Hand-Object Contact Consistency Reasoning for Human Grasps Generation

Hanwen Jiang[*]   Shaowei Liu[*]   Jiashun Wang   Xiaolong Wang
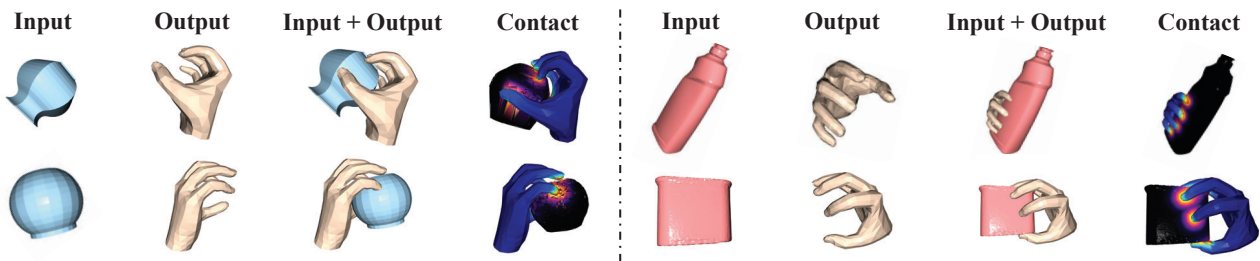UC San Diego

Figure 1: Generated human grasp on in-domain and out-of-domain objects. Object and hand contact maps are shown in the last column. The brighter the region is, the higher the contact values between the hand and object are. Best viewed in color.

Abstract

While predicting robot grasps with parallel jaw grippers have been well studied and widely applied in robot manipulation tasks, the study on natural human grasp generation with a multi-finger hand remains a very challenging problem. In this paper, we propose to generate human grasps given a 3D object in the world. Our key observation is that it is crucial to model the consistency between the hand contact points and object contact regions. That is, we encourage the prior hand contact points to be close to the object surface and the object common contact regions to be touched by the hand at the same time. Based on the hand-object contact consistency, we design novel objectives in training the human grasp generation model and also a new self-supervised task which allows the grasp generation network to be adjusted even during test time. Our experiments show significant improvement in human grasp generation over state-of-the-art approaches by a large margin. More interestingly, by optimizing the model during test time with the self-supervised task, it helps achieve larger gain on unseen and out-of-domain objects. Project page: https://hwjiang1510.github.io/GraspTTA/.

1. Introduction

Capturing hand-object interactions has been an active field of study [48, 20, 30, 13, 2, 4, 47, 42, 57] and it has wide applications in virtual reality [21, 55], human-computer interaction [51] and imitation learning in robotics [59, 49, 39]. In this paper, we study the interactions via generation:

As shown in Fig. 1, given only a 3D object in the world coordinate, we generate the 3D human hand for grasping it. Unlike predicting robot grasps with parallel jaw grippers [33, 56, 60, 5], predicting human grasps is substantially more difficult because: (i) Human hands have a lot more degrees of freedom, which leads to much more complex contact; (ii) The generated grasp needs to be not only physically plausible but also presented in a natural way, consistent with how objects are usually grasped.

To synthesize physically plausible and natural grasp poses, recent works propose to use generative models [9, 25, 47] supervised by large-scale datasets [20, 19, 16] with grasp annotations and contact analysis on hands. Specifically, the large-scale dataset allows the model to generate realistic human grasps and the contact analysis encourages the hand contact points to be close with the object but without interpenetration. While these methods put a lot of efforts into modeling the hand and its contact points, they ignore that the object itself also has more possible contact regions that need to be reached (see contact map in Fig. 1). In fact, recent work has studied the common contact regions on objects and trained neural networks to directly predict the contact map from the 3D object model [2, 4].

In this paper, we argue that it is critical for the hand contact points and object contact regions to reach mutual agreement and consistency for grasp generation. To achieve this, we propose to unify two separate models for both the hand grasp synthesis and object contact map estimation. We show that the consistency constraint between hand contact points and object contact map is not only useful for optimizing better grasps during training time by designing new

---

[*]Equal contribution.

losses, but also provides a self-supervised task to adjust the grasp when testing on a novel object. We introduce the two components as follows.

First, we train a Conditional Variational Auto-Encoder [44] (CVAE) based network which takes the 3D object point clouds as inputs and predicts the hand grasp parameterized by a MANO model [41], namely GraspCVAE. During training the GraspCVAE, we design two novel losses with one encouraging the hand to touch the object surface and another forcing the object contact regions touched by the ground truth hand close to the predicted hand. With these two consistent losses, we observe more realistic and physically plausible grasps.

Second, given the hand grasp pose and object point clouds as inputs, we train another network that predicts the contact map on the object. We name this model the ContactNet. The key role of the ContactNet is to provide supervision to finetune GraspCVAE during test time when no ground truth is available. We design a self-supervised consistency task, which requires the hand contact points produced by the GraspCVAE to be consistent and overlapped with the object contact map predicted by the ContactNet. We use this self-supervised task to perform test-time adaptation which finetunes the GraspCVAE to generate a better human grasp. This adaptation approach can be applied on each single test instance. We emphasize that this procedure does not require any extra outside supervision and it can flexibly adapt to different inputs by resuming to the model before adaptation.

We evaluate our approach on multiple datasets include Obman [20], HO-3D [19] and FPHA [16] datasets. We show that by utilizing the novel objectives based on the contact consistency constraints in training time, we achieve significant improvements on human grasps generation against state-of-the-art approaches. More interestingly, by optimizing with the proposed self-supervised task during test time, it generalizes and adapts our model to unseen and out-of-domain objects, leading to the large performance gain.

Our contributions of this paper include: (i) Novel hand-object contact consistency constraints for learning human grasp generation; (ii) A new self-supervised task based on the consistency constraints which allows the generation model to be adjusted even during test time; (iii) Significant improvement on grasp generation for both in-domain and out-of-domain objects.

2. Related Works

Hand-object interaction. Modeling and analysing hand-object interaction is an active field of study with two main paradigms: estimating hand-object poses during interaction [16, 48, 20, 13, 36, 40, 22, 30, 18] and studying from multi-modal hand-object interaction representations [17, 37, 2, 4, 47, 46, 14]. To perform hand-object pose estimation, Tekin et al. [48] proposed a 3D detection

framework, where the hand-object poses are predicted by two output grids without explicit interaction between them. Grady et al. [18] performed optimization on estimated hand meshes with contact prediction to get physically plausible hand pose estimation.

Another paradigm of study is to analyze the forces on hand and contact regions on objects from the multi-modal data. For example, Sundaram et al. [46] introduced a scalable tactile glove, and utilized the touching information for object classification, while Glauser et al. [17] leveraged it for a more difficult hand pose estimation task. Instead of using the tactile sensors on hand, Brahmbhatt et al. [2, 4] proposed to use thermal cameras to capture object contact maps, which reflects the object common contact regions after grasping. Inspired by this work, Taheri et al. [47] further built a GRAB dataset, which not only captures the contact map from hand, but also takes the whole human body into consideration. This line of research motivates us to go beyond modeling hand-centric grasp generation, and explore the object-centric contact map by designing an object-centric loss to encourage the common contact regions on object to be touched by the hand.

Grasp generation. Generating human grasp is very challenging due to the higher degree-of-freedom of the human hand [25, 9, 47, 3, 22]. To generate a realistic grasp, Karunratanakul et al. [25] proposed an implicit representation for modeling the joint distribution of hand-object shape. Instead of implicit representation, our work is more related to work by Brahmbhatt et al. [3] which made use of the object contact maps to filter multiple generated grasps from GraspIt! [32]. However, the contact maps are taken as a constraint rather than a learning target in this grasp generation framework. In our work, we leverage the consistency between hand-object contact regions as training targets with the use of object contact map. Moreover, a self-supervised task is also designed for adjusting generated grasps using the contact maps at test-time.

Affordance prediction. Predicting scene and object affordance plays an important role in visual understanding [28, 53, 6, 10, 54, 12, 26, 15, 52, 58]. For example, Corona et al. [10] proposed a novel dataset and a generative network for learning grasps of multiple on-table objects. Zhang et al. [58] proposed a novel representation of human-scene contact for generating natural scene affordance. Inspired by these works, our goal is to generate grasps by learning the object affordance with ensuring the perceptual naturalness and physical plausibility at the same time. Different from the previous works, our method allows better generalization of grasping out-of-domain objects with the help of the proposed self-supervised task on predicting the target contact.

Learning on test instances. Improving the generalization ability of neural networks is one of the most important
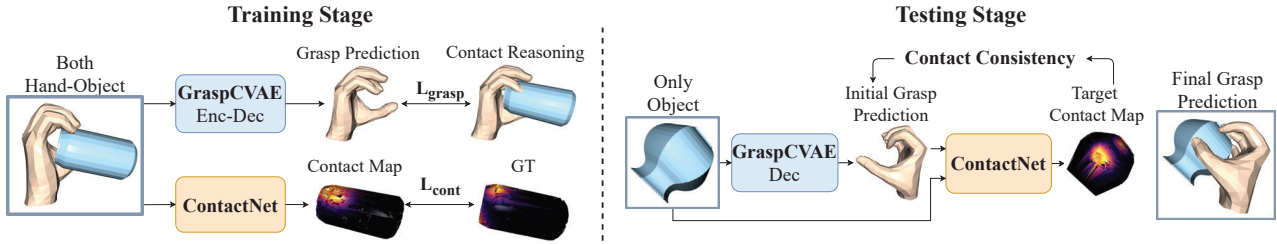
Figure 2: The *different* usage of proposed networks in training and testing. **Left**: During training, the two networks learn generating human grasps and predicting object contact map separately on ground truth data. **Right**: At test-time, the two networks are unified in a cascade manner. A initial grasp is predicted by the GraspCVAE *decoder*, and it is inputted together with the object into ContactNet for predicting a target contact map. We then leverage the contact consistency between outputs of the two networks for adjusting the initial grasp, where the target contact map serves as a self-supervision signal.

problem in machine learning [8, 31, 11, 29, 34]. Recent research has started tackling the problem by leveraging self-supervision at test-time [23, 43, 1, 45, 24, 35]. For example, Shocher *et al.* [43] proposed a self-supervised super-resolution framework where the network is only trained at test-time by up and downscaling **a single test example**. Sun *et al.* [45] extended the test-time adaptation idea to more general applications with a joint training framework of an image recognition task and a self-supervised task. At test-time, the network can be adjusted to a single test image by tuning the self-supervised objective. While this approach is intriguing, it is still unclear how the self-supervised objective can affect the main task objective. Inspired by this work, our approach also leverage self-supervision on a single instance for test-time adaptation. Different from [45], our self-supervised task is directly optimizing the main goal of generating better human grasps, which ensures the performance gain.

3. Approach

Our goal is generating hand meshes as human grasps given object point clouds as inputs. The generated hand mesh not only needs to be presented in a natural and realistic way, but it should also hold the object tightly in a physically plausible manner. We emphasize that ensuring reasonable contact between the object and synthesized hand is the key to get high-quality and stable human grasps.

To deal with this problem, we utilize both hands and object contact information and make sure they are consistent with each other, as summarized in Fig. 2. We propose two networks, a generative GraspCVAE to synthesize grasping hand mesh, and a deterministic ContactNet for modeling the contact regions on the object.

Training Stage. As shown on the left side of Fig. 2, we optimize these two networks using ground-truth supervision *separately* to learn grasp generation and predicting object contact maps. In this stage, the inputs of GraspCVAE are *both* of hand and object, and GraspCVAE learns to synthesizing grasps in the *hand reconstruction* paradigm, where *both* of the its encoder and decoder will be used. Note that this

follows the standard procedure in Conditional Variational Auto-Encoder (CVAE) [44]. To train the GraspCVAE, we propose two novel losses to ensure the hand-object contact consistency: one loss forcing the prior hand contact vertices to be close to the object surface, and another loss encouraging the object common contact regions to be touched by hand at the same time. The object and generated hand will find mutual agreement on the form of contact with the two losses during training.

Testing Stage. As shown on the right side of Fig. 2, we unify the two networks and design a self-supervised task by leveraging the consistency between their outputs. Given a test object, we first generate an initial grasp from the GraspCVAE decoder (without the encoder). Different from the training stage, the reconstruction target – grasp is not provided in testing [44]. Then, the generated grasp is forwarded together with the object to the ContactNet to predict a target contact map. Since ContactNet is trained with ground truth data, where penetration between hand-object does not exist and hand fingers are touching the object surface closely, it will model the pattern of the ideal hand-object contact. During testing, the predicted contact map from ContactNet will tend to contain the ideal contact pattern. We use the predicted contact map from the ContactNet as a target for finetuning and optimizing the grasps generated by GraspCVAE. If the grasp is predicted correctly from GraspCVAE, the object contact region from the predicted grasp should be consistent with the target object contact map. We use this consistency as a self-supervision signal to adapt grasps generated by GraspCVAE during test-time.

In the following, we will first introduce the individual framework for GraspCVAE and ContactNet, and then the test-time contact reasoning with both networks for better adaptation to the new objects.

3.1. Learning GraspCVAE

Usage and architecture The GraspCVAE is a Conditional Variational Auto-Encoder (CVAE) [44] based generative network, which uses conditional information to control generation. For GraspCVAE, the conditional information is the
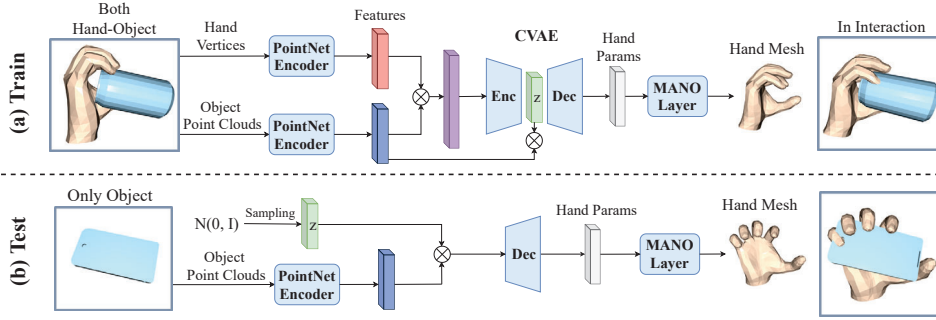
Figure 3: The architecture of GraspCVAE. (a) In training, it takes both hand-object as input to predict a hand mesh for grasping the object in a hand reconstruction manner using both of its *encoder-decoder*; (b) At test-time, its *decoder* generates grasps by conditioning only on object information as input. $\bigotimes$ means concatenation.
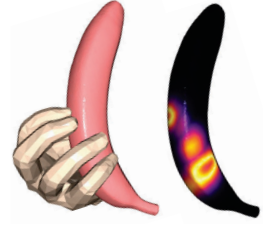

Figure 4: An example of contact map, brighter regions have larger scores. Because the MANO model does not have soft tissue, the underformable finger-tips usually penetrate into object surface slightly.

object. We follow [27, 44] to use the GraspCVAE: In training, *both* of encoder and decoder of GraspCVAE is used to learn the grasp generation task in a hand reconstruction manner by taking in *both* of hand-object as input; At test-time, *only* its decoder is used to generate human grasp of a object with *only* the 3D object as the input (without using the grasp for input). The network architecture is shown in Fig. 3.

During training, as shown in top row of Fig. 3, given two point clouds of the hand $\mathcal{P}^h \in \mathbb{R}^{778 \times 3}$ and the object $\mathcal{P}^o \in \mathbb{R}^{N \times 3}$ (where N is the number of points) as inputs, we use two separate PointNets [38] to extract their features respectively, denoted as $\mathcal{F}^h, \mathcal{F}^o \in \mathbb{R}^{1024}$. These two features are then concatenated as $\mathcal{F}^{ho}$ for the GraspCVAE encoder inputs. The outputs of the encoder are the mean $\mu \in \mathbb{R}^{64}$ and variance $\sigma^2 \in \mathbb{R}^{64}$ of the posterior Gaussian distribution $Q(z|\mu, \sigma^2)$ [27]. To reconstruct the hand, we first sample a latent code z from the distribution and the posterior distribution ensures the sampled latent code z is in correspondence with the input hand-object.

The decoder takes the concatenation of latent code z and the object feature $\mathcal{F}^o$ as input to reconstruct a hand mesh, which is represented by a differentiable MANO model [41]. The MANO model is parameterized by the shape parameter $\beta \in \mathbb{R}^{10}$ for person-specific hand shape, as well as the pose parameter $\theta \in \mathbb{R}^{51}$ for the joint axis-angles rotation and root joint translation. Given the predicted parameters $(\hat{\beta}, \hat{\theta})$ from the decoder, the MANO model forms a differentiable layer which outputs the shape of the hand with $\hat{\mathcal{M}} = (\hat{\mathcal{V}} \in \mathcal{R}^{778 \times 3}, \hat{F})$, where $\hat{\mathcal{V}}, \hat{F}$ denotes the mesh vertices and faces. Both the encoder and decoder in GraspCVAE are Multi-Layer Perceptrons (MLP).

During testing, as shown in the bottom row of Fig. 3, we only utilize the decoder from the GraspCVAE for inference. Given only the extracted object point cloud feature $\mathcal{F}^o$ and a latent code z randomly sampled from a Gaussian distribution as inputs, the decoder will generate the parameters for the MANO model which leads to the hand mesh output.

Given this architecture, we then introduce the training objectives as follows. We will first introduce the baseline objectives and then two novel losses which encourage the hand-object contact consistency.

Baseline The first objective for the baseline model is mesh reconstruction error, which is defined on both the vertices of the mesh as well as the parameters of the MANO model. We adopt the L_2 distance to compute the error. We denote the reconstruction loss between the predicted vertices and the ground-truths as $L_{\mathcal{V}} = ||\hat{\mathcal{V}} - \mathcal{P}^h||_2^2$. The losses on MANO parameters are defined in a similar way with L_θ and L_β. The reconstruction error can be represented as $L_{\mathcal{R}} = \lambda_{\mathcal{V}} \cdot L_{\mathcal{V}} + \lambda_\theta \cdot L_\theta + \lambda_\beta \cdot L_\beta$, where $\lambda_{\mathcal{V}}, \lambda_\theta$ and λ_β are constants balancing the losses.

Following the training of VAE [27], we define a loss enforcing the latent code distribution $Q(z|\mu, \sigma^2)$ to be close to a standard Gaussian distribution, which is achieved by maximizing the KL-Divergence as $L_{\mathcal{KLD}} = -KL(Q(z|\mu, \sigma^2)||\mathcal{N}(0, I))$.

We also encourage the grasp to be physically plausible, which means the object and hand should not penetrate into each other. We denote the object point subset that is inside the hand as $\mathcal{P}^o_{in}$, then the penetration loss is defined as minimizing their distances to their closest hand vertices $L_{penetr} = \frac{1}{|\mathcal{P}^o_{in}|} \sum_{p \in \mathcal{P}^o_{in}} \min_i ||p - \hat{V}_i||_2^2$. In a short summary, the loss for training the baseline is:

$$L_{base} = L_{\mathcal{R}} + \lambda_{\mathcal{KLD}} \cdot L_{\mathcal{KLD}} + \lambda_p \cdot L_{penetr} , \quad (1)$$

where $\lambda_{\mathcal{KLD}}$ and λ_p are constants balancing the losses.

Reasoning Contact in Training There are two potential challenges in the baseline framework: First, the losses in the baseline model ignore physical contact between the hand-object, which cannot ensure the stability of the grasp; Second, grasp generation is multi-modal and the ground-truth hand pose is not the only answer. To tackle these challenges, we design two novel losses from both the hand and the object aspects to reason plausible hand-object contact and find the mutual agreement between them.
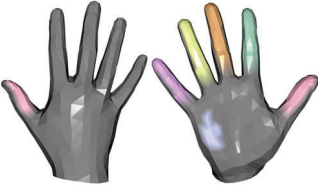
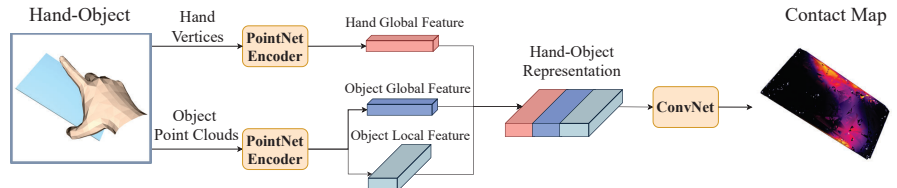Figure 5: Six hand prior contact regions are shown in color.

Figure 6: Architecture of ContactNet. It extracts local per-point feature of object point cloud, and concatenate it with global hand-object feature for predicting contact map.

Object-centric Loss. From the object perspective, there are regions that are often contacted by human hand. We encourage the human hand to get close to these regions using the object-centric loss. Specifically, from the ground-truth hand-object interaction, we can derive the object contact map $\Omega \in \mathbb{R}^N$ by normalizing the distance $\mathbf{D}(\mathcal{P}^o)$ between all object points and their nearest hand prior vertice with function $f(\cdot)$, where $f(\mathbf{D}(\mathcal{P}^o)) = 1 - 2 \cdot (\text{Sigmoid}(2\mathbf{D}(\mathcal{P}^o)) - 0.5)$. An example is shown in Fig. 4. The distances are in centermeter and contact map scores are in $[0,1]$. This normalization helps the network focus on object regions close to the hand. Then we force the object contact map $\hat{\Omega}$ computed from the generated hand to be close to the ground truth Ω, using loss

$$L_{\mathcal{O}} = ||\hat{\Omega} - \Omega||_2^2, \ \Omega = f(\mathbf{D}(\mathcal{P}^o)). \tag{2}$$

Hand-centric Loss. We define the prior hand contact vertices $\mathcal{V}^p$ as shown in Fig. 5, motivated by [20, 2]. Given the predicted locations of the hand contact vertices, we then take the object points nearby as possible points to contact. Specifically, for each object point $\mathcal{P}_i^o$, we compute the distance $\mathbf{D}(\mathcal{P}_i^o) = \min_j ||\mathcal{V}_j^p - \mathcal{P}_i^o||_2^2$, and if it is smaller than a threshold, we take it as the possible contact point on the object. Our hand-centric objective is to push the hand contact vertices close to the object as,

$$L_{\mathcal{H}} = \sum_i \mathbf{D}(\mathcal{P}_i^o), \ for \ all \ \mathbf{D}(\mathcal{P}_i^o) \leq \mathcal{T} \tag{3}$$

for all the possible contact points on the object, where $\mathcal{T} = 1 \, cm$ is the threshold. The final loss combining the two novel losses above is,

$$L_{grasp} = L_{baseline} + \lambda_{\mathcal{H}} \cdot L_{\mathcal{H}} + \lambda_{\mathcal{O}} \cdot L_{\mathcal{O}}, \tag{4}$$

where $\lambda_{\mathcal{H}}$ and $\lambda_{\mathcal{O}}$ are constants balancing the losses. Intuitively, the $L_{\mathcal{O}}$ generally answers the question *Where to grasp?* and does not specify which hand part should be close to the object contact regions. And $L_{\mathcal{H}}$ is used to find the answer of *Which finger should contact?* dynamically. During training, with the two proposed losses, the hand contact points and object contact region will reach mutual agreement and be consistent to each other for generating stable grasps.

3.2. Learning ContactNet

We propose another network, the ContactNet, to model the contact information between the hand-object as shown

in Fig. 6. The inputs are hand and object point cloud, and the output is the object contact map denoted as $\Omega^c \in \mathbb{R}^N$ for N object points. We use two PointNet encoders to extract hand and object feature maps. Since we need to predict the contact score for each point (Ω_i^c should be the score for $\mathcal{P}_i^o$), we utilize the per-point object local feature $\mathcal{F}^s \in \mathbb{R}^{N \times 64}$ of the PointNet encoder to ensure this correspondence. We also make use of hand and object global feature by first summing them then duplicate N times and concatenate it with the object local feature for a to dimension $\mathbb{R}^{N \times 1024}$ feature maps. Given these features, we apply four layers of 1-D convolution on top to regress the object contact map activated by the sigmoid function. The loss for training is the L_2 distance between the predicted contact map Ω^c and ground-truth Ω as, $L_{cont} = ||\Omega^c - \Omega||_2^2$.

During training time, the inputs for the ContactNet are directly obtained from the ground-truths.

3.3. Contact Reasoning for Test-Time Adaptation

During testing, we unify the GraspCVAE and ContactNet in a cascade manner as shown on the right side of Fig. 2. Given the object point clouds as inputs, the GraspCVAE will first generate a hand mesh $\hat{\mathcal{M}}$ as the initial grasp. We compute its object contact map $\Omega_{\hat{\mathcal{M}}}$ correspondingly. Taking both the predicted hand mesh and object as inputs, the ContactNet will predict another contact map Ω^c. If the grasp is predicted correctly, the two contact map $\Omega_{\hat{\mathcal{M}}}$ and Ω^c should be consistent. Based on this observation, we define a self-supervised consistency loss as $L_{refine} = ||\Omega_{\hat{\mathcal{M}}} - \Omega^c||_2^2$ for fine-tuning the GraspCVAE. Besides this consistency loss, we also incorporate the hand-centric loss $L_{\mathcal{H}}$ and penetration loss L_{penetr} to ensure the grasp is physically plausible. We apply the joint optimization with all three losses on a *single* test example as,

$$L_{TTA} = L_{refine} + \lambda_{\mathcal{H}} \cdot L_{\mathcal{H}} + \lambda_p \cdot L_{penetr}. \tag{5}$$

We use this loss to update the GraspCVAE *decoder*, and freeze other parts of the two networks.

4. Experiment

We show qualitative results of generated grasps from our methods, and compare the qualitative performance with other methods in Sec. 4.4. Then, we give ablation studies on the
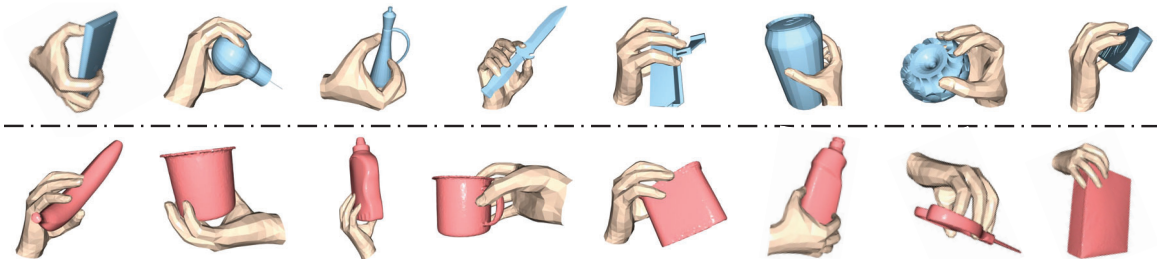
Figure 7: Visualization of generated grasps on in-domain objects from Obman dataset [20], and out-of-domain objects from HO-3D dataset [19]. **See Supplementary** for more visualization on qualitative results and diversity of generated grasps.

		Obman			HO-3D			FPHA		
		GT	GF [25]	Ours	GT	GF [25]	Ours	GT	GF [25]	Ours
Penetration	Depth $(cm)\downarrow$	**0.01**	0.56	0.46	2.94	1.46	**1.05**	**1.17**	2.37	1.58
	Volume $(cm^3)\downarrow$	**1.70**	6.05	5.12	6.08	14.90	**4.58**	**5.02**	21.9	6.37
Grasp Displace.	Mean $(cm)\downarrow$	1.66	2.07	**1.52**	4.31	3.45	**3.21**	5.54	4.62	**2.55**
	Variance $(cm)\downarrow$	± 2.44	± 2.81	$\pm\mathbf{2.29}$	± 4.42	± 3.92	$\pm\mathbf{3.79}$	± 4.38	± 4.48	$\pm\mathbf{2.22}$
Perceptual Score	$\{1,...,5\}\uparrow$	3.24	3.02	**3.54**	3.18	3.29	**3.50**	3.49	3.33	**3.57**
Contact	Ratio $(\%)\uparrow$	**100**	89.40	99.97	91.60	90.10	**99.61**	91.40	97.00	**100**

Table 1: Results on Obman [20], HO-3D [19] and FPHA datasets [16] compared with ground truth (GT) and GF [25]. Best ones are in bold.

effectiveness of proposed novel losses during training and different Test-Time Adaptation (TTA) paradigms in Sec. 4.5.

4.1. Implementation Details

We sample $N = 3000$ points on the object mesh as the input object point clouds. In training, we use Adam optimizer and $LR = 1e-4$ with 100 epochs, where the LR is reduced half when model trained $30, 60, 80, 90$ epochs. Batch size is 128. The loss weights are $\lambda_\beta = 0.1$, $\lambda_\theta = 0.1$, $\lambda_p = 5$, $\lambda_\mathcal{H} = 1500$ and $\lambda_\mathcal{O} = 100$. For Test-Time Adaptation, we use optimizer SGD with Momentum 0.8, $LR = 6.25\times10^{-6}$, same as last epochs in training. For each sample, we use batch augmentations with batch size 32. The loss weights are $\lambda_p = 5$, $\lambda_\mathcal{H} = 1$ and $\lambda_\mathcal{O} = 5$.

4.2. Datasets

Obman Dataset [20] is a synthetic dataset includes hand-object mesh pairs. The hands are generated by a non-learning based method GraspIt! [32]. 2772 object meshes covering 8 classes of everyday objects from ShapeNet [7] dataset are included. The model trained on this dataset will benefit from the diversified object models and grasp types. We train the two networks on this dataset as the initial model.

HO-3D and FPHA Dataset [16, 19] are used for evaluating the generalization ability of our proposed framework. Because only a dozen of objects are included in these two datasets, they are not suitable for training the model. These two datasets collect video sequences annotated with object-hand poses. We use the same split and data filtering of the two datasets with [25].

4.3. Evaluation Metrics

Penetration is measured by penetration **depth** and **volume** between objects and generated grasps following [20].

Grasp displacement is used to measure the stability of the grasp. We put the object and generated grasp in a simulator following [50, 20]. In general, the simulator calculates the motion of the object under the grasp. Specifically, the simulator calculates forces on fingertips, which are has a positive correlation with the penetration volume on fingertips. Then, it applies the calculated forces to hold the object against its gravity. The grasp stability is measured by the displacement of the object's center of mass during a period in the simulation. In this period, the pose and location of hand is fixed. We measure the mean and variance of the simulation displacement for all test samples. Examples with smaller simulation displacement have better grasp stability.

Perceptual score is utilized for evaluating the naturalness of generated grasps. We perform the perceptual study following [25] with Amazon Mechanical Turk.

Hand-object contact metrics are used for analyzing contact between hand-object. We calculate the sample-level hand-object contact ratio, individual object and hand contact points ratio, and the number of hand fingers contacting the object. We classify the contact status of a point by judging whether its distance to its nearest neighbor in the other point cloud is smaller than $0.5 cm$. We also calculate the object contact map score, as $s = 100 \cdot \frac{\sum \Omega}{N} \in [0, 100]$, which reflects the coverage area of grasps. Generally, larger contact areas can imply a better grasp, but this is not *strictly* correct.

4.4. Grasp Generation Performance

Qualitative results. We first visualize generated grasps for different objects. Fig. 7 shows that our framework is able to generate stable grasps with natural hand poses on both in-domain and out-of-domain objects.

Quantitative results. The evaluation results on the three datasets are shown in Table 1. We train the models on the

Loss	Penetration ↓		Grasp Displace. ↓	Contact ↑				
	Depth	Volume	Mean ± Variance	Ratio (%)	Obj Verts (%)	Hand Verts (%)	# Finger	CMap Score
$\mathcal{L}_{base}$	**0.40**	**3.00**	3.51 ± 3.70	97.69	3.58	8.20	2.97	6.42
$+\mathcal{L}_H(gt)$	0.64	4.90	1.97 ± 2.83	99.89	6.07	11.32	3.67	10.20
$+\mathcal{L}_\mathcal{H}$	0.48	4.85	1.72 ± 2.44	99.90	6.98	12.17	**3.90**	11.11
$+\mathcal{L}_\mathcal{H}+\mathcal{L}_\mathcal{O}(dist)$	0.47	4.72	1.77 ± 2.65	99.83	6.80	11.92	3.82	10.89
$+\mathcal{L}_\mathcal{H}+\mathcal{L}_\mathcal{O}$	0.48	4.92	**1.63 ± 2.43**	**99.94**	**7.16**	**12.17**	3.87	**11.24**

Table 2: Ablation study for proposed losses of GraspCVAE on the Obman [20]. To verify the effectiveness of each loss, we also compared each of them with a modified version shown in gray additionaly.

	Model		
	object-only	h-o global	h-o global-local
Error	0.161	0.148	**0.090**

Table 3: Ablation study of different ContactNet designs. The error is average contact map score absolute error of all object points between predictions and ground truth.

Obman training set, and test on Obman testset. We also test the model trained from the Obman training set extensively on **HO-3D** and **FPHA** to demonstrate the generalization ability of our method. All results are evaluated after Test-Time Adaptation (TTA). The objects in the Obman test set may overlap with its training set, while objects (with different poses) from HO-3D and FPHA are never seen in training.

On all of the three datasets, our framework shows significant improvement over the state-of-the-art approach [25] in *both* of physical plausibility, grasp stability and perceptual score. And results on HO-3D and FPHA dataset imply that our model has a much stronger cross-domain generalization ability. For instance, [25] achieves reasonably good stability on HO-3D and FPHA but suffers from huge penetration (they are correlated). However, our model performs much better on both of the two metrics with a great balance. Moreover, the perceptual scores of our framework on the three datasets are similar: 3.54 for in domain objects in Obman, 3.50 and 3.57 for out-of-domain data in HO-3D and FPHA. This shows the quality of generated grasps on out-of-domain objects are close to the in-domain objects. Besides, our results are close to or even outperform the ground truth, especially for the stability and perceptual score.

4.5. Ablation Study

We first perform ablation studies on Obman dataset [20] for evaluating the two proposed losses $\mathcal{L}_\mathcal{H}$ and $\mathcal{L}_\mathcal{O}$. We then analyse different designs of ContactNet. Finally, we compare different Test-Time Adaptation (TTA) paradigms on out-of-distribution HO-3D and FPHA [16, 19] dataset.

4.5.1 GraspCVAE Training Objectives

The results are shown in Table 2. With the hand-centric loss $\mathcal{L}_\mathcal{H}$, the simulation displacement decreases and contact metrics grows significantly while the penetration grows slightly. After adding the object-centric loss $\mathcal{L}_\mathcal{O}$, only object-related metrics, e.g. contact object vertices ratio and contact map score, and stability grows (displacement decreases). This implies that with $\mathcal{L}_\mathcal{H}$, the $\mathcal{L}_\mathcal{O}$ acts as a regularizer on the object contact regions to improves the grasp stability, which matches the design of this loss function.

We also verify the effectiveness of two losses by comparing each with a modified version. First, we can force the fingers to touch the ground truth contact regions rather than finding them dynamically with $\mathcal{L}_\mathcal{H}$. This loss is denoted as $\mathcal{L}_\mathcal{H}(gt)$. Experiments demonstrate that $\mathcal{L}_\mathcal{H}$ is better in all metrics than $\mathcal{L}_\mathcal{H}(gt)$. This implies that fitting the ground truth in the multi-solution grasp generation task may not be optimal. Second, in the loss $\mathcal{L}_\mathcal{O}$, we verify the effectiveness of using the contact map as the representation of hand-object distance. We experiment with directly minimize the residual between predicted and ground truth object-hand distances $\hat{\mathbf{D}}$ and $\mathbf{D}$ without normalizing them into contact maps. We call this loss as $\mathcal{L}_\mathcal{O}(dist)$. Experiments show that with $\mathcal{L}_\mathcal{O}(dist)$, the performance even degenerates. The reason is that the $\mathcal{L}_\mathcal{O}(dist)$ is contributed almost by hand-object point pairs with large distance, while $\mathcal{L}_\mathcal{O}$ focus more on hand vertices close to object surface with the help of normalization.

4.5.2 ContactNet Designs

We compare three different kinds of ContactNet designs, as shown in Table 3. The first model (object-only) takes solely the object as input, while the second (h-o global) and third (h-o global-local) model take in both of hand and object. The difference between the latter two is to experiment whether using object local features helps predict the contact map by maintaining point permutation information.

Without the hand as an input, predicting object contact map is a very difficult one-to-many mapping. Considering only a small part of object points are in contact, the 0.161 absolute error is actually huge. Experiments also show that without object local features, the gain from adding the hand as one of the input is trivial. With object local features, the error reduces 0.07 as a significant improvement of 50%.

4.5.3 Test-Time Adaptation (TTA) for Generalization

We compare four different TTA paradigms:

- TTA (offline): Learning-based TTA same as illustrated in Sec. 3.3, and network parameters are re-initialized before adapting each sample;
- TTA-optm (offline): Optimization-based TTA, where the MANO parameters are directly optimized;

		Penetration ↓		Grasp Displace. ↓		Contact ↑			
		Depth	Volume	Mean ± Variance	Ratio (%)	Obj Verts (%)	Hand Verts (%)	# Finger	CMap Score
HO-3D [19]	w/o TTA	**0.94**	**4.21**	4.98 ± 4.48	86.63	3.41	8.78	3.11	5.65
	TTA	1.09	4.88	3.80 ± 4.20	92.31	4.37	10.83	3.58	7.13
	TTA-optm	1.07	4.59	4.14 ± 4.31	91.45	4.32	10.97	3.68	6.78
	TTA-noise	1.12	4.98	4.22 ± 4.34	91.17	4.14	10.40	3.32	6.81
	TTA-online	1.05	4.58	**3.21 ± 3.79**	**99.61**	**4.66**	**11.55**	**3.88**	**7.80**
FPHA [16]	w/o TTA	**6.19**	**1.56**	2.93 ± 2.70	100	4.71	13.78	4.47	7.67
	TTA	6.37	1.58	**2.55 ± 2.22**	100	4.64	13.95	4.56	7.67
	TTA-online	6.31	1.69	2.77 ± 2.47	**100**	**4.83**	**14.44**	**4.73**	**7.83**

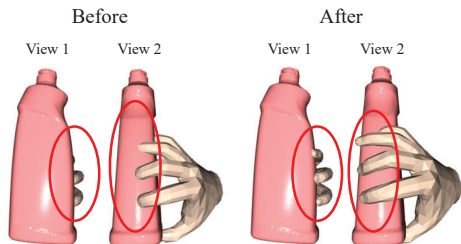Table 4: Results of different Test-Time Adaptation (TTA) methods on out-of-domain HO-3D and FPHA dataset [19, 16].

Figure 8: Visualization of grasp before and after TTA. The penetration decreases on fingertips.
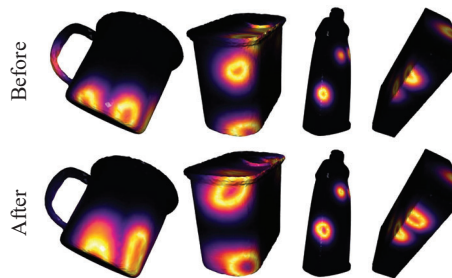

Figure 9: After TTA, object contact regions become larger, where the yellow circles on contact maps reflect the intersection rings between hand-object.

- TTA-noise (offline): Learning-based TTA. When training ContactNet, the hand parameters are injected with random noise. The method is used to compare different methods to obtain the target contact map;
- TTA-online: Learning-based TTA, the network parameters are re-initialized only *once* for each video sequence.

As shown in Table 4, on the two datasets, all TTA methods can improve the results. There are three comparisons between the different methods. First, on the HO-3D dataset, the TTA and TTA-optm achieve comparable results because they are both offline methods using the same objective function. The results of the learning-based TTA are slightly better, which can be explained by that the network parameters serve as a prior and make the adaptation more steady. Second, training with injected noise, we expect the Contact-Net can learn to predict ideal contact maps as target in TTA by "correting" the noise. However, the results deteriorate compared with the one trained on perfect ground truth data. This can be explained by: (i) Injecting noise hurts learning contact maps; (ii) It is hard to match the injected noise with the noise pattern in initially predicted grasps. Third, the online version of TTA is stronger than offline versions. Inspired by [45, 23, 35], for the online TTA, the target of the TTA can be optimized continually with the help of network parameters, and the model can fit the test distribution better. With online updating, the stability grows and penetration depth decreases simultaneously, indicating that the network leans better hand-object contact. A huge improvement in contact ratio also verifies the point. The improvement of online TTA on the FPHA dataset is not as big as on HO-3D dataset because the average video sequence length is $\frac{1}{20}$ of HO-3D, so the learning target cannot be optimized continually.

To show the effectiveness of TTA for improving both the naturalness and stability of generated grasps, we further visualize the grasps and object contact maps before and after TTA. As shown Fig. 8, after TTA, the hand penetration decreases with fingers closely contacting object surface. In Fig. 9, the object contact regions become larger, which indicates the grasps are more stable.

5. Conclusion

In this work, we propose a framework for generating human grasps given an object. To get natural and stable grasps, We reason the consistency of contact information between object and generated hand from two aspects: First, we design two novel training targets from the view of hand and object respectively, which helps them to find a mutual agreement on the form of contact. Second, we design two networks for grasp generation and predicting contact map respectively. We leverage the consistency between outputs of the two networks for designing a self-supervised task, which can be used at test-time for adapting generated grasps on novel objects. With the proposed method, we not only observe more natural and stable generated grasps, but also a strong generalization capability on cross-domain test inputs.

References

[1] David Bau, Hendrik Strobelt, W. Peebles, J. Wulff, B. Zhou, Jun-Yan Zhu, and A. Torralba. Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics (TOG)*, 38:1 – 11, 2019. 3

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, pages 8709–8719, 2019. 1, 2, 5

[3] Samarth Brahmbhatt, A. Handa, J. Hays, and D. Fox. Contact-grasp: Functional multi-finger grasp synthesis from contact. *IROS*, pages 2386–2393, 2019. 2

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. *arXiv preprint arXiv:2007.09545*, 2020. 1, 2

[5] Hanwen Cao, Hao-Shu Fang, Wenhai Liu, and Cewu Lu. Suctionnet-1billion: A large-scale benchmark for suction grasping. *arXiv preprint arXiv: 2103.12311*, 2021. 1

[6] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and J. Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 2

[7] Angel X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Qixing Huang, Zimo Li, S. Savarese, M. Savva, Shuran Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015. 6

[8] Minmin Chen, Kilian Q. Weinberger, and John Blitzer. Co-training for domain adaptation. In *NIPS*, 2011. 3

[9] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, pages 5031–5041, 2020. 1, 2

[10] Enric Corona, A. Pumarola, G. Alenyà, F. Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *CVPR*, pages 5030–5040, 2020. 2

[11] G. Csurka. Domain adaptation for visual applications: A comprehensive survey. *ArXiv*, abs/1702.05374, 2017. 3

[12] Thanh-Toan Do, Anh Nguyen, Ian Reid, D. Caldwell, and N. Tsagarakis. Affordancenet: An end-to-end deep learning approach for object affordance detection. *ICRA*, pages 1–5, 2018. 2

[13] Bardia Doosti, Shujon Naha, M. Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. *CVPR*, pages 6607–6616, 2020. 1, 2

[14] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. *CVPR*, pages 221–230, 2020. 2

[15] Kuan Fang, Te-Lin Wu, D. Yang, S. Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. *CVPR*, pages 2139–2147, 2018. 2

[16] 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. *CVPR*, pages 409–419, 2018. 1, 2, 6, 7, 8

[17] O. Glauser, Shihao Wu, Daniele Panozzo, Otmar Hilliges, and O. Sorkine-Hornung. Interactive hand pose estimation using a stretch-sensing soft glove. *ACM Transactions on Graphics (TOG)*, 38:1 – 15, 2019. 2

[18] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and C. Kemp. Contactopt: Optimizing contact to improve grasps. *ArXiv*, abs/2104.07267, 2021. 2

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and objects poses. *CVPR*, 2019. 1, 2, 6, 7, 8

[20] 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 *CVPR*, pages 11807–11816, 2019. 1, 2, 5, 6, 7

[21] Markus Höll, Markus Oberweger, C. Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. *IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 175–182, 2018. 1

[22] S. Hussain, Liu Liu, Wenqiang Xu, and Cewu Lu. Fpha-afford: A domain-specific benchmark dataset for occluded object affordance estimation in human-object-robot interaction. In *ICIP*, 2020. 2

[23] Vidit Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. *CVPR*, pages 577–584, 2011. 3, 8

[24] Zdenek Kalal. Tracking-learning-detection. *IEEE TPAMI*, 34:1409–1422, 2012. 3

[25] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *3DV*, pages 333–344, 2020. 1, 2, 6, 7

[26] D. Kim and G. Sukhatme. Semantic labeling of 3d point clouds with object affordance for robot manipulation. *ICRA*, pages 5578–5584, 2014. 2

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[28] X. Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and J. Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. *CVPR*, pages 12360–12368, 2019. 2

[29] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 3

[30] 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. 1, 2

[31] Mingsheng Long, Han Zhu, J. Wang, and Michael I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *NIPS*, 2016. 3

[32] A. Miller and P. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11:110–122, 2004. 2, 6

[33] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *ICCV*, pages 2901–2910, 2019. 1

[34] Krikamol Muandet, D. Balduzzi, and B. Schölkopf. Domain generalization via invariant feature representation. *ArXiv*, abs/1301.2115, 2013. 3

[35] R. T. Mullapudi, S. Chen, Keyi Zhang, D. Ramanan, and K. Fatahalian. Online model distillation for efficient video inference. *ICCV*, pages 3572–3581, 2019. 3, 8

[36] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE TPAMI*, 2019. 2

[37] Steffen Puhlmann, Fabian Heinemann, O. Brock, and M. Maertens. A compact representation of human single-object grasping. *IROS*, pages 1954–1959, 2016. 2

[38] C. R. Qi, H. Su, Kaichun Mo, and L. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CVPR*, pages 77–85, 2017. 4

[39] Ilija Radosavovic, Xiaolong Wang, Lerrel Pinto, and J. Malik. State-only imitation learning for dexterous manipulation. *ArXiv*, abs/2004.04650, 2020. 1

[40] Grégory Rogez, J. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. *ICCV*, pages 3889–3897, 2015. 2

[41] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245, 2017. 2, 4

[42] M. Schröder and H. Ritter. Hand-object interaction detection with fully convolutional networks. *CVPRW*, pages 1236–1243, 2017. 1

[43] Assaf Shocher, N. Cohen, and M. Irani. "zero-shot" super-resolution using deep internal learning. In *CVPR*, 2018. 3

[44] Kihyuk Sohn, H. Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 2, 3, 4

[45] Yu Sun, Xiaolong Wang, Liu Zhuang, John Miller, Moritz Hardt, and Alexei A. Efros. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 3, 8

[46] S. Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, A. Torralba, and W. Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569:698–702, 2019. 2

[47] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 1, 2

[48] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019. 1, 2

[49] Anand Thobbi and Weihua Sheng. Imitation learning of hand gestures and its evaluation for humanoid robots. *IEEE International Conference on Information and Automation*, pages 60–65, 2010. 1

[50] Dimitrios Tzionas, L. Ballan, A. Srikantha, Pablo Aponte, M. Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118:172–193, 2016. 6

[51] E. Ueda, Y. Matsumoto, M. Imai, and T. Ogasawara. A hand-pose estimation for vision-based human interfaces. *IEEE Trans. Ind. Electron.*, 50:676–684, 2003. 1

[52] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3d human motion and interaction in 3d scenes. *arXiv preprint arXiv:2012.05522*, 2020. 2

[53] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless C. Fowlkes. Geometric pose affordance: 3d human pose with scene constraints. *ArXiv*, abs/1905.07718, 2019. 2

[54] Xavier Williams and N. Mahapatra. Analysis of affordance detection methods for real-world robotic manipulation. *2019 9th International Symposium on Embedded Computing and System Design (ISED)*, pages 1–5, 2019. 2

[55] Min-Yu Wu, Pai-Wen Ting, Yahui Tang, En-Te Chou, and L. Fu. Hand pose estimation in object-interaction based on deep learning for virtual reality applications. *J. Vis. Commun. Image Represent.*, 70:102802, 2020. 1

[56] Xinchen Yan, Jasmine Hsu, M. Khansari, Yunfei Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee. Learning 6-dof grasping interaction via deep geometry-aware 3d representations. *ICRA*, pages 1–9, 2018. 1

[57] Yezhou Yang, C. Fermüller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. *CVPR*, pages 400–408, 2015. 1

[58] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J. Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. *2020 International Conference on 3D Vision (3DV)*, pages 642–651, 2020. 2

[59] T. Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *ICRA*, pages 1–8, 2018. 1

[60] Yilun Zhou and Kris Hauser. 6dof grasp planning by optimizing a deep learning scoring function. In *RSS Workshop on Revisiting Contact-Turning a Problem into a Solution*, volume 2, page 6, 2017. 1