Supplementary Material for Learning Generalizable Dexterous Manipulation from Human Grasp Affordance

We provide more details about experiment settings, motion planning implementation and more ablation results in the supplementary material. We also provide a supplementary video to better visualize the results.

1 Experiment details

The initial poses of the object should keep the object static on the table. To achieve this, We first sample random initial poses in the air and then let the object fall on the table, and use the poses when it stops moving as the initial poses of our experiment. We use 36 bottles, 39 cameras, 36 cans, 19 mugs and 24 remotes as the test data. The terms of use of ShapeNet [1] is at `https://shapenet.org/terms`. The simulator we use is MuJoCo [2], which is under Apache-2.0 License.

For cross-entropy method, we sample 200 actions each time and pick 10 elite candidates to update the μ and σ. We set the time horizon to 5 during the planning.

To learn the approximate advantage function [3] $A_\phi^{\pi_\theta}$ for demonstrations, we share the baseline function $V^{\pi_\theta}(s)$ that has been already learned to estimated A^{π_θ}. The additional model we introduce here is a value function $Q^{\pi_\theta}(s,a)$ that is used to estimate discounted reward sum of state-action pairs (s,a). The advantage function is then derived by

$$A_\phi^{\pi_\theta} = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s).$$

To make sure our experiment results are robust across categories without much parameter tuning, the experiments on all five categories share the same set of hyper-parameters. We summarize the hyper-parameters in Tab. 1.

We parameterized the value function with two separate 2-layer MLPs. For each update iteration, we collect 200 trajectories from the environments to estimate the policy gradient and update both policy and value networks.

Hyper-parameter	Value
λ_0	0.1
λ_1	0.99
λ_0'	0.01
# of trajectories for each epoch	200
initial policy log std	0.0
network architecture of MLP	$(32, 32)$
network architecture of PointNet	$(1024, 512, 32)$

Table 1: Hyper-parameters of the proposed ILAD approach for all experiments.

2 Visualization on generalizability

We execute the policies on unseen objects that are not shown during training and visualize the results in Fig. 1. For pair comparison, we fix the initial position of the object and the target position for both policies. In Fig. 1, we show that ILAD learns to hold the objects firmly even when they are

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

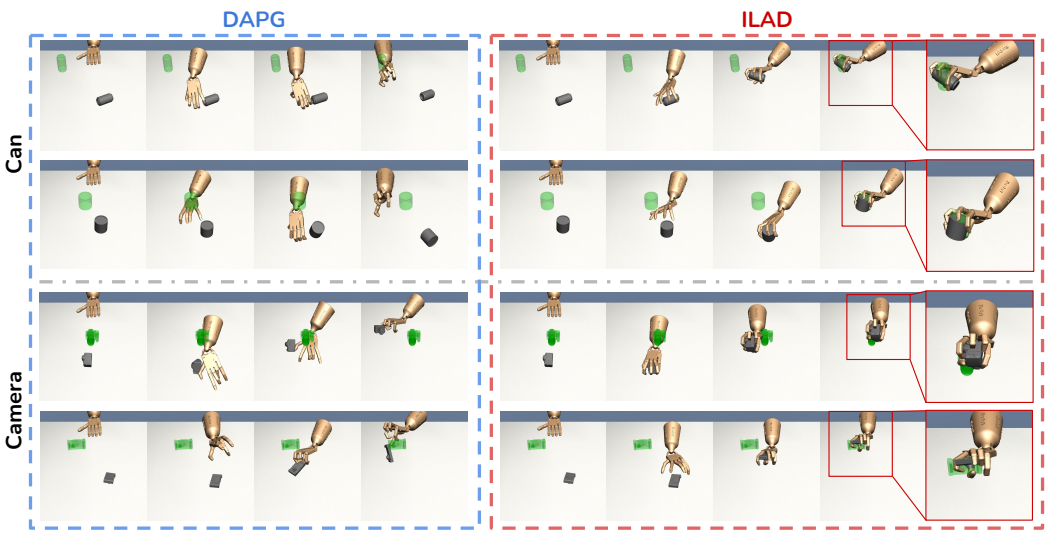

Figure 1: Comparison of the robustness on unseen can and camera objects. **Left**: policy learned by DAPG; **Right**: policy learned by ILAD. The environments in the same row share the same objects, initial position, and target position. We zoom in the last frame of our results.

not seen during training. Although DAPG achieves competitive performance in terms of average return during training, it is weak to generalize to unseen objects. It is especially challenging to grasp cylinder objects that require a specific angle and careful handling as suggested in the first row. In the fourth row, the policy is required to relocate a camera lying flat on the table. The proposed ILAD grasps the whole camera, which allows it to move the camera stably. On the other hand, DAPG only holds one side of the camera and the camera ends up being thrown away. We provide more visualization in the supplementary material.

Interval T	ILAD
$T = 10$	0.81 ± 0.16
$T = 20$	0.85 ± 0.05
$T = 50$	$\mathbf{0.99 \pm 0.01}$
$T = 70$	0.94 ± 0.03
$T = 110$	0.95 ± 0.05
$T = 150$	0.93 ± 0.03
$T = 200$	0.91 ± 0.05
$T = 400$	0.88 ± 0.02
no joint learning	0.65 ± 0.24

Table 2: The success rate of ILAD on unseen bottle objects. The performance is evaluated via 100 trials for five seeds.

3 Ablation study

Joint learning interval. To further study the influence of the joint learning interval T, we conduct experiments on the bottle category with more values of T. As suggested in Tab. 2, the value of T that obtains the highest success rate is $T = 50$. Furthermore, it is observed that even with $T = 400$, which indicates that the model are tuned with only three joint-learning during the whole training process, its success rate (0.88 ± 0.02) still outperforms the method without joint learning (0.65 ± 0.24) by a large margin.

Comparison with rapidly-exploring random trees (RRT). To illustrate the proposed demonstration generation pipeline is able to efficiently generate demonstrations for learning, we further compare the motion planning part with rapidly-exploring random trees (RRT) [4] and some variants of our proposed method. We evaluate the results on the Bottle category. Table 3 shows that the demon-

Demonstration	Bottle
RRT	0.13 ± 0.08
$\delta = 0.06$ w/o grasp poses	0.01 ± 0.01
$\delta = 0.1$ w/ grasp poses	0.40 ± 0.35
$\delta = 0.06$ w/ grasp poses	$\mathbf{0.65 \pm 0.24}$

Table 3: Demonstration quality ablation. The numbers represent the average success rate with five distinct random seeds. The policies are evaluated on unseen bottle objects and they are trained with the same imitation learning approach.

strations generated by RRT is able to capture the accuracy to the final grasp poses to some extent and makes it outperform the variant of our method that does not use the generated grasp poses. But our full pipeline still ourperforms a large margin because it uses a reward to balance reaching the target pose and preventing the object from moving during the reaching process. However, the design of RRT makes it fail to find such a balance. For implementation details of RRT, we still use GraspTTA [5] to generate target grasp pose. We set 10,000 nodes in the tree, and set step size $\epsilon = 0.01$ and probability of sampling $\beta = 0.5$.

References

[1] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *IROS*, 2012.

[3] L. C. Baird III. Advantage updating. Technical report, 1993.

[4] S. M. LaValle et al. Rapidly-exploring random trees: A new tool for path planning. 1998.

[5] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. *ICCV*, 2021.