Supplementary materials for “Target-referenced Reactive Grasping for Dynamic Objects”

Jirong Liu[1,2], Ruo Zhang[2], Hao-Shu Fang[2], Minghao Gou[2], Hongjie Fang[2],
Chenxi Wang[2,3], Sheng Xu[2], Hengxu Yan[2], Cewu Lu[1,2]†
[1]Shanghai Qi Zhi institute, [2]Shanghai Jiao Tong University, [3]Flexiv Robotics, LTD
jirong@sjtu.edu.cn, {ruozhang0608, fhaoshu}@gmail.com,
{gmh2015, galaxies, wcx1997, xs1020, hengxuyan, lucewu}@sjtu.edu.cn

1. Grasp Features Representation

We regard our problem as finding strong matches between two sets of grasp features. Therefore the first step is to extract distinctive features for each grasp. After the grasp detector, the input point cloud of size $N \times 3$ will be downsampled to seed points of size $M \times 3$ and grasp poses of size $M \times 12$ will be generated. We consider that two grasps can be distinguished by geometric and visual properties of the contact area between grasps and objects. Therefore we adopt cylinder grouping to crop the point cloud along the grasping direction. Here, $\lfloor M/8 \rfloor$ points of size are grouped for each grasp, resulting in vectors of size $M \times (\lfloor M/8 \rfloor) \times (3+3)$ which represent coordinates and pixel values of each point inside the cropped local region. After that we use multi-layer perceptrons (MLP), followed by max-pooling and mean-pooling, to map points and pixels to geometric features and visual features respectively. Both geometric and visual features have the size $M \times C$. Additionally, the backbone of the grasp detector provides seed point features of size $M \times C$. We also add global features of size $M \times C$ as mentioned. All these features are concatenated.

2. Network and Training

For grasp detector, we sample 1024 points from the scene point cloud which are used to produce 1024 grasps afterwards. In the correspondence estimation part, in order to extract the grasp features, for each grasp we group 256 points from the input point cloud using a cylinder with radius of 0.05 meter and height of 0.06 meter. The correspondence estimation module consists of 4 layers for both self and cross attention. The refinement network uses a similar structure but only with 8 self-attention layers. The feature dimension C is set to 256 and sequence length L is set to 5 during training. MLP layers of size (512, 256, 128, 9) are used for final refinement prediction.

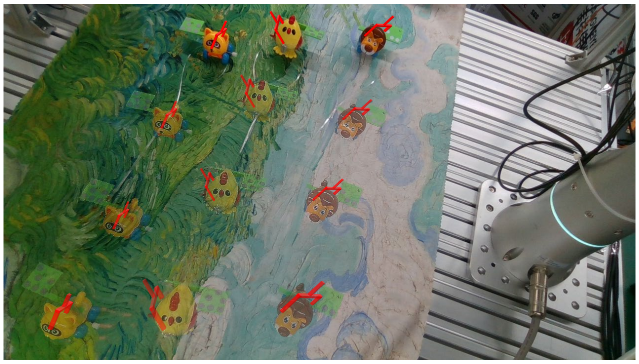

Figure 1. An example of Moving GraspNet.

Furthermore, our pipeline uses pre-trained GraspNet baseline for grasp detection, which is frozen during training. The remaining parts are trained on 8 NVIDIA A100 GPUs using Adam optimizer. The model is trained for 50 epochs and the learning rate is set to 5^{-4}. We apply step learning rate decay policy and multiply the learning rate by 0.5 per 10 epochs. The batch size is set to 2. We clip the gradient if its norm is greater than 10 as well.

3. Moving GraspNet

We collect a test set of moving objects which is named as Moving GraspNet. This test set contains 30 scenes and over 7000 frames in total. For each time-step, we record the RGBD images and object poses. To be specific, RGBD images are collected using Intel RealSense D415/D435 depth cameras and object poses are collected using calibration markers. We further project 10 object-level grasp poses to the objects according to their 6D poses. As a result, scene-level grasp poses can be annotated for each time-step as tar-

† Cewu Lu is the corresponding author, a member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China

gets for tracking. The 10 grasp poses are picked by their force-closure scores. Fig.1 shows an example of motion sequence from the test set. Three objects are presented in this scene. We annotate a ground-truth grasp at each timestep.

4. MGTA

We propose a metric for our task named as Multiple Grasp Tracking Accuracy ($\mathbf{MGTA}$) which is a modified version of the widely-used Multiple Object Tracking Accuracy ($\mathbf{MOTA}$) in multi-object tracking. $\mathbf{MGTA}$ can be defined as:

$$\mathbf{MGTA} = 1 - \frac{\sum_t FN_t + FP_t + IDSW_t}{\sum_t GT_t}, \qquad (1)$$

where FN, FP, IDSW, GT, t denote false negative, false positive, ID switching, ground truth and timestep respectively. The value of $\mathbf{MGTA}$ is in $(-\infty, 1]$. $\mathbf{MGTA}$ can be negative if the number of FN, FP ans IDSW is larger than the number ground truth grasps. Note that the only change compared with $\mathbf{MOTA}$ is the similarity measure which is the basis of classifying tracking results as well. In multi-object tracking, similarity measure is usually based on IoU whereas in our setting the similarity is computed using the grasp distance as mentioned in the paper.

5. Object List

Here we list the objects' IDs, names, colors, servo types and sizes in Tab.1.

Object ID	Name	Color	Servo	Size
1	Cat	White/Brown	Electric	Small
2	Caterpillar	Green	Electric	Medium
3	Owl	Grey	Electric	Medium
4	Dinosaur	Green	Clockwork	Medium
5	Beetle	Black/Brown	Electric	Medium
6	Dinosaur	Orange	Clockwork	Medium
7	Chicken	Yellow	Clockwork	Small
8	Chicken	White	Clockwork	Small
9	Rabbit	White	Electric	Large
10	Dinosaur	White	Electric	Large
11	Sheep	White/Green	Electric	Small
12	Dog	Golden	Electric	Large
13	Dinosaur	White	Electric	Large
14	Dinosaur	White	Electric	Large
15	Sheep	White/Blue	Electric	Small
16	Sheep	Purple	Electric	Small
17	Cat	White	Electric	Small
18	Cat	Grey	Electric	Small
19	Sheep	White/Red	Electric	Small
20	Clown	Red	Electric	Large

Table 1. Details of objects used in the experiments.