O2O-Afford: Annotation-Free Large-Scale Object-Object Affordance Learning
Kaichun Mo¹, Yuzhe Qin², Fanbo Xiang², Hao Su², Leonidas Guibas¹
¹ Stanford University  ² UC San Diego

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.

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

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

<table>
<thead>
<tr>
<th>Task</th>
<th>F-score (%)</th>
<th>AP (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>81.4 / 90.0</td>
<td>91.1 / 95.2</td>
</tr>
<tr>
<td>B-PosNor</td>
<td>78.5 / 77.5</td>
<td>90.5 / 84.2</td>
</tr>
<tr>
<td>B-Bbox</td>
<td>73.6 / 71.4</td>
<td>79.5 / 76.7</td>
</tr>
</tbody>
</table>

- Qualitative Results over Synthetic Data

- Qualitative Results over Real-world Data

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.