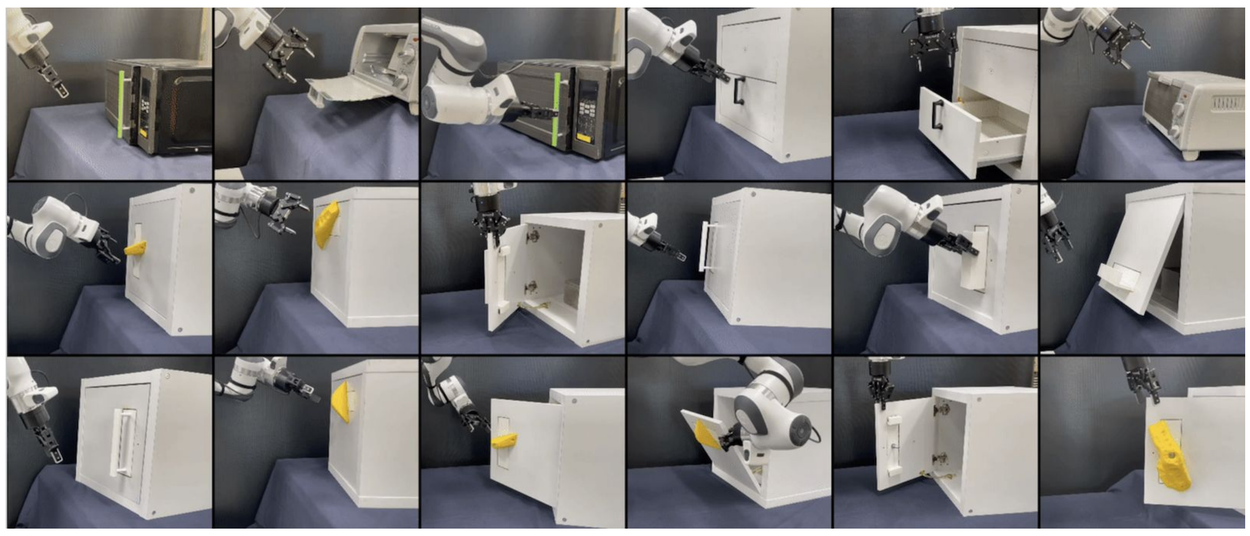
AO-Grasp: Articulated Object Grasp Generation

Carlota Parés Morlans*[1], Claire Chen*[1], Yijia Weng[1], Michelle Yi[1], Yuying Huang[1], Nick Heppert[2], Linqi Zhou[1], Leonidas Guibas[1], and Jeannette Bohg[1]

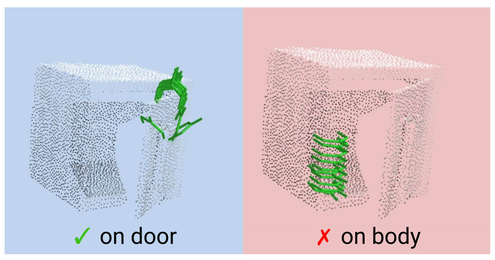
Stanford University, CA, USA[1], University of Freiburg, Germany[2]



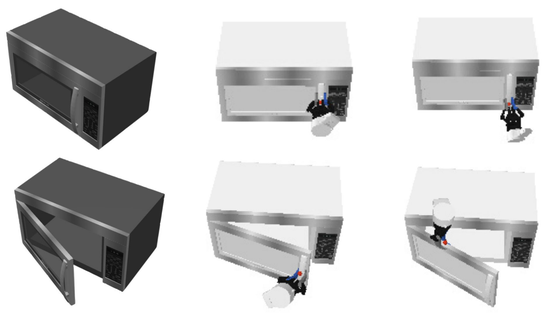
Problem

Grasping articulated objects presents **two unique challenges** compared to grasping non-articulated objects:

1. Grasps must be **stable** and **actionable**



2. Different **joint configurations** have different graspable regions



We consider the first crucial but challenging step determining **how a robot can grasp articulated objects to enable downstream tasks.**

AO-Grasp Dataset

48K stable and actionable grasps on synthetic articulated objects from PartNet-Mobility.[3,4]
5 object categories (`box`, `dishwasher`, `microwave`, `safe`, `trashcan`).

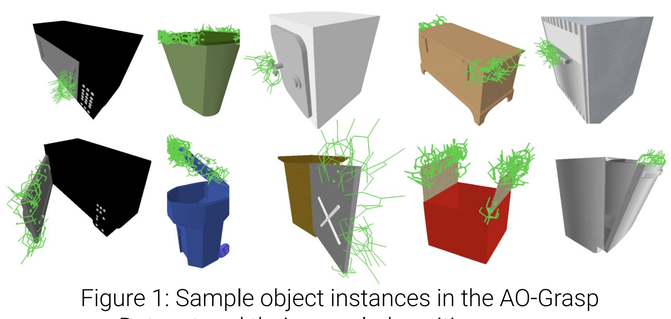


Figure 1: Sample object instances in the AO-Grasp Dataset and their sampled positive grasps.

Category	All					
# Instances	61	9	17	11	11	13
Closed State	6323	516	1396	1546	372	2493
Open State	41954	8091	8020	8022	6152	11669
Total	48277	8607	9416	9568	6524	14162

Table 1: Number of instances and grasps for each category in the AO-Grasp Dataset.

AO-Grasp Predictor

For each point in a partial point cloud, we predict a grasp-likelihood score that signifies **how likely** that **point** will **afford a stable and actionable grasp**.
We use **two training strategies to improve generalization** to new views and objects, namely (a) Siamese PointNet++ and (b) Pseudo Ground Truth Heatmaps.
We **combine a hardest contrastive loss and the mean squared error** between per-point predicted scores and pseudo ground truth heatmap labels to learn generalizable feature encodings. We set $\lambda_{HC} = 3$ and $\lambda_{MSE} = 1$.

$$\mathcal{L}_{\text{total}} = \lambda_{HC}\mathcal{L}_{HC} + \lambda_{MSE}\mathcal{L}_{MSE} \quad (1)$$

HC: Hardest contrastive
MSE: Mean squared error

Given a good grasp point, an object's **local geometry** is the **most important** factor in determining a suitable **grasp orientation**. As such, we leverage predictions from CGN[2].

Figure 2: AO-Grasp Overview. a) Siamese PointNet++ architecture: we find positive and negative correspondences between two different object views to train the network with a hardest contrastive loss. (b) Supervision labels: from sparse collected data to pseudo ground truth dense heatmaps. (c) Grasp proposal generation. From segmented partial point clouds to actionable grasp poses.

Experimental Results

Simulation Evaluation

AO-Grasp compared to baselines

Compared to baselines Contact-GraspNet (CGN)[1] and Where2Act (W2A)[2],

AO-Grasp achieves **higher grasp success rates**

	All	Train Categories						Test Categories					
	All states	Closed state						Open states					
Model	All n = 3740	All 800	80	240	80	80	320	All 980	210	210	140	140	280
AO-Grasp (Ours)	44.8	42.6	61.3	47.1	62.5	11.3	37.5	53.9	59.5	53.8	79.3	42.9	42.5
CGN [2]	31.5	21.5	31.2	4.17	37.5	0	33.4	45.9	49.5	57.6	65.0	47.1	24.3
W2A [1]	2.46	0.38	0	1.25	0	0	0	2.65	1.43	1.90	4.29	4.29	2.50

	All	Test Categories				
		Closed state		Open states		
Model	All 56		8	8	All 64	8	8	8
AO-Grasp (Ours)	33.9	35.8	33.3	50.0	27.8	55.7
CGN [2]	18.1	5	21.8	40.8	22.2	45.6
W2A [1]	0.28	0	0.36	6.81	2.78	7.86

Table 2: Simulation grasp success rates (%) for the top-10 grasps generated by AO-Grasp, CGN, and W2A baselines. Results are broken down by train/test categories, joint state, and category (denoted by icons).

AO-Grasp's predicts **more accurate actionability** heatmaps

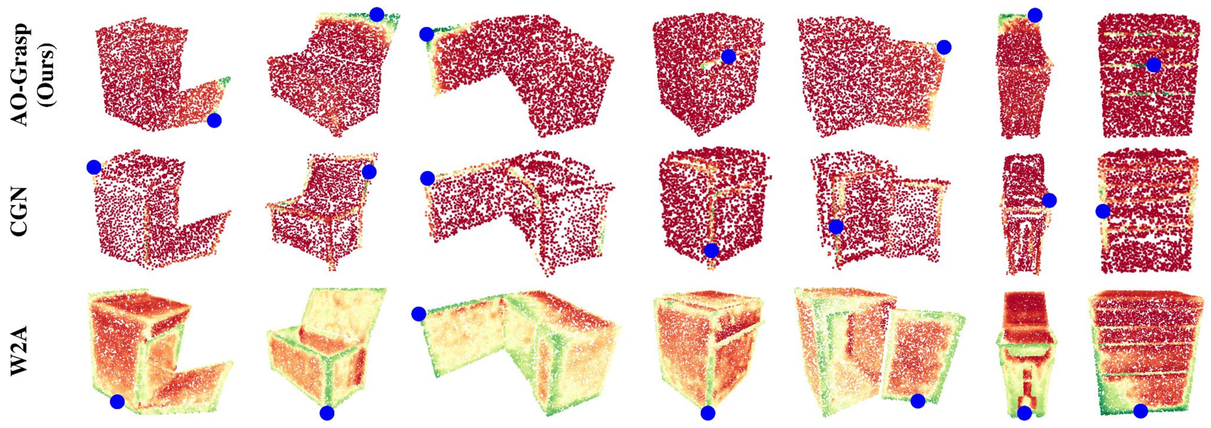


Figure 3: A comparison of grasp-likelihood heatmaps between AO-Grasp and baselines CGN and W2A, where green denotes higher scores and top-1 proposals are highlighted with blue dots. Note that no segmentation mask are required. Both baselines propose non-actionable points more often than AO-Grasp.

AO-Grasp is **more robust to varying camera viewpoints**

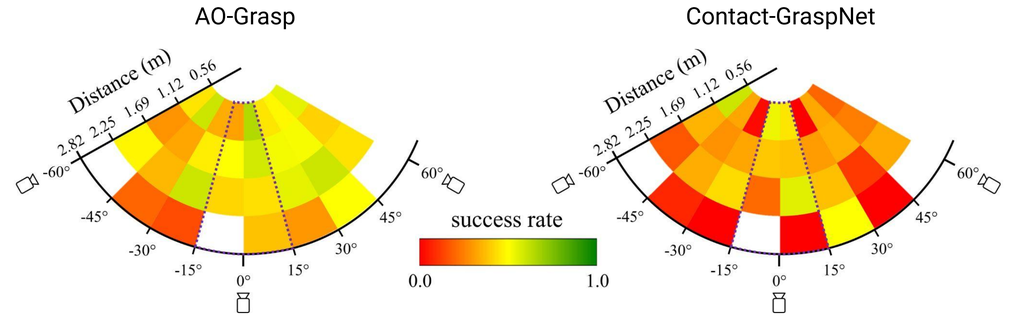


Figure 4: A breakdown of AO-Grasp's and CGN's success rates by camera distance and angle to object.

AO-Grasp ablations

Both *pre-training PointNet++ on* ***viewpoint-independent point correspondences*** and *supervising on* ***dense pseudo-ground truth heatmaps*** **improve** overall **performance**, especially on unseen test categories.

		All categories	Train categories		Test categories	
		All states	Closed	Open	Closed	Open
PT PN++	Dense heatmap	n = 3740	800	980	1080	880
✓	✓	44.8	42.6	53.9	33.9	50.0
✗	✓	42.0	38.6	54.4	26.5	49.9
✗	✗	37.9	41.6	61.8	11.1	40.8

Table 3: Simulation grasp success rates (%) for AO-Grasp ablations on pre- training (PT) PointNet++ (PN++) and training on our dense pseudo ground truth heatmaps

Real-world Evaluation

Zero-shot sim-to-real transfer

We conduct a **quantitative evaluation** of AO-Grasp and CGN **on 120 scenes** of real-world objects with **varied local geometries** and **articulation axes**, in different **joints states**, and captured from different **viewpoints.**
AO-Grasp outperforms baseline Contact-GraspNet on **real-world** articulated objects.

	All states	Closed state								Open states								
Model	All n = 120	All 56	8	8	8	8	8	8	8	All 64	8	8	8	8	8	8	8	8
AO-Grasp (Ours)	67.5	57.1	87.5	62.5	75.0	37.5	37.5	62.5	37.5	76.6	100	100	37.5	87.5	62.5	87.5	50.0	87.5
CGN	33.3	10.7	0	12.5	0	0	0	25.0	37.5	53.1	62.5	50.0	37.5	87.5	62.5	62.5	0	62.5

Table 4: Real-world success rates (%) for AO-Grasp and the baseline Contact-GraspNet[2]

Acknowledgments

This work was supported by Toyota Research Institute.

References

[1] Mo, Kaichun, et al. "Where2act: From pixels to actions for articulated 3d objects." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021.

[2] Sundermeyer, Martin, et al. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

[3] Mo, Kaichun, et al. "Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019..

[4] Xiang, Fanbo, et al. "Sapien: A simulated part-based interactive environment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.