Appendix

In this Appendix, we provide the detailed setting of our grasping experiment environment in V-REP (Section A), as well as the real-world experiment using UR5 arm.

A Grasp Experiment in Simulator

We use V-REP to set up the experimental environment of grasping tasks, which contains the UR5 arm, the RG2 gripper, a table, a box, the objects to grasp, and one depth camera. We show the grasp experiment environment in figure 1.

We produce the point cloud from the depth camera as the input and uses GDN (Jeng et al., 2020) to find the grasps based on the point cloud. We perform a single object grasping experiment in three clutter levels, free, touching, stacked, with 4,6,8 objects in the scene, respectively. An object in the YCB (Calli et al., 2015) set is randomly selected and placed on a table, and then the robot tries to grasp the object. If the robot successfully grasps the object from the table to the box, it counts one success. We do 11 trials and calculate the average success rate for each object.

Table 1 demonstrates the results of the single object grasping task in terms of success rate with 2D box and 3D Mask in different clutter level. Using a 3D mask rate than a 2D box as an input can get a higher average gripping success rate. The difference increases in a more dense cluttered scene, suggesting that with more accurate segmentation in 3D spatial environment is relatively unaffected in the cluttered and occluded environment.

B Real-World Experiment

We use UR5 robotic arm to conduct the real-world experiment and equip Intel Realsense D415 to obtain the RGB and depth information. Additionally, we utilize PyRobot (Murali et al., 2019) to high-level interact with ROS Kinetic to control the robotic arm, and we adopt rapidly exploring random tree as our planning strategy to manipulate the movement of arm. For the real-world environment, we first set up a table with black table cloth, and put some objects on it. Then, ask the user to give a referring expression to the system, and it would identify the target object and grasp it.

	top			bottom		
	F	T	S	F	T	S
2D	65.9	40.9	28.4	72.7	51.5	21.6
3D	72.7	48.5	36.4	75.0	56.1	31.8
Δ	6.8	7.6	**8.0**	2.3	4.5	**10.2**

Table 1: 2D (box) and 3D (mask) Grasp Experiment on different clutter level (%), F: Free, T: Touching, S: Stacked.

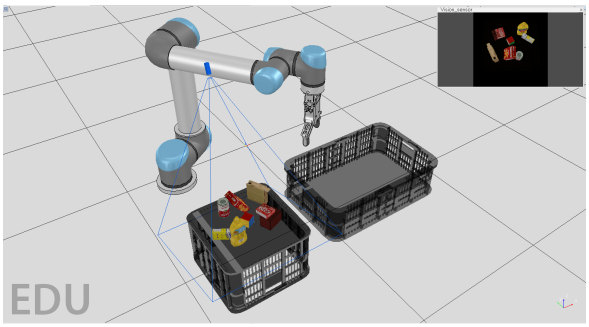

Figure 1: Grasp Experiment Environment.

References

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. 2015. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE.

Kuang-Yu Jeng, Yueh-Cheng Liu, Zhe Yu Liu, Jen-Wei Wang, Ya-Liang Chang, Hung-Ting Su, and Winston H. Hsu. 2020. Gdn: A coarse-to-fine (c2f) representation for end-to-end 6-dof grasp detection.

Adithyavairavan Murali, Tao Chen, Kalyan Vasudev Alwala, Dhiraj Gandhi, Lerrel Pinto, Saurabh Gupta, and Abhinav Gupta. 2019. Pyrobot: An open-source robotics framework for research and benchmarking. *arXiv preprint arXiv:1906.08236*.