Hand-Object Contact Consistency Reasoning for Human Grasps Generation

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

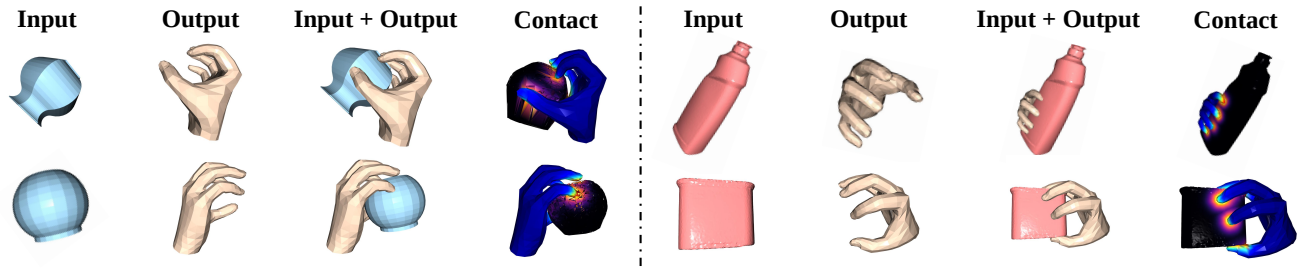

Figure 1: Generated human grasp on in-domain and out-of-domain objects. Object and hand contact maps are shown in the last column. The brighter the region, the higher the contact value between the hand and object. Best viewed in color.

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.

1. Introduction

Capturing hand-object interactions has been an active field of study [20, 8, 12, 5, 1, 2, 19, 17, 26] and it has wide applications in virtual reality [9, 24], human-computer interaction [23] and imitation learning in robotics [27, 21, 15]. 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 [13, 25, 28, 3], 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 [4, 10, 19] supervised by large-scale datasets [8, 7, 6] 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 [1, 2].

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

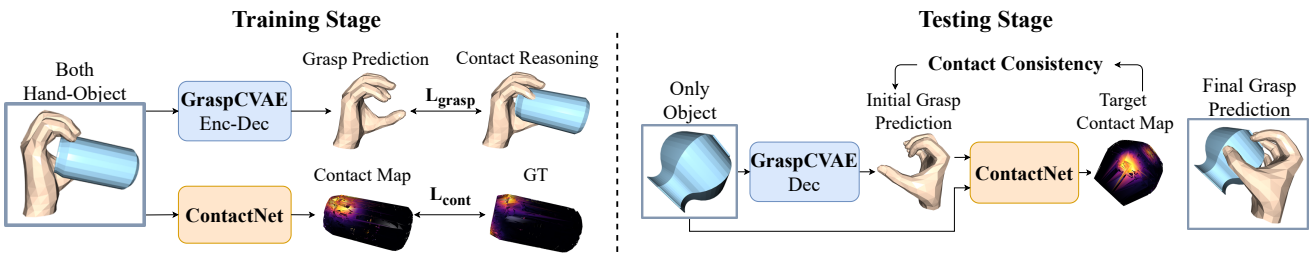

Figure 2: The *different* usage of proposed networks in training and testing. **Left**: During training, the two networks learn generating human grasps and predicting object contact map separately on ground truth data. **Right**: At test-time, the two networks are unified in a cascade manner for adaptation on novel test objects.

components as follows.

First, we train a Conditional Variational Auto-Encoder [18] (CVAE) based network which takes the 3D object point clouds as inputs and predicts the hand grasp parameterized by a MANO model [16], 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 [8], HO-3D [7] and FPHA [6] 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.

2. Approach

We emphasize that ensuring reasonable contact between the object and synthesized hand is the key to get high-quality and stable human grasps. 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. To train the GraspCVAE, we propose two novel losses to ensure the hand-object contact consistency, which will be introduced in Sec. 2.1.

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. Then, the generated grasp is forwarded together with the object to the ContactNet to predict a target contact map, which is used for finetuning the initial grasp. 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.

2.1. Learning GraspCVAE

Usage and architecture The GraspCVAE is a Conditional Variational Auto-Encoder (CVAE) [18] based generative network, using conditional information to control generation. The conditional information of GraspCVAE is the object.

During training, as shown in top row of Fig. 3, given point clouds of the hand and the object, we use two separate PointNets [14] to extract their features. Then they are concatenated as the CVAE encoder input. And the CVAE [18] learns to reconstruct the parameters of MANO model [16], by conditioning on the object information. The output of the MANO layer is the reconstructed hand mesh.

During testing, as shown in the bottom row of Fig. 3, we only utilize the decoder from the GraspCVAE for inference, where the latent code z is randomly sampled from a Gaussian distribution.

Baseline The first objective for the baseline model is mesh reconstruction error, which is defined on both the vertices $\mathcal{V}$ of the mesh as well as the parameters (β, θ) of the MANO model. We adopt the L_2 distance to compute the error

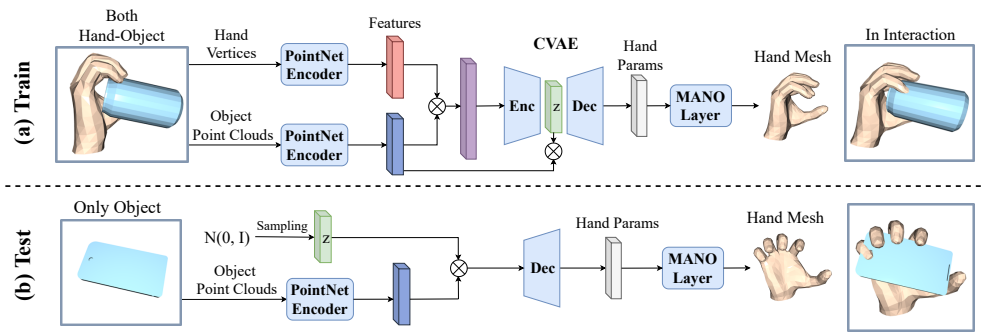

Figure 3: The architecture of GraspCVAE. (a) In training, it takes both hand-object as input to predict a hand mesh for grasping the object in a hand reconstruction manner using both of its *encoder-decoder*; (b) At test-time, its *decoder* generates grasps by conditioning only on object information as input. ⊗ means concatenation.

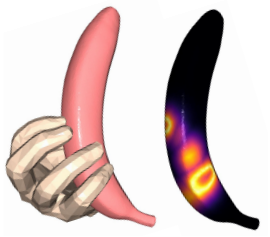

Figure 4: An example of contact map, brighter regions have larger scores. Because the MANO model does not have soft tissue, the underformable fingertips usually penetrate into object surface slightly.

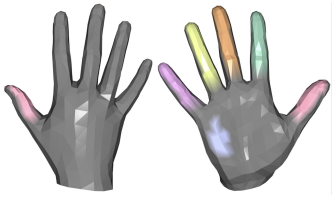

Figure 5: Six hand prior contact regions are shown in color.

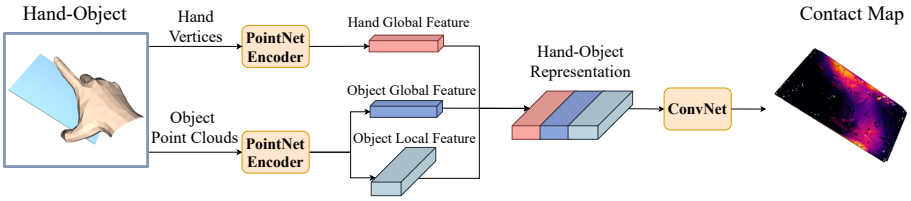

Figure 6: Architecture of ContactNet. It extracts local per-point feature of object point cloud, and concatenate it with global hand-object feature for predicting contact map.

as$L_{\mathcal{R}} = \lambda_{\mathcal{V}} \cdot L_{\mathcal{V}} + \lambda_{\theta} \cdot L_{\theta} + \lambda_{\beta} \cdot L_{\beta}$, where $\lambda_{\mathcal{V}}$, λ_{θ} and λ_{β} are constants balancing the losses.

Following the training of VAE [11], we define a loss enforcing the latent code distribution 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 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.

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 $f(\mathbf{D}(\mathcal{P}^o)) = 1 - 2 \cdot (Sigmoid(2\mathbf{D}(\mathcal{P}^o)) - 0.5)$. An example is shown in Fig. 4. 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 $L_{\mathcal{O}} = ||\hat{\Omega} - \Omega||_2^2$, $\Omega = f(\mathbf{D}(\mathcal{P}^o))$.

Hand-centric Loss. We define the prior hand contact vertices $\mathcal{V}^p$ as shown in Fig. 5, motivated by [8, 1]. 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 $\mathcal{P}^o_i$, we compute the distance $\mathbf{D}(\mathcal{P}^o_i) = \min_j ||\mathcal{V}^p_j - \mathcal{P}^o_i||_2^2$, 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, $L_{\mathcal{H}} = \sum_i \mathbf{D}(\mathcal{P}^o_i)$, $for\ all\ \mathbf{D}(\mathcal{P}^o_i) \leq \mathcal{T}$ 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,

$$L_{grasp} = L_{baseline} + \lambda_{\mathcal{H}} \cdot L_{\mathcal{H}} + \lambda_{\mathcal{O}} \cdot L_{\mathcal{O}} , \quad (2)$$

where $\lambda_{\mathcal{H}}$ and $\lambda_{\mathcal{O}}$ are constants balancing the losses.

2.2. 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. Since we need to predict the contact score for each point, we utilize the per-point

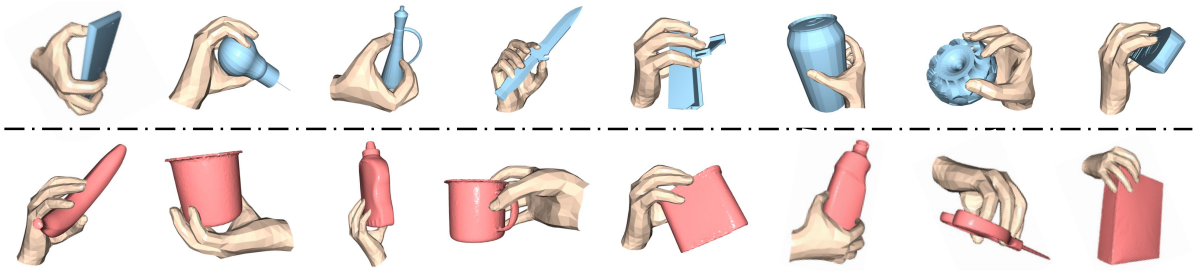

Figure 7: Visualization of generated grasps on in-domain objects from Obman dataset [8], and out-of-domain objects from HO-3D dataset [7].

		Obman			HO-3D			FPHA		
		GT	GF [10]	Ours	GT	GF [10]	Ours	GT	GF [10]	Ours
Penetration	Max Depth (cm) ↓	0.01	0.56	**0.46**	2.94	1.46	**1.05**	1.17	2.37	**1.58**
	Volume (cm^3) ↓	1.70	6.05	**5.12**	6.08	14.90	**4.58**	5.02	21.9	**6.37**
Simulation Stability	Mean (cm) ↓	1.66	2.07	**1.52**	4.31	3.45	**3.21**	5.54	4.62	**2.55**
	Variance (cm) ↓	± 2.44	± 2.81	**± 2.29**	± 4.42	± 3.92	**±3.79**	± 4.38	± 4.48	**± 2.22**
Perceptual Score	$\{1, ..., 5\}$ ↑	3.24	3.02	**3.54**	3.18	3.29	**3.50**	3.49	3.33	**3.57**
Contact	Ratio (%) ↑	100	89.40	**99.97**	91.60	90.10	**99.61**	91.40	97.00	**100**

Table 1: Results on Obman [8], HO-3D [7] and FPHA datasets [6] compared with ground truth (GT) and GF [10]. Best results are in bold and blue compared without and with GT.

object local feature of the PointNet encoder to ensure this correspondence. After concatenating all the features, we apply 1-D convolutions to regress the object contact map Ω^c. The loss is $L_{cont} = ||\Omega^c - \Omega||_2^2$.

2.3. 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 object and predicted hand mesh 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,

$$L_{TTA} = L_{refine} + \lambda_{\mathcal{H}} \cdot L_{\mathcal{H}} + \lambda_p \cdot L_{penetr}. \quad (3)$$

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

3. Experiment

We conducted experiments of three public datasets, the Obman [8], HO-3D[7] and FPHA[6] dataset. We only train the model on the Obman and test on all of the three dataset. We evaluate the grasps by physical penetration [8], grasp stability [22, 8], perceptual score [10] and hand-object contact metrics (contact ratio, contact finger number, etc).

	Penetration ↓		Simu Disp. ↓	Contact ↑
	Depth	Volume	Mean ± Variance	Ratio (%)
w/o TTA	**0.94**	**4.21**	4.98 ± 4.48	86.63
TTA	1.05	4.58	**3.21 ± 3.79**	**99.61**

Table 2: Abaltion of Test-Time Adaptation (TTA) on HO-3D dataset [7].

3.1. Grasp Generation Performance

Qualitative results Visualization results are shown in Fig. 7. From the visualization, our proposed 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. On all of the three datasets, our framework shows significant improvement over the state-of-the-art approach [10]. And results on HO-3D and FPHA dataset imply that our model has a much stronger cross-domain generalization ability. Besides, our results are close to or even outstrip the ground truth, especially for the stability and perceptual score.

3.2. Ablation Study

We validate the effectiveness of test-time adaptation (TTA) on out-of-distribution HO-3D dataset. As shown in Table. 2, with TTA, the simulation displacement decreases and the hand-object contact ratio increases significantly, which demonstrate that the grasp is much more stable. Moreover, the huge improvement of the steadiness is not at the expense of sacrificing the penetration metrics. The slightly bigger penetration between hand-object is reasonable considering the significant improvement of steadiness.

References

[1] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019. 1, 3

[2] 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

[3] 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

[4] 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. 1

[5] 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

[6] 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, 4

[7] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and objects poses. *CVPR*, 2019. 1, 2, 4

[8] 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, 3, 4

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

[10] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *arXiv preprint arXiv:2008.04451*, 2020. 1, 4

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[12] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 1

[13] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2901–2910, 2019. 1

[14] C. R. Qi, H. Su, Kaichun Mo, and L. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 77–85, 2017. 2

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

[16] 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

[17] M. Schröder and H. Ritter. Hand-object interaction detection with fully convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1236–1243, 2017. 1

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

[19] 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. 1

[20] 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

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

[22] Dimitrios Tzionas, L. Ballan, A. Srikantha, Pablo Aponte, M. Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016. 4

[23] 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

[24] 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

[25] 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. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9, 2018. 1

[26] Yezhou Yang, C. Fermüller, Y. Li, and Y. Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 400–408, 2015. 1

[27] T. Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2018. 1

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