This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Abstract

*This paper tackles category-level pose estimation of articulated objects in robotic manipulation tasks and introduces a new benchmark dataset. While recent methods estimate part poses and sizes at the category level, they often rely on geometric cues and complex multi-stage pipelines that first segment parts from the point cloud, followed by Normalized Part Coordinate Space (NPCS) estimation for 6D poses. These approaches overlook dense semantic cues from RGB images, leading to suboptimal accuracy, particularly for objects with small parts. To address these limitations, we propose a single-stage **Net**work, CAP-Net, for estimating the 6D poses and sizes of **C**ategorical **A**rticulated **P**arts. This method combines RGB-D features to generate instance segmentation and NPCS representations for each part in an end-to-end manner. CAP-Net uses a unified network to simultaneously predict point-wise class labels, centroid offsets, and NPCS maps. A clustering algorithm then groups points of the same predicted class based on their estimated centroid distances to isolate each part. Finally, the NPCS region of each part is aligned with the point cloud to recover its final pose and size. To bridge the sim-to-real domain gap, we introduce the RGBD-Art dataset, the largest RGB-D articulated dataset to date, featuring photorealistic RGB images and depth noise simulated from real sensors. Experimental evaluations on the RGBD-Art dataset demonstrate that our method significantly outperforms the state-of-the-art approach. Real-world deployments of our model in robotic tasks underscore its robustness and exceptional sim-to-real transfer capabilities, confirming its substantial practical utility. Our dataset, code and pre-trained models are available on the project page [2].*

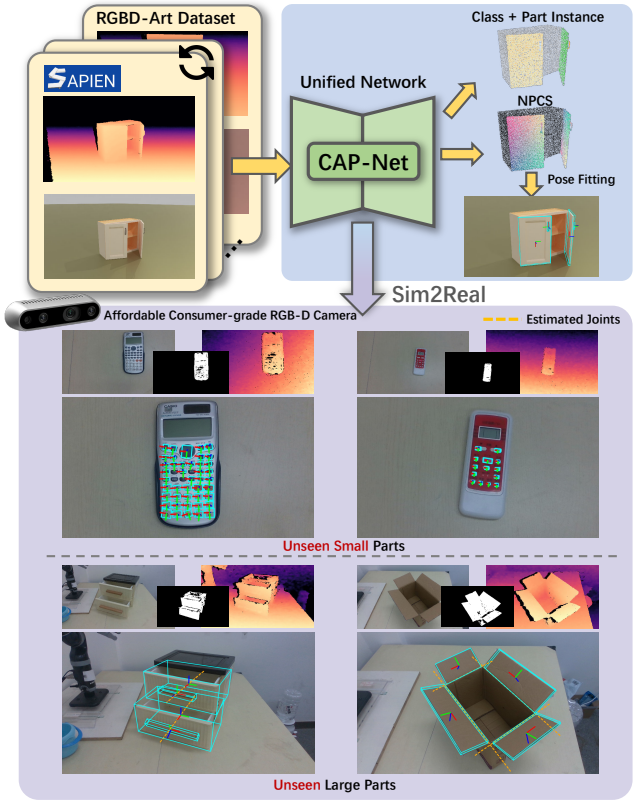

Figure 1. **Overview.** CAP-Net is a unified approach for estimating the 6D pose and size of all articulated parts from RGB-D images, requiring only object-level masks instead of part-level ones. The realistic training images in our RGBD-Art dataset allow this synthetic-trained model to effectively adapt to real-world visual perception tasks for robotic manipulation using an affordable RealSense camera.

1. Introduction

Accurately estimating object state is essential for robots to perform diverse grasping and manipulation tasks [2, 15, 29, 39, 48, 49], as shown in Fig. 1. Recent works [4, 11, 34] have advanced state estimation for rigid objects, but non-rigid objects like garments [5, 45], fluids [17, 24], and artic-

*indicates equal contributions.

*Prof. Yanwei Fu is also with Institute of Trustworthy Embodied Al, and the School of Data Science, Fudan University

[2]Webpage. https://shanehuanghz.github.io/CAPNet

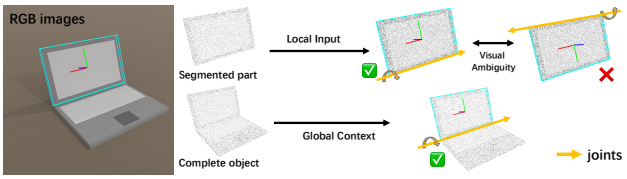

Figure 2. Comparison of complete-object versus part-based approaches for pose estimation: Recent "first-segment-then-pose" methods, like GAPartNet [10], rely on processing segmented parts in isolation. This approach can lead to a loss of global context, making it difficult to accurately identify the pose due to visual ambiguity in the input.

ulated parts [14, 19] remain challenging due to their complex properties. Articulated parts are particularly difficult because of their rigid, jointed parts, where inaccurate perception can risk damaging delicate joints. Therefore, we focus on improving the perception and estimation of articulated parts to enhance robotic manipulation capabilities.

However, significant challenges remain in accurately perceiving articulated parts, including: 1) *Intra-category part variations.* Novel articulated objects often lack specific 3D CAD models, requiring robust intra-category generalization. For example, estimating the handles of different bucket types necessitates finding shared representations that generalize across various instances within a category. 2) *Cross-category contextual variations.* Articulated parts exhibit diverse contextual configurations across object categories. Unlike category-level rigid object pose methods [4, 16, 51], which focus on single, consistent shapes, articulated objects involve multiple kinematic parts with unique contexts. This variability complicates pose and size estimation across different object categories. 3) *Sim-to-real domain gap.* Annotating articulated parts in real-world settings is costly, leading to reliance on synthetic data with more affordable annotations [10, 14, 19, 40]. However, most synthetic datasets lack photorealistic RGB images and realistic depth data that simulate real sensor capture, creating a sim-to-real gap when transferring pose estimation to real-world applications.

Prior point-based methods [10, 14, 19] tackle articulated object pose estimation using Normalized Part Coordinate Space (NPCS) as a standardized frame for category-level mapping from the camera frame. These approaches rely on geometric features but often ignore crucial semantic cues from RGB images, which are vital for precise pose identification. Intra-category methods [14, 19] lack cross-category generalization, while GAPartNet [10] extends to cross-category objects but employs a two-stage process—first segmenting parts, then estimating NPCS individually. This cascading approach can accumulate segmentation errors and lose contextual information. For instance, seeing only a laptop's top part makes it challenging to distinguish the front from the back, while the full object provides cues to resolve this ambiguity (Fig. 2).

To address these challenges, this paper presents a unified, single-shot RGB-D **Net**work for 6D pose and size estimation of **C**ategorical **A**rticulated **P**arts (CAP-Net), tailored for robotic manipulation tasks. (1) To handle intra-category part variations, we utilize a Normalized Part Coordinate Space (NPCS) [10, 14] for each part category. (2) To address cross-category contextual variations, we incorporate category-agnostic features from a pre-trained visual backbones [9, 25] for part segmentation across unseen object categories. Additionally, our model jointly learns semantic segmentation, centroid offsets, and NPCS mapping. Semantic supervision differentiates part classes, while centroid offsets help distinguish instances within categories, enabling precise localization of novel objects. (3) To bridge the sim-to-real gap, we introduce the RGB-D Realistic-Rendering Articulated Object (RGBD-Art) dataset, including photorealistic RGB images and realistic depth images that simulate sensor noise, enabling the adaptation of synthetically trained models to real-world scenarios.

Typically, given RGB-D images, we use the pretrained SAM2 [28] encoder to extract dense features, while FeatUp [9] refines SE(3)-consistent and category-agnostic local semantic features from DINOv2 [25, 50, 53] to high resolution. These fine-grained semantic features are crucial for part-level detection and pose representation learning. The dense semantic features are fused with the geometric point cloud in a point-wise manner, similar to [33]. Using PointNet++ [27], we optimize rigid part segmentation, instance centroid offsets, and dense NPCS coordinate predictions in an end-to-end manner, enhancing overall pose estimation performance. Finally, we align the NPCS regions of each segmented instance with the real point cloud using the Umeyama algorithm [31] to recover the final pose and size.

In summary, the main contributions of this paper are: (1) We introduce the first RGB-D method to estimate semantic labels, point-wise instance centers, and NPCS representations simultaneously in an end-to-end, single-shot process using whole-object input. This enables accurate perception of unseen articulated parts from a single RGB-D image in real-world settings. (2) We present a large-scale, realistic RGB-D dataset with detailed annotations for articulated pose estimation, effectively bridging the domain gap between synthetic and real-world images at both photometric and geometric levels. (3) Our method and dataset facilitate sim-to-real transfer to a physical robot, allowing accurate visual perception of articulated components, especially small parts, using only low-cost RGB-D sensors. This capability supports precise manipulation of target components, highlighting its practical utility in robotic applications.

2. Related Work

Part Instance Segmentation. Part instance segmentation in 3D point clouds is challenging due to the irregular and sparse nature of the data. Existing methods, such as PointNet++ [27] and SGPN [37], perform segmentation by learning point-wise features using deep learning. More recent methods, like SoftGroup [32], PointGroup [13], and AutoPart [20], focus on segmenting instances using geometric cues from the point cloud. GAPartNet [10] also leverages geometric features to learn domain-invariant representations for generalizing across unseen categories. In contrast, our approach incorporates semantic features from RGB images into the point cloud, enhancing segmentation accuracy and improving generalization across different categories and contextual variations.

Category-level Rigid Object Pose Estimation. Rigid object pose estimation deals with objects having a fixed shape, requiring a single static pose in 3D space. Some methods [3, 30, 34–36] estimate pose and size from single-view RGB(-D) images, like NOCS [34]. Other depth-based approaches, such as FS-Net [4], SAR-Net [16], HS-Pose [54], and GenPose [51, 52], estimate pose by purely learning the object's geometry. Unlike these methods, which are limited to rigid objects, our approach focuses on estimating the pose and size of objects with multiple movable parts. Rigid object methods estimate the entire object without needing to consider global context, relying instead on specifically designed metrics [23, 43] to address symmetry-related ambiguities. In contrast, our method incorporates global context to resolve visual ambiguity and handle the complexities of pose estimation for objects with movable components.

Category-level Articulated Object Pose Estimation. Articulated object pose estimation [10, 14, 18, 19, 40] focuses on objects made of multiple movable parts, requiring the estimation of both the overall pose and the configuration of individual components. To improve accuracy and generalization, the Articulation-aware Normalized Coordinate Space Hierarchy (ANCSH) [14] represents articulated objects within a category. While notable works like [12, 47] address 5-DoF motion axis estimation, they fall short of providing full 6-DoF pose and 3D size estimation, which are crucial for precise robotic manipulation. Additionally, [47] relies on per-part segmentation masks, which are more difficult to obtain than whole-object masks. In contrast, our approach uses readily available object-level masks to estimate full 6-DoF pose and 3D size. The AKB-48 [18] dataset provides real-world 3D articulated models with a pipeline for part reconstruction and pose estimation. GAPartNet [10] uses a two-stage approach for domain-generalizable 3D part segmentation and pose estimation, but it suffers from error accumulation and relies on high-quality point clouds from industrial cameras. In contrast, our approach improves pose and size estimation accuracy by incorporating semantic cues from RGB images, which is robust to the noise depth from consumer-grade camera.

Articulated Object Dataset. Large-scale 3D object datasets like ShapeNet [1], Objaverse [6, 7], and OmniObject3D [41] treat objects as rigid bodies without part definitions, which is inadequate for fine-grained robotic tasks such as opening bottle caps or pressing buttons. These tasks necessitate part-level perception, prompting the creation of 3D part-wise datasets [22, 38] for part-based object recognition and manipulation [26, 44, 46]. Further research, such as GAPartNet [10], uses public repositories like PartNet-Mobility [42] and AKB-48 [18] to generate point clouds with part annotations. Meanwhile, ReArtMix [19] produces semi-authentic RGB-D images using mixed reality techniques and self-scanning models called ReArt-48. These datasets lack photorealistic RGB images and depth maps that mimic real sensor noise, creating a sim-to-real gap. Our RGBD-Art dataset addresses this by offering realistic RGB images and depth maps that include sensor noise.

3. RGBD-Art Dataset

3.1. Definition of Poses and Joints

Each part category is canonically oriented and normalized to the Normalized Part Coordinate Space (NPCS) [10, 14, 34], ensuring a consistent definition of pose and joint parameters. This definition simplifies the network's task, as it only needs to estimate the pose without separately estimating joint parameters. In real-world robotics, once the object category and 6D pose are established, predefined joint parameters can be easily queried from the estimated oriented bounding box for manipulation, as shown in [10].

3.2. Realistic RGB-D Rendering

Recent datasets like CArt [14], GAPartNet [10], and ReArtMix [19] have driven progress in articulated object pose estimation but still have key limitations. (1) Their synthetic depth images are idealized and lack the realistic noise in real-world sensors like RealSense or Kinect, creating a sim-to-real domain gap when training solely on synthetic data. GAPartNet attempts to address this by using an industrial RGB-D camera for high-quality depth images in real-world applications, but these cameras are costly and have low frame rates, limiting their use on household robotic platforms. (2) Additionally, the non-photorealistic RGB images in these datasets create a significant domain gap due to pixel-level distribution differences, causing most works [10, 14, 19] to prioritize geometric cues from point clouds while neglecting valuable RGB features.

To address limitations in existing datasets, we adopt the techniques in [42] to synthesize depth images with realistic sensor noise patterns, simulating an active stereo depth camera similar to the RealSense D415. For RGB images,

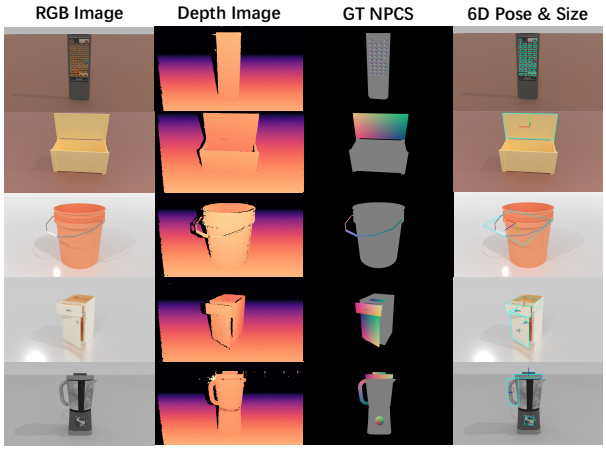

Figure 3. Exemplars of our RGBD-Art dataset. We show the photo-realistic RGB image, realistic depth images and corresponding ground-truth annotations of NPCS map, 6D pose and size.

we employ ray-tracing to achieve photorealism, incorporating domain randomization to vary lighting, textures, and object materials. With these realistic RGB-D images, we can readily obtain accurate annotations for part segmentation and pose. Figure 3 shows the the exemplars of our rendering RGB-D images and the generated ground-truth labels. We also compare our RGBD-Art with the existing synthetic dataset as in Tab. 1.

Table 1. Comparison with other synthetic articulated object datasets. 'P-RGB' represents photorealistic RGB images, 'R-D' denotes more realistic depth of objects, and 'BG' indicates datasets that use real images solely for the background.

Datasets	Obj.	Img.	Anno.	P-RGB	R-D
ReArtMix [19]	48	100K	-	✔(BG)	✘
GAPartNet [10]	1166	37K	272K	✘	✘
RGBD-Art(Ours)	1045	63K	408K	✔	✔

Data statistics. Building on the objects used in the GAPartNet dataset [10], we incorporate existing resources from PartNet-Mobility [42] that provide URDF models of articulated objects with unified annotations for each part type. AKB-48 [18] is not used due to its lack of 3D object shapes, which could impact the quality of pose annotations and rendered images. Our dataset contains 9 categories of articulated part types, including: *line fixed handle*, *round fixed handle*, *hinge handle*, *hinge lid*, *slider lid*, *slider button*, *slider drawer*, *hinge door*, and *hinge knob*. Each instance contains multiple classes of parts. To enhance realism, we place each instance on a flat surface or a desk, simulating typical backgrounds found in household settings. For each instance, we place it in two random backgrounds and, for each background, randomly sample 60 camera views to render RGB-D images. In total, we generate 63K images, complete with semantic labels and pose annotations.

4. Articulated Part Pose Estimation Method

Task Formulation. Given the RGB-D image patch $(\mathcal{I}, \mathcal{D})$, where $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ represents the RGB image and $\mathcal{D} \in \mathbb{R}^{H\times W}$ represents the depth map of the target object, our goal is to estimate the semantic class label c_i, the normalized object part coordinate maps m_i, and the centroid offsets o_i for each point i in the depth back-projected point cloud $\mathcal{P}$, where $\mathcal{P} \in \mathbb{R}^{N\times 3}$ and N is the number of points in the point cloud.

We start by using the semantic labels to group different part classes. Next, we cluster the centroid offsets o_i to distinguish between different part instances that share the same semantic label. We align the estimated NPCS maps of k-th part instance S_k with the corresponding 3D points to compute the pose and size parameters $\{\mathbf{R}_k, \mathbf{t}_k, \mathbf{s}_k\}$, where $\mathbf{R}_k \in \mathrm{SO}(3)$, $\mathbf{t}_k \in \mathbb{R}^3$, and $\mathbf{s}_k \in \mathbb{R}^3$. Here, $\mathrm{SO}(3)$ represents the Lie group of 3D rotation.

Architecture Overview. As shown in Fig. 4, our network first utilizes the pretrained SAM2 [28] and FeatUp [9] encoders to extract semantic features from RGB image patches $\mathcal{I}$, which are then concatenated with the corresponding point cloud $\mathcal{P}$ (Sec. 4.1). We then apply the PointNet++ [27] backbone to further extract geometric features, which are passed to three modules for multi-task learning. These modules include a semantic part learning module $\mathbf{M}_{\text{sem}}$ (Sec. 4.2), a centroid offset learning module $\mathbf{M}_{\text{inst}}$ (Sec. 4.3), and an NPCS learning module $\mathbf{M}_{\text{npcs}}$ (Sec. 4.4), which simultaneously estimate semantic class labels, centroid offsets, and NPCS coordinates for each point.We further cluster points based on their centroid offsets to group those belonging to the same instance, after which part labels are assigned to each instance to filter and locate the NPCS maps for each part. Finally, transformations and scale estimations between the actual points and the NPCS maps produce the 6-DoF pose and 3D size of each part.

4.1. Feature Extraction and Fusion

Unlike previous point-based methods [10, 14, 40], our approach integrates RGB images to provide semantic cues for pose estimation. Since our task differs from previous tasks like pose estimation or two-stage articulated pose estimation, we require the network to capture fine-grained features to distinguish multiple small parts within the complete object. Thus, SAM vision encoder provides strong dense feature representations as its excellent open-vocabulary instance segmentation capabilities. Meanwhile, FeatUp extracts SE(3)-consistent and category-agnostic local semantic features similar to DINOv2 [25, 50, 53], but preserve higher resolution details.

Particularly, our network begins by utilizing the pretrained backbone of SAM2 [28] alongside the FeatUp [9] encoder to extract features from the RGB image patch $\mathcal{I}$ separately as illustrated in Fig. 4.Specifically, the SAM en-

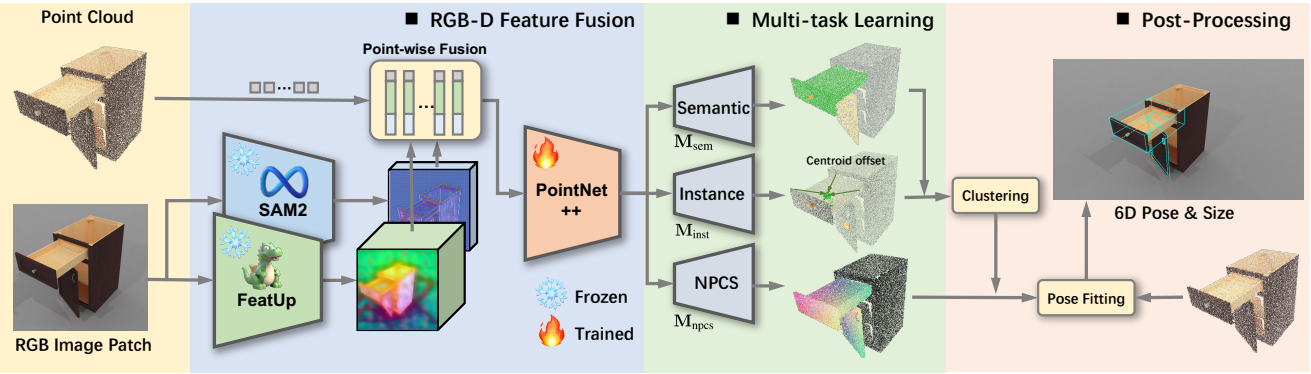

Figure 4. **Architecture overview.** CAP-Net uses pretrained vision backbones, SAM2 [28] and FeatUp [9], to extract dense semantic features, which are then fused with the point cloud in a point-wise manner. The enriched point cloud features are passed into PointNet++ [27] for further processing. These features are then used by three parallel modules to predict semantic labels, centroid offsets, and NPCS maps. A clustering algorithm groups points with the same semantic label based on centroid distances to isolate each possible part. Finally, an alignment algorithm matches the predicted NPCS map with the real point cloud to estimate each part's pose and size.

coder transforms an image of size $H \times W \times 3$ into a feature representation of size $H \times W \times 96$, while the FeatUp encoder maps the same image into a $H \times W \times 384$ embedding space. Next, we concatenate these feature maps, resulting in an embedding where each pixel corresponds to the 480-dimensional vectors that represent the semantic information of the RGB image at the corresponding location.

Subsequently, we concatenate each RGB feature vector with its corresponding 3D point in a point-wise manner, similar to [33]. The 3D points, now enriched with feature vectors, are processed using the pre-trained feature extractors from PointNet++ [27] to obtain the densely fused RGB-D features. We then employ three distinct decoders to estimate the corresponding outputs, as illustrated below.

4.2. Semantic Part Learning

To handle object categories with multiple parts, GAPartNet [10] employs a single network to segment each possible part from the complete point cloud. However, this approach loses the global context needed for accurate pose identification and introduces segmentation errors when each segmented part is processed independently by subsequent networks. To address these limitations, similar to [14, 19], we estimate all semantic labels directly from the input points. Unlike these methods, which assign each part instance a unique class label, even within the same category. They predefine semantic labels for each part, limiting their networks to single-category objects with fixed part counts. In contrast, our approach separately estimates class labels and instance labels (Sec. 4.3), enabling generalization across categories with variable part structures and counts.

Typically, we introduce a point-wise part semantic segmentation module, $\mathbf{M}_{\text{sem}}$, into the network to predict semantic labels for each point based on extracted RGB-D features. This module integrates dense semantic and geometric cues to streamline the segmentation task, with supervision provided by Focal Loss $FL(\cdot)$, denoted as $\mathcal{L}_{\text{sem}}$:

$$\mathcal{L}_{\text{sem}} = \frac{1}{N}\sum_{i=1}^{N} FL(\hat{c}_i, c_i), \tag{1}$$

where $\hat{c}_i$ is the predicted label for the i-th point, and c_i is the ground-truth class label.

4.3. Centroid Offset Learning

Considering that multiple part instances can share the same semantic label within a complete object, we design the centroid offset Llearning module $\mathbf{M}_{\text{inst}}$ to predict the centroid of each part instance, thereby enabling the network to differentiation between instances. This module predicts the Euclidean translation offset Δo_i of each point to the associated part instance center. The learning process for Δo_i is guided by an L1 loss:

$$\mathcal{L}_{\text{inst}} = \frac{1}{N}\sum_{i=1}^{N} ||\Delta o_i - \Delta \hat{o}_i|| \cdot \mathbb{I}(p_i \in S_k) \tag{2}$$

where N represents the total number of seed points on the object's surface, and Δo_i is the ground truth translation offset from seed p_i to the instance center. The indicator function $\mathbb{I}$ specifies whether point p_i belongs to the particular instance S_k.

Such optimization encourages each point to estimate the potential center of its corresponding part instance, ensuring that estimated centers belonging to the same instance are in close proximity. This facilitates the identification of each instance. Specifically, we further complete the instance segmentation process by first using the semantic labels to filter out background points, and then applying the DBSCAN [8] algorithm to cluster the estimated centers of foreground points, thereby distinguishing each instance.

4.4. NPCS Learning for Pose and Size Estimation

The semantic instance segmentation of the point cloud is handled by the semantic head, $\mathbf{M}_{\text{sem}}$, and the center offset head, $\mathbf{M}_{\text{inst}}$. Additionally, we incorporate an NPCS module, $\mathbf{M}_{\text{npcs}}$, which learns the mapping $\mathcal{P} \rightarrow \mathcal{P}_{\mathbb{C}} \in \mathbb{R}^{N\times 3}$, where $\mathcal{P}$ represents the observed object point cloud, and $\mathcal{P}_{\mathbb{C}}$ denotes the canonical-space point cloud. The mapping task is formulated as a classification problem by discretizing the NPCS coordinates into 32 bins per axis, which has proven more effective than regression for reducing the solution space [34]. We supervise NPCS learning with a softmax cross-entropy loss $SCE(\cdot)$, denoted as $\mathcal{L}_{\text{npcs}}$:

$$\mathcal{L}_{\text{npcs}} = \frac{1}{3N}\sum_{n=1}^{3}\sum_{i=1}^{N} SCE(\hat{m}_i^n, m_i^n), \tag{3}$$

where m_i^n and $\hat{m}_i^n$ are the ground truth and predicted discrete NPCS labels along each axis n.

Remark. Some classes exhibit continuous symmetry, so we apply the solution from [10, 34] for loss convergence. However, GAPartNet resolves ambiguities in parts like hinged lids or doors by tolerating 180° mirror symmetry due to its reliance on local part inputs. In contrast, our method leverages global context, enabling accurate pose direction distinction without relying on symmetry tolerance.

Multi-task Loss. We supervise the learning of modules $\mathbf{M}_{\text{sem}}$, $\mathbf{M}_{\text{inst}}$, and $\mathbf{M}_{\text{npcs}}$ jointly using a multi-task loss: $\mathcal{L} = \lambda_1\mathcal{L}_{\text{sem}} + \lambda_2\mathcal{L}_{\text{inst}} + \lambda_3\mathcal{L}_{\text{npcs}}$, where λ_1, λ_2, and λ_3 are the task-specific weights. Experimental results from the ablation study in Sec. 5.2 show that jointly training these tasks improves the final performance.

Using the predicted discrete correspondences, we recover the 6D object pose and size through post-processing pose fitting. This is achieved by first applying RANSAC for outlier elimination, followed by the Umeyama algorithm [31] to estimate the transformation parameters $\{\hat{\mathrm{s}}, \hat{\mathbf{R}}_k, \hat{\mathbf{t}}_k\}$ that align the predicted canonical-space point cloud with the estimated segmented part point cloud.

5. Experiment

Evaluation Metric. Following previous work [10], we evaluate (1) 3D semantic instance segmentation using AP50, which measures average precision at a 50% IoU threshold. We also use AP, the average precision across IoU thresholds from 50% to 95% in 5% increments, for overall performance assessment. For (2) pose estimation, we assess performance using metrics such as average rotation error $R_e(^\circ)$, translation error $T_e(cm)$, scale error $S_e(cm)$, and translation error along the interaction axis $d_e(cm)$. Additionally, we measure 3D Intersection over Union (3D mIoU) and calculate accuracy percentages for thresholds of 5° and $5cm$, as well as 10° and $10cm$.

Baselines. We use point-based methods such as PointGroup [13], SoftGroup [32], AutoGPart [20], and the state-of-the-art GAPartNet [10] as baseline methods. Note that PointGroup, SoftGroup, and AutoGPart have been adapted for part-based 3D semantic instance segmentation and pose estimation, similar to [10].

Implementation Details. We train our model on 4 L20 GPUs for a total of 100 epochs, using a batch size of 4. The initial learning rate is set to 0.001, using a warm-up scheduler for gradual increase at the start of training. Input images are cropped and resized to 640×640 resolution, and point clouds are randomly sampled to 24,576 points before being processed by the network. The hyperparameters of multi-task loss are empirically set to $\lambda_1 = 17.5$, $\lambda_2 = 125$, and $\lambda_3 = 0.15$ for semantic segmentation, instance offset learning, and NPCS mapping tasks, respectively.

5.1. Comparison to Baselines

We compared the accuracy of our method against baseline methods in part segmentation, pose and size estimation, with the results summarized in Tab. 2 and Tab. 3. While our method primarily focuses on pose and size estimation, we also evaluated 3D instance segmentation against baseline methods to highlight the importance of semantic cues from RGB images. The comparison results in Table 2 show that our method outperforms all point-based approaches [10, 20, 32], achieving an average precision (AP50) of 53.58 in seen object setting and 19.17 in unseen object setting, compared to other state-of-the-art methods. This suggests that incorporating RGB semantic cues is essential for part segmentation, especially for small components like buttons on a remote. Due to noise in depth data, distinguishing each instance using geometric cues can be difficult, but RGB images provide fine details that facilitate the identification of each part.

In the pose estimation tasks, the results in Tab. 3 demonstrate a significant improvement in pose accuracy with our method compared to the previous state-of-the-art method, GAParNet, particularly in the $\boldsymbol{Re}$ and $\mathbf{A}_5$ metric, confirming the accuracy of the estimated poses. GAParNet does not effectively resolve ambiguities in parts like hinged lids or doors; it only uses a relaxed metric that permits 180° mirror symmetry due to its reliance on local part inputs. In our strict evaluation, we remove this tolerance, leading to significant rotation errors in its two-stage method. This two-stage method struggles to distinguish poses based on individual segmented parts, as shown in Fig. 2. This highlights the importance of considering global context.

We also present qualitative results in Fig. 5, which demonstrate the sim-to-real capability of our method, even with noisy depth data from a consumer-grade RealSense camera in real-world settings. In contrast, the point-based method GAPartNet suffers a drop in accuracy. Unlike the

Table 2. Results of part segmentation on seen object categories and unseen object categories in terms of per-part-class AP50 (%), average AP50 (%) Ln.=Line. F.=Fixed. Rd.=Round. Hl.=Handle. Ld.=Lid. Bn.=Button. Dw.=Drawer. Dr.=Door. Kb.=Knob. PG=PointGroup [13]. SG=SoftGroup [32]. AGP=baseline modified from AutoGPart [20].

		Ln.F.Hl.	Rd.F.Hl.	Hg.Hl.	Hg.Ld.	Sd.Ld.	Sd.Bn	Sd.Dw.	Hg.Dr.	Hg.Kb.	Avg.AP50
Seen (%)	SG [32]	0.02	0.07	0.0	10.34	6.24	5.56	14.32	3.51	0.0	4.44
	AGP [20]	1.23	0.15	0.01	13.24	10.21	7.29	16.45	7.41	0.019	6.22
	GAPartNet [10]	3.97	0.26	0.0	25.94	18.41	12.07	26.34	15.15	0.038	11.35
	Ours	**55.30**	**16.88**	**70.71**	**76.05**	**93.76**	**44.06**	**58.09**	**51.05**	**16.23**	**53.58**
Unseen (%)	SG [32]	0.32	5.2	0.0	3.34	0.0	11.5	1.2	6.20	1.12	3.21
	AGP [20]	1.9	0.0	0.0	9.3	0.01	7.8	10.02	6.1	1.5	4.07
	GAPartNet [10]	2.27	0.19	0.0	10.94	0.02	10.42	18.32	14.07	3.1	6.59
	Ours	**28.88**	**0.925**	**0.67**	**51.69**	**1.23**	**28.53**	**20.47**	**24.05**	**18**	**19.38**

two-stage method GAParNet, our single-shot approach considers the entire object input rather than just the segmented part. This helps to resolve visual ambiguity in pose identification, as shown in Fig 2 and Fig. 5. We also present comparison results for indoor large scene parts from the real-world OPD [12] and MultiScan [21] datasets. For more results, please refer to the Appendix.

Table 3. Results of part pose estimation on different evaluation metric. PG=baseline modified from PointGroup [13]. AGP=baseline modified from AutoGPart [20].

Method	$R_e \downarrow$	$T_e \downarrow$	$S_e \downarrow$	mIoU $\uparrow$	$A_5 \uparrow$	$A_{10} \uparrow$
PG [13]	89.30	0.091	0.057	18.41	0.54	1.21
AGP [20]	99.40	0.099	0.061	20.10	0.53	1.32
GAPartNet [10]	83.3	0.061	0.043	39.53	0.71	1.40
Ours	**10.39**	**0.055**	**0.026**	**56.23**	**33.91**	**58.44**

5.2. Ablation study

To evaluate the effectiveness of each component in our framework, we conduct an ablation study to evaluate the contributions of multi-task training and RGB features incorporation. The results of these evaluations are summarized in Tab. 4 and Tab. 5.

Importance of multi-task learning. We trained three individual sub-modules of our method, each supervised by a separate task. After training, we combined the results from the independently trained heads to estimate the final pose and size, and compared this performance with that of the model trained using multi-task learning. As shown in Table 4, the multi-task model consistently outperforms the individual models across all metrics. These results show that multi-task learning enhances pose and size estimation accuracy. Furthermore, the end-to-end optimization of the three sub-tasks improves performance, unlike a cascaded approach that may lead to error accumulation.

Importance of dense RGB features. To assess the contribution of RGB features, we start with a baseline method using the PointNet++ backbone, which relies solely on point cloud data. We then sequentially incorporate vision backbones, such as SAM2 [28], FeatUp [9], and DINOv2 [25]. The comparison results in Tab. 5 demonstrate that the SAM2 encoder notably improves accuracy. FeatUp, which provides high-resolution features, outperforms DINOv2 and further enhances performance. The full model leverages both feature types for optimal results, with RGB features enhancing the segmentation of small parts that are difficult to achieve with point cloud data alone.

Table 4. Comparison of model via multi-task training and individual training under different evaluation metric.

Method	$R_e \downarrow$	$T_e \downarrow$	$S_e \downarrow$	mIoU $\uparrow$	$A_5 \uparrow$	$A_{10} \uparrow$
Individual	15.30	0.063	0.036	51.75	27.41	50.86
Multi-task	**10.39**	**0.055**	**0.026**	**56.23**	**33.91**	**58.44**

Table 5. Comparison of variants of model and our CAP-Net. 'PNet' indicates network modified from PointNet++ [27].

Method	$R_e \downarrow$	$T_e \downarrow$	$S_e \downarrow$	mIoU $\uparrow$	$A_5 \uparrow$	$A_{10} \uparrow$
PNet	21.30	0.062	0.060	41.09	16.50	38.87
PNet + SAM	12.20	0.063	0.027	48.79	23.68	47.90
PNet + FeatUp	20.03	0.058	0.054	43.52	21.44	42.81
PNet + DINOv2	25.12	0.062	0.060	40.10	13.68	34.36
Full Model	**10.39**	**0.055**	**0.026**	**56.23**	**33.91**	**58.44**

5.3. Robotic Experiment

Hardware Settings. Our algorithm is deployed on a PC workstation equipped with an Intel i9-13900K CPU and an NVIDIA RTX 6000 Ada Generation GPU to provide visual perception of the target objects. For executing the grasping and manipulation tasks, we utilize the Kinova Gen2 6-DoF robotic arm. This robotic arm features three under-actuated fingers, each of which can be individually controlled. The RealSense D435 captures RGB-D images from a tripod positioned opposite the robot workspace and is calibrated to the robotic base frame.

Manipulation Strategy. Using the NPCS representation, we obtain the joint or prismatic axis and predefined grasp poses in the NPCS frame. By aligning the NPCS with the real-world point cloud, we transform these actionable axes and grasp poses to the camera frame. With camera calibration to the robot's base frame, this allows easy transformation of grasp poses to the robot frame for motion planning.

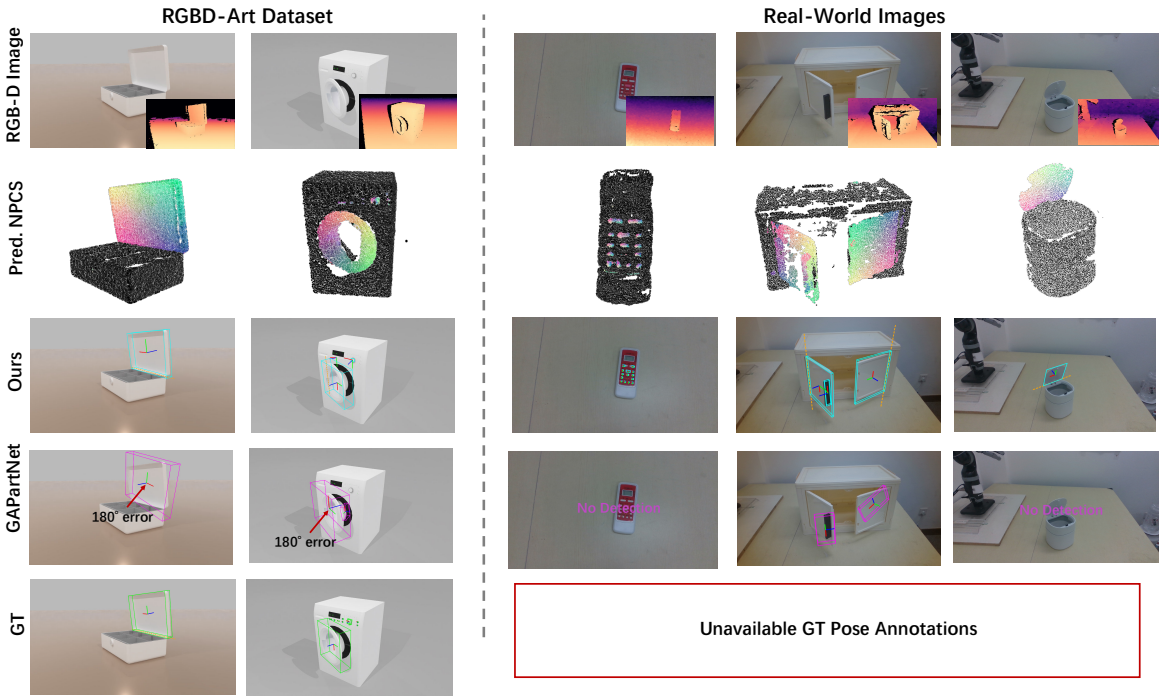

Figure 5. Qualitative results from the RGBD-Art dataset and real-world images captured using the RealSense D435 camera. We showcase the RGB-D images, the estimated NPCS for each component of our method, and the pose and size results. Additionally, we provide comparisons with the baseline method GAPartNet [10] and ground truth annotations.

Detailed motion policies for each category of part refers to our supplemental material.

Task Description. To assess the sim-to-real capability of our method and evaluate its robustness and generalizability, we deployed our algorithm onto a real KINOVA robotic arm. To ensure the representativeness of our experiments, we selected three distinct part classes: drawer, hinge lid, and hinge handle. The corresponding tasks involved pulling the drawer, lifting the lid, and raising the handle.

Evaluation Metric. Depending on the specific experimental task, different metrics are used. For the drawer task, the robot arm successfully completed the task by pulling the drawer out by 0.2 meters. For the hinge handle task, success was defined by rotating the handle 30 degrees around its axis. Similarly, for the hinge lid task, the robot arm successfully completed the task by rotating the lid 50 degrees around its axis.

Results. The success rate of manipulating articulated objects in real-world robotic experiments is summarized in Table 6. The results show that our lightweight model competes effectively with the baseline method, GAParNet. Our single-shot approach accurately generates poses that guide the robot in interacting with objects not seen during the training stages, demonstrating the utility of our method in robotic applications. More results please refers to our supplemental material and video.

Table 6. Success Rate of Robot Manipulation.

	Hinge Handle	Drawer	Hinge Lid	Total
GAPartNet [10]	1/10	2/10	2/10	5/30
Ours	**9/10**	**10/10**	**9/10**	**28/30**

6. Conclusion

We introduce CAP-Net, a unified model for estimating the 6D pose and size of articulated parts at the category level. Unlike multi-stage methods, CAP-Net uses a single-stage framework on object-level point clouds for end-to-end part pose and size estimation. Experiments on our RGBD-Art dataset and real-world data validate its synthetic-to-real capability, and robotic tests demonstrate its effectiveness on unseen articulated objects. However, our method relies on depth for pose fitting, and missing surface points can affect pose estimation accuracy. While off-the-shelf depth-completion methods can address this issue, it is beyond the scope of this paper and will be explored in future work.

Acknowledgment. This work is supported in part by NSFC Project (62176061), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), Shanghai Technology Development and Entrepreneurship Platform for Neuromorphic and AI SoC, and the Science and Technology Commission of Shanghai Municipality(No. 24511103100). Yanwei Fu and Xiangyang Xue are the corresponding authors.

References

[1] 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. 3

[2] Chilam Cheang, Haitao Lin, Yanwei Fu, and Xiangyang Xue. Learning 6-dof object poses to grasp category-level objects by language instructions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8476–8482. IEEE, 2022. 1

[3] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 3

[4] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 1, 2, 3

[5] Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3324–3333, 2021. 1

[6] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023. 3

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 5

[9] 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. 2, 4, 5, 7

[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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7081–7091, 2023. 2, 3, 4, 5, 6, 7, 8

[11] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022. 1

[12] 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. 3, 7

[13] 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. 3, 6, 7

[14] 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, 3, 4, 5

[15] Haitao Lin, Chilam Cheang, Yanwei Fu, and Xiangyang Xue. I know what you draw: Learning grasp detection conditioned on a few freehand sketches. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 8417–8423. IEEE, 2022. 1

[16] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2022. 2, 3

[17] Haitao Lin, Yanwei Fu, and Xiangyang Xue. Pourit!: Weakly-supervised liquid perception from a single image for visual closed-loop robotic pouring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 241–251, 2023. 1

[18] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 3, 4

[19] Liu Liu, Han Xue, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31:1072–1083, 2022. 2, 3, 4, 5

[20] 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. 3, 6, 7

[21] 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. 7

[22] 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. 3

[23] Ningkai Mo, Wanshui Gan, Naoto Yokoya, and Shifeng Chen. Es6d: A computation efficient and symmetry-aware 6d pose regression framework. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 6718–6727, 2022. 3

[24] Gautham Narasimhan, Kai Zhang, Ben Eisner, Xingyu Lin, and David Held. Self-supervised transparent liquid segmentation for robotic pouring. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 4555–4561. IEEE, 2022. 1

[25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 4, 7

[26] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3215, 2021. 3

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 3, 4, 5, 7

[28] 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. 2, 4, 5, 7

[29] Qiang Sun, Haitao Lin, Ying Fu, Yanwei Fu, and Xiangyang Xue. Language guided robotic grasping with fine-grained instructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1319–1326. IEEE, 2023. 1

[30] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020. 3

[31] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 2, 6

[32] 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. 3, 6, 7

[33] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 2, 5

[34] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1, 3, 6

[35] Tianyu Wang, Yifan Li, Haitao Lin, Xiangyang Xue, and Yanwei Fu. Wall-e: Embodied robotic waiter load lifting with large language model. *arXiv preprint arXiv:2308.15962*, 2023.

[36] Tianyu Wang, Haitao Lin, Junqiu Yu, and Yanwei Fu. Polaris: Open-ended interactive robotic manipulation via syn2real visual grounding and large language models. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9676–9683. IEEE, 2024. 3

[37] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 3

[38] 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. 3

[39] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. *arXiv preprint arXiv:2312.08344*, 2023. 1

[40] 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/CVF International Conference on Computer Vision*, pages 13209–13218, 2021. 2, 3, 4

[41] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 3

[42] 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/CVF conference on computer vision and pattern recognition*, pages 11097–11107, 2020. 3, 4

[43] 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. 3

[44] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 3

[45] Han Xue, Wenqiang Xu, Jieyi Zhang, Tutian Tang, Yutong Li, Wenxin Du, Ruolin Ye, and Cewu Lu. Garmenttracking: Category-level garment pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21233–21242, 2023. 1

[46] Kaizhi Yang and Xuejin Chen. Unsupervised learning for cuboid shape abstraction via joint segmentation from point clouds. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 3

[47] Vicky Zeng, Tabitha Edith Lee, Jacky Liang, and Oliver Kroemer. Visual identification of articulated object parts.

In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2443–2450. IEEE, 2021. 3

[48] Chuanrui Zhang, Yonggen Ling, Minglei Lu, Minghan Qin, and Haoqian Wang. Category-level object detection, pose estimation and reconstruction from stereo images. *arXiv preprint arXiv:2407.06984*, 2024. 1

[49] Jinyu Zhang, Yongchong Gu, Jianxiong Gao, Haitao Lin, Qiang Sun, Xinwei Sun, Xiangyang Xue, and Yanwei Fu. Lac-net: Linear-fusion attention-guided convolutional network for accurate robotic grasping under the occlusion. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10059–10065. IEEE, 2024. 1

[50] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4

[51] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[52] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2025. 3

[53] Ruida Zhang, Ziqin Huang, Gu Wang, Chenyangguang Zhang, Yan Di, Xingxing Zuo, Jiwen Tang, and Xiangyang Ji. Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation. *arXiv preprint arXiv:2409.15727*, 2024. 2, 4

[54] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17163–17173, 2023. 3