# You Only Estimate Once: Unified, One-stage, Real-Time Category-level Articulated Object 6D Pose Estimation for Robotic Grasping

Jingshun Huang[1*] Haitao Lin[2*] Tianyu Wang[1] Yanwei Fu[1] Yu-Gang Jiang[1] Xiangyang Xue[1]

*Abstract*— **This paper addresses the problem of category-level pose estimation for articulated objects in robotic manipulation tasks. Recent works have shown promising results in estimating part pose and size at the category level. However, these approaches primarily follow a complex multi-stage pipeline that first segments part instances in the point cloud and then estimates the Normalized Part Coordinate Space (NPCS) representation for 6D poses. These approaches suffer from high computational costs and low performance in real-time robotic tasks. To address these limitations, we propose YOEO, a single-stage method that simultaneously outputs instance segmentation and NPCS representations in an end-to-end manner. We use a unified network to generate point-wise semantic labels and centroid offsets, allowing points from the same part instance to vote for the same centroid. We further utilize a clustering algorithm to distinguish points based on their estimated centroid distances. Finally, we first separate the NPCS region of each instance. Then, we align the separated regions with the real point cloud to recover the final pose and size. Experimental results on the GAPart dataset demonstrate the pose estimation capabilities of our proposed single-shot method. We also deploy our synthetically-trained model in a real-world setting, providing real-time visual feedback at 200Hz, enabling a physical Kinova robot to interact with unseen articulated objects. This showcases the utility and effectiveness of our proposed method [2].**

## I. INTRODUCTION

Accurately estimating the state information of objects is crucial for robots before undertaking motion planning in various grasping and manipulation tasks [1]–[4], as shown in Fig 1. Recent research [5] has made significant progress in estimating the state of rigid bodies from single images. However, estimating the state information of non-rigid bodies remains challenging due to their complex physical properties. For example, recent works have explored the perception of garments [6], [7], fluids [8], [9], and articulated objects [10], [11]. Among these, articulated objects pose a unique challenge due to their multiple rigid kinematic parts, making their perception and manipulation particularly complex. Inaccurate perception of articulated objects can lead to the robot damaging delicate joints, unlike with liquids and garments, which are less susceptible to damage due to their flexible nature. In this work, we focus on advancing the perception and estimation of articulated objects to enhance robotic manipulation capabilities.

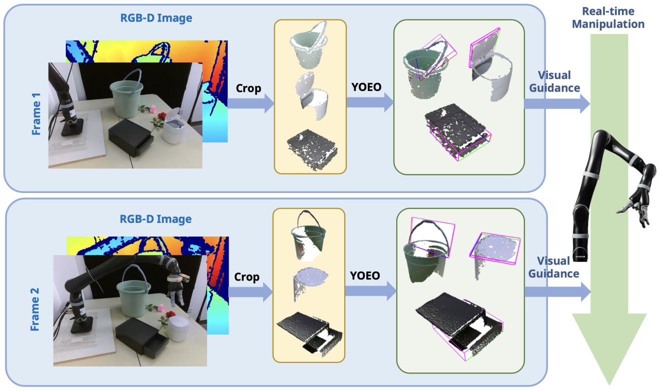


Fig. 1: **Overview.** We propose a unified, single-stage method for articulated object 6D pose estimation named YOEO, which enables real-time robotic manipulation.

However, there are still significant challenges in perceiving articulated parts. These challenges include: 1) *Intra-category part variations.* Novel articulated objects often lack exact 3D CAD models, necessitating intra-category generalization. For instance, estimating the handles of different bucket types requires finding shared representations that can generalize across various instances within a category, as illustrated in Fig. 2 (a). 2) *Cross-category context variations.* Articulated parts of a category exhibit vast variations in part contexts across different object categories. Unlike category-level rigid object pose methods [3], [12], [13], which deal with single, consistent shapes, articulated objects have multiple kinematic parts leading to diverse contexts. Thus, even parts from the same instance can be assembled differently with other rigid parts. For example, a hinged lid can be part of a laptop or a bin, as shown in Fig. 2 (b). This variability complicates the estimation of the pose and size of target parts across different categories of objects. These challenges highlight the need for advanced methods to accurately perceive and manipulate articulated objects, accommodating both intra-category part variations and cross-category context differences.

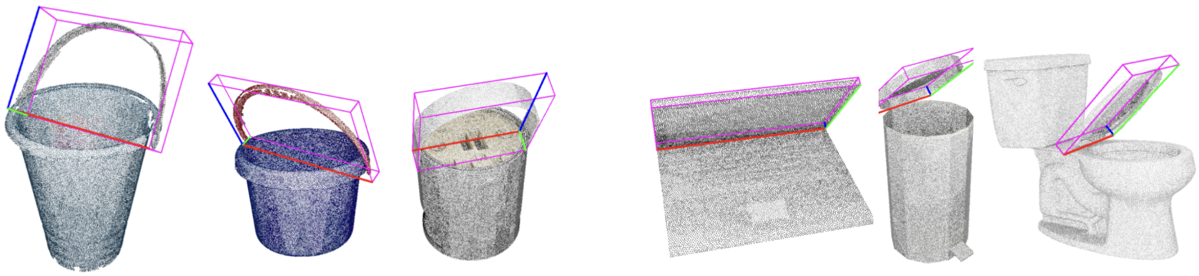
To tackle these challenges, previous methods [11], [14] propose Normalized Part Coordinate Space (NPCS) representation to provide a normalized space for canonical part references. This representation maps instances within the same category into a canonical space, facilitating the learning of category-level mapping from the camera frame to a shared frame. For example, Li et al. [11] address intra-category articulated pose estimation, but their method does not generalize to cross-category objects. Additionally,

$*$ indicates equal contribution.

[1]Fudan University. jshuang23@m.fudan.edu.cn.

[2]Tencent Robotics X Lab, Shenzhen, China.

[2]Project webpage. https://shanehuanghz.github.io/YOEO/



(a) Intra-category part variations

(b) Cross-category context variations

Fig. 2: (a) Illustration of challenge posed by intra-category part variations. Estimating the handles of different bucket types requires finding shared representations that can generalize across various instances within a category. (b) Illustration of challenge posed by cross-category context variations, exemplified by the variability of a hinge lid, which can be part of a laptop or a bin.This variability complicates the estimation of the pose and size of target parts across different categories of objects.

GAPartNet [14] learns domain-invariant features to facilitate the cross-category generalization. However, their method is a two-stage process, first segmenting the parts and then estimating the NPCS for each part individually. This cascaded pipeline accumulates segmentation errors, thereby reducing the accuracy of the NPCS estimation.

To address these challenges, this paper presents a simple yet efficient single-shot pipeline that You Only Estimate Once (YOEO), which provides a *unified, one-stage, real-time category-level* articulated object 6D Pose estimation for robotic grasping. Particularly, (1) To tackle the challenges of intra-category part variations, we represent each category of part in NPCS similar to works [11], [14]. This standardized space normalizes the position and orientation of parts, establishing a consistent reference frame for objects within the same category. Consequently, this facilitates accurate 6D pose and size estimation for unseen parts. (2) To address cross-category context variations, we jointly model semantic understanding and instance centroid offset while learning NPCS mapping. Semantic supervision enables the model to learn to distinguish between part classes belonging to the same object, thereby developing a more unified feature representation for each part class. Then, centroid offset learning enables the model to distinguish between multiple part instances within the same part classes. This enables the method to localize segments with similar category-level features in novel objects, even when these objects have different contextual parts, thus can generalizable to novel objects.

Typically, given that the point cloud of an articulated object contains multiple kinematic parts, we employ the unified network RandLA-Net [15] to learn object features, facilitating simultaneous optimization of rigid part semantic segmentation, dense coordinate predictions in each NPCS map, and instance centroid offsets. This end-to-end optimization process enhances the performance of each output, thereby improving the accuracy of pose estimation. Subsequently, part semantic segmentation and instance centroid offsets are used to filter and cluster for instance segmentation. Once each instance is obtained, we further extract the region of estimated NPCS and register it with the point cloud using the Umeyama algorithm [16], allowing the calculation of the final part pose and size.

In summary, the main contributions of this paper are as follows: (1) We introduce a synthetic-to-real pipeline designed to perceive previously unseen articulated object instances from a single depth input in real-world settings. (2) We propose an end-to-end unified network that concurrently estimates semantic labels, instance centroid offsets, and NPCS representations. This holistic optimization approach improves the accuracy of NPCS estimation and facilitates the generation of accurate 6D poses for each part, even amidst noisy depth data. (3) Our method is deployed within a real-time robotic system, enabling the visual perception of articulated objects at a rate of 200Hz. Furthermore, it guides the robot in real-time manipulation of the target part, demonstrating practical applicability in dynamic environments.

## II. Related Work

**3D Part-wise Objects Assets.** The task of 3D part-wise object representation and manipulation has gained significant attention in robotics and computer vision. Large-scale 3D datasets are the cornerstone of research in this field, such as ShapeNet [5], Objaverse [17], [18], OmniObject3D [19], etc. However, merely having a holistic perception of objects is insufficient. For fine-grained robotic manipulation (e.g., opening bottle caps, pressing buttons and opening refrigerators), a focused perception of object parts is often required, which necessitates the support of 3D part-wise datasets. Many previous works [20]–[23] have abstracted the shapes of 3D objects and decoupled the parts to construct datasets, thereby promoting a series of studies on part-wise object perception [24]–[26]. Furthermore, a dataset capable of supporting cross-category domain-generalizable object perception, GAPartNet [14], with rich part annotations, offering valuable guidance. Specifically, our model was primarily trained and tested on [14], and demonstrated its ability to estimate the part poses of the articulated objects.

**Part Instance Segmentation and Clustering from Point Cloud Observations.** Part instance segmentation in 3D point cloud is a challenging task due to the irregular and sparse nature of the data. Existing methods such as PointNet++ [27]

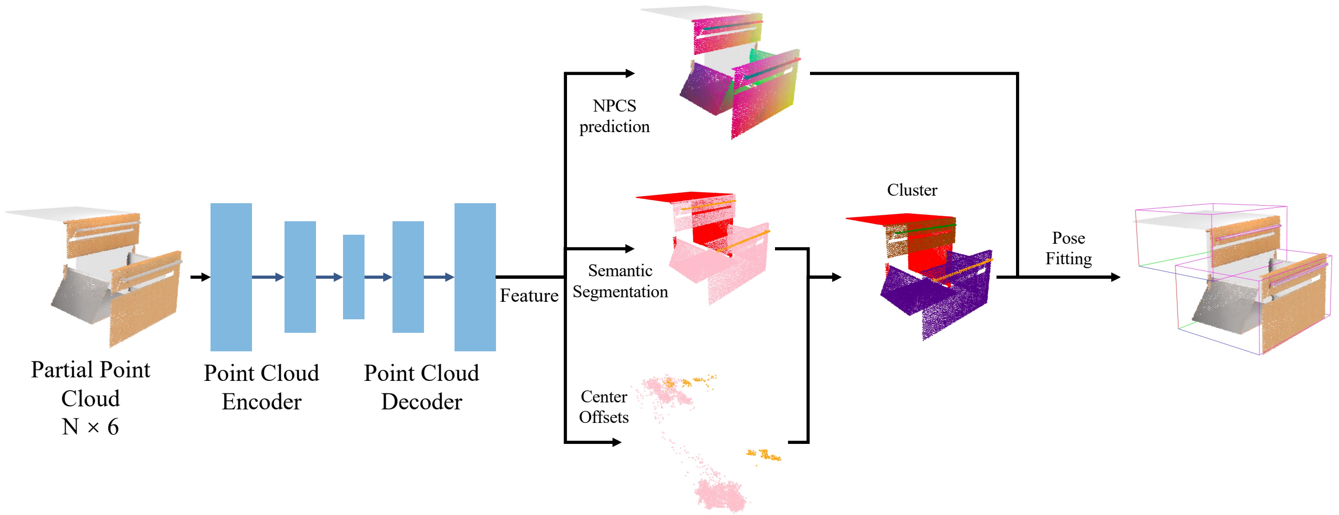


Fig. 3: **Architecture overview.** The Feature Extraction module extracts the per-point feature from an partial point cloud. They are fed into three parallel modules to predict the NPCS maps, semantic labels and the offsets to centroids of each point. A clustering algorithm is then applied to distinguish different instances with the same semantic label and points on the same instance. Finally, an aligning algorithm is applied to the predicted npcs map and real point cloud to estimate 6DoF pose parameters.

and SGPN [28] utilize deep learning techniques to segment instances by learning point-wise features. However, these methods typically require a separate clustering step to group points into instances, which can be computationally expensive. Recent advancements like VoteNet [29] and 3D-SIS [30] have improved clustering efficiency but still involve multi-stage processes. Our approach integrates instance segmentation and clustering [31], [32] within a unified network, leveraging point-wise centroid offsets to facilitate efficient and accurate segmentation. This end-to-end learning framework not only simplifies the pipeline but also improves the segmentation quality and speed.

**Category-level Rigid Object Pose Estimation.** Rigid object pose estimation deals with objects that maintain a fixed shape and structure, necessitating the determination of a single, static pose in 3D space. Some works [3], [12], [33]–[37] estimate the pose and size from single view RGB-D images. For example, NOCS [37] Some point-based methods like FS-Net [13], SAR-Net [3] and GenPose [12] focus on estimated the pose by learning the geometry shape of the instances. However, these methods are only suitable for rigid bodies, limiting their potential to be extended for perceiving complex objects composed of multiple movable parts.

**Category-level Articulated Object Pose Estimation.** Conversely, articulated object pose estimation addresses objects composed of multiple interconnected parts that can move relative to each other, requiring the estimation of both the overall pose and the configuration of individual movable components.To enhance accuracy and generalization across unseen articulated objects, Articulation-aware Normalized Coordinate Space Hierarchy (ANCSH) [11] was proposed to represent different articulated objects in a given category. [38] uses interactive learning to segment articulated objects into parts, discovering structures effectively and generalizing to unseen categories. AKB-48 [39] project offers a comprehensive Articulated object Knowledge Base with 2,037 real-world 3D models, supported by a fast modeling pipeline. GAPartNet [14] introduces a two-stage method for domain-generalizable 3D part segmentation and pose estimation by learning domain-invariant features. However, this two-stage pipeline has slow inference speed and tends to accumulate errors from the segmentation stage.

## III. Method

**Task Formulation.** Given the point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$ of the articulated object, our task is to estimate the semantic labels $C_i$, the normalized object part coordinate maps $M_i$, and the centroid offsets $O_i$ for the $i$-th point. We first utilize the semantic labels to cluster different part classes. Subsequently, we cluster based on the centroid offsets $O_i$ to differentiate between different part instances that share the same semantic label. Utilizing the normalized object part coordinate maps $M_i$, we determine the NPCS map for each part instance. Once the point cloud $\mathcal{P}$ is available, we register the estimated NPCS map with the corresponding points to calculate the transformation parameters $\{s, R, t\} \in SIM(3)$, where $s \in \mathbb{R}$, $R \in SO(3)$, and $t \in \mathbb{R}^3$. $SIM(3)$ is the Lie group of 3D similarity transformations.

**Architecture Overview.** As shown in Fig.3, our network processes the input point cloud, which is obtained from the output of the segmentation model. Here, we use Grounding-DINO [40],which is a vision-language model that detects and segments objects based on textual descriptions, to generate the input point cloud. The network then estimates semantic class labels, NPCS maps, and centroid offsets for each point simultaneously. Clustering based on these centroids groups

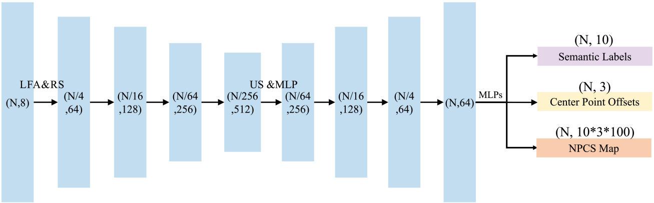


Fig. 4: **The detailed architecture of our YOEO.** FC: Fully Connected layer, LFA: Local Feature Aggregation, RS: Random Sampling, MLP: shared Multi-Layer Perceptron, US: Up-sampling.

points belonging to the same instance. Part labels are then assigned to each instance to locate the filtered NPCS maps of each part. Finally, the transformations and scales between the actual points and the estimated NPCS maps yield the 6-DoF pose and 3D size of each part.

**Details of the Network.** Specifically, our point-based method consists of an encoder and decoder module. The details of each module are shown in Figure 4. We use RandLA-Net [15] for feature extraction from the point cloud. The extracted features from the network are fed into the following modules: semantic segmentation, center point offset prediction, and NPCS Map prediction, all of which are composed of shared MLPs.

### A. Semantic Part Learning

To handle object categories with multiple parts, previous methods utilize existing grouping architectures to process features extracted from the backbone network, using a post-processing module to obtain part segmentation masks. Tian et al. [41] build the pose estimation models with the segmentation masks as input to simplify the problem. The pose estimation problem is divided into two stages, part segmentation and part pose estimation, which are trained separately. However, by incorporating the part pose estimation problem into the Semantic Part Learning module, the NPCS learning module, and the Center Offset Learning module, we hypothesize that these three tasks can enhance each other's performance through parallel training. Our ablation study confirms this hypothesis. Firstly, the semantic segmentation module forces the model to extract global and local features on each instance to distinguish different part classes, which allows the NPCS learning module to focus more on the part. Secondly, the semantic segmentation provides distinct semantic labels, which helps the Centroid Offset Learning module more accurate in distinguishing the centroids of different parts, as different semantic labels correspond to different centroids.

Based on this observation, we introduce a pointwise part semantic segmentation module $\mathcal{M}_s$ into the network and jointly optimize it with module $\mathcal{M}_c$ and $\mathcal{M}_n$ . Specifically, the semantic segmentation module $\mathcal{M}_s$ predicts semantic labels for each point by using the extracted features. The supervision for this module is provided using Focal Loss [42].

$$L_{\text{semantic}} = -\alpha(1-q_i)^{\gamma}\log(q_i) \quad \text{where} \quad q_i = c_i \cdot l_i \tag{1}$$

Here, $\alpha$ and $\gamma$ are the balancing and focusing parameter, respectively; $c_i$ is the predicted confidence for the $i$-th point belonging to a specific class, and $l_i$ is the one-hot encoded ground truth class label.

### B. Centroid Offset Learning

Considering that there can be multiple part instances with the same semantic label in an object, we design the Centroid Offset Learning module to predict the centroid of each instance to distinguish between them. It utilizes the per-point feature to predict the Euclidean translation offset $\Delta x_i$ to the associated object center. The learning process of $\Delta x_i$ is guided by an L1 loss:

$$L_{\text{center}} = \frac{1}{N}\sum_{i=1}^{N} ||\Delta x_i - \Delta x_i^*||\mathbb{I}(p_i \in I) \tag{2}$$

In this equation, $N$ represents the total number of seed points on the object's surface, and $\Delta x_i^*$ 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.

### C. NPCS learning for pose and size estimation

For the Normalized Part Coordinate Space Map Learning module, we aim to learn a mapping $\Phi : \mathcal{P}_o \rightarrow \mathcal{P}_{\mathbb{C}}$, where $\mathcal{P}_o$ represents the observed object point cloud and $\mathcal{P}_{\mathbb{C}}$ represents the canonical space point cloud. Both $\mathcal{P}o$ and $\mathcal{P}_{\mathbb{C}}$ consist of 3 channels, representing the 3D coordinates. $\Phi(\cdot)$ is constructed using a PointNet-like architecture for its lightweight design and computational efficiency [43]. The learning task is formulated as a classification problem by discretizing the coordinates $p_{\mathbb{C}}^i$ into 100 bins for each of the three axes (x, y, and z). For each region filtered by the predicted part segmentation mask $C_i$, we use the Softmax cross-entropy loss, as it has proven to be more effective than regression in reducing the solution space [37]. In addition to the predicted dense correspondence, the 6D object pose $\xi_o \in \{SE(3)\}$ is also recovered. This is computed using *RANSAC* for outlier elimination and the *Umeyama algorithm* [44] to determine the transformation parameters $\{s, R, t\} \in SIM(3)$ from the predicted canonical space point cloud $\mathcal{P}_{\mathbb{C}}$ to the observed object segment point cloud $\mathcal{P}_o$, ensuring that the rotation component is orthonormal.

### D. Grasping, Manipulation Strategy and Motion policy

Utilizing the NPCS representation, we possess information about the joint or prismatic axis in the NPCS frame, along with predefined category-level grasp poses. By aligning the NPCS with the real-world point cloud through registration, we can transform both the actionable axis and predefined grasp poses from the NPCS frame to the camera frame. In real robot experiments, the camera is calibrated to the robot's base frame, enabling a straightforward transformation of these elements into the robot frame for motion planning.

We also define category-level motion policies within the NPCS framework. During actual manipulation, aligning the NPCS with the real-world point cloud allows us to transform

TABLE I: **Results of Part Pose Estimation in terms of $R_e$ (°), $T_e$ (cm), $S_e$ (cm), mIoU=3D mIoU (%), A$_5$=5°5cm accuracy (%), A$_{10}$=10°10cm accuracy (%), Param (millions) and Speed (Hz).** PG=baseline modified from PointGroup [45]. AGP=baseline modified from AutoGPart [46].

| Method | $R_e$ ↓ | $T_e$↓ | $S_e$↓ | mIoU ↑ | A$_5$ ↑ | A$_{10}$ ↑ | Param.(M)↓ | Speed (Hz)↑ |
|---|---|---|---|---|---|---|---|---|
| PG [45] | 14.3 | 0.034 | 0.039 | 49.4 | 24.4 | 47.0 | / | / |
| AGP [46] | 14.4 | 0.036 | 0.039 | 48.7 | 26.8 | 49.1 | / | / |
| GAPartNet [14] | 9.9 | **0.024** | **0.035** | 51.2 | 28.3 | 53.1 | 7.9 | 20 |
| Ours | **9.0** | 0.11 | 0.036 | **57.6** | **30.4** | **54.4** | **1.9** | **200** |

motion policies from the NPCS frame to metric space. This approach ensures that our system not only adheres to the theoretical framework but also adapts effectively to real-world physical constraints, such as variations in object size and position.

## IV. Experiment

**GAPartNet Dataset.** The GAPartNet dataset is a comprehensive resource designed to facilitate research in articulated object manipulation. It encompasses 9 distinct classes of parts, each accompanied by detailed semantic labels and pose annotations. The dataset includes a total of 8,489 part instances derived from 1,166 objects, which span 27 diverse object categories. On average, each object within the dataset has 7.3 functional parts, highlighting the complexity and variety of the dataset. A notable characteristic of GAPartNet is its extensive cross-category representation: each class of parts appears in objects from at least 3 different object categories, and on average, a single part class is represented across 8.8 object categories. This diverse cross-category distribution is pivotal for establishing a robust benchmark for evaluating and enhancing generalizable part recognition and pose estimation methods.

**Evaluation Metric.** We evaluate part pose estimation performance using metrics such as average rotation error $R_e(°)$, translation error $T_e$(cm), scale error $S_e$(cm), and translation error of the part interaction axis $d_e$(cm). Specifically, we follow the standards of the GAPartNet Dataset, where the scales of all objects are normalized to a range of 0 to 1 cm. Additionally, we measure 3D Intersection over Union (3D mIoU) and accuracy percentages for specific thresholds: $5°, 5$cm and $10°, 10$cm. Furthermore, the parameters of the networks and inference speeds, which are calculated from feeding object point clouds to get part poses, are considered.

### A. Comparison to Baselines

We compared the pose estimation accuracy, inference speed, and model parameters with the baseline methods, and the summarized results are presented in Table I. Our method demonstrates significantly improved pose accuracy compared to the previous state-of-the-art method, GAParNet, particularly in the mIoU metric, validating the accuracy of the estimated poses. In comparison to the two-stage method GAParNet, our approach requires fewer parameters and achieves faster inference speeds, thereby reducing computational cost and enabling deployment on devices with limited computational resources. We also visualize the qualitative results in Fig. 5 and Fig. 6.

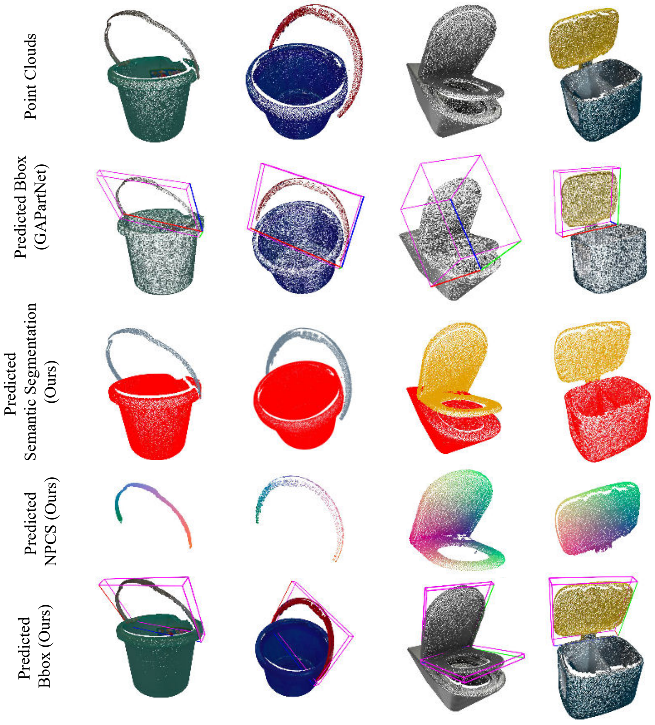

Point Clouds
Predicted Bbox (GAPartNet)
Predicted Semantic Segmentation (Ours)
Predicted NPCS (Ours)
Predicted Bbox (Ours)

Fig. 5: Qualitative results on the GAPartNet dataset. The left two columns illustrate the intra-category results for hinge handles within the bucket category. The right two columns display the cross-category results for hinge lids across toilet and box categories.

### B. Ablation study

We trained each of the three prediction heads individually by freezing the other two, repeating the process three times, once for each head. Then, we combined the individually trained heads and compared the results to those obtained from co-training. The results clearly support our conjecture that combining the prediction heads enhances the overall performance of our model compared to training them separately.

As shown in Table II, parallel (co-) training consistently improves performance across all metrics. The rotation error ($R_e$) drops from 19.6 to 9.0, while the translation error ($T_e$) and scale error ($S_e$) are reduced from 0.14 to 0.11 and from 0.041 to 0.036, respectively. 3D Intersection over Union (3D mIoU) also improves significantly, increasing from 52.3%

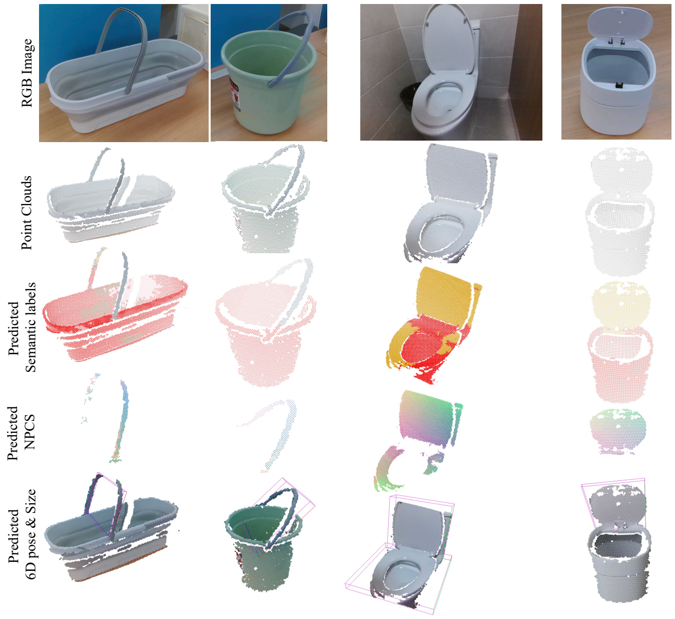


Fig. 6: Qualitative results of the real-world perception by using our YOEO method. We captured the object's RGB images and point cloud using [47]. The left two columns illustrate the intra-category results for hinge handles within the bucket. The right two columns show the cross-category results for hinge lids across toilet and box classes.

in individual training to 57.6% in parallel training. These results indicate that parallel training leads to more accurate pose and size estimation.

TABLE II: Ablation Study: Individual vs. Parallel Training

| Method | $R_e$ ↓ | $T_e$↓ | $S_e$↓ | mIoU ↑ | $A_5$ ↑ | $A_{10}$ ↑ |
|---|---|---|---|---|---|---|
| Ind. Training | 19.6 | 0.14 | 0.041 | 52.3 | 23.9 | 52.4 |
| Para. training | **9.0** | **0.11** | **0.036** | **57.6** | **30.4** | **54.4** |

### C. 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. To execute 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. A MantisVision camera [47] is used to capture RGB-D images of the scene and is mounted on a tripod positioned opposite the robot workspace. The camera is calibrated to the robotic base frame.

**Task Description.** To assess the sim-to-real capability of our method and evaluate its robustness and generalizability, we deployed our algorithm on a real robotic arm, specifically the KINOVA robot 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.

TABLE III: Robot Manipulation Success Rate.

| | hinge handle | drawer | hinge lid | Total |
|---|---|---|---|---|
| GAPartNet [14] | 7/10 | **7/10** | 6/10 | 20/30 |
| Ours | **9/10** | 5/10 | **8/10** | **22/30** |

**Evaluation Metric.** Depending on the specific experimental task, different metrics were used. For the drawer task, the robot arm successfully completed the task by pulling the drawer out 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 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 III. 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.

## V. Conclusion

We present YOEO, a lightweight model for real-time category-level articulated object 6D pose estimation. Unlike multi-stage methods, YOEO employs a single-stage framework directly on the point cloud, enabling end-to-end part pose estimation. It efficiently combines instance segmentation and NPCS representations, utilizing accurate point offset calculations and clustering for precise NPCS region alignment. Experiments on the GAPart dataset and real-world data demonstrate its real-time synthetic-to-real pose estimation capability. Robotic experiments on Kinova Gen 2 further showcase its proficiency with unseen articulated objects.

**Limitations**. Our method for articulated object pose estimation faces two challenges: suboptimal performance on smaller objects and inaccuracies with metallic surfaces due to poor point cloud quality. Future work will integrate RGB information for improved precision.

## VI. Acknowledgment

This work is supported in part by NSFC Project (62176061), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), and Shanghai Technology Development and Entrepreneurship Platform for Neuromorphic and AI SoC.

## References

[1] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[2] H. Lin, C. Cheang, Y. Fu, and X. Xue, "I know what you draw: Learning grasp detection conditioned on a few freehand sketches," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8417–8423.

[3] H. Lin, Z. Liu, C. Cheang, Y. Fu, G. Guo, and X. 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*, 2022, pp. 6707–6717.

[4] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," *arXiv preprint arXiv:2312.08344*, 2023.

[5] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[6] C. Chi and S. Song, "Garmentnets: Category-level pose estimation for garments via canonical space shape completion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3324–3333.

[7] H. Xue, W. Xu, J. Zhang, T. Tang, Y. Li, W. Du, R. Ye, and C. Lu, "Garmenttracking: Category-level garment pose tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 233–21 242.

[8] H. Lin, Y. Fu, and X. 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*, 2023, pp. 241–251.

[9] G. Narasimhan, K. Zhang, B. Eisner, X. Lin, and D. Held, "Self-supervised transparent liquid segmentation for robotic pouring," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 4555–4561.

[10] L. Liu, H. Xue, W. Xu, H. Fu, and C. Lu, "Toward real-world category-level articulation pose estimation," *IEEE Transactions on Image Processing*, vol. 31, pp. 1072–1083, 2022.

[11] X. Li, H. Wang, L. Yi, L. J. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3706–3715.

[12] J. Zhang, M. Wu, and H. Dong, "Generative category-level object pose estimation via diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. 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*, 2021, pp. 1581–1590.

[14] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. 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*, 2023, pp. 7081–7091.

[15] Q. Hu, B. Yang, L. Xie, S. Rosa, Y. Guo, Z. Wang, N. Trigoni, and A. Markham, "Randla-net: Efficient semantic segmentation of large-scale point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 108–11 117.

[16] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

[17] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. Vander-Bilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 142–13 153.

[18] M. Deitke, R. Liu, M. Wallingford, H. Ngo, O. Michel, A. Kusupati, A. Fan, C. Laforte, V. Voleti, S. Y. Gadre *et al.*, "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] T. Wu, J. Zhang, X. Fu, Y. Wang, J. Ren, L. Pan, W. Wu, L. Yang, J. Wang, C. 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*, 2023, pp. 803–814.

[20] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. 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*, 2019, pp. 909–918.

[21] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.

[22] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2553–2560.

[23] Y. Li, U. Upadhyay, H. Slim, A. Abdelreheem, A. Prajapati, S. Pothigara, P. Wonka, and M. Elhoseiny, "3d compat: Composition of materials on parts of 3d things," in *European Conference on Computer Vision*. Springer, 2022, pp. 110–127.

[24] D. Paschalidou, A. Katharopoulos, A. Geiger, and S. 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*, 2021, pp. 3204–3215.

[25] C. Xu, Y. Chen, H. Wang, S.-C. Zhu, Y. Zhu, and S. Huang, "Partafford: Part-level affordance discovery from 3d objects," *arXiv preprint arXiv:2202.13519*, 2022.

[26] K. Yang and X. Chen, "Unsupervised learning for cuboid shape abstraction via joint segmentation from point clouds," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–11, 2021.

[27] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[28] W. Wang, R. Yu, Q. Huang, and U. Neumann, "Sgpn: Similarity group proposal network for 3d point cloud instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2569–2578.

[29] Z. Ding, X. Han, and M. Niethammer, "Votenet: A deep learning label fusion method for multi-atlas segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*. Springer, 2019, pp. 202–210.

[30] J. Hou, A. Dai, and M. Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4421–4430.

[31] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 632–11 641.

[32] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3003–3013.

[33] C. Cheang, H. Lin, Y. Fu, and X. Xue, "Learning 6-dof object poses to grasp category-level objects by language instructions," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8476–8482.

[34] Q. Sun, H. Lin, Y. Fu, Y. Fu, and X. Xue, "Language guided robotic grasping with fine-grained instructions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1319–1326.

[35] T. Wang, H. Lin, J. Yu, and Y. 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)*. IEEE, 2024, pp. 9676–9683.

[36] T. Wang, Y. Li, H. Lin, X. Xue, and Y. Fu, "Wall-e: Embodied robotic waiter load lifting with large language model," *arXiv preprint arXiv:2308.15962*, 2023.

[37] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. 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*, 2019, pp. 2642–2651.

[38] S. Y. Gadre, K. Ehsani, and S. Song, "Act the part: Learning interaction strategies for articulated object part discovery," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 752–15 761.

[39] L. Liu, W. Xu, H. Fu, S. Qian, Q. Yu, Y. Han, and C. Lu, "Akb-48: A real-world articulated object knowledge base," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 809–14 818.

[40] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, and J. Zhu, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint*

*arXiv:2303.05499*, 2023. [Online]. Available: https://arxiv.org/abs/2303.05499

[41] M. Tian, M. H. Ang, and G. H. Lee, "Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 530–546.

[42] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[43] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[44] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[45] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4867–4876.

[46] X. Liu, X. Xu, A. Rao, C. Gan, and L. Yi, "Autogpart: Intermediate supervision search for generalizable 3d part segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 624–11 634.

[47] "Mantisvision camera," https://www.mantis-vision.com.cn/.