RISE: 3D Perception Makes Real-World Robot Imitation Simple and Effective

Chenxi Wang[1], Hongjie Fang[2], Hao-Shu Fang[2], Cewu Lu[2]

***Abstract*— Precise robot manipulations require rich spatial information in imitation learning, which remains a challenge in both 2D and 3D based policies. To tackle this problem, we present RISE, an end-to-end baseline for real-world imitation learning, which predicts continuous actions directly from single-view point clouds. It compresses the point cloud to tokens with a sparse 3D encoder. After adding sparse positional encoding, the tokens are featurized using a transformer. Finally, the features are decoded into robot actions by a diffusion head. Trained with 50 demonstrations for each real-world task, RISE surpasses currently representative 2D and 3D policies by a large margin, showcasing significant advantages in both accuracy and efficiency. Project website: rise-policy.github.io.**

I. Introduction

Recent work has made significant strides in imitation learning in an end-to-end fashion [1, 3, 5, 6, 23], which opens new possibilities for addressing complex manipulation tasks and drives research in the field of manipulation [16].

Spatial information is crucial for precise manipulations. Image-based imitation learning tends to learn implicit spatial representations from fixed camera views [1, 3, 5, 10, 19, 23]. Many of these approaches utilize distinct image encoders for each view and increase the number of cameras to enhance stability and precision, consequently increasing the number of network parameters and computational overhead.

Recently, imitation learning based on point clouds is drawing increasing interest in our community [2, 7, 8, 9, 13, 17, 21, 22]. Most of the 3D-based methods learn to predict the next keyframe as opposed to continuous actions, which often struggle with tasks involving frequent contacts and abrupt environmental changes. Meanwhile, addressing the annotation of keyframes at scale for real-world data necessitates additional manual effort.

In this work, we propose an end-to-end imitation baseline, **RISE**, a method leveraging 3D perception to make real-world robot imitation simple and effective. RISE takes point clouds from a single-view RGB-D camera as input directly, and outputs continuous action trajectories.

We test RISE in 6 real-world tasks, where all the objects are randomly arranged throughout the entire workspace. Trained on 50 demonstrations for each task, RISE significantly outperforms other representative methods and keeps stable when the number of objects increases. We also find that RISE is more robust to environmental disturbance, which enhances the error tolerance of real-world deployment.

[1] Shanghai Noematrix Intelligence Technology Ltd.
[2] Shanghai Jiao Tong University.
Cewu Lu is the corresponding author.
Author e-mails: chenxi.wang@noematrix.cn, galaxies@sjtu.edu.cn, fhaoshu@gmail.com, lucewu@sjtu.edu.cn

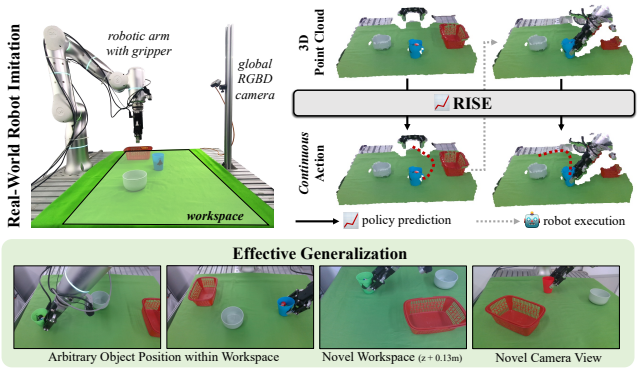


Fig. 1: RISE focuses on real-world robot imitation settings with a noisy single-view partial point cloud as input, and outputs continuous robot actions. While simple, it shows effective generalization ability across object locations, novel workspaces and camera views.

II. Method

Given a point cloud $\mathcal{O}^t$ as the observation at time t, RISE aims to predict the next n-step robot actions $\mathcal{A}^t$, which contain the translations, rotations and widths of the gripper. RISE is decomposed into three functions: a sparse 3D encoder $h_{\mathrm{E}} : \mathcal{O}^t \to \mathcal{F}_{\mathrm{P}}^t$, a transformer $h_{\mathrm{T}} : \mathcal{F}_{\mathrm{P}}^t \to \mathcal{F}_{\mathrm{A}}^t$ and an action decoder $h_{\mathrm{D}} : \mathcal{F}_{\mathrm{A}}^t \to \mathcal{A}^t$, where $\mathcal{F}_{\mathrm{P}}^t$ and $\mathcal{F}_{\mathrm{A}}^t$ denote the features of point clouds and actions respectively.

A. Modeling Point Clouds using Sparse 3D Encoder

The sparse 3D encoder h_{E} adopts a shallow ResNet architecture [11] built on sparse convolution [4]. Such design saves computation and inherits the core advantage of conventional convolution. By h_{E}, the voxelized point cloud $\mathcal{O}^t$ is encoded to sparse point features $\mathcal{F}_{\mathrm{P}}^t$ in an efficient way, avoiding redundant computing on huge empty space. $\mathcal{F}_{\mathrm{P}}^t$ is then fed into the transformer h_{T} as sparse tokens.

B. Transformer with Sparse Point Tokens

We adopt transformer [20] to implement the mapping from point features $\mathcal{F}_{\mathrm{P}}^t$ to action features $\mathcal{F}_{\mathrm{A}}^t$. Sparse positional encoding is employed for point tokens. Let (x, y, z) be the coordinate of the point token P with d-dimension, the position of P is defined as $pos = [pos_x, pos_y, pos_z]$ with

$$pos_k = \frac{k}{v} + c, \quad k \in \{x, y, z\} \tag{1}$$

where c and v are fixed offsets, and $[\cdot]$ stands for vector concatenation. The encoding dimension along each axis is $d_x = d_y = \lfloor d/3 \rfloor$, $d_z = d - d_x - d_y$. The position encoding

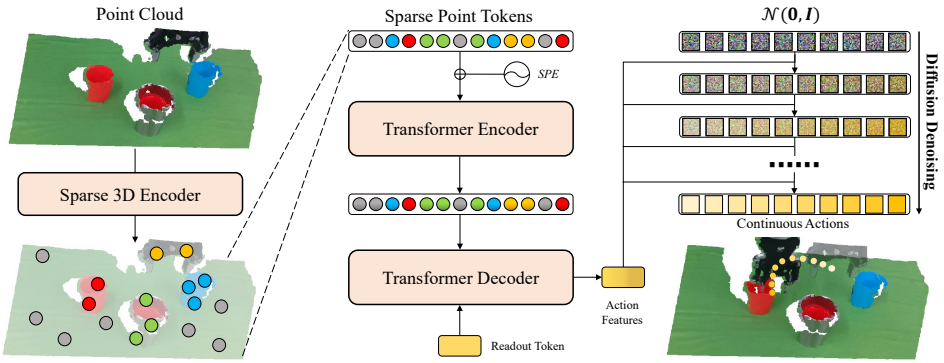


Fig. 2: Overview of RISE architecture. The input of RISE is a noisy point cloud captured from the real world. A 3D encoder built with sparse convolution is employed to compress the point cloud into tokens. The tokens are fed into the transformer encoder after adding sparse positional encoding. A readout token is used to query the action features from the transformer decoder. Conditioned on the action features, the Gaussian samples are denoised into continuous actions iteratively using a diffusion head.

of P is computed by $SPE = [SPE^x, SPE^y, SPE^z]$ where

$$\begin{cases} SPE^k_{(pos,2i)} = \sin \dfrac{pos_k}{10000^{2i/d_k}} \\ SPE^k_{(pos,2i+1)} = \cos \dfrac{pos_k}{10000^{2i/d_k}} \end{cases}, \quad k \in \{x, y, z\} \quad (2)$$

With the help of sparse positional encoding, we effectively capture intricate 3D spatial relationships among unordered points, which enables seamless embedding of the 3D features into conventional transformers.

C. Diffusion as Action Decoder

The action decoder h_D is implemented as a denoising process by diffusion [3, 12, 15]. Conditioning on $\mathcal{F}^t_{\mathrm{A}}$, h_{D} denoises the Gaussian noises $\mathcal{N}(0, \sigma^2 I)$ to actions $\mathcal{A}^t$ iteratively. We use the DDIM scheduler [18] to accelerate the inference speed in real-world experiments.

RISE adopts a unified action representation in camera coordinates which is composed of translations, rotations, and gripper widths. We opt for absolute position for translation and 6D representation [24] for rotation.

III. Experiments

A. Setup

Tasks. We designed 6 tasks for the experiments in Fig. 3.

Hardware. We use a Flexiv Rizon robotic arm with a Dahuan AG-95 gripper for interacting with objects. Two Intel RealSense D435 RGB-D cameras are installed for scene perception (One global, one inhand).

Baselines. We employ two representative image-based policies as our baselines: ACT [23] and Diffusion Policy [3]. We also evaluate a keyframe-based 3D policy Act3D [7], the current state-of-the-art policy on RLBench [14].

Protocols. For 3D perception, only the global camera is used to generate a noisy single-view partial point cloud; while for image-based policies, both cameras are used for a better understanding of spatial geometries. We gathered 50 expert demonstrations for each task for training, and each policy was tested for 20 consecutive trials. During evaluations, objects in the task are randomly initialized within the robot workspace of approximately 50cm × 70cm. The evaluation time limit for each task is sufficient for each method to accomplish the task.

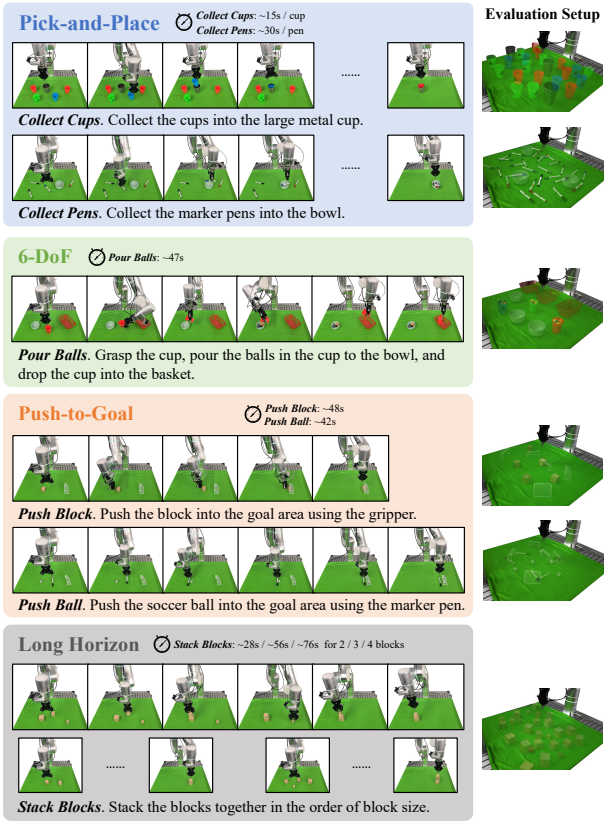


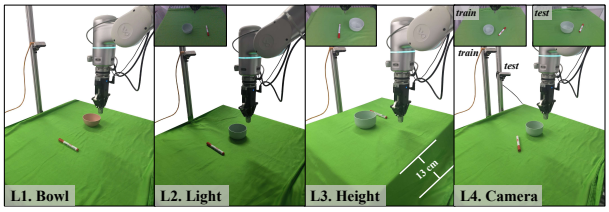
Fig. 3: Definition of the tasks in the experiments. During evaluation, each task is randomly initialized within the robot workspace. For each task, only 3 to 5 evaluation setups are depicted for clarity.

B. Results

Pick-and-place tasks are crucial in robotics, focusing on precise object manipulations and efficient policy generalization. The evaluation in Fig. 4 for ***Collect Cups*** and ***Collect Pens*** reveals RISE consistently outperforming all baselines, demonstrating its ability to not only predict the translation part but also accurately forecast planner rotation. We also discover that Act3D performs comparably to image-based baselines. Moreover, given that Act3D requires specially designed motion planners for more complicated actions and cannot provide immediate responses to sudden changes in the

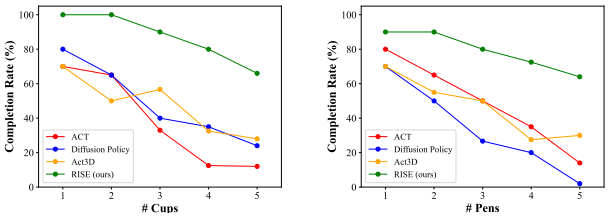
Method	*Pour Balls*					*Push Block*	*Push Ball*	*Stack Blocks*		
	SR (%)			CR (%)		SR (%)	SR (%)	CR (%)		
	Grasp	Pour	Place	Overall	If Poured			1 block	2 blocks	3 blocks
ACT [23]	30	30	0	13.0	43.3	-	-	60.0	25.0	10.0
Diffusion Policy [3]	55	55	35	30.5	55.5	50	30	70.0	25.0	16.7
RISE (ours)	**80**	**80**	**70**	**49.0**	**61.3**	**55**	**60**	**80.0**	**75.0**	**30.0**

TABLE I: Experimental results of the ***Pour Balls***, ***Push Block***, ***Push Ball*** and ***Stack Blocks*** task, where SR denotes success rate and CR represents completion rate.



Method	Completion Rate (%)				
	Original	Disturbance			
		Bowl	Light	Height	Camera
ACT [23]	80	70 ↓10	40 ↓40	0 ↓80	0 ↓80
Diffusion Policy [3]	70	50 ↓20	30 ↓40	0 ↓70	0 ↓70
Act3D [7]	70	40 ↓30	60 ↓10	50 ↓20	10 ↓60
RISE (ours)	**90**	**80** ↓10	**80** ↓10	**80** ↓10	**50** ↓40

TABLE II: Generalization test setup and experimental results of the ***Collect Pens*** task with 1 pen (10 trials).



(a) ***Collect Cups*** **(b)** ***Collect Pens***

Fig. 4: Experimental results of the *pick-and-place* tasks.

Method	3D	# Cameras	Completion Rate (%)
ACT [23]		2	12
	✓	1	32 ↑20
Diffusion Policy [3]		2	24
	✓	1	36 ↑12
Act3D [7]	✓	1	28
RISE (ours)	✓	1	**66**

TABLE III: Effectiveness test of 3D perception on the ***Collect Cups*** task with 5 cups (10 trials).

environment, we therefore only employ ACT and Diffusion Policy as baselines in our subsequent experiments.

The *6-DoF* ***Pour Balls*** task assesses robot policies' capability in forecasting actions involving complex spatial rotations, unlike the simpler planner rotations in *pick-and-place* tasks. Tab. I presents the experimental results. RISE demonstrates superior learning of actions with intricate spatial rotations compared to image-based policies, as evidenced by higher action success rates. Moreover, its precision in pouring positions leads to increased task completion rates, highlighting the effectiveness of 3D perception in capturing accurate spatial object relationships.

For effective task completion, robot policies must promptly respond to environmental changes and adapt to object movements. We designed *push-to-goal* tasks, ***Push Block*** and ***Push Ball*** (Fig. 3), to assess this ability. Evaluation results in Tab. I show RISE slightly surpassing Diffusion Policy in the ***Push Block*** task, while significantly outperforming Diffusion Policy in the ***Push Ball*** task, demonstrating its adeptness in 3D perception for object positioning adjustments and swift policy action modifications.

Long-horizon tasks are essential in robotics, revealing how errors accumulate over extended actions and showcasing a policy's robustness and adaptability. Hence, we introduced the ***Stack Blocks*** task to evaluate this aspect, especially since block stacks are prone to toppling as they grow. Tab. I shows the experimental results. Initially, with just two blocks, all policies performed similarly. However, as the block count increased, RISE notably outperformed the baselines, demonstrating its strong adaptability to *long-horizon* tasks and ability to effectively control accumulated errors.

C. Generalization Test

We assess the generalization abilities of different methods using the ***Collect Pens*** task with 1 pen under various environmental disturbances detailed in Tab. II. The results in Tab. II indicate that image-based policies achieve decent L1 and some L2-level generalizations but fall short in L3 and L4-level generalizations involving spatial transformations. Act3D, as a 3D policy, shows good generalization up to L3-level disturbances but struggles significantly in L4-level tests. On the contrary, RISE demonstrates strong generalization across all testing levels, even excelling in the challenging L4-level tests involving camera view changes.

D. Effectiveness of 3D Perception

We explore how 3D perception enhances the performance of policies on the ***Collect Cups*** task with 5 cups. We replace the image encoder of the image-based policies ACT and Diffusion Policy with the sparse 3D encoder used in RISE. We observe a significant improvement after applying 3D perception even with fewer camera views, surpassing the 3D policy Act3D, as shown in Tab. III, which reflects the effectiveness of our 3D perception module.

IV. Conclusion

We present RISE, an efficient end-to-end policy utilizing 3D perception for real-world robot manipulation. RISE significantly outperforms representative 2D and 3D policies in multiple tasks, demonstrating great advantages in both accuracy and efficiency. We hope our baseline inspires the integration of 3D perception into real-world policy learning.

REFERENCES

[1] Anthony Brohan et al. "RT-1: Robotics Transformer for Real-World Control at Scale". In: *Robotics: Science and Systems*. 2023.

[2] Shizhe Chen et al. "PolarNet: 3D Point Clouds for Language-Guided Robotic Manipulation". In: *Conference on Robot Learning*. PMLR. 2023, pp. 1761–1781.

[3] Cheng Chi et al. "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion". In: *Robotics: Science and Systems*. 2023.

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. "4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3075–3084.

[5] Hao-Shu Fang et al. "RH20T: A Robotic Dataset for Learning Diverse Skills in One-Shot". In: *IEEE International Conference on Robotics and Automation (ICRA)*. 2024.

[6] Hongjie Fang et al. "Low-cost exoskeletons for learning whole-arm manipulation in the wild". In: *IEEE International Conference on Robotics and Automation*. 2024.

[7] Théophile Gervet et al. "Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation". In: *Conference on Robot Learning*. PMLR. 2023, pp. 3949–3965.

[8] Ankit Goyal et al. "RVT: Robotic View Transformer for 3D Object Manipulation". In: *Conference on Robot Learning*. PMLR. 2023, pp. 694–710.

[9] Pierre-Louis Guhur et al. "Instruction-Driven History-Aware Policies for Robotic Manipulations". In: *Conference on Robot Learning*. PMLR. 2022, pp. 175–187.

[10] Huy Ha, Pete Florence, and Shuran Song. "Scaling Up and Distilling Down: Language-Guided Robot Skill Acquisition". In: *Conference on Robot Learning*. PMLR. 2023, pp. 3766–3777.

[11] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[13] Stephen James et al. "Coarse-to-Fine Q-Attention: Efficient Learning for Visual Robotic Manipulation via Discretisation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 13739–13748.

[14] Stephen James et al. "RLBench: The Robot Learning Benchmark & Learning Environment". In: *IEEE Robotics and Automation Letters* 5.2 (2020), pp. 3019–3026.

[15] Michael Janner et al. "Planning with Diffusion for Flexible Behavior Synthesis". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 9902–9915.

[16] Rouhollah Rahmatizadeh et al. "Vision-Based Multi-Task Manipulation for Inexpensive Robots using End-to-End Learning from Demonstration". In: *IEEE International Conference on Robotics and Automation*. IEEE. 2018, pp. 3758–3765.

[17] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. "Perceiver-Actor: A Multi-Task Transformer for Robotic Manipulation". In: *Conference on Robot Learning*. PMLR. 2022, pp. 785–799.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. "Denoising Diffusion Implicit Models". In: *The International Conference on Learning Representations*. 2021.

[19] Octo Model Team et al. *Octo: An Open-Source Generalist Robot Policy*. 2023.

[20] Ashish Vaswani et al. "Attention is All You Need". In: *Advances in Neural Information Processing Systems* 30 (2017).

[21] Zhou Xian et al. "ChainedDiffuser: Unifying Trajectory Diffusion and Keypose Prediction for Robotic Manipulation". In: *Conference on Robot Learning*. PMLR. 2023, pp. 2323–2339.

[22] Yanjie Ze et al. "GNFactor: Multi-Task Real Robot Learning with Generalizable Neural Feature Fields". In: *Conference on Robot Learning*. PMLR. 2023, pp. 284–301.

[23] Tony Z Zhao et al. "Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware". In: *Robotics: Science and Systems*. 2023.

[24] Yi Zhou et al. "On the Continuity of Rotation Representations in Neural Networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5745–5753.