = \|\epsilon - \epsilon_\theta(x_t, z, t)\|^2. \tag{7}$$

Please note that the whole diffusion model is trained with the estimator in an end-to-end fashion, which takes the Pointnet++ feature predicted by the estimator as input. Specifically, in our diffusion model, ϵ_θ is implemented in MLP. We first concatenate Pointnet++ feature z, contact map x_t and the time embedding in current step t. After that, we pass the concatenated feature into MLP. Then we use the predicted noise to compute affordance map in the next step. Same to [3], we also employ multiple generations to boost performance. Finally, the total loss of the whole pipeline can be written as,

$$L = L_{pose} + \lambda_{adv} L_{adv} + \lambda_{diff} L_{diff}, \tag{8}$$

where λ_{diff} and λ_{adv} are hyperparameters balancing the contact diffusion model loss and the adversial training loss.

4.4. Test Time Adaptation

Once we learn the **Articulation Discriminator** and the **Contact Diffusion Model**, we can improve the initial object and hand pose estimation with test time adaption. For optimizing object pose during test time, we fix the parameters of the discriminator $\mathcal{D}$ and use it to calculate the adversarial loss and back propagate the gradients to object pose estimator $\mathcal{E}$ to boost object pose estimation. For optimizing hand pose, we employ contact diffusion model to estimate the contact region and obtain the contact point set $C \in \mathbb{R}^{K\times 3}$ where K is the number of contact points. We then optimize the MANO [60] parameters of the hand which is initialized by the FrankMocap [61] hand pose estimator. Specifically, we minimize the chamfer distance between the hand vertices $V \in \mathbb{R}^{N\times 3}$ where N is the number of vertices and the contact points,

$$L_{CD} = \frac{1}{N}\sum_{v\in V}\min_{c\in C}\|v-c\|_2 + \frac{1}{K}\sum_{c\in C}\min_{v\in V}\|c-v\|_2. \tag{9}$$

5. Experiments

5.1. Datasets

We train our model on ContactArt and test on HOI4D [47], RBO [48], BMVC [49] respectively. **HOI4D** [47] is a large-scale hand-object interaction dataset where we can evaluate both object and hand pose estimation. We use 4 categories for evaluation: safe, trashcan, laptop, and drawer. We use 6000 frames for each category as the test set. We also perform experiments with finetuning on it. We find the model trained on ContactArt then finetuned on HOI4D will benefit from ConatactArt and achieve better performance than training from scratch. **RBO** [48] is a collection of RGB-D video sequences. There is no annotation of hand pose in RBO, we perform object pose estimation comparisons. We evaluate on 3 categories: laptop, microwave and drawer. **BMVC** [49] includes video sequences recording articulated object with a moving camera. There is no human manipulating the object, we evaluate object pose estimation on laptop following CAPTRA [78].

5.2. Metrics and Methods for Comparison

Metrics for comparison. For category-level articulated object pose estimation, we evaluate the following metrics: 5°5cm: percentage of results with rotation error smaller than 5° and translation error smaller than 5cm, mIoU: the average 3D intersection over union of ground-truth and predicted bounding boxes, R_{err}: rotation error in degrees, T_{err}: translation error in centimeters. For hand pose estimation, we report mean per vertex position error (MPVPE) and mean per joint position error (MPJPE).

Methods for comparison. We compare our method with two state-of-the-art image-based pose estimation works ANCSH [45] and ReArtNocs [46], and a tracking method CAPTRA [78], we provide the initial pose estimated by our method for fair comparisons. All the methods are trained on ContactArt. We also compare first training on ContactArt then finetuning on HOI4D (named Finetune) with training on HOI4D from scratch (named HOI4D*).

5.3. Object Pose Estimation Comparison

We summarize the quantitative articulated object pose estimation results on HOI4D in Tab. 2. Compared with the other methods, ours has the lowest average rotation and translation error, the highest mIoU and 5°5cm. Although CAPTRA leverages the temporal information, our method

Category	Metric	Ansch	ReArtNocs	CAPTRA	Ours	HOI4D*	Finetune
Laptop	$5^{\circ}5$cm$\uparrow$	10.54	10.60	16.35	**18.65**	61.95	**62.50**
	mIoU$\uparrow$	47.8	49.52	51.57	**52.45**	65.21	**66.45**
	$R_{err}\downarrow$	24.06	23.40	18.52	**17.73**	6.24	**5.20**
	$T_{err}\downarrow$	23.35	22.41	19.75	**18.91**	7.68	**7.17**
Trashcan	$5^{\circ}5$cm$\uparrow$	0	0	**3.05**	2.70	22.8	**24.2**
	mIoU$\uparrow$	38.38	39.30	41.50	**41.95**	63.65	**64.95**
	$R_{err}\downarrow$	26.93	25.57	21.97	**21.43**	7.05	**5.98**
	$T_{err}\downarrow$	36.72	36.65	**30.70**	30.75	7.75	**7.22**
Safe	$5^{\circ}5$cm$\uparrow$	1.66	5.30	4.65	**8.43**	32.05	**33.40**
	mIoU$\uparrow$	46.20	47.05	46.83	**47.96**	62.31	**63.50**
	$R_{err}\downarrow$	18.63	16.91	16.43	**16.24**	5.74	**5.11**
	$T_{err}\downarrow$	17.52	16.51	16.55	**15.66**	6.85	**6.50**
Cabinet	$5^{\circ}5$cm$\uparrow$	0.50	0	0.16	**1.00**	22.07	**23.87**
	mIoU$\uparrow$	49.83	49.90	49.76	**50.40**	64.43	**66.71**
	$R_{err}\downarrow$	18.94	19.52	19.87	**17.84**	6.46	**5.75**
	$T_{err}\downarrow$	23.33	23.28	24.23	**22.60**	6.03	**5.83**
Average	$5^{\circ}5$cm$\uparrow$	3.17	3.97	6.05	**7.69**	34.72	**35.99**
	mIoU$\uparrow$	45.55	46.44	47.41	**48.19**	63.90	**65.40**
	$R_{err}\downarrow$	22.14	21.35	19.20	**18.31**	6.37	**5.51**
	$T_{err}\downarrow$	25.23	24.71	22.81	**21.98**	7.08	**6.68**

Table 2. Quantitative comparison of articulated object pose estimation on HOI4D [47]. Our method and dataset could both improve the estimation performance on different categories and metrics.

Category	Metric	Ansch	ReArtNocs	CAPTRA	Ours
BMVC Laptop	$5^{\circ}5$cm$\uparrow$	1.45	0.75	4.02	**4.70**
	mIoU$\uparrow$	54.32	54.93	60.25	**61.22**
	$R_{err}\downarrow$	26.72	24.04	19.08	**17.30**
	$T_{err}\downarrow$	18.58	18.05	12.34	**11.45**
RBO Laptop	$5^{\circ}5$cm$\uparrow$	23.01	29.12	33.12	**33.83**
	mIoU$\uparrow$	49.85	51.11	51.77	**52.95**
	$R_{err}\downarrow$	11.39	11.56	10.89	**10.76**
	$T_{err}\downarrow$	9.60	8.95	7.12	**6.65**
RBO Microwave	$5^{\circ}5$cm$\uparrow$	43.02	47.51	55.31	**57.66**
	mIoU$\uparrow$	69.21	70.65	71.45	**73.05**
	$R_{err}\downarrow$	7.28	6.56	10.89	**4.69**
	$T_{err}\downarrow$	5.20	4.87	4.91	**4.70**
RBO Drawer	$5^{\circ}5$cm$\uparrow$	22.02	26.98	**33.79**	28.23
	mIoU$\uparrow$	53.93	54.83	55.01	**56.34**
	$R_{err}\downarrow$	8.27	8.34	8.01	**7.85**
	$T_{err}\downarrow$	14.93	13.45	13.10	**12.60**

Table 3. Quantitative comparison of articulated object pose estimation on BMVC [49] and RBO [48] datasets. Given to the articulation prior the model learnt, our method can achieve the best performance on most categories.

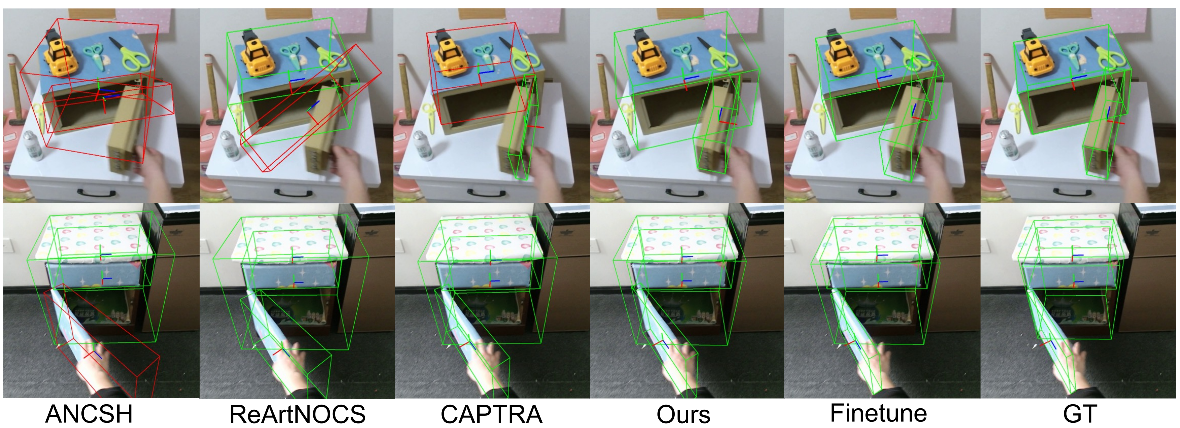


Figure 5. Qualitative comparison of object pose estimation. We use red box to indicate error larger than 5° or 5 cm. Image-based baselines fail to get an accurate pose. And our method also performs better than the tracking-based method [78].

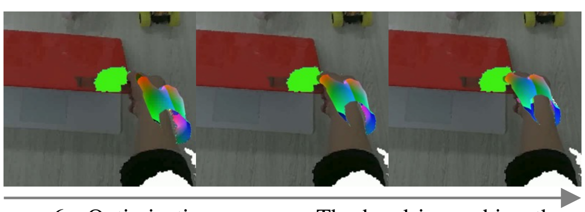


Figure 6. Optimization process. The hand is reaching the predicted contact map and getting to the correct pose.

Metric	Frankmocap	Affine	Regression	Ours
MPVPE$\downarrow$	71.6	54.1	52.4	**49.9**
MPJPE$\downarrow$	64.3	45.9	44.8	**41.9**

Table 4. Quantitative comparison of hand pose estimation. Utilizing contact map can largely reduce the estimation error compared with Frankmocap. And among all the ablative baselines, our MLP-based contact diffusion model achieves the best performance.

still outperforms it. The last two columns in Tab. 2 shows results of training from scratch and finetuning respectively. We observe finetuning performs better than training from scratch for all the metric, which demonstrated ContactArt can serve as a "prior" for pose estimation and could be used as a warm start for other datasets with smaller size.

We visualize the qualitative comparisons of articulated object poses and bounding boxes in Fig. 5. Following [45], we utilize $10^{\circ}10$cm as a threshold and use red color to indicate the results larger than this threshold and green to indicate the one within it. We observe that our method could achieve the best performance compared with the other methods. Two image-based estimation baselines fail to estimate right 3D poses when testing on challenging layouts or camera view. Our method is the most accurate and robust one even compared with tracking-based method.

We summarize the quantitative comparisons on BMVC and RBO in Tab. 3. Our method outperforms the baselines on all the categories across all the metrics. Though our model is trained on ContactArt with rich hand-object interaction but it still works well on BMVC.

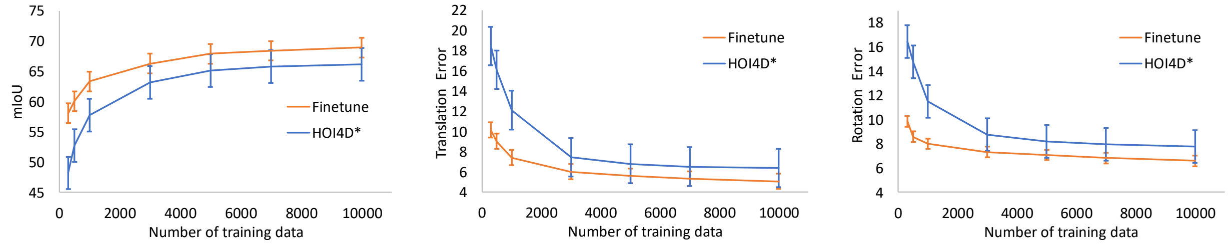


Figure 7. Ablation on the amount of training data compared between finetuning (Finetune) and training from scratch (HOI4D*). Finetuning can achieve better performance with much less data. We also report the standard deviation.

Metric	CAPTRA		Ours		HOI4D*		Finetune	
	w/o tta	w/ tta	w/o tta	w/ tta	w/o tta	w/ tta	w/o tta	w/ tta
5°5cm ↑	16.35	**16.70**	18.00	**18.65**	62.05	**62.50**	61.55	**61.95**
mIoU↑	51.50	**51.65**	52.15	**52.45**	66.25	**66.45**	**65.25**	65.20
R_{err} ↓	**18.22**	18.3	18.26	**17.73**	5.54	**5.20**	6.42	**6.24**
T_{err} ↓	19.75	**19.60**	19.10	**18.90**	7.35	**7.20**	7.75	**7.70**

Table 5. Evaluation on test time adaption. TTA can almost benefit all the methods. HOI4D* denotes train on HOI4D from scratch.

5.4. Hand Pose Estimation Comparison

For hand pose estimation, we design an effective post process specifically designed for hand-object interaction. We take an off the shelf hand estimation method Frankmocap [61] and use our test time adaption to improve the hand estimation results. We use variants of the backbone for our contact diffusion model as baselines: Regression, where we use 1D Convolution network to decode the Pointnet++ feature and directly regress the contact map; Affine, where we change the MLP architecture of diffusion model to ConcatSquash layers [26]. We report MPJPE and MPVPE in Tab. 4. All the three methods which leverage contact map to optimize hand outperforms Frankmocap. For MPVPE, our method largely decreases from 71.6 to 49.9 and decreses from 64.3 to 41.9 for MPJPE. Our contact diffusion model can successfully learn the contact prior during the interaction and anchor hand to a more reasonable pose in space. Among the three contact map estimator, our method which employs MLP-based diffusion model performs best. We also visualize the optimization process in Fig. 6.

5.5. Ablation Study

Ablation on the amount of training data. Our ContactArt could serve as a warm start before training on other datasets. To prove this, we perform ablation study training with different amount of training data on HOI4D. We compare Finetune and HOI4D* on training with 300, 500, 1k, 3k, 5k, 7k and 10k images. We compare the mIoU, rotation error and translation error in Fig. 7. In general, finetuning is always better than training from scratch with the same amount of data and has much lower standard deviation. The less data we give, the larger improvements our model can achieve. We can also observe that finetuning with 300, 1k and 3k images are better than training with 1k, 3k and 10k images respectively. We only need one-third of the data to achieve comparable performance if using ContactArt as a

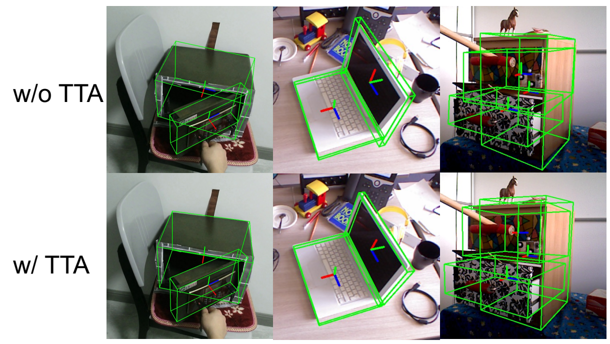


Figure 8. Comparison between our method with and without TTA. TTA can help get a more natural layout.

warm start. This is of great importance when we only have a small amount of in-the-wild data. One can first train on ContactArt and quickly adapt to new test scenarios.

Ablation on Test time adaption (TTA). Our articulation discriminator is highly scalable and can be plugged in any pose estimation method, such as CAPTRA. We select four methods, CAPTRA; our method trained on ContactArt; HOI4D*; and Finetune. We add the discriminator to each method and apply the TTA to each. We compare the one with and without TTA on laptop of HOI4D. We report the results in Tab. 8. we observe that the articulation discriminator and TTA improve all the methods. The test time adaption mechanism enables estimator to adapt to various test scene since the discriminator can learn an invariant prior of the articulation structure, specifically the layout of the bounding boxes. In Fig. 8, we visualize the results of our method with and without TTA, our articulation could learn a good prior and improve the layout naturalness of each articulated part. After TTA, the drawers are parallel and the laptop's screen and keyboard are at similar size.

6. Conclusion

In this paper, we present ContactArt, an interaction- and contact-rich and easily scalable dataset. We further propose to use a discriminator as an articulation prior to improve the articulated object pose estimation. We introduce a contact diffusion model to estimate the contact map between the hand and articulated objects, which can be utilized to optimize the hand pose estimation. Extensive experiments demonstrate the effectiveness of our ContactArt dataset, the articulation prior and the contact prior.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3

[2] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[3] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models, 2021. 3, 6

[4] Dafni Antotsiou, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[5] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, pages 640–653. Springer, 2012. 3

[6] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2021. 3

[7] Emad Barsoum, John Kender, and Zicheng Liu. HP-GAN: probabilistic 3d human motion prediction via GAN. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1418–1427. Computer Vision Foundation / IEEE Computer Society, 2018. 3

[8] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 3

[9] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393. IEEE, 2019. 3

[10] 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 *European Conference on Computer Vision*, pages 361–378. Springer, 2020. 3

[11] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 3

[12] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 3

[13] 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. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021. 3

[14] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20577–20586, 2022. 3

[15] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction, 2021. 3

[16] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 3

[17] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 3

[18] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 3

[19] Guanglong Du, Ping Zhang, Jianhua Mai, and Zeling Li. Markerless kinect-based hand tracking for robot teleoperation. *International Journal of Advanced Robotic Systems*, 9(2):36, 2012. 3

[20] Guang-Long Du, Ping Zhang, Li-Ying Yang, and Yan-Bin Su. Robot teleoperation using a vision-based manipulation method. In *2010 International Conference on Audio, Language and Image Processing*, pages 945–949. IEEE, 2010. 3

[21] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *arXiv preprint arXiv:2204.13662*, 2022. 3

[22] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 4

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[24] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2

[25] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 3

[26] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continu-

ous dynamics for scalable reversible generative models. *International Conference on Learning Representations*, 2019. 8

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 3

[28] Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José M. F. Moura. Adversarial geometry-aware human motion prediction. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV*, volume 11208 of *Lecture Notes in Computer Science*, pages 823–842. Springer, 2018. 3

[29] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018. 3

[30] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010. 3

[31] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 3

[32] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170. IEEE, 2020. 3

[33] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 3

[34] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2, 3

[35] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 2, 6

[36] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021. 3

[37] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 3

[38] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 3

[39] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. 3

[40] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 3

[41] Jonathan Kofman, Siddharth Verma, and Xianghai Wu. Robot-manipulator teleoperation by markerless vision-based hand-arm tracking. *International Journal of Optomechatronics*, 1(3):331–357, 2007. 3

[42] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3980–3987. IEEE, 2019. 3

[43] Vineet Kosaraju, Amir Sadeghian, Roberto Martín-Martín, Ian Reid, Hamid Rezatofighi, and Silvio Savarese. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[44] Shuang Li, Xiaojian Ma, Hongzhuo Liang, Michael Görner, Philipp Ruppel, Bin Fang, Fuchun Sun, and Jianwei Zhang. Vision-based teleoperation of shadow dexterous hand using end-to-end deep neural network. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 416–422. IEEE, 2019. 3

[45] Xiaolong Li, He Wang, Li Yi, Leonidas Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 4, 6, 7

[46] Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31:1072–1083, 2022. 1, 2, 4, 6

[47] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 3, 4, 6, 7

[48] Roberto Martín-Martín, Clemens Eppner, and Oliver Brock. The rbo dataset of articulated objects and interactions, 2018. 2, 4, 6, 7

[49] Frank Michel, Alexander Krull, Eric Brachmann, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Pose estimation of kinematic chain instances via object co-

ordinate regression. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015. 2, 4, 6, 7

[50] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding, 2018. 4

[51] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13001–13011, 2021. 2

[52] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 3

[53] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1898–1912, 2019. 3

[54] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *2011 International Conference on Computer Vision*, pages 2088–2095. IEEE, 2011. 3

[55] Paschalis Panteleris, Nikolaos Kyriazis, and Antonis A Argyros. 3d tracking of human hands in interaction with unknown objects. In *BMVC*, pages 123–1, 2015. 3

[56] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4, 5

[57] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *arXiv preprint arXiv:2204.12490*, 2022. 3

[58] Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, pages 570–587. Springer, 2022. 3

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[60] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 2, 4, 6

[61] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 6, 8

[62] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019. 3

[63] Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2021. 3

[64] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 3

[65] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 3

[66] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: learning a robotic hand imitator by watching humans on youtube. *arXiv preprint arXiv:2202.10448*, 2022. 3

[67] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 2, 3

[68] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[69] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. 2

[70] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[71] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[72] 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 *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 3

[73] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020. 3

[74] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 3

[75] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object

coordinate space for category-level 6d object pose and size estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 4

[76] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 3

[77] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 2

[78] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J. Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 13209–13218, October 2021. 1, 2, 4, 6, 7

[79] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion models for implicit image segmentation ensembles, 2021. 3

[80] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020. 3

[81] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 3

[82] Vicky Zeng, Tabitha Edith Lee, Jacky Liang, and Oliver Kroemer. Visual identification of articulated object parts. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2443–2450. IEEE, 2021. 2

[83] Hao Zhang, Zi-Hao Bo, Jun-Hai Yong, and Feng Xu. Interactionfusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019. 3

[84] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. Manipnet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics (ToG)*, 40(4):1–14, 2021. 3

[85] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 40(3):1–12, 2021. 3

[86] Xiaoshuai Zhang, Rui Chen, Fanbo Xiang, Yuzhe Qin, Jiayuan Gu, Zhan Ling, Minghua Liu, Peiyu Zeng, Songfang Han, Zhiao Huang, et al. Close the visual domain gap by physics-grounded active stereovision depth sensor simulation. *arXiv preprint arXiv:2201.11924*, 2022. 4

[87] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 3

[88] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2018. 4

[89] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021. 3