ICAF-4: An Integrated Framework of Category-level Articulated Object Perception and Manipulation for Embodied Intelligence

Wenbo Xu[1]*
2023170714@mail.hfut.edu.cn
Li Zhang[2]*
zanly20@mail.ustc.edu.cn
Qiankun Li[2]
qklee@mail.ustc.edu.cn
Qi Wu[3]
robotics_qi@sjtu.edu.cn
Lin Yuanbo Wu[4]
L.y.wu@swansea.ac.uk
Liu Liu[1]†
liuliu@hfut.edu.cn

[1] Hefei University of Technology
Hefei, CN
[2] University of Science and Technology of China
Hefei, CN
[3] Shanghai Jiao Tong University
Shanghai, CN
[4] Swansea University
Swansea, UK

Abstract

Articulated objects are common in human's daily life. Current research on articulated objects often emphasizes visual understanding of articulations rather than high-level functional manipulation tasks from a single RGB-D or point cloud observation. In this paper, to study the problem of **C**ategory-level **V**isually **A**rticulated object **P**erception task (**C-VAP**), we propose an **I**ntegrated **C**ategory-level visual **A**rticulated object perception **F**ramework, namely **ICAF-4**. Given the RGB and depth information as input, the ICAF-4 is capable of end-to-end processing of four mainstream tasks for articulated objects: object detection, part segmentation, pose estimation and manipulation. To support the C-VAP task, we re-annotate the rich functional grasping affordance and grasp poses by an automatic annotation generation way for two popular articulation benchmarks, ArtImage and ReArtMix, covering object-level and scene-level datasets. Accompanying the datasets, our ICAF-4 takes the part segmentation branch, pose estimation branch and manipulation prediction branch into a single forward pass. To boost the manipulation learning performance, we propose an anchor-based grasp pose estimation strategy where the "anchor" poses serve as references at multiple sizes and the grasp pose can be learned by the anchor selection and refinement process. Experiments demonstrate the superior performance of our ICAF-4 on integrating these visual tasks for articulation perception. Our code and dataset are available in https://github.com/xwb0117/ICAF-4.

1 Introduction

Articulated objects are ubiquitous in our daily lives, spanning from small table-scale objects (*e.g.*, eyeglasses) to large-size objects (*e.g.*, dishwashers). The manipulation of articulated objects often involves specific semantic actions, like opening a drawer using its handle or activating a switch. Therefore, the comprehension of articulated objects from visual observations is instrumental for embodied intelligence, which necessitates precise object manipulation within both simulated and real-world environments [2, 4], such as advanced visual perception [12], sophisticated motion planning [27], and adaptive learning capabilities [28]. Despite progress made on articulation problems, most existing studies concentrate primarily on lower-level visual tasks, such as part segmentation [17], pose estimation [11, 13, 14], and motion prediction [15], rather than addressing high-level robot functional manipulation tasks. In this paper, we aim to investigate the problem of Category-level Visual Articulated object Perception task (**C-VAP**), which considers predicting the visual perception as well as manipulation from a single RGB-D image. To achieve this goal, several major problems need to be addressed:

(i) Function Gap in Different Tasks. When tackling perceptual tasks, robots often decompose them into several sub-tasks and optimize them progressively, layer by layer. However, this approach can easily lead to the accumulation of perceptual errors. **(ii) Richness of Grasp Pose.** Prior arts [18, 27] on grasp pose modeling tend to generate a single grasp pose for each pixel/point, which cannot contribute a complete manipulation annotation for training. **(iii) Inaccurate Grasp Pose Regression.** Current methods [3, 16, 23] prefer to conduct a direct regression of the grasp poses from visual inputs, resulting in poor performance of manipulation learning.

In this paper, we propose **ICAF-4**, a **I**ntegrated **C**ategory-level visual **A**rticulated object perception **F**ramework for addressing C-VAP task. Our approach leverages both RGB and depth information to simultaneously tackle four tasks: articulated object detection, part segmentation, pose estimation and manipulation prediction, as shown in Fig. 1. Specifically, given an RGB-D image as input, our ICAF-4 first utilizes an object detector to locate the precise 2D bounding box. Next, the targeted object is back-projected into a point cloud and processed by three parallel modules for predicting part segmentation, 6D pose, and manipulation prediction (grasping affordance and grasp pose). In pose estimation module, we define the part-level pose and object-level pose to describe the transformation state from the perspective of the parts and the overall object. To avoid the drawback of direct grasp pose regression, we propose a series of "anchor" poses that serve as references at multiple sizes and our network predicts the final grasp pose by matching the closest anchor and refining the grasp pose.

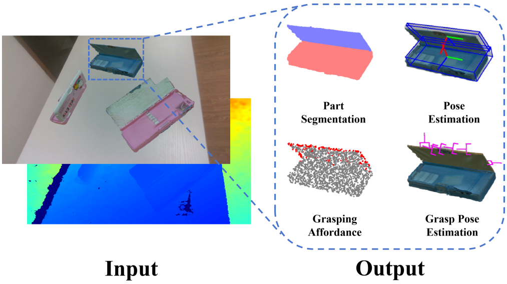

Figure 1: Given an RGB-D image, our **ICAF-4** aims to achieve four mainstream articulation tasks in an integrated framework: object detection, part segmentation, pose estimation and manipulation prediction (grasping affordance and grasp pose).

We validate our ICAF-4 on two articulation datasets (ArtImage [26] and ReArtMix [12]

). To obtain the training samples, we re-annotate these datasets with rich functional grasping affordance and grasp poses by an automatic annotation generation way. To validate the generalizability of our approach, we also test ICAF-4 on the real world scenarios showcasing significant success in articulated object manipulation. In summary, the contributions of our paper are:

(i) An integrated framework ICAF-4 for articulated object perception and manipulation is proposed, which integrates multiple visual tasks. **(ii) Rich grasp pose annotation for articulated objects**, we present a method for generating complete grasp poses for each point/pixel. The re-annotated datasets support the ICAF-4 network training. **(iii) Anchor based grasp pose alignment.** We propose to predict grasp pose through an anchor matching and refinement way. Compared with the regression strategy, our method significantly boosts the grasp prediction performance.

2 Related Work

2.1 Category-level Articulation Understanding

Category-level Articulation Understanding aims to grasp the underlying principles and distinctions within each category for deeper analysis. In recent years, research on category-level articulated object pose estimation tasks has been extremely active, attracting widespread attention in the field. Unlike instance-level pose estimation, which predicts the 3D rotation and translation of a 3D articulated object model [9, 22], the goal of category-level pose estimation is to predict the poses of unseen objects within the same category of articulated objects. The initial approach for category-level pose estimation was introduced by NOCS [24]. Later on, the method was extended to the task of estimating part-level poses by A-NCSH [10], which generalized the concept of normalized coordinate symbols to joints. Moreover, AKB48 [11] proposed a complete pipeline for robotic manipulation utilizing estimated joint poses.

Despite progress, priors are limited to a single target of articulated objects. In contrast, our work builds upon category-level pose estimation tasks and further explores how to effectively apply estimated pose information for manipulating and grasping articulated objects.

2.2 Articulated Object Manipulation

Interacting with articulated objects [1, 20] has emerged as a prominent topic in the embodied AI community, aming to gathering rich perceptual information, including shape, texture, and motion details. Previous approaches to articulated object manipulation have predominantly focused on imitation learning [6, 7, 25], leveraging demonstrations from experts to learn manipulation policies. However, these methods suffer from limitations in collecting diverse demonstrations, which can be time-consuming and costly in practice. As a pioneer work, Where2art [18] introduced dense visibility maps as actionable visual representations, revealing action probabilities at each point on the 3D surface of articulated objects. Building upon Where2art's framework, GAMMA [27] extends the approach by learning articulation modeling and grasp pose affordance from diverse articulated objects across different categories.

Building upon this, our approach extends the diversity of actionable possibilities at each point, significantly increasing potential action opportunities, thus providing robots with greater flexibility and adaptability in grasping and manipulating articulated objects.

3 Problem Statement

To achieve a comprehensive algorithm for C-VAP task, our key idea is to output the 4 targets in an end-to-end way. Here, we formulate a new paradigm for C-VAP task with a novel integrated framework named ICFA-4. Specifically speaking, given a single RGB-D image I as input, our ICFA-4 conducts the predictions under unknown CAD models for (1) V Detected Instances (partial observation $\{p_i\}_{i=1}^{N} = \mathcal{P} \in \mathbb{R}^{N\times 3}$) as well as their corresponding categories $\mathcal{C} = \{c^v\}_{v=1}^{V}$. (2) Semantic Segmentation at part-level (*i.e.*, $\mathcal{P} = \{\delta_k\}_{k=1}^{K}$, where δ_k represents for k-th rigid part). (3) 6D Pose Estimation, including global 6D pose $T = \{R, \mathbf{t}\}$ and per-part pose estimation $T^{(k)} = \{R^{(k)}, \mathbf{t}^{(k)}\}$. Concretely, we predict: *i*) object-level NOCS map $\mathcal{P}'_{obj}$ describes *global* pose ($'$ is used to define the coordinates in NOCS). *ii*) part-level NOCS map $\mathcal{P}'_{part}$ describes pose of each *rigid* part. Afterward, the global pose T as well as per-part pose $T^{(k)}$ part-level will be recovered, individually. (4) Manipulation Prediction. Given the partial observation $\mathcal{P}$ with N points, we predict the grasping affordance $G = \{G_i\}_{i=1}^{N}$, where $G_i \in \{0, 1\}$ (discrete score) indicates the grasp result, grasp pose $Q = \{q_j\}_{j=1}^{J}$, where q_j represents the manipulation pose and J is the total number of attempts.

4 Grasping Affordance and Grasp Pose Generation

To automatically generate the rich grasping affordance and grasp poses, we adopt an interactive strategy that allows a robot gripper to interact with each point of the articulated object in a simulated environment. By observing the actions of the gripper and the state of articulated objects, we can learn relevant information about grasping affordance and grasp poses. For grasp pose generation, previous pixel-based manipulation generation methods usually annotate only one interactive grasp pose at each pixel [18, 27]. To overcome this constraint, we attempt to generate rich and complete grasp poses. During the interaction process of the robotic gripper with the articulated object, we represent the gripper's grasp pose by a rotation matrix $R \in \mathbb{R}^{3\times 3}$. To visually inspect the gripper's grasp, we further decompose the rotation matrix R into three independent direction vectors ($\alpha \in \mathbb{R}^{3\times 1}, \beta \in \mathbb{R}^{3\times 1}, \gamma \in \mathbb{R}^{3\times 1}$), which are mutually orthogonal in the euclidean space(see Fig. 2 (a)). It can be re-formulated as Equation 1.

$$R = [\alpha, \beta, \gamma].T \tag{1}$$

where α signifies the gripper's interaction direction with the grasp point (typically opposite to the normal vector of the grasp point). β represents the gripper's movement direction and γ is the movement direction of the gripper. Thus, the grasp pose is generated by finding all the possible $\{\alpha, \beta, \gamma\}$ combinations that achieve successful object manipulation. Here, we define the "manipulation success" as the gripper can grasp the target point and move the part through 50% of its motion range.

To generate the rich and complete grasp poses, the interactive attempts follow two steps: **1) Rotation around the grasp point**(see Fig. 2 (b)). A random α is firstly selected within the hemisphere opposite to the normal vector of the grasp point. Then we choose another vector that is not collinear with it, and the cross product of these two vectors yields a rotation axis perpendicular to the α. Next, we select an appropriate rotation angle within the hemisphere range with m angles, and rotating the α around the rotation axis by the chosen angle leads

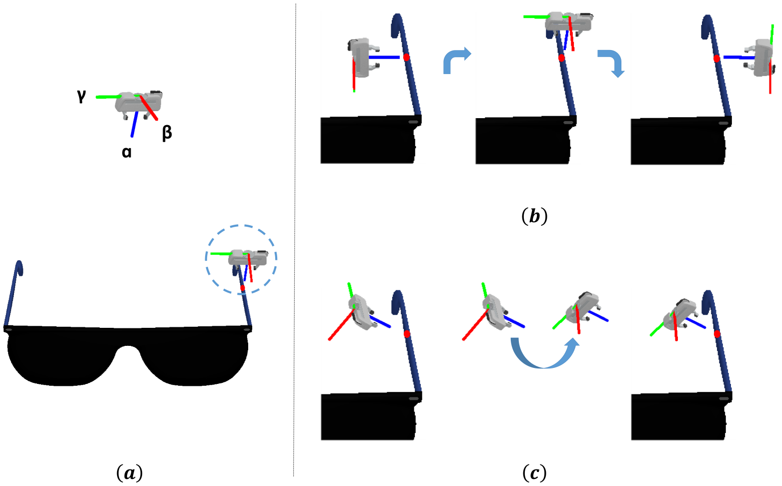

Figure 2: (a) The three direction vectors of the gripper. (b) Rotation around the grasp point. (c) Rotation around the gripper.

to m different α orientations. **2) Rotation around the gripper**(see Fig. 2 (c)). Given the α, we build a discrete and orthogonal space for sampling β and γ to alter the rotation of the gripper itself since the β and γ are precisely the two mutually perpendicular direction vectors in the plane vertical to the α. After the dense direction vector sampling, we can obtain nearly 200 grasp poses for each point. Meanwhile, we record the points that contain at least one successful manipulation as grasping affordance. By combining the two steps, we re-annotate the ArtImage [26] and ReArtMix [12] by adding the grasping affordance and grasp pose annotations. The new datasets will support our ICAF-4 network training and validation.

5 ICAF-4 Architecture

The overall pipeline of ICAF-4 framework is shown in Fig. 3. Our ICAF-4 is capable of performing four visual perception and manipulation tasks with the input of RGB-D image. Subsequently, we detail the learning pipeline of each module in ICAF-4.

5.1 Articulated Object Detection and Part Segmentation

The ICAF-4 initiates with object detection that identifies articulated objects within a scene and discerns them from other objects. We utilize Mask R-CNN [8] as the detector to achieve this goal. Using 2D bounding box extracted from the object detection module, we back-project the image into point cloud $\mathcal{P}$ using depth information. To accurately separate part boundaries for segmentation, $\mathcal{P}$ is processed through a point cloud backbone (PointNet++ [19]) and a three-layer MLP to output K channels for part segmentation δ_k. The training loss function is cross-entropy.

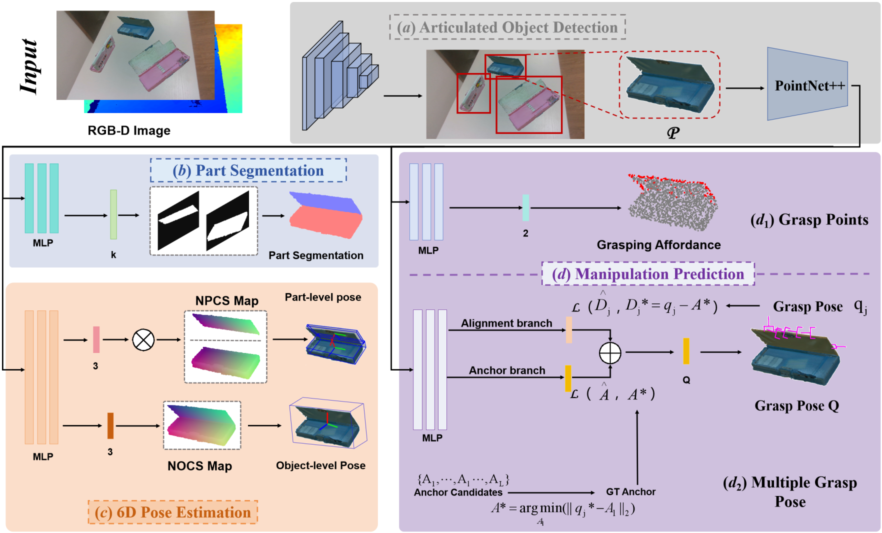

Figure 3: **The pipeline of our ICAF-4 framework.** Taking the partial observation as input, ICAF-4 can conduct category-level articulated object detection, segmentation, pose estimation and manipulation tasks in an end-to-end way. It comprises 4 key components: **(a) Articulated Object Detection**. Articulations will be detected. **(b) Part Segmentation**. Semantic segmentation at part-level. **(c) 6D Pose Estimation**. Part-level and Object-level Pose estimations are conducted. **(d) Manipulation Prediction**. Grasp region is segmented from $\mathcal{P}$ and multiple grasp poses are predicted.

5.2 Pose Estimation

To accurately estimate 6D pose, this module is inspired by the NOCS map [10, 24] for prediction. In ICAF-4, we conduct two types of poses for each articulated object, *i.e.*, part-level pose and object-level pose. Specifically, part-level pose is defined as the transformation between each part in its local canonical space and camera space. Exploiting the definition of NOCS map, the part-level pose $\mathcal{P}'_{part}$ for k-th part can be calculated by normalizing $P^{(k)}$ within the center of the part. In contrast, the object-level pose $\mathcal{P}'_{obj}$ is defined as the transformation between the part in the whole object space and camera space. Thus, the NOCS map for object-level pose is calculated by normalizing $P^{(k)}$ using the object center point.

To train the pose estimation module, we build a three-layer MLP at the end of the backbone network and output two parallel branches with $3 \times K$ and $3 \times K$ channels for predicting part-level and object-level NOCS map respectively. Each 3-channel aims to predict the 3-dimensional NOCS coordinates at k-th part. The final NOCS map will be masked by multiplying the predicted part segmentation from the former module. To obtain the per-part pose $T^{(k)}$, we use Umeyama algorithm to build an energy function for optimization and RANSAC for outlier removal similar to [10].

5.3 Manipulation Prediction

Manipulation prediction module aims to predict grasping affordance $G = \{G_i\}_{i=1}^{N}$ and grasp poses $Q = \{q_j\}_{j=1}^{J}$. In detail, we model the grasping affordance as a region segmentation task that predicts whether the point can be grasped or not so the grasp region is defined as the aggregation of all satisfied graspable points. Formally, we also utilize a three-layer MLP to conduct the region segmentation with cross-entropy loss function. During inference, the graspable points will be the points whose grasping affordance scores are greater than 0.5.

Another branch is to predict the grasp poses $Q = \{q_j\}_{j=1}^{J}$. We propose an anchor based grasp pose rather than using direct regression. This strategy consists of three steps:

(1) **Anchor Candidates Generation.** For each graspable point, we generate anchor candidates $\mathcal{A} = \{A_l\}_{l=1}^{L}$ (L is the total number of anchor candidates), which are sampled from 0° to 360° uniformly around *xyz* directions. Totally, the anchor number is determined by the sample interval size and dense anchors might influence the network performance. An effectiveness experiment is conducted to discuss this point in Sec. 6.2.

(2) **Anchor-based Grasp Pose Alignment.** After the grasp pose anchor generation, we align each target grasp pose q_j with all the anchor candidates $\mathcal{A}$. The aligned anchor A^* is defined as the closest rotation degree distance between q_j and A_l:

$$A^* = \arg\min_{A_l}(\|q_j - A_l\|_2) \tag{2}$$

After all the grasp poses Q matches the corresponding anchors A, we generate a grasp mask array $M^{(Q)} = \{M_l^{(Q)}\}_{l=1}^{L}$ to identify the matched anchor and non-matched anchor, which is defined as:

$$M_l^{(Q)} = \begin{cases} 1, & \text{if } \sigma(A_l, Q) \textit{ is True} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\sigma(A_l, Q)$ is the matching function that indicates there is a q_j is aligned with the anchor A_l. For each element $M_l^{(Q)}$, it indicates A_l is the matching anchor if $M_l^{(Q)} = 1$.

(3) **Anchor-based Grasp Pose Refinement.** When each grasp pose q_j is aligned with a corresponding anchor, we utilize the refinement mechanism to obtain the predicted grasp pose. Specifically, the network outputs $5L$ channels at the end of point cloud backbone, in which L channels are used to classify the indexes of matched anchors M_l and $4L$ channels are used to predict the pose distance D_j^* between q_j and A_l as $D_j^* = q_j - A_l$. We adopt cross-entropy loss and L2 loss functions to train the two branches. Finally, the predicted grasp pose q_j can be obtained by:

$$q_j = M_l * A_l + D_j^* \tag{4}$$

6 Experiments

6.1 Experimental Settings.

Datasets and Metric. We conduct experiments on ArtImage [26] and ReArtMix [12]. For ArtImage, we report the mIoU for Part Segmentation and Affordance, pose error (rotation and translation error) for 6D Pose Estimation, and rotation error for Grasp Pose. For ReArtMix, we use the mean average precision (mAP) to report experimental results, which are evaluated under both IoU and pose error (rotation and translation error). Note that we report the mean error of all the parts for each metric on 6D Pose Estimation.

Implementation Details. During the training process, we preprocessed the data by downsampling the input point clouds to 2048 points before feeding them into the network

Table 1: Comparison with state-of-the-art on ArtImage [26]. Note that we report mIoU for Part Segmentation, mean rotation error (°) and mean translation error (m) calculated by all the parts for 6D Pose Estimation, mIoU for Affordance, and rotation error (°) for Grasp Pose. The up or down arrows indicate higher or lower values corresponding to better results.

Category	Method	Part Segmentation (↑)	6D Pose Estimation				Manipulation Prediction	
			Part-level (↓)		Object-level (↓)		Affordance (↑)	Grasp Pose (↓)
Eyeglasses	FPFH [21]	0.34	32.7	0.367	32.4	0.296	0.34	38.1
	DAISY [5]	0.37	31.5	0.344	31.6	0.286	0.39	36.8
	AH-Where [10, 18]	0.93	16.4	0.229	-	-	0.96	37.1
	ICFA-4 (**Ours**)	**0.97**	**4.1**	**0.100**	**4.2**	**0.107**	**0.98**	**16.9**
Scissors	FPFH [21]	0.37	27.1	0.107	27.1	0.096	0.37	30.0
	DAISY [5]	0.40	26.5	0.101	26.8	0.089	0.41	28.6
	AH-Where [10, 18]	**0.85**	**2.5**	**0.030**	-	-	**0.85**	27.6
	ICFA-4 (**Ours**)	0.84	6.1	0.044	**5.8**	**0.038**	0.84	**17.1**
Laptop	FPFH [21]	0.45	30.8	0.240	31.6	0.207	0.45	27.8
	DAISY [5]	0.51	30.2	0.229	29.3	0.187	0.45	27.3
	AH-Where [10, 18]	0.90	4.4	0.049	-	-	0.90	27.8
	ICFA-4 (**Ours**)	**0.95**	**2.2**	**0.035**	**2.1**	**0.037**	**0.95**	**14.7**
Dishwasher	FPFH [21]	0.60	30.9	0.379	30.6	0.229	0.60	25.0
	DAISY [5]	0.61	28.9	0.358	28.8	0.201	0.67	23.8
	AH-Where [10, 18]	0.95	4.4	0.091	-	-	0.95	24.8
	ICFA-4 (**Ours**)	**0.98**	**2.3**	**0.076**	**2.1**	**0.062**	**0.98**	**15.1**
Drawer	FPFH [21]	0.43	30.0	0.395	29.4	0.235	0.46	25.2
	DAISY [5]	0.48	28.3	0.369	28.8	0.212	0.47	24.6
	AH-Where [10, 18]	0.68	3.3	0.108	-	-	0.87	25.6
	ICFA-4 (**Ours**)	**0.69**	**2.7**	**0.098**	**2.5**	**0.089**	**0.89**	**14.3**

for training. When training the PointNet++ [19], we used the Adam optimizer with an initial learning rate of 0.001 and a weight decay rate of 0.0001. The learning rate decay step size was set to 20, and the learning rate decay factor was set to 0.5. All the experiments are implemented on four NVIDIA GeForce RTX 4090 GPUs with 24GB memory.

Baselines. For a fair comparison, we introduce two types of baselines, *i.e.*, classical feature extraction for training and deep learning-based method. The former is more interpretable, while the latter can iterate over a larger search space to obtain the optimal results. Classical feature extraction for training include DAISY [5], and FPFH [21]. However, since there is no existing integrated framework, we zip the ANCSH [10] technique and Where2art [18] technique into a framework named *AH-Where*. Concretely, we use the former for 6D pose estimation and the latter for manipulation prediction.

6.2 Experiments on ArtImage Dataset

Comparison with state-of-the-art on ArtImage can be seen in Tab. 1. Compared with classical feature extraction for training (*i.e.*, FPFH [21] and DAISY [5]), our ICAF-4 demonstrates better performance in all the metrics. For example, considering the category *Eyeglasses*, our method achieves significant performance improvement (**0.98** ours *vs* **0.39** in DAISY [5] for affordance). Compared with deep learning based methods, our method can still maintain the same performance advantages. Concretely, regarding pose estimation, our ICAF-4 achieves **4.1**° and **0.100**m error, outperforming the **16.4**°, **0.229**m in AH-Where. Also, we achieve a higher affordance (**0.98**) and a lower grasp pose error (**16.9**°). It proves that the anchor-based training strategy facilitates the multi grasp poses alignment.

Additionally, we show the qualitative results on ArtImage in Fig. 4. We can observe that our method could generate more grasp poses, effectively extending the diversity of actionable possibilities at each point. We believe this can facilitate the downstream tasks, such as Interactive perception, and imitative learning for embodied AI.

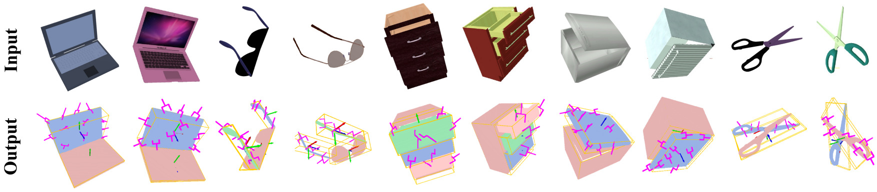

Figure 4: **Qualitative results on ArtImage.** Note that partial grasp poses are visualized for a better view.

Effectiveness of Anchor. To further investigate the effectiveness under different anchor size, we conduct the effectiveness experiment on ArtImgae, which ranges from 10, 20, 30, and 45. Quantitative results can be seen in Tab. 2. We can observe that ICAF-4 achieves similar performance when the anchor size ranges from 10 to 45, which doesn't work as expected: as the anchor size increases, so does the performance. This confirms the robustness of our anchor-based mechanism in another way. Moreover, compared to the direct grasp pose regression, our method achieves better performance (Ours **14.7**° vs direct regression **27.8**° of *Latptop*). It can be concluded that the anchor-based training strategy can effectively train the manipulation of articulated objects compared to the prior arts.

Table 2: **Effectiveness of Anchor** test on ArtImage [26]. we report mIoU for Affordance, and rotation error (°) for Grasp Pose . "-" means the direct grasp pose regression.

Category	Metric	Anchor Size				
		-	10	20	30	45
Eyeglasses	Affordance	0.96	0.97	0.98	0.97	**0.98**
	Grasp Pose	37.1	17.5	17.2	17.1	**16.9**
Scissors	Affordance	0.85	0.84	0.84	0.85	**0.85**
	Grasp Pose	27.6	17.9	17.4	17.2	**17.1**
Laptop	Affordance	0.90	0.93	0.94	0.95	**0.95**
	Grasp Pose	27.8	15.2	14.9	14.7	**14.7**
Dishwasher	Affordance	0.95	0.97	0.96	0.97	**0.98**
	Grasp Pose	24.8	15.6	15.3	15.1	**15.1**
Drawer	Affordance	0.87	0.85	0.86	0.87	**0.89**
	Grasp Pose	25.6	14.7	14.7	14.4	**14.3**

6.3 Experiments on ReArtMix Dataset

Quantitative results of ReArtMix are reported in Tab. 3. In detail, our method achieves high mAP under most of the metrics, especially for the manipulation Prediction (*e.g.* **0.99** and **0.98** on the grasping affordance of *Cutter* and *Drawer*). We conjecture that our optimization between each branch is well-handled. Qualitative results can be seen in Fig. 5. It can be concluded that our method can perform well in this scenario.

Table 3: Quantitative results on ReArtMix. We report mAP@IoU$_{75}$ for Part Segmentation, mAP@(5°, 5 *cm*) within all the parts' mean error for 6D Pose Estimation, mAP@IoU$_{75}$ and mAP@IoU$_{95}$ for Affordance, and mAP@(5°) and mAP@(10°) for Grasp Pose.

Category	Part Segmentation	6D Pose Estimation		Manipulation Prediction			
		Part-level	Object-level	Affordance		Grasp Pose	
Box	0.98	0.91	0.89	0.98	0.78	0.22	0.42
Cutter	0.99	0.67	0.66	0.99	0.84	0.24	0.44
Drawer	0.98	0.93	0.91	0.98	0.87	0.23	0.44
Scissor	0.99	0.80	0.76	0.99	0.93	0.24	0.43
Stapler	0.99	0.62	0.60	0.99	0.98	0.24	0.45

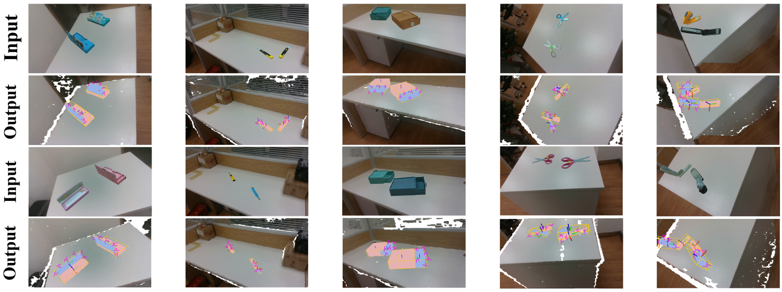

Figure 5: **Qualitative results on ReArtMix.** Note that partial grasp poses are visualized for a better view.

6.4 Demonstrations on the Real World

To assess the generalization capability of our proposed ICFA-4 framework, we conduct experiments aiming to evaluate its performance in real-world scenarios. Qualitative results can be seen in Fig. 6. It can be observed that our method can perform well in grasping articulated objects accurately under real-world conditions.

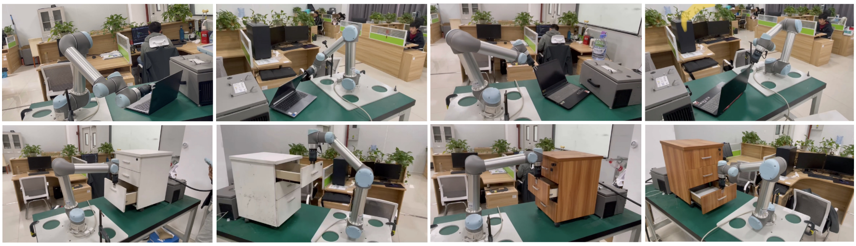

Figure 6: **Qualitative results on real-world scenarios.**

7 Conclusion

This paper targets at solving the Category-level Visually Articulated object Perception task thus proposes an integrated framework namely ICAF-4 that is capable of an end-to-end processing of four major articulation related tasks in a single forward pass: object detection, part segmentation, pose estimation and manipulation prediction. To support the network training, we provide rich and complete grasp annotations for two popular benchmarks ArtImage and ReArtMix. During the grasp pose prediction branch, we propose an anchor-based strategy to boost the prediction performance. We evaluate our method on the synthetic datasets and real world scenarios with exceptional performance in integrating these tasks for articulated object perception and manipulation.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 62302143 and 62372150, and Anhui Provincial Natural Science Foundation under Grant 2308085QF207, and was partially funded by Beijing XiaoYu Intelligence Manufacturing Ltd Inc, Beijing China.

References

[1] Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. Sim-to-real transfer for vision-and-language navigation, 2020.

[2] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020.

[3] Hu Cheng, Yingying Wang, and Max Q-H Meng. Grasp pose detection from a single rgb image. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4686–4691. IEEE, 2021.

[4] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

[5] Fariborz Ghorbani, Hamid Ebadi, Amin Sedaghat, and Norbert Pfeifer. A novel 3-d local daisy-style descriptor to reduce the effect of point displacement error in point cloud registration. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:2254–2273, 2022.

[6] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023.

[7] Pierre-Louis Guhur, Shizhe Chen, Ricardo Garcia Pinel, Makarand Tapaswi, Ivan Laptev, and Cordelia Schmid. Instruction-driven history-aware policies for robotic manipulations. In *Conference on Robot Learning*, pages 175–187. PMLR, 2023.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[9] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017.

[10] 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.

[11] 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.

[12] 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.

[13] Liu Liu, Jianming Du, Hao Wu, Xun Yang, Zhenguang Liu, Richang Hong, and Meng Wang. Category-level articulated object 9d pose estimation via reinforcement learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 728–736, 2023.

[14] Liu Liu, Qi Wu, Zhendong Xue, Sucheng Qian, and Rui Li. Reaper: Articulated object 6d pose estimation with deep reinforcement learning. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 21–25. IEEE, 2023.

[15] Liu Liu, Anran Huang, Qi Wu, Dan Guo, Xun Yang, and Meng Wang. Kpa-tracker: Towards robust and real-time category-level articulated object 6d pose tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3684–3692, 2024.

[16] Qingkai Lu and Tucker Hermans. Modeling grasp type improves learning-based grasp planning. *IEEE Robotics and Automation Letters*, 4(2):784–791, 2019.

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

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

[19] 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.

[20] Santhosh K. Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration, 2020.

[21] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE International Conference on Robotics and Automation*, pages 3212–3217, 2009. doi: 10.1109/ROBOT.2009.5152473.

[22] 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.

[23] Chenxi Wang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Jin Gao, and Cewu Lu. Graspness discovery in clutters for fast and accurate grasp detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15964–15973, 2021.

[24] 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.

[25] Josiah Wong, Albert Tung, Andrey Kurenkov, Ajay Mandlekar, Li Fei-Fei, Silvio Savarese, and Roberto Martín-Martín. Error-aware imitation learning from teleoperation data for mobile manipulation. In *Conference on Robot Learning*, pages 1367–1378. PMLR, 2022.

[26] Han Xue, Liu Liu, Wenqiang Xu, Haoyuan Fu, and Cewu Lu. Omad: Object model with articulated deformations for pose estimation and retrieval. *arXiv preprint arXiv:2112.07334*, 2021.

[27] Qiaojun Yu, Junbo Wang, Wenhai Liu, Ce Hao, Liu Liu, Lin Shao, Weiming Wang, and Cewu Lu. Gamma: Generalizable articulation modeling and manipulation for articulated objects. *arXiv preprint arXiv:2309.16264*, 2023.

[28] Andy Zeng, Shuran Song, Stefan Welker, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning, 2018.