Interaction Prediction and Monte-Carlo Tree Search for Robot Manipulation in Clutter

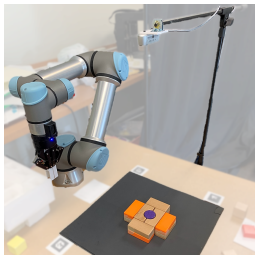
Baichuan Huang*, Shuai D. Han*, Abdeslam Boularias*, Jingjin Yu*
*Department of Computer Science, Rutgers University
Email: baichuan.huang@rutgers.edu

***Abstract*—We considered two related robot manipulation problems: clutter removal and object retrieval in clutter, while emphasizing action efficiency and exploring synergy between pushing and grasping. For clutter removal, we propose a Deep Interaction Prediction Network (DIPN) for predicting interactions between robot gripper and objects. Together with a grasp network for predicting grasping success, DIPN generates intelligent pushes for solving clutter removal problem efficiently. For the second, more complex object retrieval problem, we combine DIPN and Monte-Carlo Tree Search to generate smart multi-step push predictions that yields the best target object pose for grasping. We call this fusion of methods the Visual Foresight Tree (VFT). Experiments in simulation and using real robot show that our methods significantly outperform the previous state-of-the-art, and produce human-like performance. More detail can be found at https://arxiv.org/pdf/2011.04692.pdf and https://arxiv.org/pdf/2105.02857.pdf**

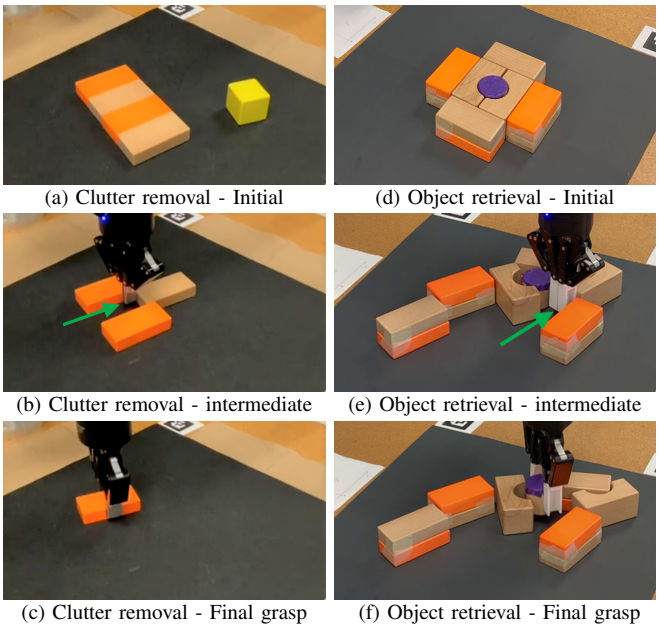
I. Preliminaries

We study two tasks that are common in everyday environments, clutter removal as shown in Fig. 1a-1c and object retrieval as shown in Fig. 1d-1f. The synergy between prehensile and non-prehensile actions are essential and has drawn increasing interest in recent years, for example, visual-pushing-grasping [1] for clutter removal and goal-oriented push-grasping [2] for object retrieval in clutter. For such tasks, we aim to realize intelligent, human-like performance with high levels of action efficiency.

The idea behind this work is the ability to "imagine" the outcome of an action before taking it so that the robot can make informed decisions. For this purpose, we propose a Deep Interaction Prediction Network (DIPN) which can predict the poses of the objects after an arbitrary push action. The clutter removal task can be solved efficiently by direct integration of DIPN with a deep Grasp Network (GN) as shown in Fig. 2. Object retrieval from clutter is harder, as selecting the shortest sequence of actions involves a huge combinatorial search space. Together with DIPN, GN, and Monte-Carlo Tree as shown in Fig. 3, we propose Visual Foresight Tree (VFT) to retrieve the target object with a minimal number of actions.



Our experimental setup is shown on the right and a similar setup is used in simulation. We have a planar workspace and a camera on top for state observation. In this study, clutter removal is the task that all objects in the workspace should be grasped as shown in the left column of Fig. 1. Whereas the object retrieval from the clutter asks the robot only to grasp the target as shown in the right column of Fig. 1.



(a) Clutter removal - Initial (d) Object retrieval - Initial
(b) Clutter removal - intermediate (e) Object retrieval - intermediate
(c) Clutter removal - Final grasp (f) Object retrieval - Final grasp

Fig. 1: (a-c) are the clutter removal task, where all objects are removed. (d-f) are the object retrieval from clutter task, where the purple cylinder is the target to be retrieved.

Our simulation and real robot experiments demonstrate that DIPN can accurately predict objects' poses given a push action. With that, our systems exceed the state-of-art [1, 2] for clutter removal and object retrieval, and produce human-like performance.

II. DIPN+GN for Clutter Removal

The Grasp network (GN) is a DQN [3] adapted from [1] where we use ResNet + FPN [4] as the backbone and image-based pre-training [5]. GN takes an RGB-D image as input. It outputs a pixel-wise grasp reward, where each pixel represents the center of a grasp action.

The Deep Interaction Prediction Network (DIPN) takes an RGB-D image, a push action and object segmentation from Mask R-CNN [6] as input. It outputs translations and rotations for each object in 2D space as the prediction. DIPN uses ResNet as the backbone to encode image information and linear layers to encode the action and position of objects. We adopt the design philosophy from [7] to process interactions

between objects by using two linear layers explicitly. One is the *direct transformation* module that captures the direct interaction between the gripper and objects. The other one is the *interactive transformation* module that computes the interaction between each pair of objects. All hidden internal transformations are combined with encoded global information to produce the final transformations for each object.

Clutter removal is solved by integrating GN and DIPN as a one-step-ahead planning algorithm as shown in Fig. 2. The RGB-D image is the input to GN for estimating grasp rewards. If the maximum grasp reward is higher than a threshold, then the robot grasps directly. If not, a push action is needed. A set of randomly sampled push actions are passed to DIPN along with object information. DIPN outputs a set of transformations to the corresponding push action. For each push action, transformations are applied to segments of objects to generate a synthetic image. Finally, the synthetic images are sent to GN to choose the best action.

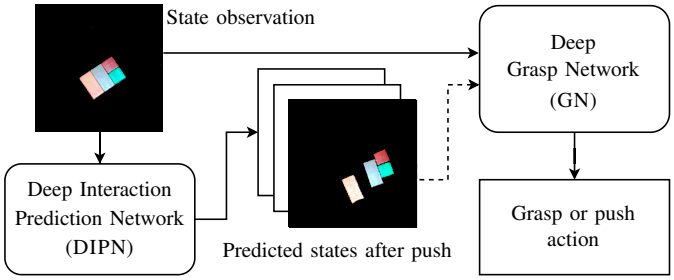


Fig. 2: System architecture of DIPN+GN.

GN is trained in the real world, whereas the DIPN is trained in simulation and used in real experiments directly for this clutter removal task. We compared our method with VPG [1] for both random and hard instances as shown in Table. I. Our method is more action efficient for solving clutter removal and has a much higher grasp and completion success rate. *Completion* means the rate of the robot completely removes all objects. *Grasp success* is defined as the total number of objects grasped divided by the total number of grasp actions. *Action efficiency* is calculated as the total number of objects grasped divided by the total number of push and grasp actions. Example instances can be found in Fig. 4.

	Method	Completion	Grasp Success	Action Efficiency
Rand	VPG[1]	80.0%	79.0%	67.9%
	DIPN+GN (ours)	**100**%	**94.0**%	**98.2**%
Hard	VPG[1]	64.0%	69.0%	47.8%
	DIPN+GN (ours)	**98.0**%	**89.9**%	**78.2**%

TABLE I: Clutter removal metric mean on a real system.

III. VFT for Object Retrieval in Clutter

Object retrieval requires longer-horizon planning, as a single push may not increase the grasp rewards at all, so one-step planning would be the same as a random push policy. A Monte-Carlo Tree Search (MCTS) comes to play and is fused with GN and DIPN as Visual Foresight Trees (VFT) as shown in Fig. 3. GN first estimates the input image to decide whether the target object is directly graspable. If not, then the MCTS module explores possible push actions using DIPN as the state transition function to expand the tree. GN is also used to obtain the reward value for each search node and as the reward and terminal function. Finally, the action associated with the best child node of the root will be executed.

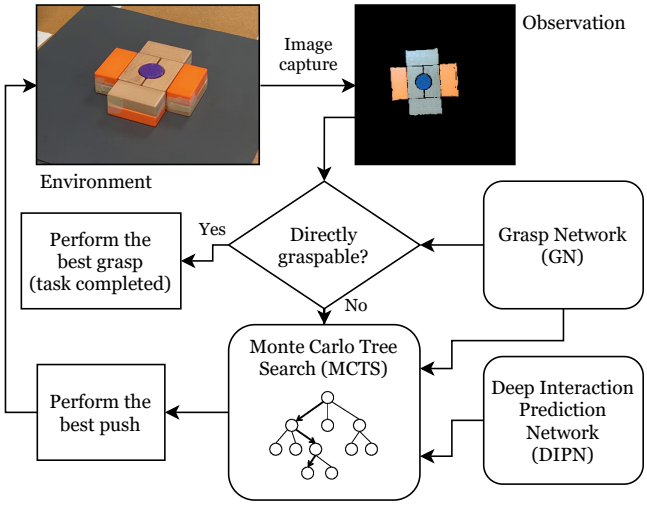


Fig. 3: System architecture of VFT.

GN and DIPN are both trained in simulation and used directly in real experiments for object retrieval tasks. We first compared our VFT with go-PG [2] in simulation. Table. II shows the result of simple instances from [1] but with a given target object as a retrieval task. Our method solves tasks with a near-optimal number of actions. Even our one-step ahead DIPN+GN uses a fewer number of actions comparing to [2].

	Completion	Grasp Success	Num. of Actions
go-PG [2]	99.0%	90.2%	2.77
DIPN+GN (ours)	100%	100%	2.30
VFT (ours)	**100**%	**100**%	**2.00**

TABLE II: Simulation metric mean for simple instances in [2]

As DIPN+GN already outperforms [2], we evaluate our methods on 22 hard instances on a real robot for further study. We use the experimental results from [2] on its four hard instances for comparison. Results in Table. III are consistent with simulation. Example instances can be found in Fig. 4.

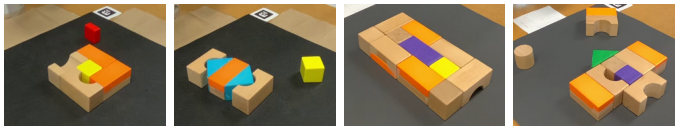


Fig. 4: Example instances. First two are for clutter removal problem. Last two are for object retrieval problem, where the target is in purple.

	Method	Completion	Grasp success	Num. of Actions
22 Hard	DIPN+GN (ours)	100%	97.0%	4.78
	VFT (ours)	**100**%	**98.5**%	**2.65**
4 Hard	go-PG [2]	95.0%	86.6%	4.62
	DIPN+GN (ours)	100%	100%	4.00
	VFT (ours)	**100**%	**100**%	**2.60**

TABLE III: Real system metric mean for hard instances.

We also explored our methods on unseen daily-life objects

such as soapboxes, water bottles. The result is consistent with our trained objects as segmentation can be obtained accurately.

REFERENCES

[1] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4238–4245. IEEE, 2018.

[2] Kechun Xu, Hongxiang Yu, Qianen Lai, Yue Wang, and Rong Xiong. Efficient learning of goal-oriented push-grasping synergy in clutter. *arXiv preprint arXiv:2103.05405*, 2021.

[3] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[4] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *CoRR*, abs/1612.03144, 2016.

[5] Lin Yen-Chen, Andy Zeng, Shuran Song, Phillip Isola, and Tsung-Yi Lin. Learning to see before learning to act: Visual pre-training for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7286–7293. IEEE, 2020.

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[7] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. Visual interaction networks: Learning a physics simulator from video. In *Advances in neural information processing systems*, pages 4539–4547, 2017.