LEARNING OBJECT AFFORDANCE WITH CONTACT AND GRASP GENERATION (SUPPLEMENTARY MATERIALS)

Anonymous authors
Paper under double-blind review

A APPENDIX

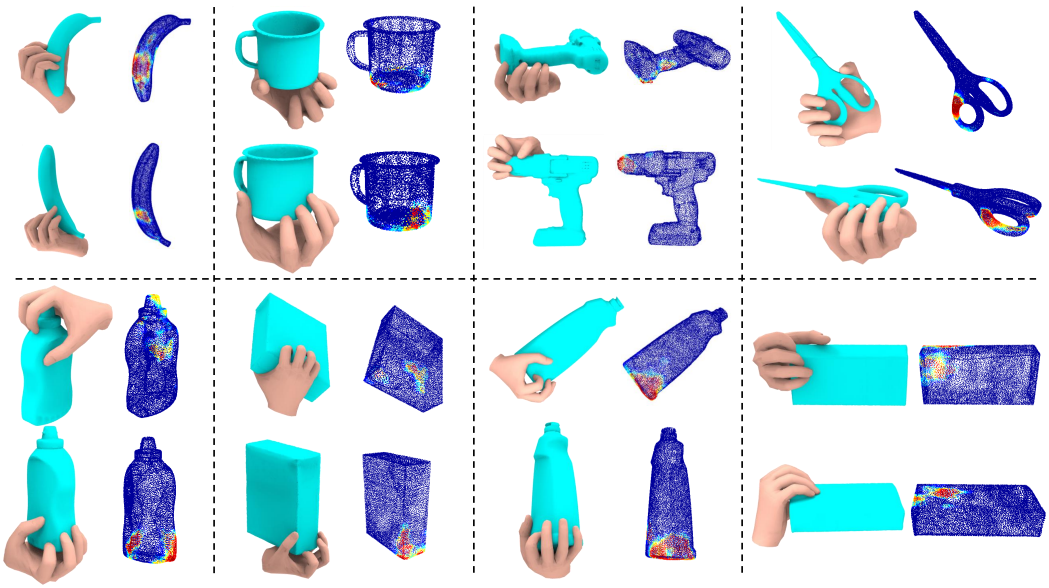

Figure 1: Generated contacts and grasps on out-of-domain HO-3D objects. We generate 2 grasps for each object. For each result, we visualize both the generated contact map (right) and the corresponding grasping pose (left).

A.1 EVALUATION ON THE HO-3D DATASET

HO-3D We further validate our framework on the HO3D dataset (Hampali et al., 2020), which is a real-world datset and includes 10 objects. Due to the limitation of number of objects and grasp types, we follow the work (Karunratanakul et al., 2020) and use the dataset for testing the models trained on Obman. Examples of generated grasps of our method are shown in Figure 1.

Local Refinement Details For the inference refinement on the HO3D, we adopt Adam optimizer with a learning rate of 1.00×10^{-4} to update the GraspNet decoder (MLP) parameters. The number of steps is 150.

Table 1 compares the quality of generated grasps of our method to that of the baselines and existing works on HO3D. Our method improves the Param2Mesh baseline on the penetration metrics and the contact ratio significantly. It outperforms the state-of-the arts on all metrics except the penetration volume. Specifically our method achieves the best performance on simulation (2.46 cm) and C-Rate (97.81 $\%$) metrics. Notice that the contact ratio of all methods in Table 1 is not so good as that of testing results on Obman in the main paper, which mainly results from the object size difference of these two datasets. The method proposed in Jiang et al. (2021) has difficulty in generalizing to HO3D while our method adapts to the new objects of larger size better regarding the contact ratio metric.

Table 1: Comparison on HO3D

Methods	Penetration, ↓		Simulation, ↓		C-Rate(↑)	S-Rate(↑)
	Dep	Vol	Mean	Var		
GT	2.94	6.08	4.31	±4.42	91.60	-
GraspField (Karunratanakul et al., 2020)	1.46	14.90	3.45	±3.92	90.10	-
Jiang et al. (2021) (w/o refine)	0.94	**4.21**	4.98	±4.48	86.63	-
Jiang et al. (2021)	1.05	4.88	3.80	±4.20	92.31	-
Param2Mesh (Baseline)	1.50	17.70	2.67	**±2.58**	92.78	13.97
Ours (w/o refine)	0.90	6.86	**2.46**	±2.63	**97.81**	34.22
Ours	**0.89**	5.08	2.53	±2.66	96.19	**40.56**

A.2 More Visualization on the ContactPose Dataset

In Figure 2, we present the generated results on objects from the ContactPose (Brahmbhatt et al., 2020) training group. We can observe that the some generated contacts and predicted grasps are corresponding with the object affordance.

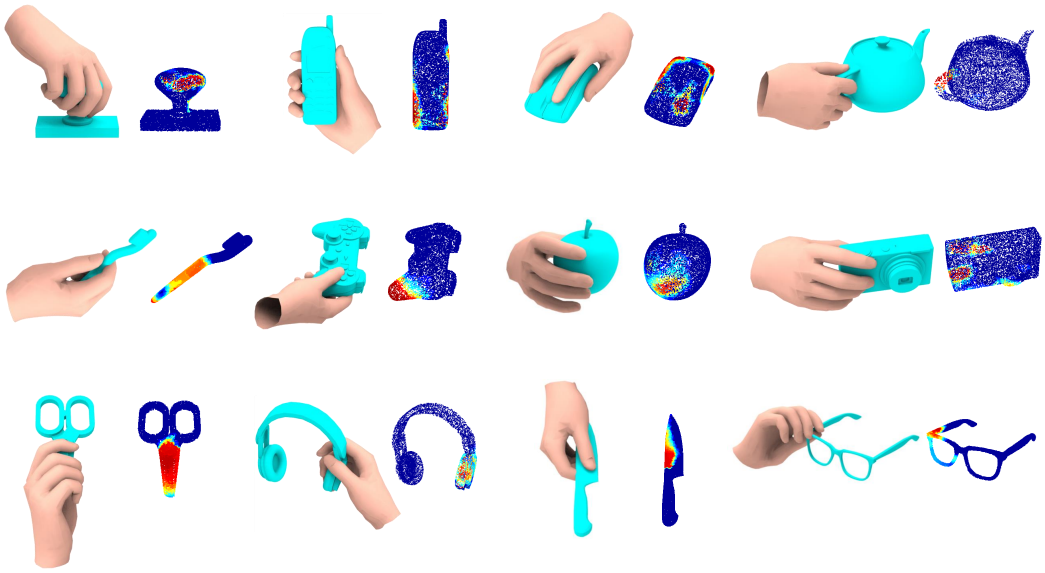

Figure 2: Generated contacts and grasps on ContactPose objects from the training group.We select 12 out of 21 objects for the visualization. For each object, we visualize both the generated contact map (right) and the corresponding grasping pose (left). Note that the generated grasps is more realistic and reasonable.

A.3 The Mathematical Descriptions of Metrics

Simulation Displacement. We choose this metric by following (Hasson et al., 2019) as it provides a way to measure the physical realism of generated grasps and we can compare with (Jiang et al., 2021; Taheri et al., 2020) on this metric. The simulation environment (PyBullet physics engine) checks the force closure. If the force closure is not met, objects fall and cause the displacement of objects after placing generated grasps. Therefore, the displacement metric actually has the force-closure metric included. The displacement is defined as $Dist = f(x, x')$ if $f(x, x') < 0.1$, else $Dist = 0.1$, where $f(x, x') = \|x - x'\|_2$. x, x' represent the initial and end position of object center.

Penetration volume (P_v) and depth (P_d) follows the definition in (Hasson et al., 2019),To calculate the penetration, we first voxelize a grasp and an object, then a sign distance field (SDF) is constructed for the object (positve inside the object and negative outside). For each voxel of the grasp, if its value in the object SDF is positive, it is an intersected voxel.$P_v = \mu \sum_{i=0}^{N} v_i, v_i \in V$, where V denotes

the intersection voxel set between the grasp and the object, v_i represents the i-th voxel.μ denote unit size of voxel and N denotes the number of voxels. The penetration depth is the maximum distance of the penetration, defined as$P_d = \underset{v_i \in V}{max} SDF(v_i)$, where $SDF(v_i)$ denotes the distance value of an intersected voxel.

Contact Rate (C_r). We define a contact between the object and the hand when any point on the surface of the hand is on or inside the surface of the object (Karunratanakul et al., 2020).The formuation is defined as:

$$C_r = \frac{1}{N_s} \sum_{i=0}^{N_s} x_i, \qquad x_i = \begin{cases} 1, & (max(C'') \neq 0) \wedge (P_v > 0) \\ 0, & else \end{cases}, \tag{1}$$

where N_s denotes the number of generated samples.x_i denotes the i-th sample.

Grasp Success Rate (S_r) is a comprehensive metrics. It define a grasp with both less penetration and simulation displacement as positive sample. The definition of S_r is:

$$S_r = \frac{1}{N_s} \sum_{i=0}^{N_s} x_i, \qquad x_i = \begin{cases} 1, & (P_v < \eta) \wedge (Dist < \alpha) \\ 0, & else \end{cases}, \tag{2}$$

where $\eta = 5(cm^3)$ and $\alpha = 2.0(cm)$ denote threshold of penetration volume and simulation displacement respectively.

A.4 The extension of our work to robotics applications

The highlight of our work is that using generated contact map to get a reasonable grasp pose, which can be easily extended to robotics applications. In (Mandikal & Grauman, 2021), the contact maps are used for input of reinforcement learning (RL) for grasping and defining better rewards. In (Mandikal & Grauman, 2022),the human poses reconstructed from videos are retargeted to the robotic hands by simply keeping the joint rotations (or keeping the task vector space) of the generated grasps. The resulting robotic hand poses are used as rewards for RL for refinement.

Based on these works, we can extend our work to robotic grasping as the following steps: (1) Retargeting human poses to posed for robotic hands. (2) Using the generated contact maps as the input for defining more detailed hand object features and better rewards for exploration under the reinforcement learning framework. (3) Using the retargeted robotic hand poses for rewards and more efficient sampling of reinforcement learning.

References

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*, pp. 361–378, 2020.

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*, pp. 3196–3206, 2020.

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*, pp. 11807–11816, 2019.

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*, pp. 11107–11116, 2021.

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*, pp. 333–344, 2020.

Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *IEEE International Conference on Robotics and Automation*, pp. 6169–6176, 2021.

Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pp. 651–661, 2022.

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*, pp. 581–600, 2020.