MobileAfford: Mobile Robotic Manipulation through Differentiable Affordance Learning

Yu Li[*4] **Kai Cheng**[*2] **Ruihai Wu**[*1,5]
Yan Shen[1,5] **Kaichen Zhou**[3] **Hao Dong**[†1,5]
[1]CFCS, School of CS, PKU [2]School of EECS, PKU [3]University of Oxford [4]BUPT
[5]National Key Laboratory for Multimedia Information Processing, School of CS, PKU

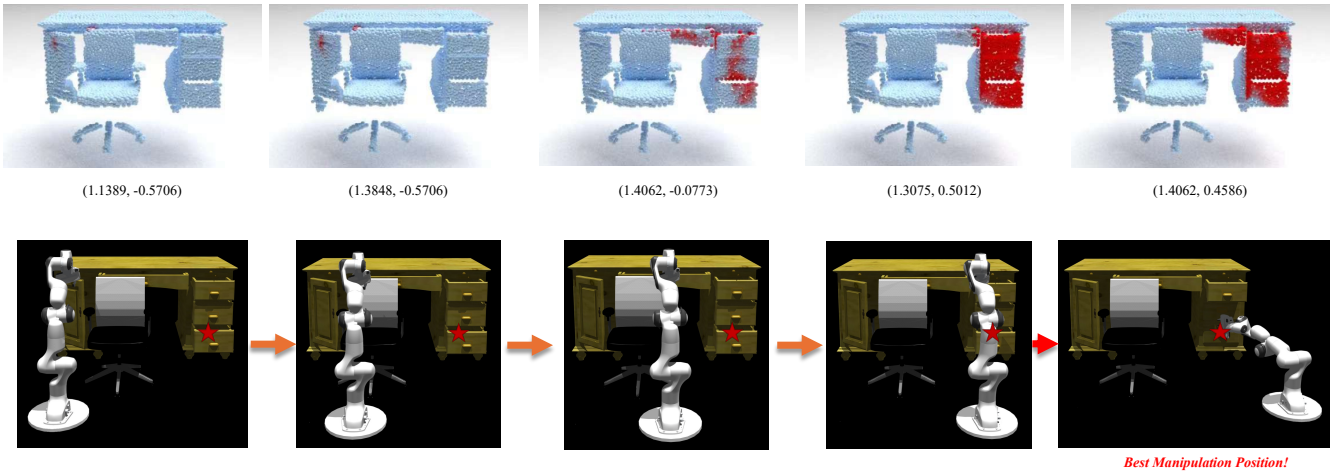

Fig. 1: **Mobile Robotic Manipulation through Differentiable Affordance Learning**. Our proposed affordance is conditioned on robot positions, which is also differentiable and thus can serve as an objective for optimization. For heatmap color denotation, from **Blue** to **Red**, the per-point color refers to the target affordance value. The red in the first RGB sub-figure star denotes the target manipulation point.

***Abstract*— Mobile manipulation in diverse environments is essential yet challenging for robotic home assistants and flexible production. Point-level affordance, which predicts the per-point actionable score and thus proposes the best point to interact with, has demonstrated excellent performance and generalization capabilities in static manipulation. However, whether such actionable priors can be directly used for mobile manipulation remains untested. In this paper, we present a comprehensive differentiable-affordance-based learning framework, *MobileAfford*, which uses only visual input to guide the whole motion and manipulation process. We unify the motion and manipulation process for known and unknown objects in arbitrary environments into trajectory and target affordance optimization. We demonstrate the applicability of the framework in various experiments, including pushing and pulling known and unknown articulated objects on movable robot platforms. Experiment results showcase the state-of-the-art effectiveness of our approach.**

I. INTRODUCTION

We, humans, spend little effort finding optimal paths and interacting with various scene objects to accomplish everyday tasks in our daily lives. However, such capability of mobile manipulation is extremely challenging for intelligent robots to achieve due to the exceptionally high complexity in obstacle-existed motion space coupled with rich 3D object space.

There have been tremendous research endeavors studying generalizable perception for static manipulation in computer vision and robotics. Many great advances in 3D representation learning for various tasks, *e.g.*, 3D shape recognition [3], pose estimation [21], [25] and semantic segmentation [17], [20] have laid a solid foundation for generalizable 3D perception. It then results in the emergence of object-centric 3D manipulation, among which object-centric *affordance learning* occurs and serves as a bridge between 3D geometry learning and robotic skill learning [15], [24], [28], [23], [26], [8], [4]. Such learned visual actionable affordance essentially predicts the action likelihood for accomplishing a certain downstream manipulation task at each point on a 3D input geometry like a point cloud.

Though showing promising generalization capability, the application scenarios of these object-centric representation learning methods are mostly limited to simple static manipulation tasks, *e.g.*, single-gripper pushing/pulling [15], [24], [23], [26], flying-manipulator collaboration [28], [8]. Recent works [22], [4] have proposed a novel robot-object-environment handshaking framework for whole-body manipulation, which essentially predicts the affordance maps

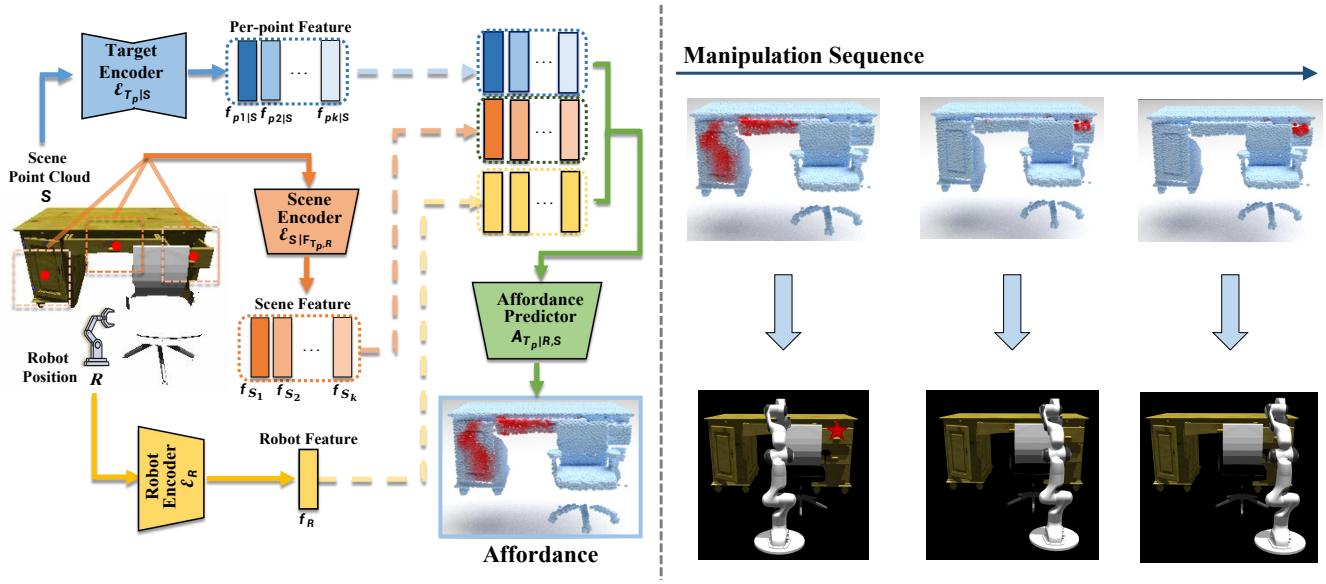

Fig. 2: **The Whole Framework.** In the **left**, we learned the environment-aware affordance. In the **right**, we use environment-aware affordance to guide the mobile manipulation to avoid collisions with occluders. For **point-level affordance scores**, from **Blue** to **Red**, the color means from not affordable to totally affordable. The red in the first RGB sub-figure star denotes the target manipulation point. The trajectory of our manipulation sequence is guided by trajectory affordance optimization.

for a position-fixed robot arm instead of flying grippers. Such robot-object-environment coupled actionable affordance, trained across diverse 3D shape geometry (*e.g.*, refrigerators, microwaves), diverse occlusion environments (*e.g.*, various combinations of occluders like chairs, boxes), and diverse robot positions (*e.g.*, in the front or right to the target) for a specific downstream manipulation task (*e.g.*, pushing and pulling), is proven to generalize to novel unseen objects (*e.g.*, tables), environments (*e.g.*, new combinations of occluders like bottles, buckets) and robot positions (*e.g.*, in the left to the target). However, even though such affordance can be coupled with robot positions, its application does not go beyond static manipulation and thus the robot's position - affordance correlation is not fully utilized. As we can observe from Figure 1, the affordance is dependent on the robot's position for mobile manipulation. How to design a framework to utilize the affordance curve for motion strategy remains intriguing.

In this paper, we study the mobile manipulation task from the perspective of representation learning and investigate learning mobile manipulation strategy through conditional visual actionable affordance over the 2D base trajectory and the 3D target. We propose a novel framework *MobileAfford* to tackle the mobile manipulation problem. At the core of our design, *MobileAfford* treats the mobile manipulation strategy as an optimization problem over the 2D position and 3D target conditioned affordance. Concretely, we design a differentiable affordance learning framework to guide the motion and manipulation process. The conditional affordance is differentiable and thus the motion and manipulation process can both be unified through affordance optimization. Such affordance representation inherits the generalization capability of previous works [4] and can be applied to known and unknown objects in arbitrary environments.

We evaluated the proposed method on four mobile manipulation tasks: pushing without occlusions, pushing with occlusions, pulling without occlusions and pulling with occlusions. We set up our benchmark for experiments using assets from Partnet-Mobility dataset [18] and ShapeNet [3] dataset, with IssacGym [12] as our simulation environment. Quantitative comparisons against baseline methods prove the effectiveness of our proposed framework. Qualitative results further show that our learned affordance is reasonably convex and can serve as the objective for various mobile manipulation tasks. To summarize, we make the following contributions:

- We present a novel framework *MobileAfford* to learn differentiable actionable affordance for mobile manipulation;
- We propose a unified perspective for representing both motion strategy and manipulation action through conditional affordance optimization over the 2D trajectory and the 3D target;
- We set up a benchmark built upon IssacGym [12] using Partnet-Mobility dataset [18] and ShapeNet [3] dataset for generalizable mobile manipulation tasks;
- We show qualitative results and quantitative comparisons against the baselines to validate the effectiveness of our proposed approach.

II. RELATED WORK

A. Agent-centric Motion Planning

Berenson, et.al. [1] proposed the sampling-based planner based on rapidly exploring random trees (RRTs) for

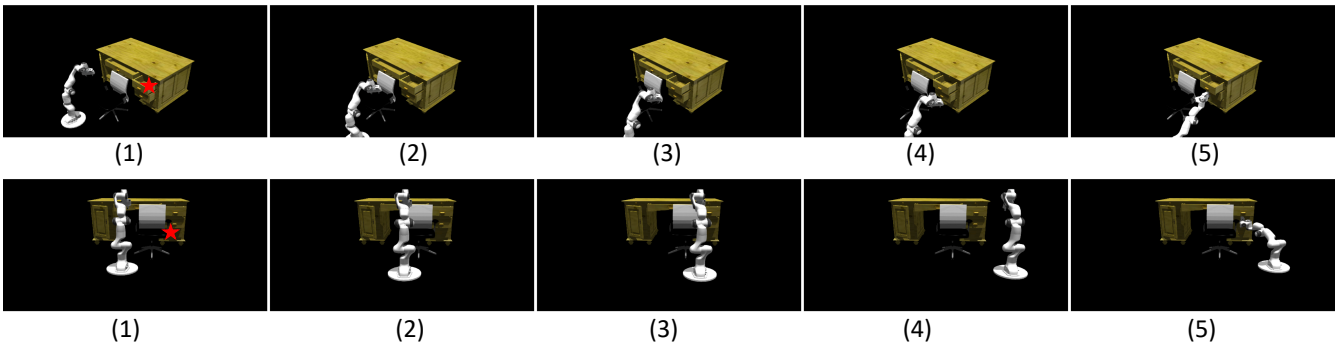

Fig. 3: **Visualization of Manipulation Trajectory Guided by Our Proposed Approach.** The red star in each first sub-figure denotes the target manipulation point.

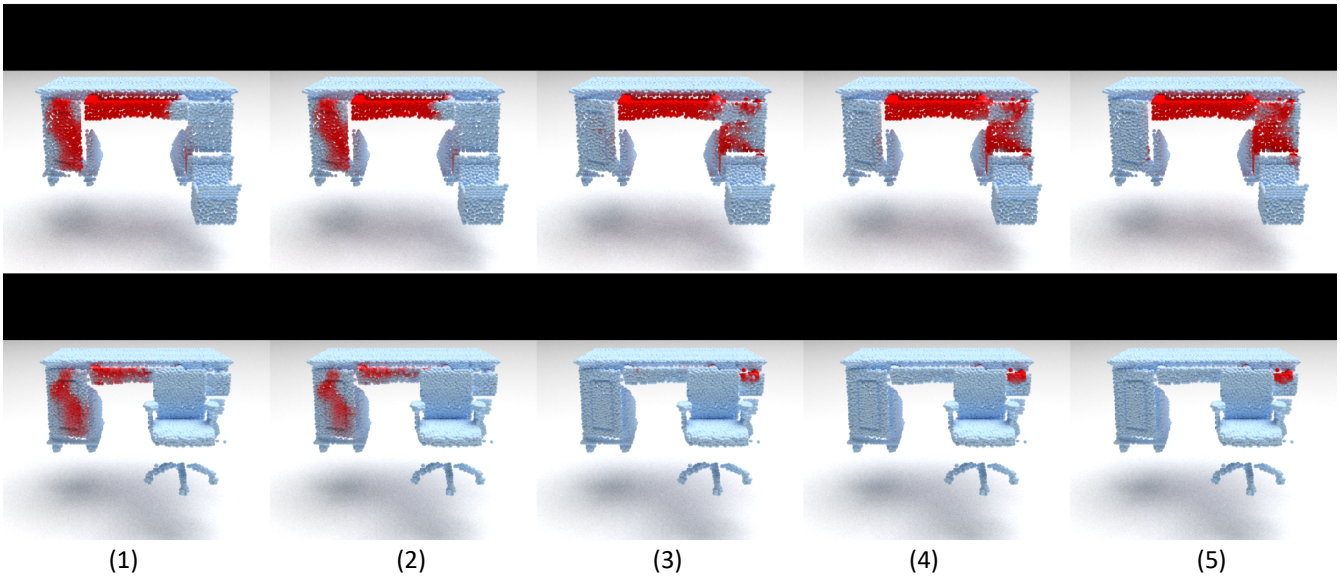

Fig. 4: **Visualization of Our 3D Target Affordance Maps.** In each row, the robot is positioned from left to right.

manipulation under end-effector pose constraints. However, it suffers from shortcomings such as incorporating non-holonomic constraints and dynamics. Previous works like [2], [5], [14] incorporate the object's kinematic model as a task space constraint and apply sampling-based planners with IK to find feasible solutions. However, these methods are often computationally expensive, require offline planning, and may suffer from the shortcomings of IK when approaching singularities.

B. Object-centric and Conditional Affordance Learning

Gibson [9] first proposed the affordance as a kind of representation that indicates possible ways for robots to interact with the target and complete tasks. Many works study object-centric affordance for the classic grasping task in robotics [13], [19], [6], [11], [10], [27], while there exist many current works on point-level affordance indicating object geometrics for articulated object manipulation [15], [24], [23], [7], [8], dual-gripper collaboration [28] and object to object interaction [16]. Except for these object-centric affordance works, recent works like [22], [4] have proposed conditional affordance for whole-body manipulation, which essentially predicts the affordance maps conditioned on different robot positions and occlusion environments. However, even though such affordance can be conditioned on robot positions and acquire great generalization capabilities, its application does not go beyond static manipulation.

III. METHOD

As shown in Figure 2, our method mainly consists of 3 modules, the Motion Planning Module $\mathscr{P}$, the Affordance Prediction Module $\mathscr{A}$ and the Optimization Module $\mathscr{O}$.

In the initial state, the robot will perceive the partial 3D point cloud as input and feed it through $\mathscr{A}$ to get an initial conditional affordance map. This affordance map is then fed into $\mathscr{O}$ to conduct iterative closed-loop affordance optimization, with the trajectory and target affordance value as the objective and robot position as the condition. The optimization trajectory for the robot position forms our motion strategy, which is shown in Figure 5. Besides, to first approach the object from far away and finally finish the manipulation actions, we also need $\mathscr{P}$ to guide the robot's

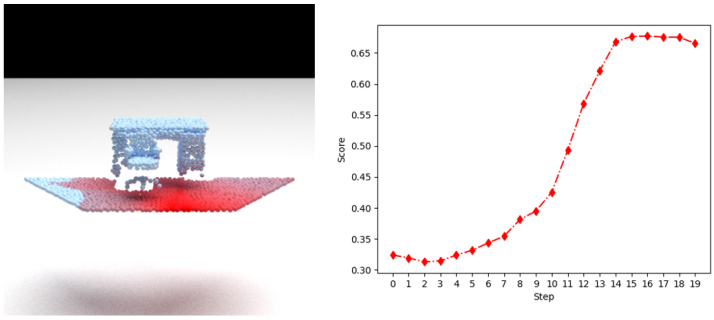

Fig. 5: **Visualization of Our 2D Trajectory Optimization.** The left subfigure demonstrates the 2D affordance map about the robot's different positions in the motion trajectory. The right subfigure shows a typical trajectory affordance curve of our proposed method.

whole-body motion.

A. Motion Planning Module

We design a mobile MPC (Model Predictive Controller) for our 9-DoF mobile panda arm as the motion planner $\mathscr{P}$.

B. Affordance Prediction Module

We use the conditional affordance network from EnvAwareAfford [4] as our predictor $\mathscr{A}$, which is shown in Figure 2.

C. Optimization Module

We perform sampling-based affordance optimization $\mathscr{O}$ to obtain our motion strategy. Concretely, at each iteration, our model predicts the affordance for the present robot position. Our model then samples N directions with M strides as our sampled next position. Our predictor $\mathscr{A}$ then inferences over these $N \times M$ positions, the results of which are then fed into $\mathscr{O}$ to find the optimal one.

D. Planning-Optimization Adaptation

The affordance predictor $\mathscr{A}$, when applied to a robot positioned at a considerable distance initially, tends to output an affordance map filled entirely with zeros, which lacks the necessary gradient to guide the optimization process effectively. Thus we use the motion planner $\mathscr{P}$ to first approach the object for a certain distance to obtain a meaningful 2D trajectory affordance gradient for optimization $\mathscr{O}$.

When the final conditional affordance value reaches its optimum, the optimizer $\mathscr{O}$ will stop and the motion planner $\mathscr{P}$ will output the manipulation strategy.

IV. EXPERIMENT

A. Task and Environment Settings

We evaluated our proposed method on two mobile manipulation tasks: pushing without occlusions, pushing with occlusions. In all tasks, a mobile robotic arm is required to accomplish a specific manipulation goal with different objects.

TABLE I: Quantitative results in no occluder, one occluder, multi-occluder scenes.

Method	No occluder	One occluder	Multiple occluders
Heuristic	46.7	42.8	35.7
Ours	**70.2**	**64.3**	**52.6**

Our offline data is collected in the simulator IssacGym [12], using assets from PartNet-Mobility, a 3D articulated object dataset [18] and ShapeNet, a comprehensive rigid 3D shape dataset [3]. These offline data only contain single-position interaction results for the training of our affordance predictor $\mathscr{A}$.

B. Baselines and Ablation

We benchmarked our method against another algorithm, which is briefly described in the following:

- **Heuristic**: In this method, we move the robot directly in the front of the target point, and then manipulate it.

C. Quantitative and Qualitative Results

Table I shows our large-scale evaluation results in simulation over different tasks. As the heuristic method does not include the procedure of optimization procedure, our environment-aware-affordance-based method outperforms this baseline by quite a large margin.

Figure 3 shows that the mobile manipulation of our method successfully helps in avoiding collisions with occlusion objects and manipulating the target point.

V. CONCLUSIONS

In this paper, we proposed a novel framework *MobileAfford* for learning differentiable conditional actionable affordance, which unified the motion and manipulation process for known and unknown objects in arbitrary environments through affordance optimization. We set up large-scale benchmarks in IssacGym [12] for four mobile manipulation tasks using the PartNet-Mobility [18] and ShapeNet [3] datasets. Results proved the effectiveness of our approach and its superiority over the baselines.

REFERENCES

[1] Dmitry Berenson, Siddhartha Srinivasa, and James Kuffner. Task space regions: A framework for pose-constrained manipulation planning. *IJRR*, pages 1435–1460, 2011.

[2] Felix Burget, Armin Hornung, and Maren Bennewitz. Whole-body motion planning for manipulation of articulated objects. In *ICRA*, pages 1656–1662, 2013.

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[4] Kai Cheng, Ruihai Wu, Yan Shen, Chuanruo Ning, Guanqi Zhan, and Hao Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] S. Chitta, B. Cohen, and M. Likhachev. Planning for autonomous door opening with a mobile manipulator. In *ICRA*, pages 1799–1806, 2010.

[6] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020.

[7] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15752–15761, October 2021.

[8] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. End-to-end affordance learning for robotic manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2023.

[9] James J Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[10] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 5(2):3352–3359, 2020.

[11] Mia Kokic, Johannes A. Stork, Joshua A. Haustein, and Danica Kragic. Affordance detection for task-specific grasping using deep learning. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 91–98, 2017.

[12] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021.

[13] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[14] Mayank Mittal, David Hoeller, Farbod Farshidian, Marco Hutter, and Animesh Garg. Articulated object interaction in unknown scenes with whole-body mobile manipulation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1647–1654, 2022.

[15] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[16] 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.

[17] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2019.

[18] Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Luis Montesano and Manuel Lopes. Learning grasping affordances from local visual descriptors. In *2009 IEEE 8th international conference on development and learning*, pages 1–6. IEEE, 2009.

[20] Manolis Savva, Angel X Chang, and Pat Hanrahan. Semantically-enriched 3d models for common-sense knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–31, 2015.

[21] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[22] Liquan Wang, Nikita Dvornik, Rafael Dubeau, Mayank Mittal, and Animesh Garg. Self-supervised learning of action affordances as interaction modes. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7279–7286. IEEE, 2023.

[23] Yian Wang, Ruihai Wu, Kaichun Mo, Jiaqi Ke, Qingnan Fan, Leonidas Guibas, and Hao Dong. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. *European conference on computer vision (ECCV 2022)*, 2022.

[24] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. VAT-mart: Learning visual action trajectory proposals for manipulating 3d ARTiculated objects. In *International Conference on Learning Representations*, 2022.

[25] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[26] Zhenjia Xu, He Zhanpeng, and Shuran Song. Umpnet: Universal manipulation policy network for articulated objects. *IEEE Robotics and Automation Letters*, 2022.

[27] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3750–3757. IEEE, 2018.

[28] Yan Zhao, Ruihai Wu, Zhehuan Chen, Yourong Zhang, Qingnan Fan, Kaichun Mo, and Hao Dong. Dualafford: Learning collaborative visual affordance for dual-gripper object manipulation. *International Conference on Learning Representations (ICLR)*, 2023.