7th Conference on Robot Learning (2023)
Atlanta, USA

PolarNet: 3D Point Clouds for Language-Guided Robotic Manipulation

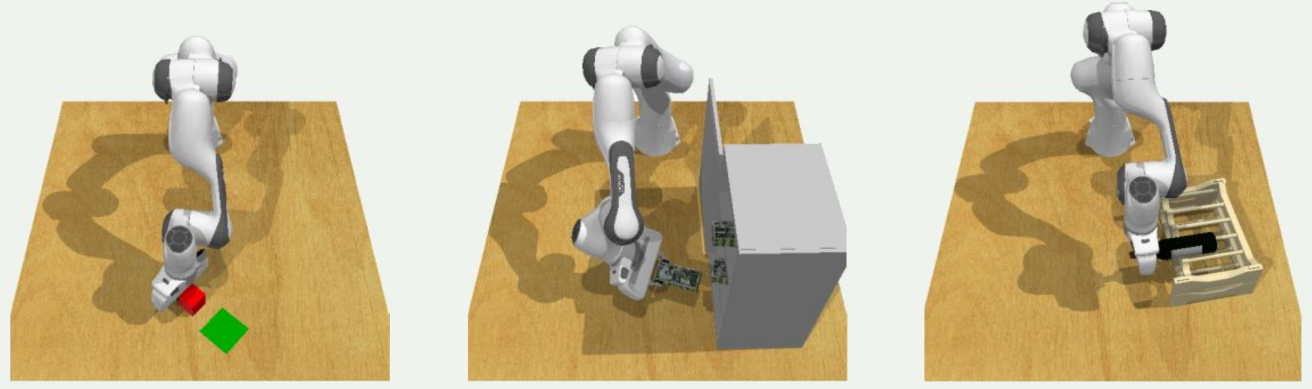
Shizhe Chen, Ricardo Garcia*, Cordelia Schmid, Ivan Laptev*

Inria Paris, École normale supérieure, PSL

Motivation

Goal: Train a robot to follow language instructions to perform various manipulation tasks

Slide block | **Take money** | **Stack wine**



"slide the block towards the green target" | "take the money off of the first shelf of the safe" | "slide the bottle onto the wine rack"

Dominant approaches based on **2D representations**:

- \+ Benefit from pretrained 2D vision models.
- \- Hard to address visual occlusion with multi-view cameras.

We propose using **3D point cloud representations**:

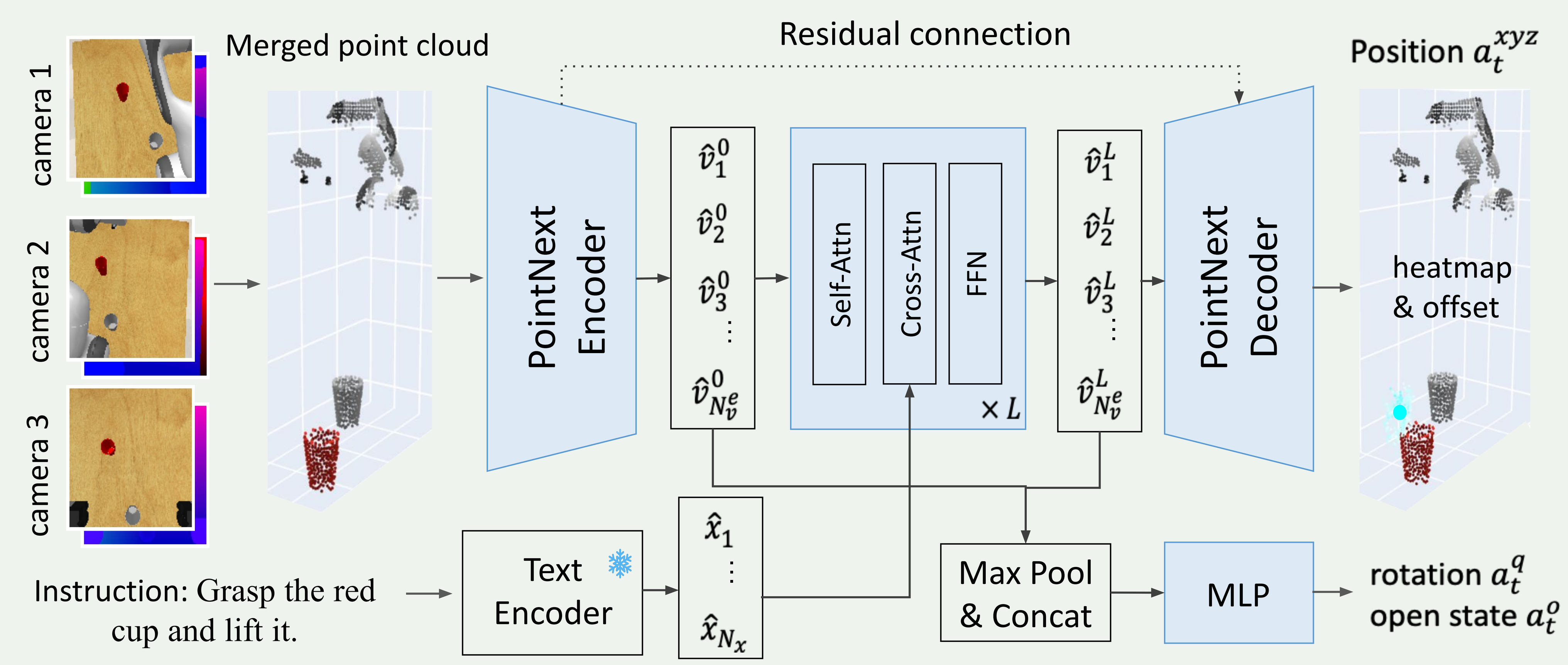
- \+ Natural way to merge multi-view observations.
- \+ Geometric structure: easy to select relevant point via preprocessing.
- \+ Accurate 3D localization.
- \- Need special models to efficiently process them.
- \- Multiple design choices.

Contribution

- Systematically explore the designs of 3D inputs: 3D features, coordinate systems, point removal.
- Efficiently predict 7 DoF actions given the point cloud and instruction using a light-weighted PointNext encoder-decoder and multimodal transformer
- Outperform state-of-the-art methods and achieve promising real world results

di.ens.fr/willow/research/polarnet/

PolarNet architecture



Point cloud design

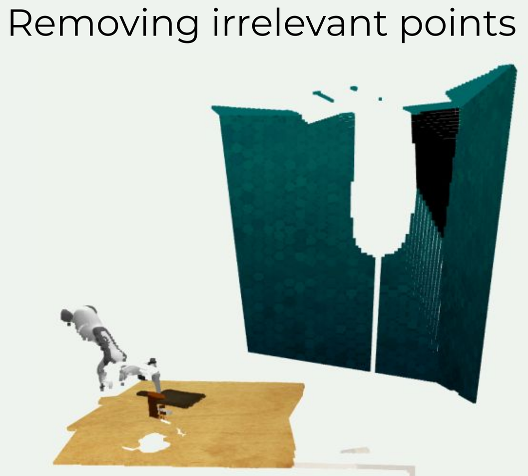
Training: 100 demonstrations per task.
Evaluation: 500 unseen episodes per task.
Metric: Success Rate (SR).

Three setups:
- Single-task (10 / 74 tasks)
- Multi-task (10 / 74 tasks)
- Multi-task multi-variation (18 tasks - 249 variations)

Point cloud preprocessing

Coord origin	Remove Table	Remove Background	Avg.
Center	✓	✓	92.1 ±2.0
Gripper	×	×	81.6 ±3.2
	×	✓	89.9 ±2.8
	✓	✓	**92.1** ±0.4

- Gripper and center coordinate frames perform similarly.
- Removing irrelevant points is highly effective.



Raw point cloud

Background removal

Background and table removal

Camera views

Left	Right	Wrist	Avg.
✓	×	×	37.6 ±4.8
×	✓	×	48.0 ±4.5
×	×	✓	35.0 ±5.5
✓	✓	×	67.0 ±4.7
✓	×	✓	80.2 ±3.0
×	✓	✓	76.6 ±5.6
✓	✓	✓	**92.1** ±0.4

- Single camera insufficient due to occlusions.
- Wrist camera alone performs worst but more complementary to the other two cameras.

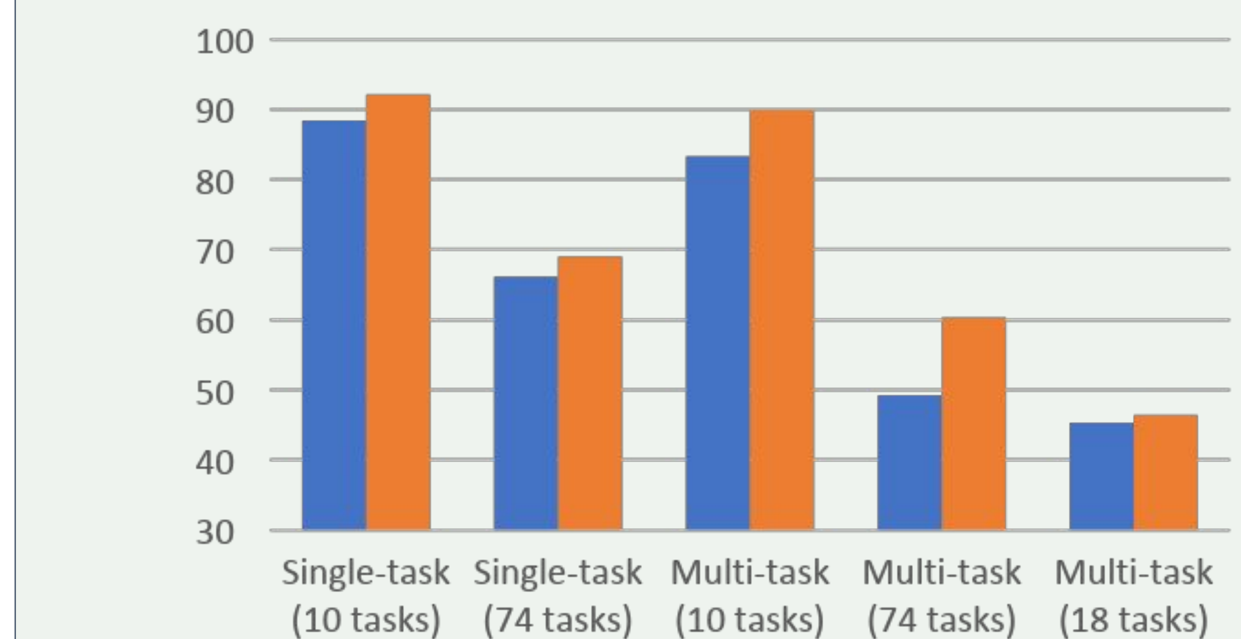
Point cloud representations

RGB	Normal	Height	Avg.
×	×	×	72.1 ±4.4
✓	×	×	91.3 ±1.6
✓	✓	×	90.3 ±3.1
✓	×	✓	91.5 ±1.4
✓	✓	✓	**92.1** ±0.4

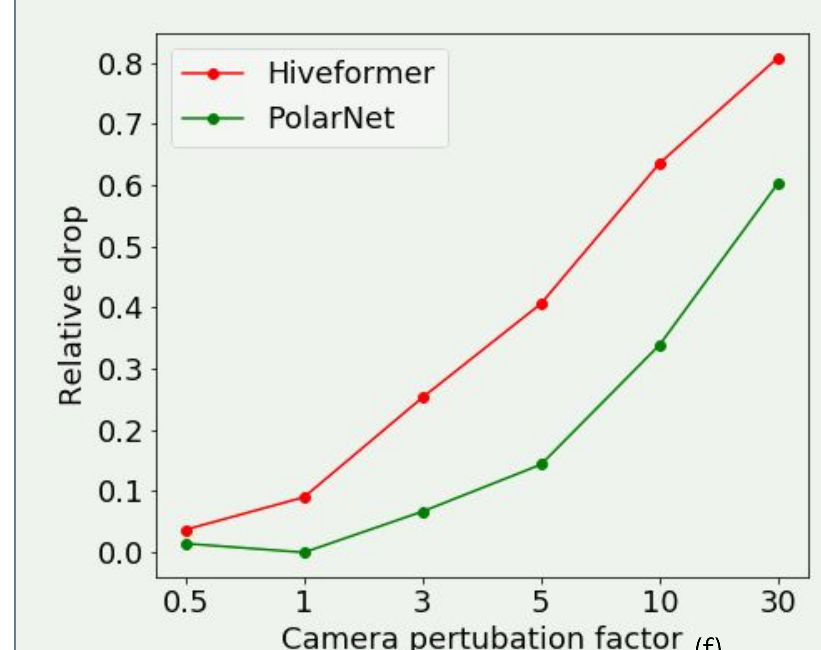
- Vanilla point cloud with only XYZ perform the worst.
- RGB color important to distinguish colors.
- Height relative to the table slightly improve results.
- Improvement from normal is less stable.

State-of-the-art comparison

Comparison to Hiveformer [1] (state-of-the-art method based on 2D representations) :



Robustness of viewpoint variances:



Training
Fixed viewpoints.

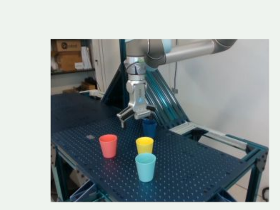
Evaluation
Randomly shifted viewpoints:
- Position: ± f cm
- Rotation: ± 5f degrees

[1] Instruction-driven history-aware policies for robotic manipulations, P.-L. Guhur etc., CoRL 2022

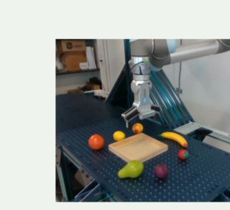
Real robot experiments

Policy pretrained on simulation and finetuned on real robot data. Policy shows promising results on 7 different tasks:

Task	PolarNet
Stack cup	8/10
Put fruit in box	8/10
Put plate on table	3/10
Open drawer	9/10
Put item in drawer	4/10
Put item in cabinet	4/10
Hang mug	6/10
Average	6/10

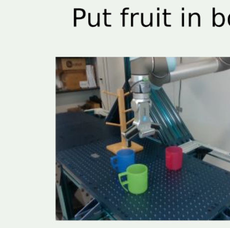


Stack cup



Put fruit in box

Open drawer



Hang mug

Put item in drawer

Put item in cabinet

Put plate on table