Mapping Beyond Geometry

Jie Hu

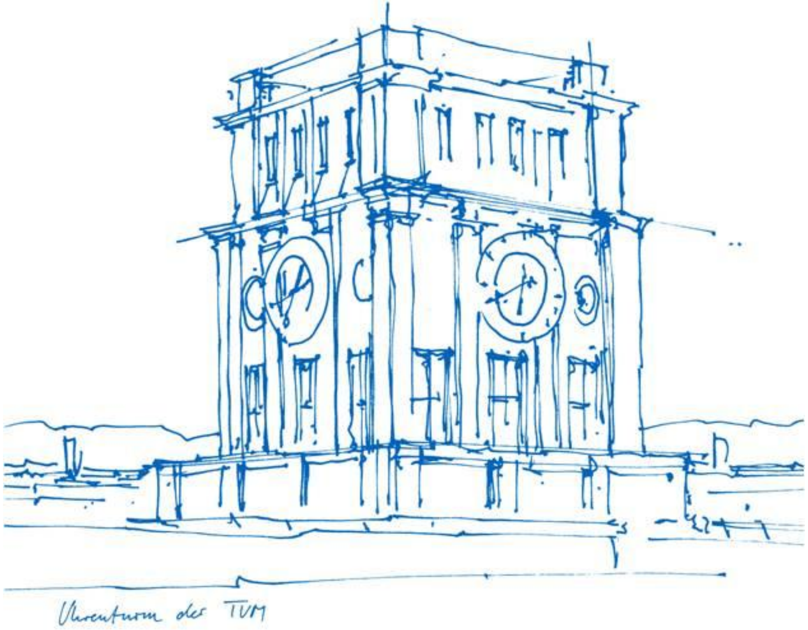
Supervisor: Hanzhi Chen

Technical University of Munich

TUM School of CIT

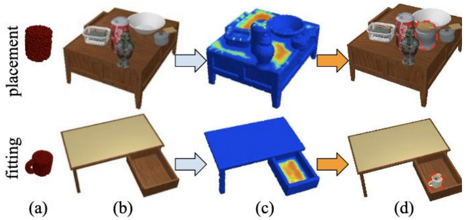
Robot Perception & Intelligence Seminar

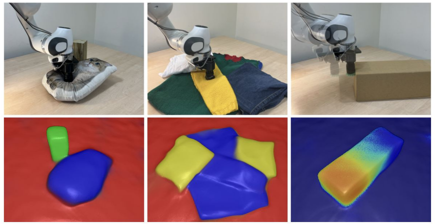
Garching Forschungszentrum, 03. December 2024



Overview

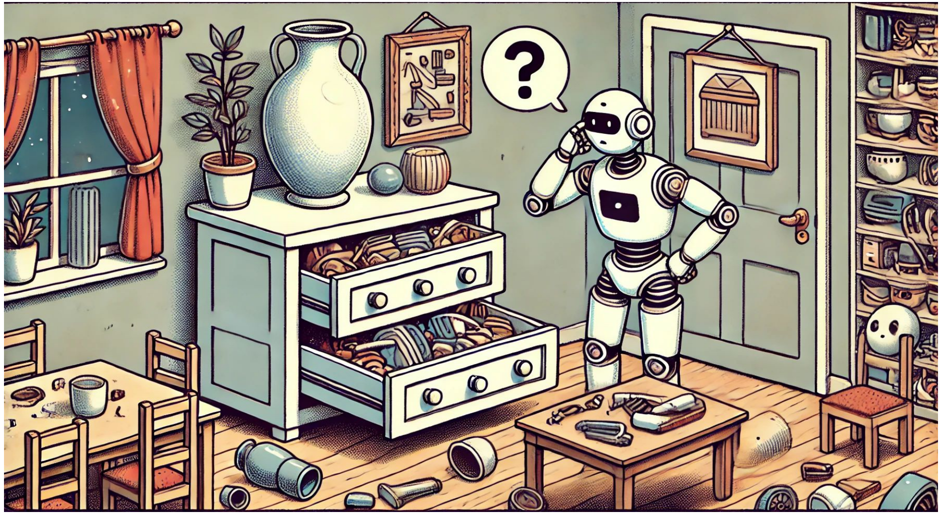
- Motivation
- Related Works
- Discussions
- Future work
- Summary





Rigidity classification | Material classification | Force distribution

Motivation: Mapping Beyond Geometry



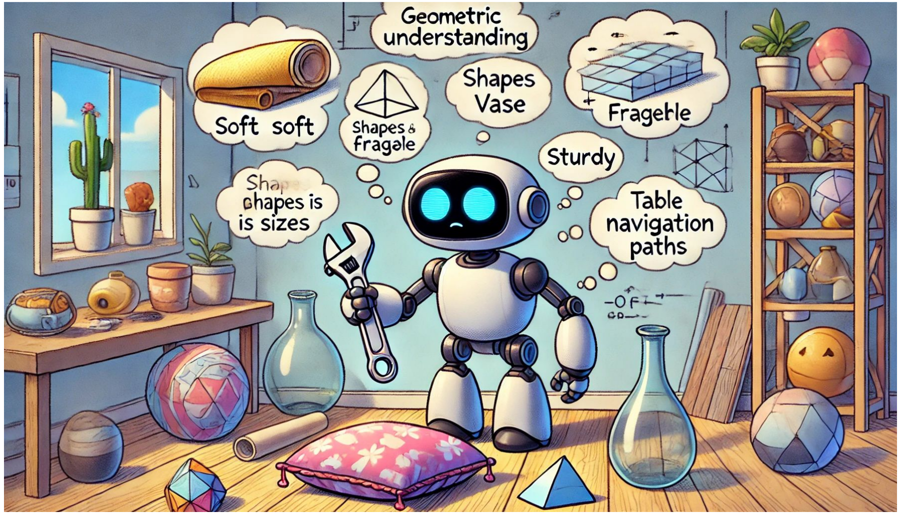
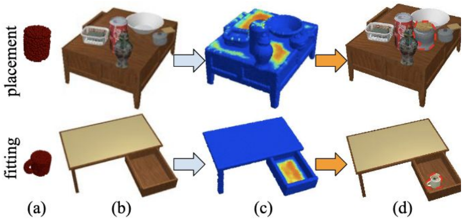


*Image generated with DALL·E by OpenAI

Mapping:

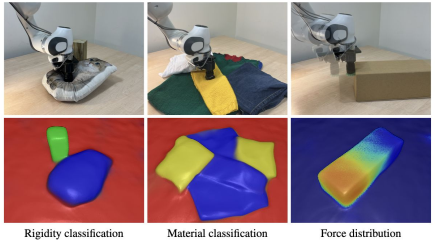
Affordance



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]UCSD

Physical Properties

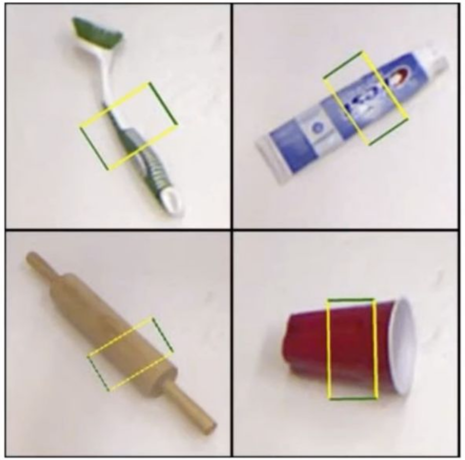


Real-time Mapping of Physical Scene Properties with an Autonomous Robot Experimenter

Iain Haughton[1] Edgar Sucar[2] Andre Mouton[1] Edward Johns[3] Andrew J. Davison[2]
[1] Dyson Technology Ltd.
[2] Dyson Robotics Lab, Imperial College
[3] Robot Learning Lab, Imperial College
iain.haughton@dyson.com

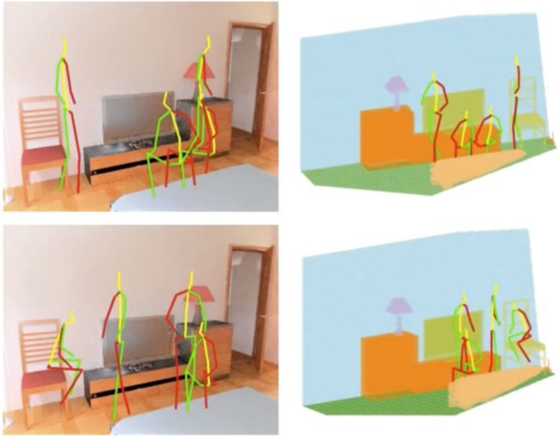
Agent-Object Affordance Learning

Robot-Object



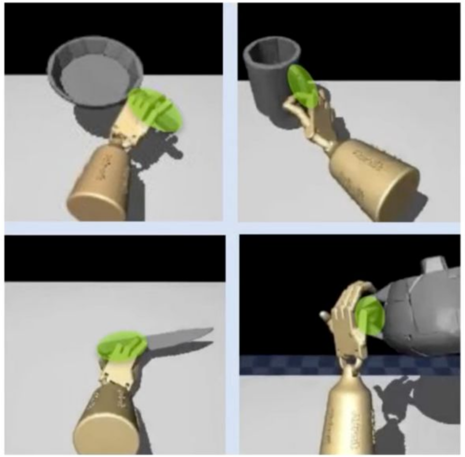
Redmon, Joseph, and Anelia Angelova. "Real-time grasp detection using convolutional neural networks." ICRA 2015

Human-Object



Li, Xueting, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. "Putting humans in a scene: Learning affordance in 3d indoor environments." CVPR 2019

Hand-Object

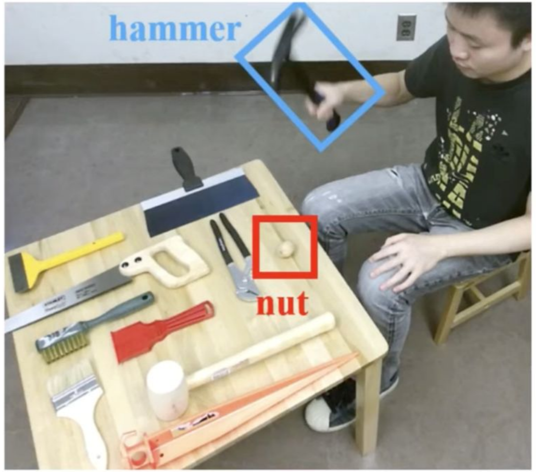


Mandikal, Priyanka, and Kristen Grauman. "Learning Dexterous Grasping with Object-Centric Visual Affordances.", ICRA 2021

Object-Object Affordance Learning

Small-scale & Require Human Annotation or Demonstration

Sun, Yu, Shaogang Ren, and Yun Lin. "Object–object interaction affordance learning." *Robotics and Autonomous Systems*, 2014

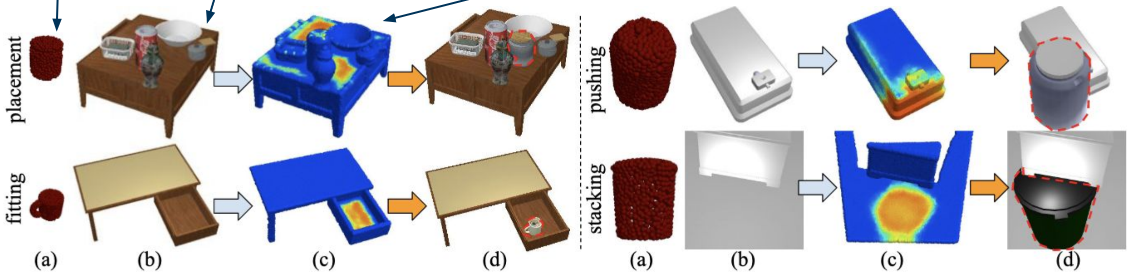


Zhu, Yixin, Yibiao Zhao, and Song Chun Zhu. "Understanding tools: Task-oriented object modeling, learning and recognition." CVPR 2015

O2O-Afford: Annotation-Free Large-Scale Object-Object Affordance Learning (Mo et al., 2021)

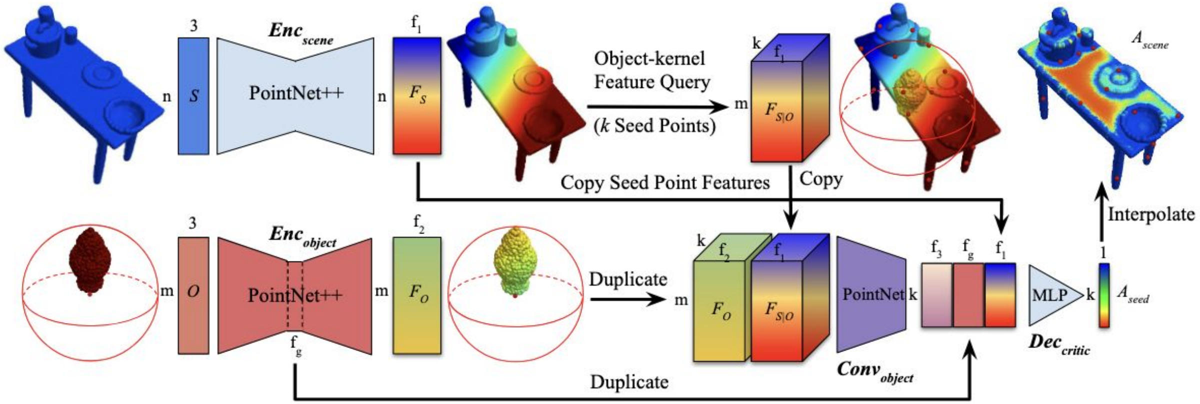
Input: Task (Acting object, Scene)

Output: Object-Object interaction priors **(O2O priors)**



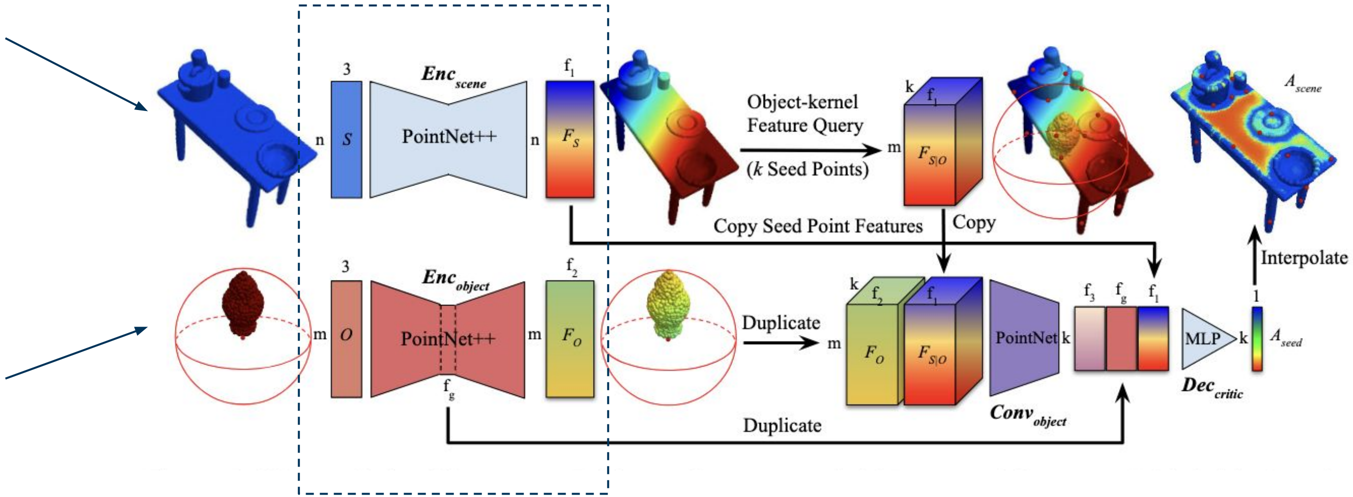
Method

A **Unified Framework** for Diverse Object-object Interaction Tasks



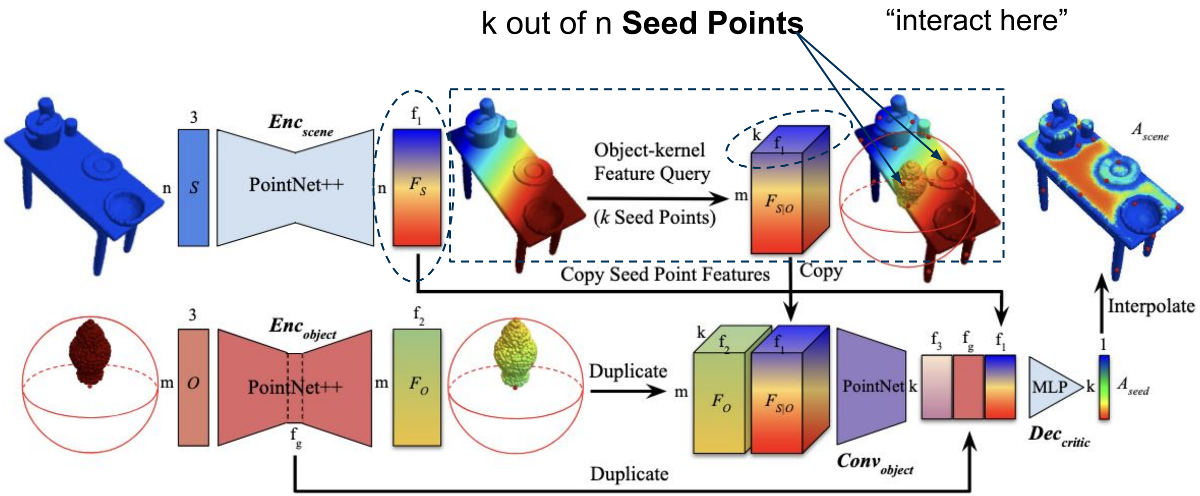
Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

Method



Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

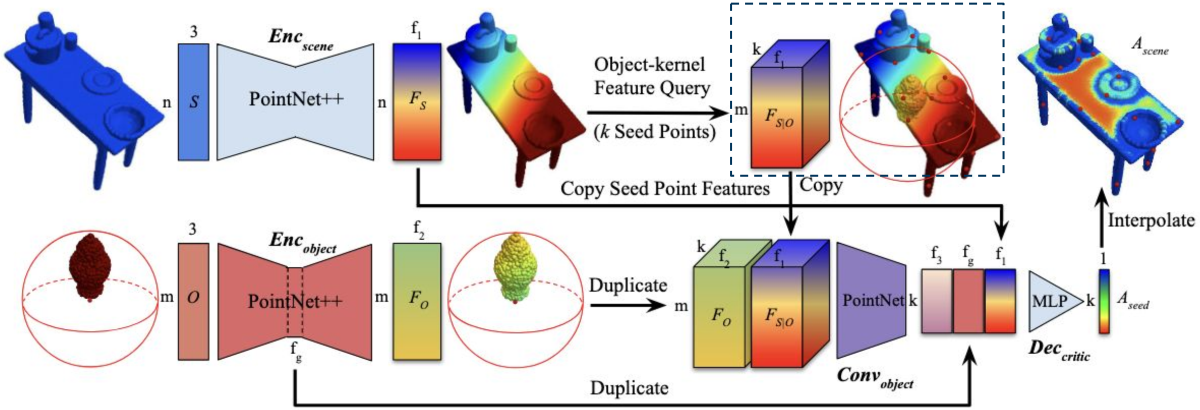
Method



Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

Method

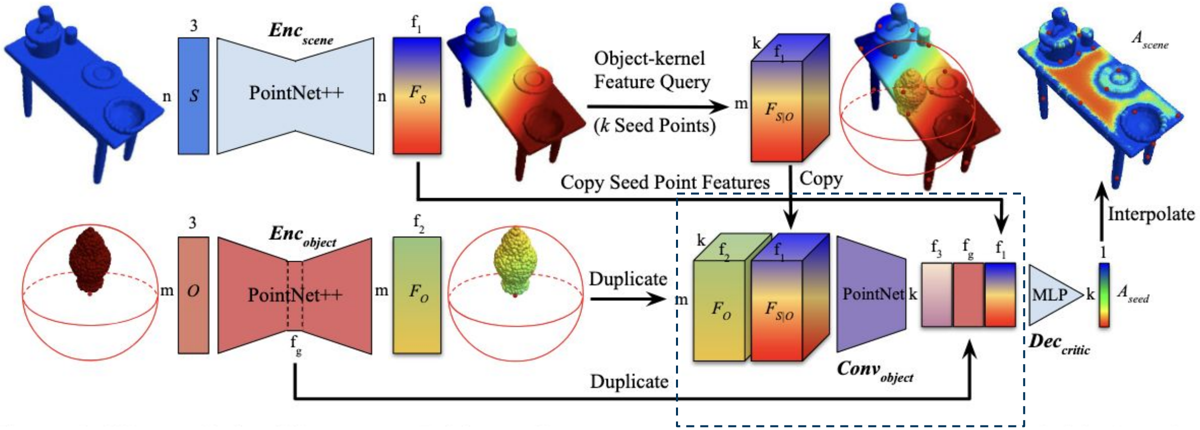
“Object Kernel Point Convolution”



Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

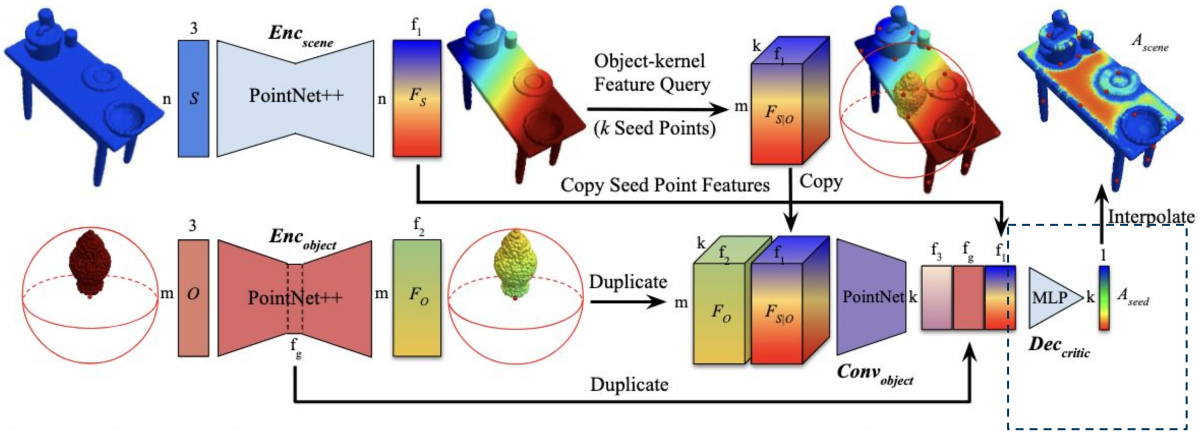
Method

“Object Kernel Point Convolution”



Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

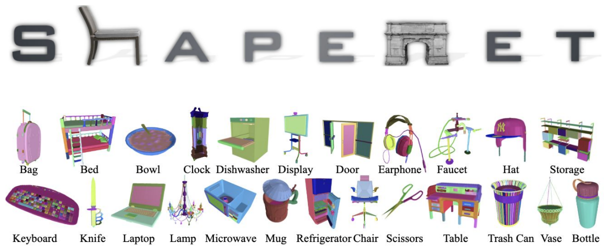
Method



"Final Affordance labeling"

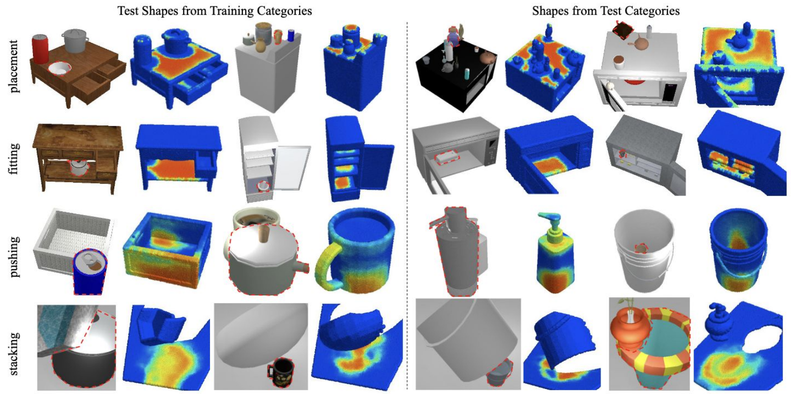
Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

Experiments



- Large-scale object data
- 4 Tasks

Results



Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. CoRL 2021.

Results

		F-score (%)	AP (%)
placement	B-PosNor	62.1 / 81.7	60.5 / 78.2
	B-Bbox	80.9 / **90.6**	90.5 / 94.5
	B-3Branch	63.8 / 77.1	69.8 / 82.3
	Ours	**81.4** / 90.0	**91.1 / 95.2**
fitting	B-PosNor	45.4 / 59.3	46.8 / 66.7
	B-Bbox	69.5 / 79.5	80.1 / 80.6
	B-3Branch	48.2 / 56.9	47.1 / 60.7
	Ours	**73.6 / 80.3**	**80.1 / 86.3**

		F-score (%)	AP (%)
pushing	B-PosNor	31.9 / 34.9	37.0 / 35.5
	B-Bbox	33.2 / 35.0	39.2 / 37.6
	B-3Branch	35.2 / 36.6	42.2 / 36.4
	Ours	**35.5 / 40.3**	**46.9 / 43.1**
stacking	B-PosNor	79.3 / 77.9	79.9 / 76.5
	B-Bbox	85.7 / 83.2	87.7 / 87.2
	B-3Branch	87.3 / 84.8	90.8 / 88.2
	Ours	**89.6 / 87.5**	**91.7 / 90.8**

Conclusion

Strengths & Limitations:

+ Self-supervised -> Annotation-free
+ Simulation-based -> Large-Scale
+ Generalizes

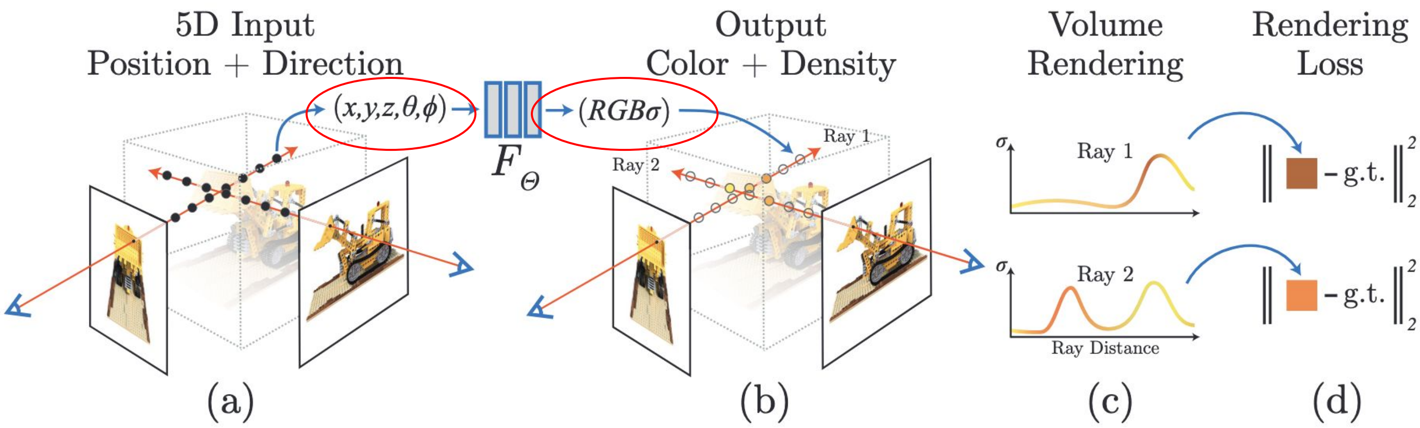
- Task specific
- Only 4 Object-Object interaction included
- Assumes uniform density for all objects

Future Works:

- Joint training across tasks
- Extend to more O2O interactions
- Extract more semantic-rich point features

Detour: implicit mapping

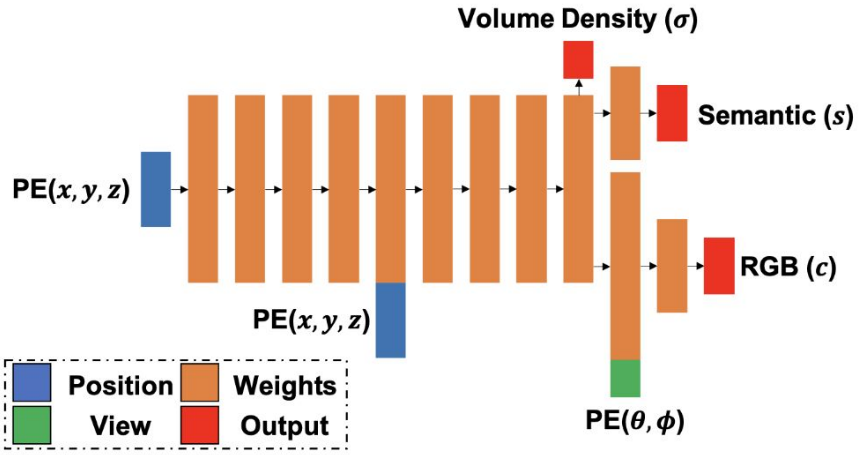
Neural Radiance Field (NeRF) (Ben et al., 2021)



Ben et al., Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 2021.

Detour: implicit mapping

Semantic-NeRF (Zhi et al,. 2021)



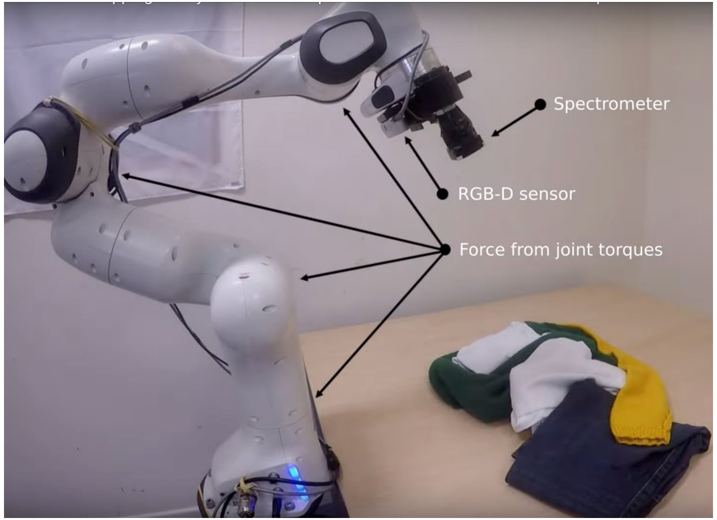
Zhi et al., Inplace scene labelling and understanding with implicit scene representation. CVPR 2021.

Real-time Mapping of Physical Scene Properties with an Autonomous Robot Experimenter (Haughton et al., 2022)

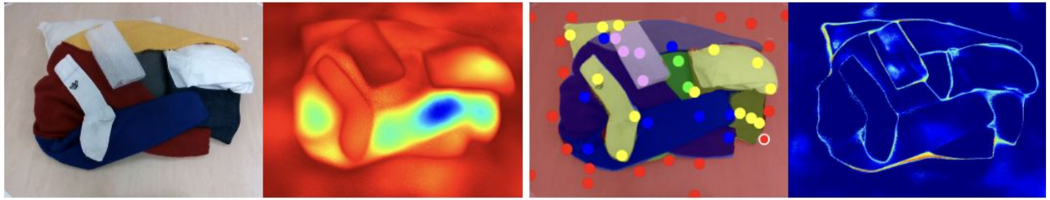
Rigidity classification Material classification Force distribution

Haughton et al., Real-time mapping of physical scene properties with an autonomous robot experimenter, 2022.

Method



- Scene Exploration
- Entropy-Guided Interaction Selection.
- Autonomous Robot Experimentation.
- Semantic Map Optimization.



Initial keyframe and uncertainty map   Final segmentation and uncertainty map

Haughton et al., Real-time mapping of physical scene properties with an autonomous robot experimenter, 2022.

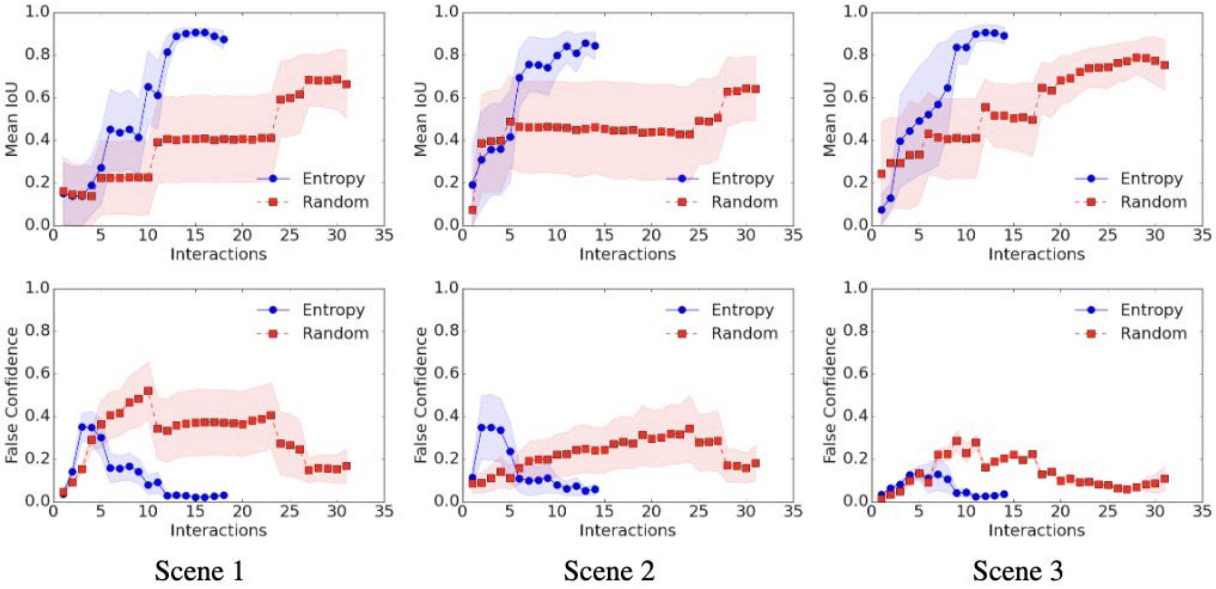
Experiments & Results

Segmentation	Example	Ours	Mask R-CNN	UCN + RICE
Material	Scene 1	0.91 ± 0.02	0.92 ± 0.02	0.90 ± 0.02
Material	Scene 2	0.89 ± 0.03	0.56 ± 0.11	0.56 ± 0.10
Rigidity	Scene 3	0.91 ± 0.04	0.92 ± 0.02	0.91 ± 0.02

Scene 1 – material   Scene 2 – material   Scene 3 – rigidity

Haughton et al., Real-time mapping of physical scene properties with an autonomous robot experimenter, 2022.

Experiments & Results



Haughton et al., Real-time mapping of physical scene properties with an autonomous robot experimenter, 2022.

Conclusion

Strengths & Limitations:

+ Fully autonomous
+ Learning from scratch
+ Novel scene properties

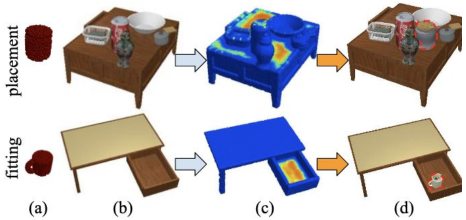
- Still suffer from catastrophic forgetting.
- Sensor quality matters
- Range limited by robot's kinematics

Future Works:

- Higher-resolution sensors
- Extend to more physical properties

Comments:

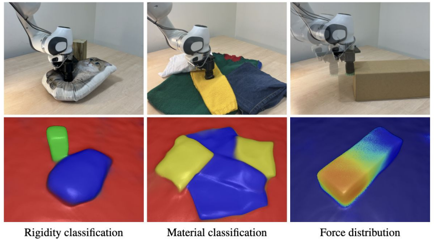
Affordance



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]UCSD

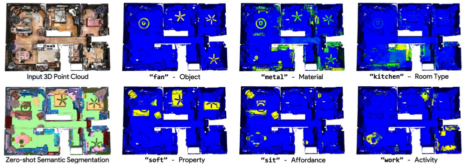
Physical Properties



Real-time Mapping of Physical Scene Properties with an Autonomous Robot Experimenter

Iain Haughton[1] Edgar Sucar[2] Andre Mouton[1] Edward Johns[3] Andrew J. Davison[2]
[1] Dyson Technology Ltd.
[2] Dyson Robotics Lab, Imperial College
[3] Robot Learning Lab, Imperial College
iain.haughton@dyson.com

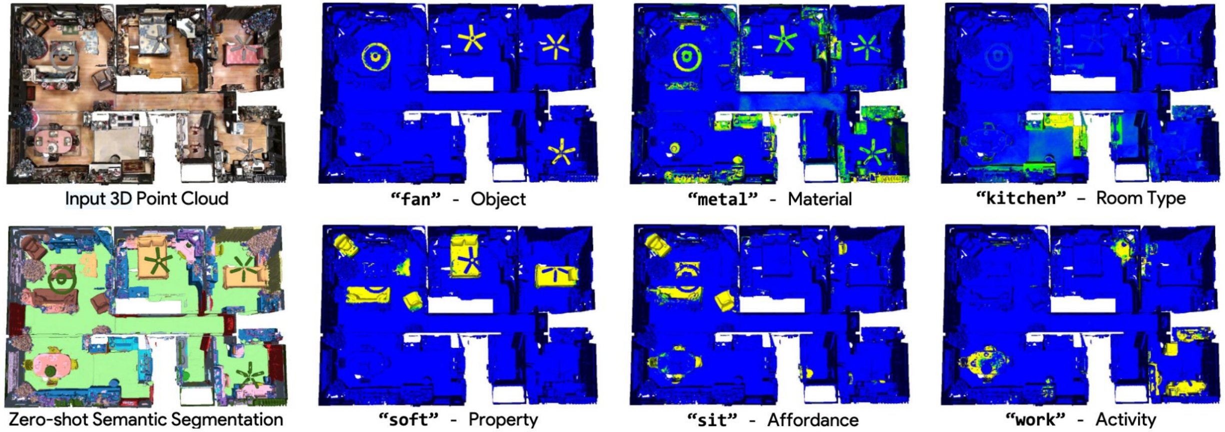
Semantics



OpenScene: 3D Scene Understanding with Open Vocabularies

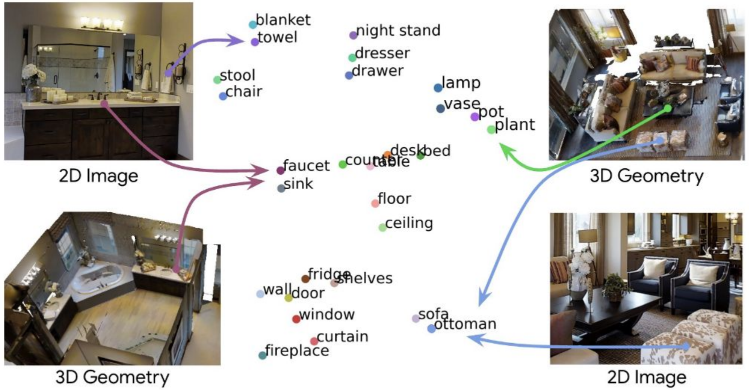
Songyou Peng[1,2,3] Kyle Genova[1] Chiyu "Max" Jiang[4] Andrea Tagliasacchi[1,5]
Marc Pollefeys[2] Thomas Funkhouser[1]
[1] Google Research [2] ETH Zurich [3] MPI for Intelligent Systems, Tübingen [4] Waymo LLC [5] Simon Fraser University

OpenScene: 3D Scene Understanding with Open Vocabularies (Peng et al. 2023)

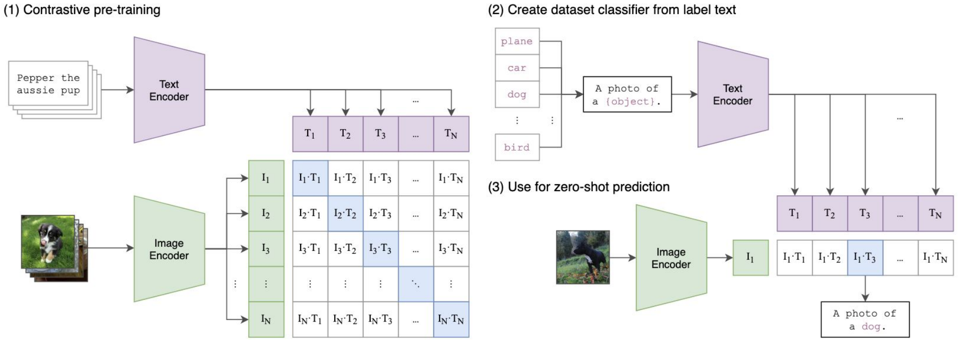


Peng et al., Openscene: 3d scene understanding with open vocabularies. CVPR 2023

Idea

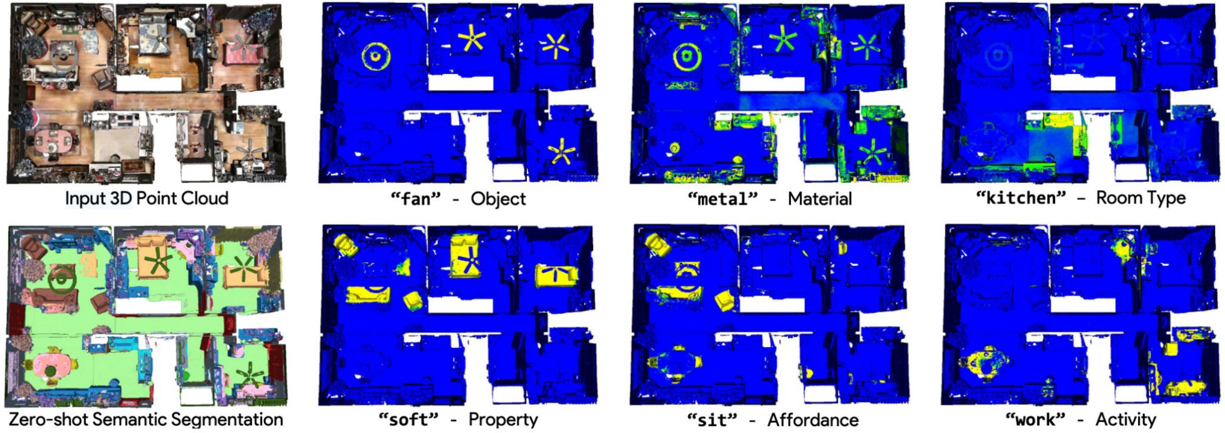


Peng et al., Openscene: 3d scene understanding with open vocabularies. CVPR 2023



Radford et al., Learning transferable visual models from natural language supervision, PMLR 2021

Results



Peng et al., Openscene: 3d scene understanding with open vocabularies. CVPR 2023

From my ADL4CV project, 2024

Discussion

Comment: **Do you think LLM will eliminate the need for the direct mappings discussed before?**

My personal take: **No.**

Instead, we can integrate them.

Summary

O2O-Afford:

Affordance

Large-scale simulation

supervised by simulated interaction results

Real-Time Mapping:

Physical Properties

Sparse interactions

supervised by experiment results

Summary

Mapping beyond geometry is essential for **robotic perception and intelligence**.
It enables capabilities such as:

- Identifying physical properties such as rigidity and material.
- Predicting how objects will react to actions like pushing, lifting, or stacking.

It helps to bridge the gap between **perception** (seeing the world) and **action** (interacting with it)

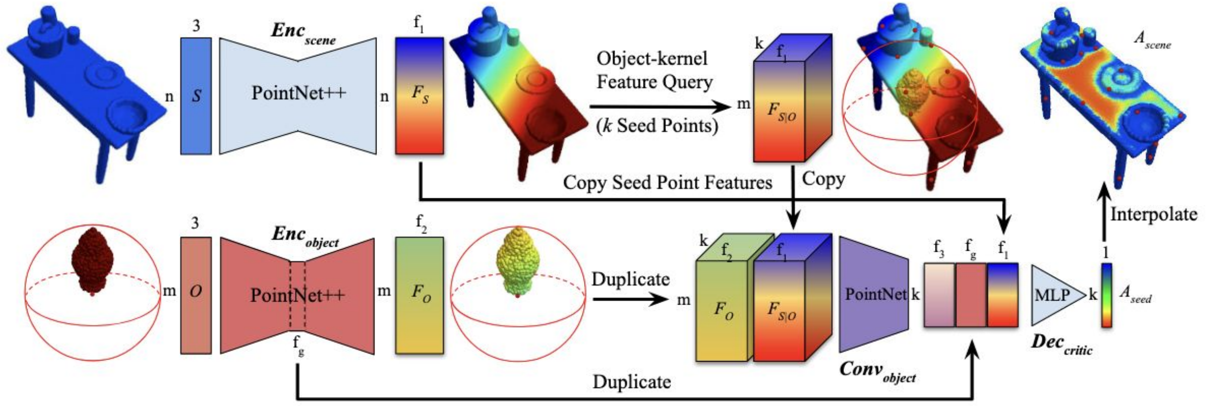
Thank you for listening!

Questions?

References:

[1] Iain Haughton, Edgar Sucar, Andre Mouton, Edward Johns, and Andrew J.Davison. Real-time mapping of physical scene properties with an autonomous robot experimenter, 2022.
[2] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2O-Afford: Annotation-free large-scale object-object affordance learning. In Conference on Robot Learning (CoRL), 2021.
[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65(1):99–106, 2021.
[4] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. KinectFusion: Real-time dense surface mapping and tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed
and Augmented Reality, ISMAR ’11, pages 127–136, Washington, DC, USA,2011.
[5] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6229–6238, 2021.
[6] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. Inplace scene labelling and understanding with implicit scene representation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 15838–15847, 2021.
[7] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Interactive neural scene labelling. arXiv preprint arXiv:2111.14637, 2021.

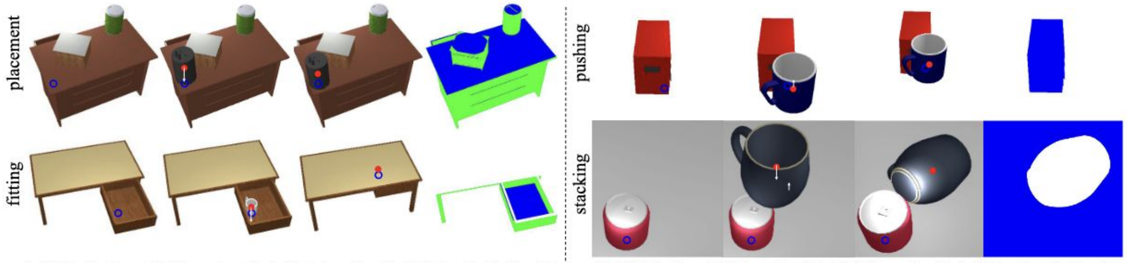
O2O-Afford



$$F_{S|o} = \frac{\sum_{l=1}^{t} w_l F_{S|e_l}}{\sum_{l=1}^{t} w_l}, w_l = \frac{1}{\|o - e_l\|_2}, l = 1, 2, \cdots, t,$$

O2O-Afford

initialization

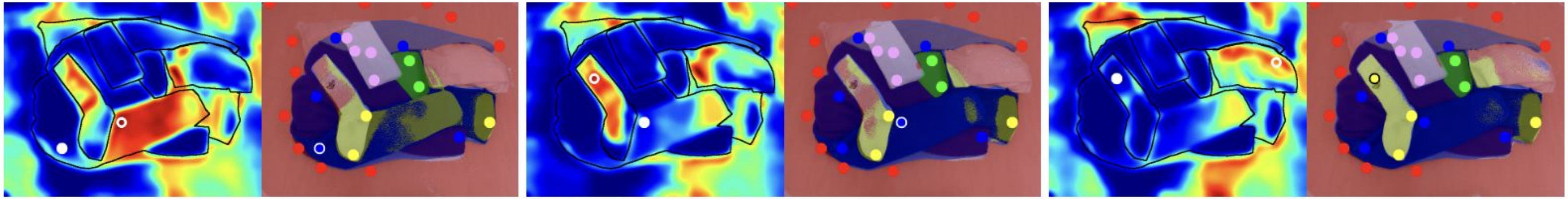


(a) Initialization (b) Interaction Start (c) Interaction End (d) Applicable/Possible (a) Initialization (b) Interaction Start (c) Interaction End (d) Applicable/Possible

shapes

Train-Cats	All	Basket	Bottle	Bowl	Box	Can	Pot
Train-Data	867	77	16	128	17	65	16
Test-Data	281	43	44	5	18	5	
		Mug	Fridge	Cabinet	Table	Trash	Wash
		134	34	272	70	25	13
		46	9	73	25	10	3
Test-Cats	All	Bucket	Disp	Jar	Kettle	Micro	Safe
Test-Data	637	33	9	528	26	12	29

Real-Time Mapping



Interaction 25 | Interaction 26 | Interaction 27

$$u_S = -\sum_{c=1}^{C} \hat{\mathbf{S}}_c [u, v] \log \left(\hat{\mathbf{S}}_c [u, v] \right)$$