O2O-Afford: Annotation-Free Large-Scale Object-Object Affordance Learning

Kaichun Mo[1], Yuzhe Qin[2], Fanbo Xiang[2], Hao Su[2], Leonidas Guibas[1]

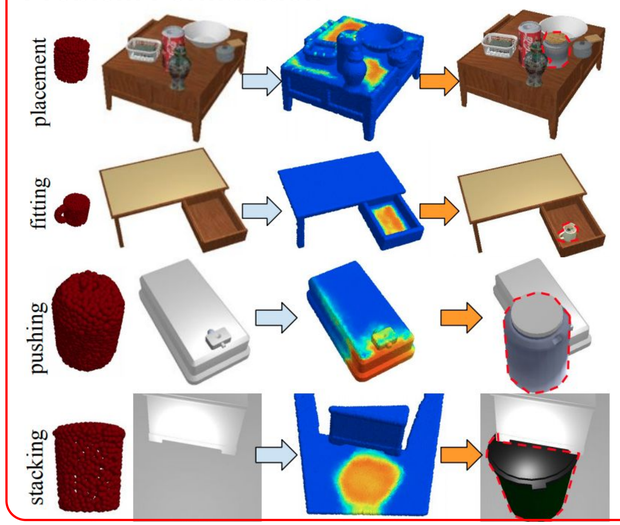
[1] Stanford University [2] UC San Diego

CoRL 2021

Abstract

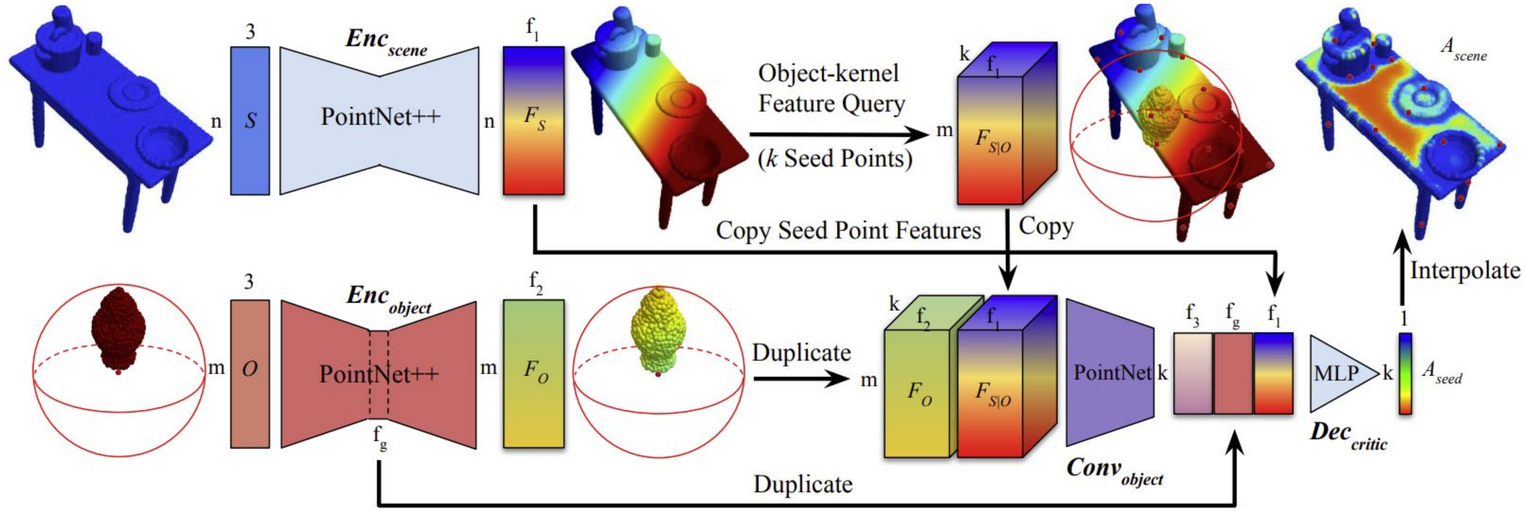
Contrary to the vast literature in modeling, perceiving, and understanding agent-object (e.g., human-object, hand-object, robot-object) interaction in computer vision and robotics, very few past works have studied the task of object-object interaction, which also plays an important role in robotic manipulation and planning tasks. There is a rich space of object-object interaction scenarios in our daily life, such as placing an object on a messy tabletop, fitting an object inside a drawer, pushing an object using a tool, etc. In this paper, we propose a unified affordance learning framework to learn object-object interaction for various tasks. By constructing four object-object interaction task environments using physical simulation (SAPIEN) and thousands of ShapeNet models with rich geometric diversity, we are able to conduct large-scale object-object affordance learning without the need for human annotations or demonstrations. At the core of technical contribution, we propose an object-kernel point convolution network to reason about detailed interaction between two objects. Experiments on large-scale synthetic data and real-world data prove the effectiveness of the proposed approach.

Problem Formulation



*Given a **3D scene** and a **3D object**, **where** could the 3D object **interact** with the 3D scene to accomplish the **task**, defined by **task-specific success metrics**?*

Network Architecture



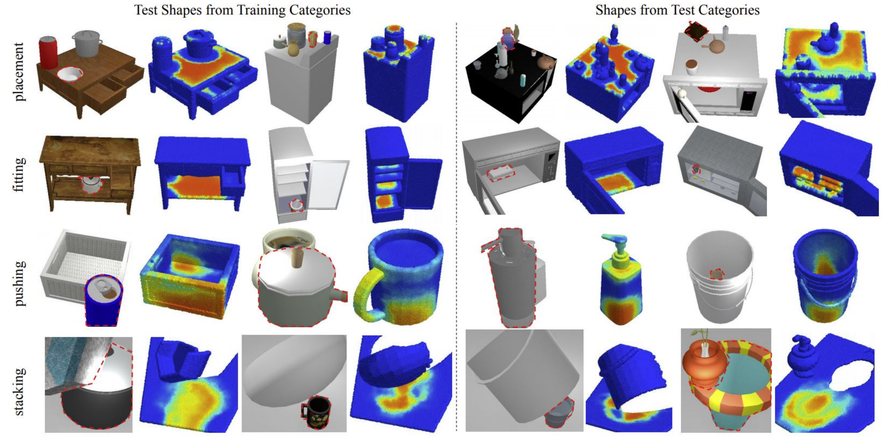
Network Architecture. Taking as inputs a partial 3D scan of the scene S (dark blue) and a complete 3D point cloud of acting object O (dark red), our network learns to extract per-point features on both inputs, correlate the two point cloud feature maps using an object-kernel point convolution, and finally predict a point-wise affordance heatmap over the scene point cloud.

Results and Visualization

- Quantitative Evaluations and Comparisons to Baselines

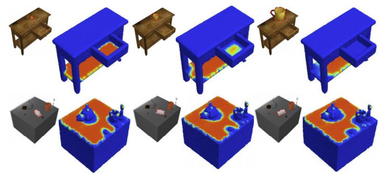
		F-score (%)	AP (%)			F-score (%)	AP (%)
placement	B-PosNor	62.1 / 81.7	60.5 / 78.2	pushing	B-PosNor	31.9 / 34.9	37.0 / 35.5
	B-Bbox	80.9 / **90.6**	90.5 / 94.5		B-Bbox	33.2 / 35.0	39.2 / 37.6
	B-3Branch	63.8 / 77.1	69.8 / 82.3		B-3Branch	35.2 / 36.6	42.2 / 36.4
	Ours	**81.4** / 90.0	**91.1 / 95.2**		Ours	**35.5 / 40.3**	**46.9 / 43.1**
fitting	B-PosNor	45.4 / 59.3	46.8 / 66.7	stacking	B-PosNor	79.3 / 77.9	79.9 / 76.5
	B-Bbox	69.5 / 79.5	80.1 / 80.6		B-Bbox	85.7 / 83.2	87.7 / 87.2
	B-3Branch	48.2 / 56.9	47.1 / 60.7		B-3Branch	87.3 / 84.8	90.8 / 88.2
	Ours	**73.6 / 80.3**	**80.1 / 86.3**		Ours	**89.6 / 87.5**	**91.7 / 90.8**

- Qualitative Results over Synthetic Data



- Qualitative Results over Real-world Data

- Result Analysis



Acknowledgements

This research was supported by NSF grant IIS-1763268, NSF grant RI-1763268, a grant from the Toyota Research Institute University 2.0 program, a Vannevar Bush faculty fellowship, and gift money from Qualcomm. This work was also supported by AWS Machine Learning Awards Program.