Mastering Robot Control through Point-based Reinforcement Learning with Pre-training

Extended Abstract

Yihong Chen*
Tsinghua University
chenyiho21@mails.tsinghua.edu.cn

Cong Wang*†
Fuxi Robotics in Netease
wangcong09@corp.netease.com

Tianpei Yang
University of Alberta
tpyang@tju.edu.cn

Meng Wang
Fuxi Robotics in Netease
wangmeng02@corp.netease.com

Yingfeng Chen
Fuxi Robotics in Netease
chenyingfeng1@corp.netease.com

Jifei Zhou
Fuxi Robotics in Netease
zhou_jf@zju.edu.cn

Chaoyi Zhao
Netease Fuxi AI Lab
zhaochaoyi@corp.netease.com

Xinfeng Zhang
Netease Fuxi AI Lab
zhangxinfeng01@corp.netease.com

Zeng Zhao
Netease Fuxi AI Lab
zengzhao_wl@163.com

Changjie Fan
Fuxi Robotics in Netease
fanchangjie@corp.netease.com

Zhipeng Hu
Fuxi Robotics in Netease
zphu@corp.netease.com

Rong Xiong
Zhejiang University
rxiong@zju.edu.cn

Long Zeng†
Tsinghua University
zenglong@sz.tsinghua.edu.cn

ABSTRACT

Visual-based Reinforcement Learning (RL) has gained prominence in robotics decision-making due to its significant potential. However, the prevalent utilization of images in visual-based RL lacks explicit descriptions of object structures and spatial configurations in scenes, thereby limiting the overall efficiency and robustness of RL in robot control. Additionally, training an RL policy solely using visual observations from scratch is typically sample-inefficient, rendering it impractical for real-world application. To address these challenges, this paper proposes a novel method, called Pre-training on Point-based RL (P2RL), which takes the point cloud representations of scenes as states and preserves the intricate spatial details between objects. To further enhance efficiency, we leverage the pre-training method to bolster the perception ability of the network. Key factors in the pre-training process are systematically examined to optimize downstream RL training. Experimental results demonstrate the superior robustness and efficiency of P2RL compared to the state-of-the-art image-based RL method, especially in evaluations involving untrained scenes.

KEYWORDS

Reinforcement Learning; Pre-training; Robotics

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

ACM Reference Format:
Yihong Chen*, Cong Wang*†, Tianpei Yang, Meng Wang, Yingfeng Chen, Jifei Zhou, Chaoyi Zhao, Xinfeng Zhang, Zeng Zhao, Changjie Fan, Zhipeng Hu, Rong Xiong, and Long Zeng†. 2024. Mastering Robot Control through Point-based Reinforcement Learning with Pre-training: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Recently, Reinforcement Learning (RL) has been widely used in robotics scenes [1, 9]. Most of these methods utilize images as input [6, 12, 13] to actively perceive and gather environmental information, as images are ease of acquisition and integration into the RL pipeline. However, these image-based states fail to capture the structural and spatial information of objects and scenes, which are crucial for complex tasks. To address this limitation, point-based RL has been proposed and studied [7, 8]. However, the point cloud state consists of numerous points, making it challenging to process and further decreasing RL sample efficiency.

To enhance the sample efficiency of point-based RL in a general manner, we propose leveraging pre-training methods from the computer vision field. Our method decouples the framework into the RL Network and the Visual Perception (VP) Network. Firstly, the VP Network undergoes pre-training to enhance its perception capabilities. Subsequently, the pre-trained VP Network extracts features from the point cloud, which serves as the state representation for the RL Network. It is important to note that the policy training

* Both authors contributed equally to this work.
† Corresponding authors.

is end-to-end. We conducted experiments on various manipulation tasks, and the results demonstrate that P2RL outperforms the RL policies trained using image or vector inputs.

2 METHODOLOGY

We model the vision-based manipulation task as a Markov decision process (MDP): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where $s \in \mathcal{S}$ is a state, $a \in \mathcal{A}$ is an action, $r \in \mathcal{R}$ is the reward, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition function, and $\gamma \in [0, 1)$ is the discount factor. P2RL adopts point cloud observation $o \in \mathcal{O}$ as the substitute for the RL input s. During training, the agent aims to learn a policy π that maximizes the discounted cumulative rewards on $\mathcal{M}$: $J(\pi) = \mathbb{E}_{\pi,\mathcal{M}} \sum_{t=1}^{T} \left[\gamma^{t-1} r_t\right]$.

Overall Architecture. P2RL comprises two parts: the VP Network and the RL network. The VP Network extracts the scene state I_s from the point cloud observation o. Subsequently, the proprioceptive state I_m of the manipulator is concatenated with I_s to form the final state $I = [I_s, I_m]$, which serves as the input for the RL network to train the downstream tasks. The Proximal Policy Optimization (PPO) algorithm [11] is employed in the RL network to train the policy. To enhance the stability and speed, we froze the parameters of the Batch Normalization (BN) layer, implemented a distributed training and sampling framework for policy derivation, and incorporated gradient accumulation and automatic mixed precision techniques.

Pre-training Paradigm. During the pre-training phase, we adapt the backbone in the pre-training framework to align with the VP Network used in P2RL, ensuring that the pre-trained parameters can be loaded for subsequent RL training. To explore the impact of different pre-training methods on downstream RL training, we selected two pre-training tasks. The first task is unsupervised contrastive learning, limited only by contrastive loss, and does not require labels during training. To perform this pre-training task, we utilized the STRL framework [5]. The second pre-training task is semantic segmentation, which requires identifying different categories of points in the scene under the supervision of point-wise semantic labels, resulting in a more comprehensive understanding of the scene. We adopt the semantic segmentation process from the PyTorch implementation of PointNet++ as the pre-training framework. At the end of pre-training, the backbone parameters are saved as the pre-training model, which is loaded to initialize the VP Network for downstream RL training.

3 EXPERIMENT

We evaluate P2RL on six manipulation tasks from Robosuite [14]. To conduct pre-training, the ShapeNet [3] and Stanford Large-Scale 3D Indoor Spaces (S3DIS) [2] datasets are selected for contrastive learning and semantic segmentation, respectively. The results of the network initialized with model pre-training through **C**ontrastive **L**earning and **S**emantic **S**egmentation are abbreviated to CL and SS, respectively.

Point-based RL Results. To investigate the performance of **P2RL**, we compare the results of **P2RL** with two state-of-the-art RL methods in Robosuite: **RAPS** [4] and **MAPLE** [10]. RAPS leverages parameterized actions to learn a high-level policy with sparse rewards, using images as input. To assess the disparity between

Table 1: Success rates and standard deviations with different state-of-the-art methods (%).

	Reach	Lift	LiftMulti
RAPS	96.0±8.0	82.0±14.0	-
MAPLE	**100.0±0.0**	98.8±0.4	-
P2RL(ours)	**100.0±0.0**	97.0±6.4	77.0±16.2
	Door	Cleanup	Peg in hole
RAPS	96.0±8.0	-	-
MAPLE	97.2±0.7	**96.6±1.0**	**92.8±2.1**
P2RL(ours)	**100.0±0.0**	89.0±9.4	88.0±7.5

P2RL and the vector-based method (directly obtaining state from the environment), we also compare with the manipulation primitive-augmented RL (MAPLE) method. MAPLE enhances standard RL algorithms with a pre-defined library of behavior primitives. Table 1 demonstrates that P2RL exhibits a 7.6% higher average success rate than RAPS across the initial three tasks, thereby showcasing the superior performance of P2RL in visual-based methods. Regrettably, RAPS is not compatible with the remaining tasks. When comparing P2RL to MAPLE, P2RL achieves comparable results in most tasks. MAPLE only outperforms P2RL by 2.6% in average success, which is acceptable considering the relative ease of learning vector-based RL. Furthermore, the action space of MAPLE is task-constrained, reducing the difficulty of the learning process.

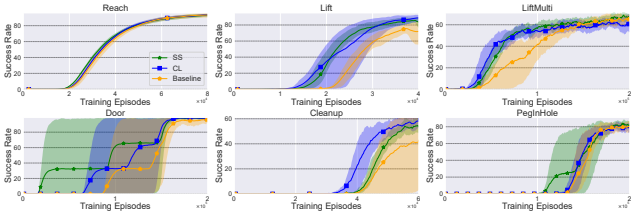

Figure 1: The RL training performance results in various downstream tasks are analyzed concerning different pre-training manners.

Pre-training Results. We examined the impact of pre-training using two distinct approaches on six diverse downstream tasks, as illustrated in Fig. 1. The pre-trained network consistently outperforms the baseline in all downstream tasks. Notably, pre-training significantly enhances the convergence speed of RL training, primarily during the initial stages. The pre-trained network exhibits significant improvements of 200%, 150%, and 150% for the Lift, LiftMulti, and Cleanup tasks, respectively. For the Door task, the pre-trained network achieves success in significantly fewer episodes on average compared to the baseline.

ACKNOWLEDGMENTS

This work is supported by the Key Research and Development Program of Zhejiang Province (No. 2022C01011) and the Guangdong Natural Science Fund-General Programme (No. 2022A1515011234).

REFERENCES

[1] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research* 39, 1 (2020), 3–20.

[2] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1534–1543.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. 2015. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012* (2015).

[4] Murtaza Dalal, Deepak Pathak, and Russ R Salakhutdinov. 2021. Accelerating robotic reinforcement learning via parameterized action primitives. *Advances in Neural Information Processing Systems* 34 (2021), 21847–21859.

[5] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. 2021. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6535–6545.

[6] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. 2018. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*. PMLR, 651–673.

[7] Kenzo Lobos-Tsunekawa and Tatsuya Harada. 2020. Point cloud based reinforcement learning for sim-to-real and partial observability in visual navigation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 5871–5878.

[8] Qingkai Lu, Yifan Zhu, and Liangjun Zhang. 2022. Excavation Reinforcement Learning Using Geometric Representation. *IEEE Robotics and Automation Letters* 7, 2 (2022), 4472–4479.

[9] Jeffrey Mahler, Florian T Pokorny, Brian Hou, Melrose Roderick, Michael Laskey, Mathieu Aubry, Kai Kohlhoff, Torsten Kröger, James Kuffner, and Ken Goldberg. 2016. Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 1957–1964.

[10] Soroush Nasiriany, Huihan Liu, and Yuke Zhu. 2022. Augmenting reinforcement learning with behavior primitives for diverse manipulation tasks. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 7477–7484.

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[12] Bohan Wu, Suraj Nair, Li Fei-Fei, and Chelsea Finn. 2021. Example-driven model-based reinforcement learning for solving long-horizon visuomotor tasks. *arXiv preprint arXiv:2109.10312* (2021).

[13] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. 2021. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15222–15232.

[14] Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. 2020. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293* (2020).