Supplementary Material

Haoxiang Ma[1,2] Modi Shi[1,2] Boyang Gao[3,4] Di Huang[1,2*]
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[4]Geometry Robotics
{mahaoxiang822,modishi,dhuang}@buaa.edu.cn, {boyang.gao}@geometryrobot.com

1. Network Details

In this section, we detail the training process of our 6-DoF grasp detection network. Given the reconstructed TSDF, we first sample a point-cloud $\in \mathbb{R}^{N\times 3}$ with normal $\in \mathbb{R}^{N\times 3}$, where N is the number of points. Subsequently, we use the sparse UNet[1] to extract point-wise features $\in \mathbb{R}^{N\times C}$ from the sampled point-cloud. Following [4], objectness and graspness score are predicted on each points. We use softmax loss $L_o$for objectness classification and smooth-l_1 loss $L_{gs}$for graspness score regression. After that, M grasp candidates are sampled from the point-cloud by Fatherst Point Sampling (FPS) where the graspness larger than a threshold. The approach directions of these candidates are supervised by the approach loss L_a, which is formulated as:

$$L_a = cos(a,\hat{a}) + \alpha \cdot (1 - \frac{\sum\limits_{\angle a,a_i<30^\circ} s_i \cdot cos(a,a_i)}{\sum\limits_{\angle a,a_i<30^\circ} cos(a,a_i)}) \quad (1)$$

where a and $\hat{a}$ are the predicted and ground-truth approach direction, $cos(,)$ is the cosine similarity and $\alpha = 0.1$. Different from previous works[2–4] which predict the approach direction score on discrete spherical regions, we formulate the approach prediction as a regression task to enable the differential learning of physical constraints. The second term of L_a imposes a regularization on the scores s of the approaches around the predicted direction, which leads to more robust approach prediction. To regress other grasp configurations, we leverage multi-scale cylinder grouping to extract local features for candidates. An operation head is used to regress the plane rotation, width, and grasp score with smooth-l_1 loss L_r,L_w,L_s. The total training loss of the 6-DoF grasp detection network can be formulated as:

$$L = L_o + L_{gs} + L_a + \beta \cdot (L_r + L_w + L_s) + \phi \cdot R \quad (2)$$

where $\beta = 0.2$ and $\alpha = 0.1$.

2. Implementation Details of Object SDF

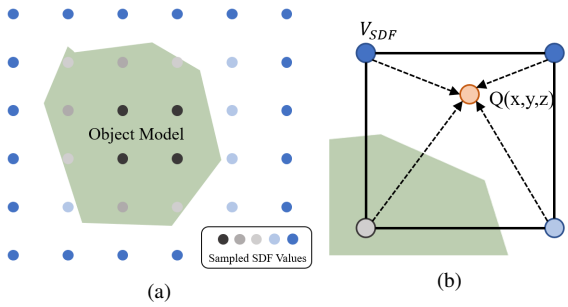

Figure 1. (a) Grid based SDF encoding (2D for visualization) and (b) the calculation of the SDF value of query position Q.

As acquiring the actual SDF of object models can be challenging, we encode them by sampling distance values of 3D grids with positions inside and outside the object model, as illustrated in Fig.1 (a). Utilizing these densely sampled grids, the SDF value of a query point Q can be determined through trilinear interpolation. Fig. 1 (b) provides a 2D visualization of this process. For the caomputation of physical constraint regularization, the surface distance and normal vector for any position near the object can be retrieved differentially from the object SDF grids.

3. Results on Kinect Scenes

The results on scenes captured by Kinect camera of GraspNet-1billion benchmark are shown in Table 1. We give a comparison with other state-of-the-art methods and our method shows significant superior performance on similar and novel objects, demonstrating the effectiveness of introducing physical and contact map prior.

*Corresponding author.

Model	Seen			Similar			Novel		
	AP	**AP**$_{0.8}$	**AP**$_{0.4}$	**AP**	**AP**$_{0.8}$	**AP**$_{0.4}$	**AP**	**AP**$_{0.8}$	**AP**$_{0.4}$
GraspNet-baseline [2]	27.10	30.22	24.75	25.20	27.80	24.44	6.45	7.99	3.89
Scale-balanced Grasp [3]	43.72	49.41	39.83	37.49	42.25	34.35	12.50	15.07	6.86
GSNet [4]	43.97	53.28	35.53	40.07	48.58	32.03	14.54	16.50	6.96
Ours Baseline	57.23	**65.45**	49.57	48.67	55.54	43.88	18.70	23.57	8.68
Ours	**57.56**	65.21	**51.82**	**52.88**	**61.23**	**47.37**	**21.59**	**27.30**	**11.56**

Table 1. Comparison with the state-of-the-art methods on Kinect scenes of GraspNet-billion benchmark.

4. Visualization of Real-world Grasping

We attach a video in supplementary material to visualize our real-world grasping system and demonstrate the effectiveness of the proposed method.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[2] Haoshu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2

[3] Haoxiang Ma and Di Huang. Towards scale balanced 6-dof grasp detection in cluttered scenes. In *Conference on Robot Learning*, 2022. 2

[4] Chenxi Wang, Haoshu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *IEEE/CVF International Conference on Computer Vision*, 2021. 1, 2