Supplemental Materials for “CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image”

Jingshun Huang[1*] Haitao Lin[1*] Tianyu Wang[1] Yanwei Fu[1] Xiangyang Xue[1] Yi Zhu[2]

[1]Fudan University [2]Huawei, Noah’s Ark Lab

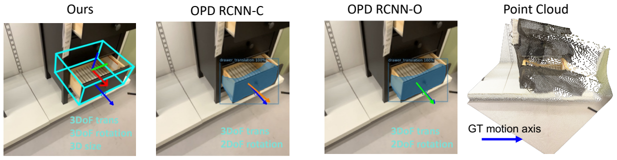
This supplemental material is organized as follows: In Section A, we compare results on real-world datasets and analyze inference speed in Section B. Section C includes additional examples from our RGBD-Art dataset. Details of the robotic experiments can be found in Section D. Section E presents and analyzes pose and size estimation results for unseen objects, small parts, and symmetric parts. Finally, we showcase more qualitative results to demonstrate the accuracy and sim-to-real adaptability of our approach in Section F.

A. Results on Real-world Dataset

OPD [3] and MultiScan [6] datasets are valuable contributions to the study of articulated objects, focusing on large scene-level parts with relatively coarse pose annotations while lacking smaller parts and fine-grained pose and size annotations. While our method should perform well to these datasets, they are not our main focus due to their limited coverage of smaller parts. We have attempted to use the MultiScan dataset, but its depth images remain inaccessible due to the ‘.zlib’ depth format, which we have been unable to decode. We provide motion axis error results on OPD-real dataset as in Tab. 1, highlighting robustness to realistic distorted depth.

Table 1. Comparison results on OPD [3] dataset.

Method	SG	AGP	GAPNet	OPD-C	OPD-O	Ours
Motion axis error↓	11.21°	12.03°	6.31°	9.06°	6.67°	**5.47°**



B. Inference Efficiency Analysis

The pre-trained vision model (SAM [7] and FeatUp [1]) slightly increases computational cost compared to the baselines (tested on an A6000 GPU), as shown in the Tab. 2. However, this cost is manageable for realistic robotics tasks, as demonstrated in the video. Inference speed is not the primary issue. During training, the backbone can preprocess images and store feature embeddings to save time. We also plan to optimize inference speed through distillation or quantization methods in future work.

Table 2. Inference speed of different methods.

Method	SG	AGP	GAPartNet	Ours
Inference (Hz)	5	7	**15**	4

C. Dataset Examples

We present additional rendered RGB-D images and their corresponding annotations from our RGBD-Art dataset in Fig. 2 and Fig. 3. The dataset is divided into two subsets: **seen** (Fig. 2) and **unseen** (Fig. 3). The **seen** subset contains objects with articulated parts similar to those in the training categories, while the **unseen** subset includes novel objects with previously unseen articulated parts that belong to the same categories.

D. Robotics Experiment Setup

Robotic Setup. We use the Kinova Gen2 6-DoF robotic arm to test our algorithm. The RealSense D435 camera captures RGB-D images of the scene and is mounted on a tripod across from the robot’s workspace. The camera is calibrated to the robotic base frame, as shown in Figure 1.

Manipulation Strategy. Similar to GAParNet [2], we adopt the manipulation stratey as follows after estimating the pose and size:

1. **Round Fixed Handle**: Approach the handle along the positive z-axis, open the gripper wider than the bounding box, and then close it to grasp.
2. **Line Fixed Handle**: Similar to the round handle, but orient the gripper’s opening perpendicular to the handle, aligning it with the y-axis of the bounding box.
3. **Hinge Handle**: Approach and grasp the hinge handle, then rotate it around the predicted axis of the revolute joint.
4. **Slider Button**: Close the gripper, approach the button from the positive z-axis, and press it.

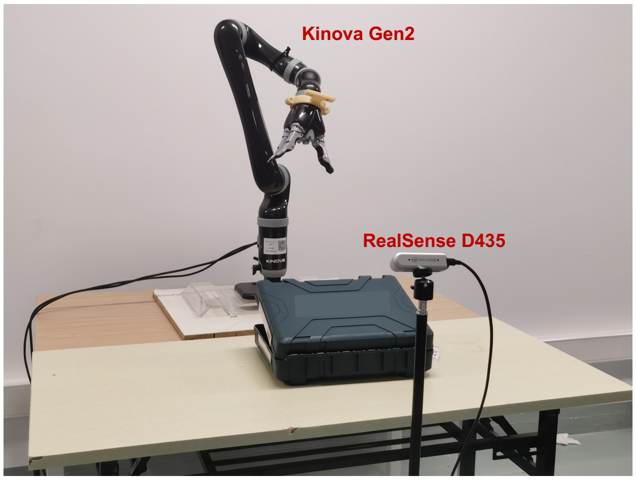


Figure 1. Robotic Setup.

5. **Slider Drawer**: Approach an open drawer along the z-axis to retrieve items, or along the x-axis to open it, typically targeting a handle on the front face.
6. **Hinge Door**: Grab the handle to open the door, rotating the gripper around the predicted shaft. If there's no handle and the door is ajar, clamp the outer edge along the y-axis to open it.
7. **Hinge Lid**: Use a similar approach as for the hinge door.

E. More Quantitative Results

Unseen Object Results. We present the pose and size results for unseen objects in our RGBD-Art dataset in Table 3. The results demonstrate that our method can generalize across object categories, effectively handling novel parts that belong to previously seen categories.

Table 3. Results of part pose estimation on unseen object categories. PG=baseline modified from PointGroup [4]. AGP=baseline modified from AutoGPart[5].

Method	$R_e\downarrow$	$T_e\downarrow$	$S_e\downarrow$	mIoU ↑	$A_5\uparrow$	$A_{10}\uparrow$
PG [4]	99.78	0.131	0.091	9.63	0.34	0.56
AGP [5]	105.62	0.125	0.088	12.54	0.37	0.74
GAPartNet [2]	90.81	0.073	0.052	30.71	0.54	1.03
Ours	**12.79**	**0.062**	**0.036**	**50.54**	**25.23**	**50.71**

Improvement on Small-part Objects. Table 4 has shown improved performance for small part classes like Hg.Kb and Sd.Bn. We also evaluate performance using an extreme challenge metric of $\frac{\text{part diameter}}{\text{object diameter}} \leq 0.1$ to further highlight small part performance. The improved performance on small parts is presented in the table below.

Improvement on Symmetric Parts. We present per-part pose results for the R_e metric ($\downarrow$), focusing on symmetric parts such as the slider button, hinge door, slider lid, and hinge lid, without symmetry tolerance. The results in Tab 5 show that our method effectively resolves visual ambiguity in rotation by incorporating global context.

Table 4. Comparison results on small parts. SG=SoftGroup [8]. AGP=baseline modified from AutoGPart [5]. '-' indicates no detection, and we show only 5 detected classes.

Small Parts(AP50↑)	Method	Ln.F.Hl.	Rd.F.Hl.	Sd.Bn	Hg.Dr.	Hg.Kb.
Seen	SG [8]	-	-	-	-	-
	AGP [5]	-	-	-	-	-
	GAPNet [2]	-	-	5.92	-	-
	Ours	**16.36**	**16.07**	**38.59**	**25.00**	**0.427**
Unseen	SG [8]	-	-	-	-	-
	AGP [5]	-	-	-	-	-
	GAPNet [2]	-	-	8.16	-	-
	Ours	**13.39**	-	**21.85**	**8.35**	**0.645**

Table 5. Comparison results on symmetric parts.

Sym.($R_e\downarrow$)	Method	Hg.Ld.	Sd.Ld.	Sd.Bn	Hg.Dr.
Seen	GAPNet	24.42	147.01	59.73	75.11
	Ours	**12.37**	**6.10**	**9.69**	**6.96**
Unseen	GAPNet	38.47	159.21	75.88	89.27
	Ours	**18.00**	**28.62**	**7.00**	**16.53**

F. More Qualitative Results.

We present additional qualitative results using our realistic RGBD-Art dataset, which mimics images captured by RealSense D415 and real-world images captured by RealSense D435. These cameras have different baselines, where the baseline of 55mm for the D415 and 50mm for the D435. The results are displayed in Fig. 4 and Fig. 5.

Results on RGBD-Art Datasets. As shown in Fig. 4, CAPNet effectively detects small parts and accurately estimates their poses and sizes by leveraging RGB image features.

Results on Real-world Images. Despite differences in camera baselines between real-world and training images, the results highlight the sim-to-real capability of our model, demonstrating cross-camera generalization and sim-to-real performance. This also underscores the value of our realism-enhanced dataset.

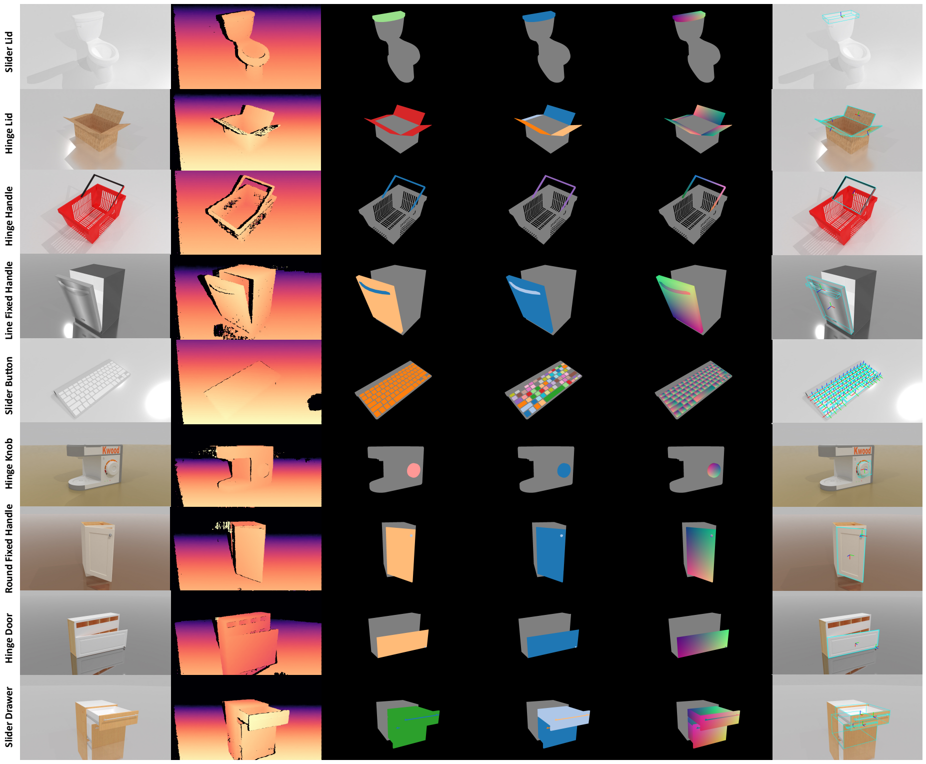


Figure 2. **Seen examples both used for training and testing in our RGBD-Art dataset.** We show the photo-realistic RGB image, realistic depth images, corresponding ground-truth annotations of semantic label, instance label, NPCS map, 6D pose and size.

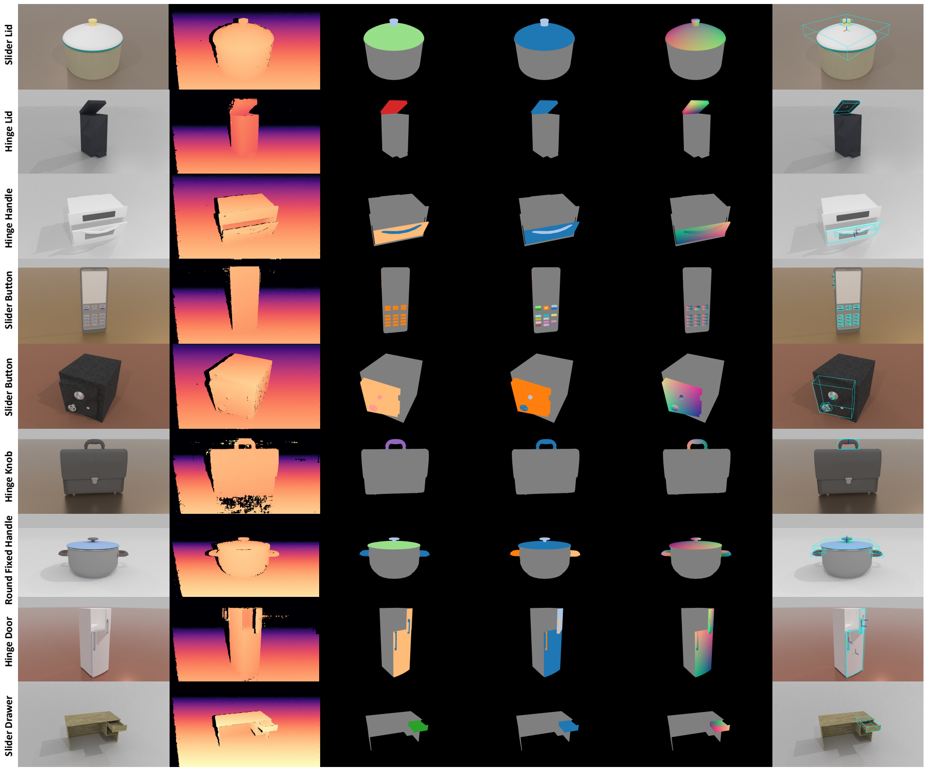


Figure 3. **Unseen examples used for testing in our RGBD-Art dataset.** We show the photo-realistic RGB image, realistic depth images, corresponding ground-truth annotations of semantic label, instance label, NPCS map, 6D pose and size.

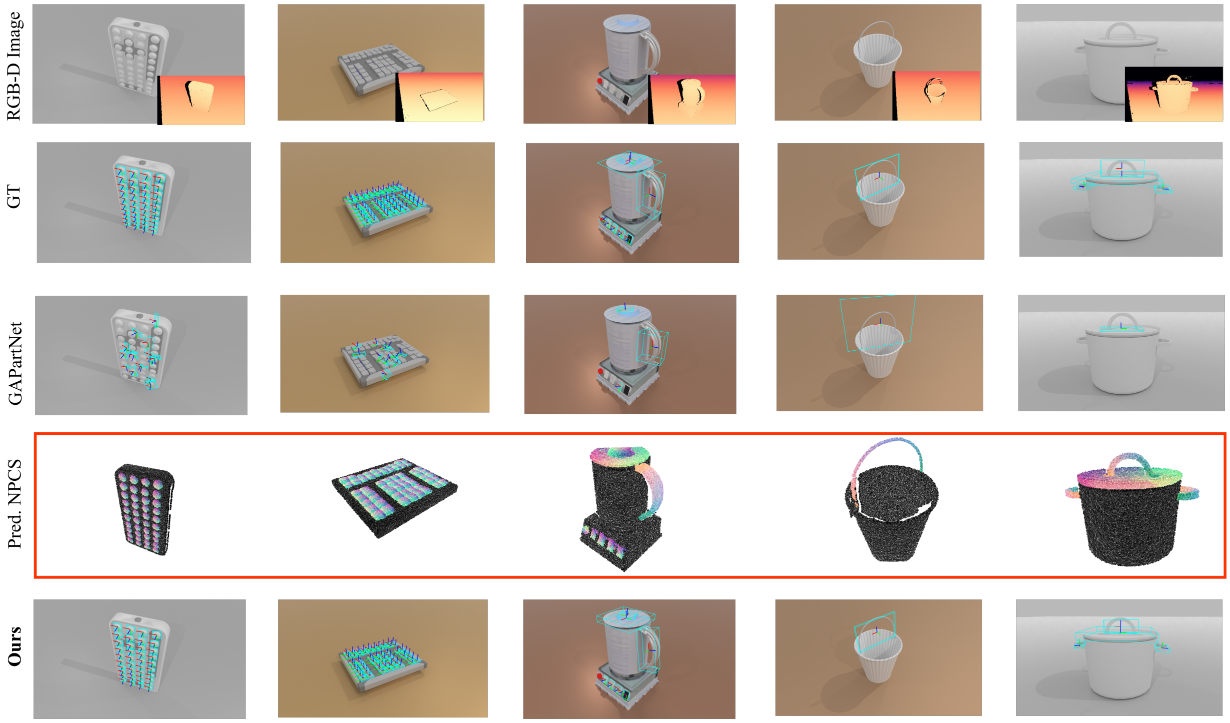


Figure 4. **Qualitative results of the RGBD-Art dataset.** We present the RGB-D images, the estimated NPCS for each component of our method, as well as the resulting pose and size estimations. Additionally, we provide comparisons with the baseline method GAPartNet [2] and ground truth annotations for qualitative evaluation .

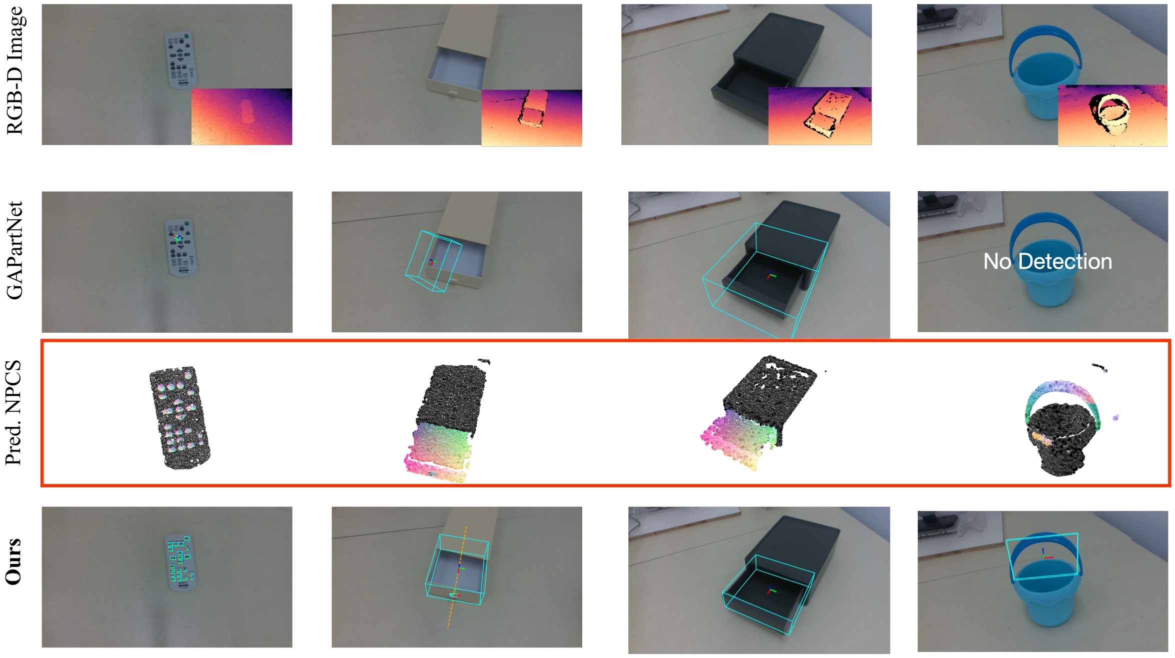


Figure 5. **Qualitative results from real-world images captured using the RealSense D435 camera.** We showcase the RGB-D images, the estimated NPCS for each component of our method, and the resulting pose and size estimations. Additionally, we provide comparisons with the baseline method GAPartNet [2] for qualitative evaluation .

References

[1] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 1

[2] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 1, 2, 4

[3] Hanxiao Jiang, Yongsen Mao, Manolis Savva, and Angel X Chang. Opd: Single-view 3d openable part detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 410–426. Springer, 2022. 1

[4] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 2

[5] Xueyi Liu, Xiaomeng Xu, Anyi Rao, Chuang Gan, and Li Yi. Autogpart: Intermediate supervision search for generalizable 3d part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11624–11634, 2022. 2

[6] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel X Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In *Advances in Neural Information Processing Systems*, 2022. 1

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[8] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 2