Learning Part-Aware Visual Actionable Affordance for 3D Articulated Object Manipulation

Yuanchen Ju [1,2*] Haoran Geng [1*] Ming Yang [3*]
Yiran Geng [1] Yaroslav Ponomarenko [1] Taewhan Kim [1] He Wang [1] Hao Dong [1†]
[1]Peking University [2]Southwest University [3]Beihang University

Abstract

Recent advancements in visual actionable affordance research have demonstrated its strong ability to manipulate various articulated objects. The point-level actionable score indicates where and how the robot interacts with the object, which is learned through self-supervised trial and error without any expert demonstration, rule-based policy, or task-specific reward design. Previous works mainly focused on object-centric visual manipulation. However, we have noticed that human-made articulated objects (e.g. handles on doors) often have salient parts designed for interaction. Selecting these parts for manipulation is crucial for the success rate of many tasks. In this work, we consider both part-level and point-level geometry information simultaneously. We first design a part selecting score to choose suitable parts for interaction. By leveraging per-part predictions and utilizing the prior information provided by these parts, we then predict the part-aware fine-grained affordance map in an SE(3) invariant manner. Thus, it will result in a significant improvement in the success rate of many long-term manipulation tasks.

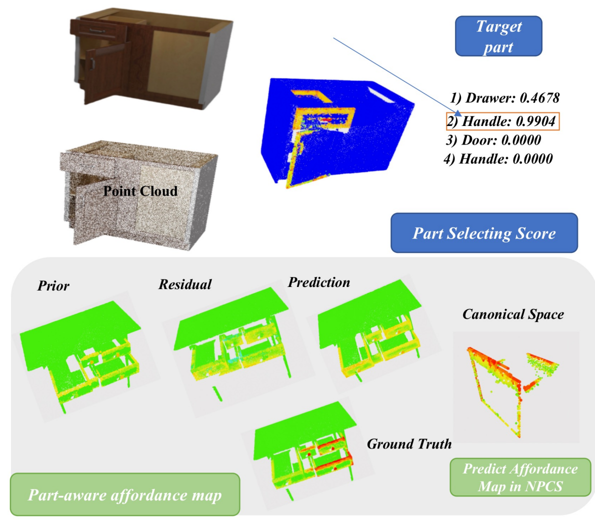

Figure 1. **Overview.** We propose to use both point-level and part-level affordance for object manipulation.

1. Introduction

3D articulated objects are common in our daily lives, and they involve sophisticated interactions by humans due to their complex structures and functionalities. Similarly, we expect modern robots to help humans perform a range of in-home activities, automatically recognizing and manipulating various objects. For example, robots can open and close articulated objects such as doors, drawers, and cabinets to complete assigned tasks.

A prevailing paradigm in existing methods for robotic manipulation involves perceiving objects' joint parameters and structures. However, using these representations as the input for the manipulation policy may neglect the geometric features of the object that are necessary for subsequent robotic manipulation tasks. For instance, the shape of a drawer handle can vary, and it may require different grabbing stances. Recent studies have suggested a solution to mitigate this problem through visual actionable affordance [32, 56, 63], which involves predicting the point-level actionability or motion trajectories on an object's surface using action primitives like pushing and pulling.

However, the current visual actionable affordance approaches for manipulating articulated objects have some flaws in their design. Our first argument is that previous pipelines did not include part conception, which could result in high actionability scores being predicted for the frame or board of an item, particularly when it comes to unseen objects. In order to achieve reasonable performance, these approaches often use ground-truth part masks during testing, which is not realistic when manipulating novel objects in the real world. Second, based on the definition of the affordance score, the prediction should remain the same

*Equal contribution.
†Corresponding author.

regardless of the object's position or orientation. However, the results of these methods show that the prediction can significantly change when the same object changes its posture. In these methods, we also observed that a high predicted affordance score at the corner of a door could make it difficult to interact with, whereas a point on the handle with a slightly lower score could make it easier. This is because these approaches can only predict affordance scores at the individual point level without considering any semantic information.

To sum up, our contributions are summarized as follows

1. We propose a part-aware approach for manipulating 3D articulated objects that does not require ground truth part masks during inference. The method employs a coarse-to-fine strategy that refines the part segmentation over multiple stages, utilizing both part-level and global information.

2. We are the first to consider part-awareness in the context of visual actionable affordance, addressing the limitations of existing methods that neglect the prior information provided by the parts, leading to ambiguity.

3. We use a subset of the large-scale part-centric interactive dataset GAPartNet and simulation environments created in IsaacGym to collect data and evaluate the proposed approach. The approach is evaluated using two-part categories (doors and drawers) to cover 7 common indoor object categories. The ablation study provides insights into the effectiveness of the proposed method.

2. Related Works

2.1. Articulated Object Manipulation

Robotics and computer vision researchers have long studied the manipulation of articulated objects. For example, there were approaches of inferring object motion and pose from visual perception [14, 19, 48, 49, 60], manipulating objects interactively and understanding scenes [12, 13, 17, 18, 27, 30, 45, 51], and using machine learning methods for 3D object manipulation and scene reasoning. [6, 28, 29, 39–41].

A vast amount of literature has shown practical methods for getting precise link poses, joint parameters, kinematic structures, and even system dynamics of 3D articulated objects. These methods use visual feature trackers, motion segmentation predictors, and probabilistic estimators. Numerous robotic planning and control techniques have also been investigated in earlier works [3, 4, 15, 42] for handling 3D articulated objects. More recent efforts have used learning techniques to improve predictions of articulated part configurations, parameters, and states [16, 23, 25, 35, 53–55, 61, 64], as well as estimation of kinematic structures [1, 47] and manipulation of 3D articulated objects using learned visual knowledge [2, 7, 9, 21, 31, 52, 59].

2.2. Visual Actionable Affordance

Affordance indicates possible ways for robots to interact with the object and environment [11]. Previous works have investigated affordance for various tasks, including robot grasping [22, 43], robot manipulation [26, 33, 38, 44, 56, 63], hand-object interaction [5, 8, 20, 26, 62] and object-object [34, 50, 65] interaction. Among these studies, many works require human annotations or demonstrations [8, 20, 22, 37], while some recent works learn affordance through trial and error without the need for human annotations [33, 38, 56, 63]. Recent studies [33, 56, 63] have proposed point-level visual actionable affordance to manipulate articulated objects. These affordances indicate every location on the object and suggest how robots can interact with them. In addition, this approach has shown promising generalizable ability over diverse shapes. Different from studies that use part information during testing, our work is a top-down method that utilizes part-level information during the training phase to suggest which part the robot should interact with and where on the object.

3. Method

Part-aware Affordance. Part-aware visual affordance learning is a framework for learning the affordances of objects by considering their parts and visual appearance. This approach combines object recognition and grasping pose prediction to identify the relevant parts of an object and the motion direction necessary for successful grasping. SE(3) invariant affordance learning is also used in part-aware visual affordance learning. By learning SE(3) invariant features, the affordance learning model can generalize to objects with different poses and orientations. We will then explain how to implement this in our pipeline in detail by introducing our point-level and part-level affordance learning module below.

3.1. Part-aware Visual Affordance Learning

Point-level Affordance Learning. Following the definition in [32], we first predict per-point affordance. Due to the lack of part information in [32], they directly collect the interaction result for each point on the object. Thanks to the rich part annotation in GAPartNet [10], we thus benefit from the part segmentation and poses. Hence, the only points under the notated actionable parts can be interacted. So in our method, we directly collect the interaction information on the points under the actionable parts. What's more, thanks to the GPU-parallel simulator IsaacGym [24], we can parallel sample each point we want, instead of a subset of all points in [32].

After data collection, we propose our part-aware point-level affordance learning module. Given a partially observed point cloud O, we first use the part segmentation and pose estimation module proposed in GAPartNet [10] to segmentation each part $\{P_i\}$ and pose $\{p_i = (t_i, R_i)\}$. For a part P_i, we first query the points O_i, which belong to this part. Then we transfer this point cloud into its estimated canonical space using the estimated pose p_i, and we get transferred part point cloud $\hat{O}_i$ for part P_i. Till now, we finish the point cloud pre-processing stage, then we use the

point-level affordance head in our pipeline to estimate the affordance score for each point in the part canonical point cloud $\hat{O}_i$. Following this process for each predicted part, we mix the results and get the first-stage per-point affordance map A_{pre}.

$$A_{\text{pre}} = \bigcup_{i=1}^{N_{\text{part}}} A_{\text{pre}}^i$$

Then we also tackle the problem that the estimated parts may be inaccurate, we thus introduce a residual affordance prediction module. Given the first-stage predicted affordance map A_{pre} and the whole point cloud O, we estimate a residual affordance score for each point in the whole point cloud and get A_{res}. Finally, we can get the predicted point-level affordance map $A = A_{\text{pre}} + A_{\text{res}}$. A is supervised by $\hat{A}$ with L2 loss.

$$\mathcal{L}_{\text{point}} = \frac{1}{N_{\text{point}}} \sum_{j=1}^{N_{\text{point}}} (A_j - \hat{A}_j)^2$$

Part-level Affordance Learning. We also innovatively define a part-level affordance, which is a score for each predicted part. The higher part-level score means it's better to interact with this part to finish the given task, *e.g.* if we want to open a door with a handle, we can interact with both the door and the handle to finish it. And if we estimate that the handle is a better one to finish, we may try to interact with and the corresponding part-level affordance should be higher.

To train the part-level affordance module, we use the collected point-level affordance map to calculate the average score for points with a score higher than a given threshold τ in a given part. And use this truncated average score as the ground truth of the part-level affordance score, which is $\{\hat{s}_i\}$

For each predicted part P_i, we estimate a part-level affordance score $\hat{s}_i$ for it. This score is also supervised by L2 loss.

$$\mathcal{L}_{\text{part}} = \frac{1}{N_{\text{part}}} \sum_{i=1}^{N_{\text{part}}} (s_i - \hat{s}_i)^2$$

We integrate the point-level and part-level affordance in our pipeline and in addition to the part segmentation and pose estimation loss $\mathcal{L}_{\text{seg}}, \mathcal{L}_{\text{pose}}$, we add our affordance loss to the pipeline, which is

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{pose}} + \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{part}}.$$

3.2. Part-aware Interaction Policy

We then introduce how we finish the object manipulation task using the predicted part-level and point-level affordance.

Our method first takes an observation O and predicts the part-level score s and affordance map A for all parts in the observation. We first select the part with the highest part-level score as the part as the target part. Then, in this part, we select a point with the highest point-level score to interact with. We follow the affordance definition to pull or push this point and finish the first interaction step.

Then, iteratively, we follow the process above several times until we finish the manipulation task or we reach a maximum number of interaction steps.

3.3. Training Data Collection

It is infeasible to collect training data from human interactions. Instead, we benefit from the physics simulator to collect the data. Thanks to the GPU-parallel simulator IsaacGym [24], we can collect interaction data in parallel. We build up an interaction environment, in which we use the parallel gripper of the Franka Robot Arm to interact with each point in a certain direction we want and see whether to part moved or not. We collect data for each point on the part and cover 68 different directions.

4. Experiments

In this section, we evaluate our method in a simulated environment qualitatively and quantitatively. We first elaborate the environment and settings in Sec. 4.1. Qualitative results show promising generalization capability.

4.1. Environment and Settings

We evaluate our method with both diverse simulation manipulation tasks and real-world robot experiments.

Simulation We set up the simulated environment with NVIDIA's IsaacGym [24], a simulator tailored towards high-performance GPU parallelization. Objects are from GAPartNet dataset [10], a large-scale part-centric dataset with rich part annotation based on PartNet- Mobility [58].

For each interaction session, we first randomly load one articulated 3D object into the environment with randomly initialized joint configurations. Then, a Franka Panda Flying gripper with 2 fingers is used as the robot actuator. There are 8 degree-of-freedom (DoF) in total (3 DoF for position, 3 DoF for orientation, and 2 DoF for the 2 fingers). We use an RGB-D camera of resolution 800×800 with a randomly sampled viewpoint in front of the object.

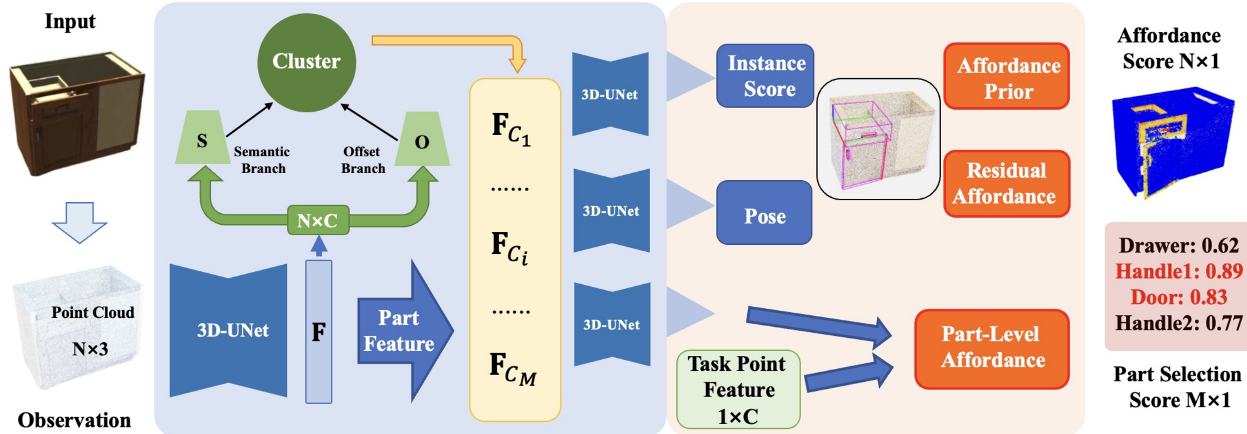

Figure 2. **Overview.** We first select the suitable part for interaction based on the part selecting score, and then predict the affordance map in NPCS. Our method takes point clouds as input, extracts point cloud features using 3D-Unet, processes them through two branches, semantic and offset,finally obtains the features of each point on the per-part through clustering. Then, after processing with 3D-Unet, pooling layer, and NPCS, we can obtain information such as pose and local affordance map prior. Simultaneously, by inputting task point feature, the model can obtain part-level affordance score.

We evaluate our method in a simulated environment with objects from GAPartNet [10] and a physics engine from Isaac Gym [24]. We use two types of parts: drawer and door, which cover 7 common categories of indoor objects.

4.2. Baselines and Metrics

Baselines. To verify the effectiveness of our method, we compare two types of baselines:

- **PPO**: We use the PPO algorithm to finish the tasks in an RL manner. We take the same observation as input as ours. We design the dense reward borrowed from ManiSkill [36].
- **PPO+BC**: We use PPO for state-based policy and collect demonstrations at the same time. Then we use behavior cloning for vision-based policy.
- **ILAD**: We follow the ILAD algorithm to finish our tasks. The setting is similar to the PPO baseline.
- **M-Where2Act**: We modify Where2Act [32] baseline. we sample data for every point on the object and boost the performance compared with the original implementation. This baseline takes the oracle part mask as input.

Metrics. Following [32], we run interaction trials in simulation and report success rates for quantitative evaluation.

4.3. Ablation Study and Analysis

To further evaluate the different components of our method, we conduct an experiment to evaluate the usage of our proposed part-level affordance in Tab.2.

Methods	Success Rate for Door and Drawer
PPO [46]	13.92
PPO+BC	44.95
ILAD [57]	37.32
M-Where2Act [32]	53.04
Ours	**59.97**

Table 1. Results

Methods	Success Rate
Ours w/o Part-level Score	55.28
Ours w/ Part-level Score	**59.97**

Table 2. Ablation

5. Conclusion

We present a novel approach that aims to improve the manipulation of articulated objects by utilizing visual actionable affordance.Our proposed framework utilizes per-part predictions and preliminary part information to overcome the limitations of existing visual actionable affordance methods. By considering the robot's perception of articulated objects at both the point-level and part-level, our framework provides a more comprehensive understanding of the object's affordances.The part-selecting score serves as an indicator of the suitability of each part for manipulation, based on its grasp ability and affordance. This approach allows the robot to identify the most optimal parts for manipulation, leading to higher task success rates. To predict affordance maps, we employ Normalized Part Coordinate Space (NPCS), which eliminates the dependence on object pose and orientation. This standardizes the reference frame for the object's parts, providing a normalized and standardized basis for more accurate predictions, which leads to better generalization to novel objects and more robust manipulation in real-world scenarios.

References

[1] Ben Abbatematteo, Stefanie Tellex, and George Konidaris. Learning to generalize kinematic models to novel objects. In *Proceedings of the 3rd Conference on Robot Learning*, 2019. 2

[2] Miguel Arduengo, Carme Torras, and Luis Sentis. Robust and adaptive door operation with a mobile robot. *arXiv e-prints*, pages arXiv–1902, 2019. 2

[3] Felix Burget, Armin Hornung, and Maren Bennewitz. Whole-body motion planning for manipulation of articulated objects. In *2013 IEEE International Conference on Robotics and Automation*, pages 1656–1662. IEEE, 2013. 2

[4] Sachin Chitta, Benjamin Cohen, and Maxim Likhachev. Planning for autonomous door opening with a mobile manipulator. In *2010 IEEE International Conference on Robotics and Automation*, pages 1799–1806. IEEE, 2010. 2

[5] 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. 2

[6] Karthik Desingh, Shiyang Lu, Anthony Opipari, and Odest Chadwicke Jenkins. Factored pose estimation of articulated objects using efficient nonparametric belief propagation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7221–7227. IEEE, 2019. 2

[7] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. *arXiv preprint arXiv:2205.04382*, 2022. 2

[8] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018. 2

[9] Peter Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 2019. 2

[10] Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, and He Wang. Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts. *arXiv preprint arXiv:2211.05272*, 2022. 2, 3, 4

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

[12] Karol Hausman, Scott Niekum, Sarah Osentoski, and Gaurav S Sukhatme. Active articulation model estimation through interactive perception. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3305–3312. IEEE, 2015. 2

[13] Sebastian Höfer, Tobias Lang, and Oliver Brock. Extracting kinematic background knowledge from interactions using task-sensitive relational learning. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4342–4347. IEEE, 2014. 2

[14] Xiaoxia Huang, Ian Walker, and Stan Birchfield. Occlusion-aware reconstruction and manipulation of 3d articulated objects. In *2012 IEEE International Conference on Robotics and Automation*, pages 1365–1371. IEEE, 2012. 2

[15] Advait Jain and Charles C Kemp. Pulling open novel doors and drawers with equilibrium point control. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 498–505. IEEE, 2009. 2

[16] Ajinkya Jain, Rudolf Lioutikov, Caleb Chuck, and Scott Niekum. Screwnet: Category-independent articulation model estimation from depth images using screw theory. *arXiv preprint arXiv:2008.10518*, 2020. 2

[17] Dov Katz, Moslem Kazemi, J Andrew Bagnell, and Anthony Stentz. Interactive segmentation, tracking, and kinematic modeling of unknown 3d articulated objects. In *2013 IEEE International Conference on Robotics and Automation*, pages 5003–5010. IEEE, 2013. 2

[18] Dov Katz, Andreas Orthey, and Oliver Brock. Interactive perception of articulated objects. In *Experimental Robotics*, pages 301–315. Springer, 2014. 2

[19] Dov Katz, Yuri Pyuro, and Oliver Brock. Learning to manipulate articulated objects in unstructured environments using a grounded relational representation. In *In Robotics: Science and Systems*. Citeseer, 2008. 2

[20] Hedvig Kjellström, Javier Romero, and Danica Kragić. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, 115(1):81–90, 2011. 2

[21] Ellen Klingbeil, Ashutosh Saxena, and Andrew Y Ng. Learning to open new doors. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2751–2757. IEEE, 2010. 2

[22] 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. 2

[23] Xiaolong Li, He Wang, Li Yi, Leonidas J Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3706–3715, 2020. 2

[24] Jacky Liang, Viktor Makoviychuk, Ankur Handa, Nuttapong Chentanez, Miles Macklin, and Dieter Fox. Gpu-accelerated robotic simulation for distributed reinforcement learning, 2018. 2, 3, 4

[25] Qihao Liu, Weichao Qiu, Weiyao Wang, Gregory D Hager, and Alan L Yuille. Nothing but geometric constraints: A model-free method for articulated object pose estimation. *arXiv preprint arXiv:2012.00088*, 2020. 2

[26] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 2

[27] Roberto Martin Martin and Oliver Brock. Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2494–2501. IEEE, 2014. 2

[28] Roberto Martın-Martın and Oliver Brock. Building kinematic and dynamic models of articulated objects with multi-

modal interactive perception. In *AAAI Symposium on Interactive Multi-Sensory Object Perception for Embodied Agents, AAAI, Ed*, 2017. 2

[29] Roberto Martín-Martín and Oliver Brock. Coupled recursive estimation for online interactive perception of articulated objects. *The International Journal of Robotics Research*, page 0278364919848850, 2019. 2

[30] Roberto Martín-Martín, Sebastian Höfer, and Oliver Brock. An integrated approach to visual perception of articulated objects. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5091–5097. IEEE, 2016. 2

[31] Mayank Mittal, David Hoeller, Farbod Farshidian, Marco Hutter, and Animesh Garg. Articulated object interaction in unknown scenes with whole-body mobile manipulation. *arXiv preprint arXiv:2103.10534*, 2021. 2

[32] Kaichun Mo, Leonidas Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4

[33] 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 (ICCV)*, pages 6813–6823, October 2021. 2

[34] 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. 2

[35] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-sdf: Learning disentangled signed distance functions for articulated shape representation. *arXiv preprint arXiv: 2104.07645*, 2021. 2

[36] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. ManiSkill: Generalizable Manipulation Skill Benchmark with Large-Scale Demonstrations. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 4

[37] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *ICCV*, 2019. 2

[38] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In *NeurIPS*, 2020. 2

[39] Urbano Miguel Nunes and Yiannis Demiris. Online unsupervised learning of the 3d kinematic structure of arbitrary rigid bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3809–3817, 2019. 2

[40] Antonio Paolillo, Anastasia Bolotnikova, Kévin Chappellet, and Abderrahmane Kheddar. Visual estimation of articulated objects configuration during manipulation with a humanoid. In *2017 IEEE/SICE International Symposium on System Integration (SII)*, pages 330–335. IEEE, 2017. 2

[41] Antonio Paolillo, Kevin Chappellet, Anastasia Bolotnikova, and Abderrahmane Kheddar. Interlinked visual tracking and robotic manipulation of articulated objects. *IEEE Robotics and Automation Letters*, 3(4):2746–2753, 2018. 2

[42] L Peterson, David Austin, and Danica Kragic. High-level control of a mobile manipulator for door opening. In *Proceedings. 2000 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2000)(Cat. No. 00CH37113)*, volume 3, pages 2333–2338. IEEE, 2000. 2

[43] Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, and Hao Su. S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In *Conference on robot learning*, pages 53–65. PMLR, 2020. 2

[44] Zengyi Qin, Kuan Fang, Yuke Zhu, Li Fei-Fei, and Silvio Savarese. Keto: Learning keypoint representations for tool manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7278–7285. IEEE, 2020. 2

[45] Tanner Schmidt, Richard A Newcombe, and Dieter Fox. Dart: Dense articulated real-time tracking. In *Robotics: Science and Systems*, volume 2. Berkeley, CA, 2014. 2

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. 4

[47] Rafał Staszak, Milena Molska, Kamil Młodzikowski, Justyna Ataman, and Dominik Belter. Kinematic structures estimation on the rgb-d images. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, volume 1, pages 675–681. IEEE, 2020. 2

[48] Jürgen Sturm, Vijay Pradeep, Cyrill Stachniss, Christian Plagemann, Kurt Konolige, and Wolfram Burgard. Learning kinematic models for articulated objects. In *IJCAI*, pages 1851–1856, 2009. 2

[49] Jürgen Sturm, Cyrill Stachniss, and Wolfram Burgard. A probabilistic framework for learning kinematic models of articulated objects. *Journal of Artificial Intelligence Research*, 41:477–526, 2011. 2

[50] Yu Sun, Shaogang Ren, and Yun Lin. Object–object interaction affordance learning. *Robotics and Autonomous Systems*, 62(4):487–496, 2014. 2

[51] Dimitrios Tzionas and Juergen Gall. Reconstructing articulated rigged models from rgb-d videos. In *European Conference on Computer Vision*, pages 620–633. Springer, 2016. 2

[52] Yusuke Urakami, Alec Hodgkinson, Casey Carlin, Randall Leu, Luca Rigazio, and Pieter Abbeel. Doorgym: A scalable door opening environment and baseline agent. *Deep RL workshop at NeurIPS 2019*, 2019. 2

[53] 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 (CVPR)*, pages 2642–2651, 2019. 2

[54] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 2

[55] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J. Guibas. Captra: Category-level pose tracking for rigid and

articulated objects from point clouds. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 13209–13218, October 2021. 2

[56] 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. 2021. 1, 2

[57] Yueh-Hua Wu, Jiashun Wang, and Xiaolong Wang. Learning generalizable dexterous manipulation from human grasp affordance. *arXiv preprint arXiv:2204.02320*, 2022. 4

[58] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[59] Zhenjia Xu, Zhanpeng He, and Shuran Song. Umpnet: Universal manipulation policy network for articulated objects. *arXiv preprint arXiv:2109.05668*, 2021. 2

[60] Jingyu Yan and Marc Pollefeys. Automatic kinematic chain building from feature trajectories of articulated objects. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 712–719. IEEE, 2006. 2

[61] Zhihao Yan, Ruizhen Hu, Xingguang Yan, Luanmin Chen, Oliver van Kaick, Hao Zhang, and Hui Huang. RPM-NET: Recurrent prediction of motion and parts from point cloud. *ACM Trans. on Graphics*, 38(6):Article 240, 2019. 2

[62] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 2

[63] Kaichun Mo* Jiaqi Ke Qingnan Fan Leonidas Guibas Hao Dong Yian Wang*, Ruihai Wu*. Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions. In *ECCV*, 2022. 1, 2

[64] Vicky Zeng, Timothy E Lee, Jacky Liang, and Oliver Kroemer. Visual identification of articulated object parts. *arXiv preprint arXiv:2012.00284*, 2020. 2

[65] Yixin Zhu, Yibiao Zhao, and Song Chun Zhu. Understanding tools: Task-oriented object modeling, learning and recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2855–2864, 2015. 2