2024 IEEE International Conference on Robotics and Automation (ICRA)
May 13-17, 2024. Yokohama, Japan

RGBManip: Monocular Image-based Robotic Manipulation through Active Object Pose Estimation

Boshi An*, Yiran Geng*, Kai Chen*, Xiaoqi Li, Qi Dou, Hao Dong

***Abstract*— Robotic manipulation requires accurate perception of the environment, which poses a significant challenge due to its inherent complexity and constantly changing nature. In this context, RGB image and point-cloud observations are two commonly used modalities in visual-based robotic manipulation, but each of these modalities have their own limitations. Commercial point-cloud observations often suffer from issues like sparse sampling and noisy output due to the limits of the emission-reception imaging principle. On the other hand, RGB images, while rich in texture information, lack essential depth and 3D information crucial for robotic manipulation. To mitigate these challenges, we propose an image-only robotic manipulation framework that leverages an eye-on-hand monocular camera installed on the robot's parallel gripper. By moving with the robot gripper, this camera gains the ability to actively perceive the object from multiple perspectives during the manipulation process. This enables the estimation of 6D object poses, which can be utilized for manipulation. While, obtaining images from more and diverse viewpoints typically improves pose estimation, it also increases the manipulation time. To address this trade-off, we employ a reinforcement learning policy to synchronize the manipulation strategy with active perception, achieving a balance between 6D pose accuracy and manipulation efficiency. Our experimental results in both simulated and real-world environments showcase the state-of-the-art effectiveness of our approach. We believe that our method will inspire further research on real-world-oriented robotic manipulation. See https://rgbmanip.github.io/ for more details.**

I. Introduction

Robotic manipulation is a field with immense potential to enhance human life. Nevertheless, realizing robust and dependable robotic manipulation in our daily lives continues to pose a substantial challenge, primarily due to the intricacies of our surroundings and the complexities in information acquisition. A critical factor in addressing this challenge lies in the precise perception and understanding of the environment by robots.

In this case, visual perception becomes a pivotal role in robotic manipulation, as it facilitates object identification, localization, pose estimation, and task planning and execution. In the quest for improved perception capabilities, researchers and engineers have commonly used RGB cameras and depth cameras as primary sources of sensory data [21], [19], [8], [10]. However, despite their respective strengths, they both come with inherent drawbacks that may limit their applicability in complex or nuanced environments. Point-cloud data obtained from depth cameras, are often sparse and may fail to capture small or intricate features of objects, particularly at greater distances [15]. Additionally, although some industrial-grade depth cameras offer higher resolutions and extended capturing distances, these improvements often come at a significant financial cost, posing challenges for academic research and large-scale deployments [12]. Moreover, depth cameras are prone to optical interference from other light sources and struggle with accurately imaging transparent and specular objects, such as glass and metal [15], [31], [6]. RGB cameras, on the other hand, are generally more price-friendly and can capture high-resolution images rich in color and texture. However, they are fundamentally limited by their inability to capture 3D spatial information directly. This absence of information can pose challenges in determining the 3D pose of an object, limiting their applicability in most manipulation tasks.

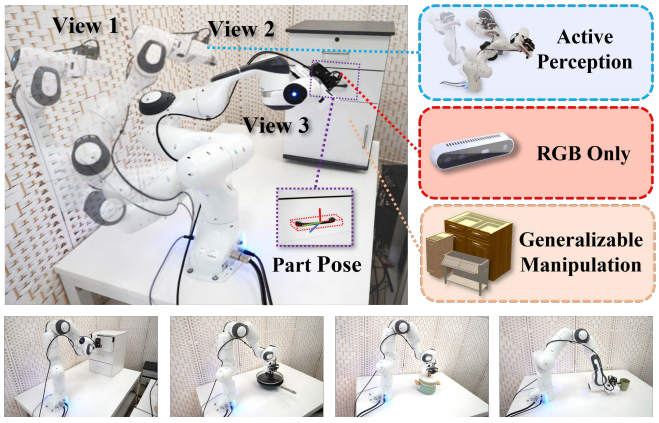

Fig. 1. An eye-on-hand camera captures multiple RGB images to estimate the object pose in the manipulation process

To empower robots perceive the environment in a way rich in both high-resolution details and 3D information, and adjust adaptively according to the environment, we propose an image-only robotic manipulation policy that utilizes a single eye-on-hand camera to actively observe the environment to finish given manipulation tasks, as illustrated in Figure I. Our approach decouples the manipulation process into three processes: 1) Global Scheduling, the first process proposes way points for the robot to explore the environment adaptively. Powered by reinforcement learning, this process enables the robot to adapt to different manipulation tasks (open door, pick mug, etc.) and gather information from different views, avoiding occlusions, making it possible for

Boshi An, Yiran Geng, Xiaoqi Li and Hao Dong are with Hyperplane Lab, School of CS, Peking University and National Key Laboratory for Multimedia Information Processing. Xiaoqi Li is also with Beijing Academy of Artificial Intelligence (BAAI). Kai Chen and Qi Dou are with Department of Computer Science and Engineering, The Chinese University of Hong Kong.

* The first three authors contributed equally.
Corresponding to hao.dong@pku.edu.cn

3D representations in the next process. 2) Active Perception, the second process takes as input RGB images from different viewpoints that are captured while the gripper is approaching the object to be manipulated and learns to estimate the 6D pose of either the entire object or a specific object part *e.g.*, the pose of a mug on the table in the mug-picking task, or the pose of a door handle in the door-opening task. This process paves the road for the third process. 3) Manipulation, when the gripper grasped the object, we use a control-based approach to manipulate the object. To ensure high accuracy and reliability on different tasks, a closed-loop impedance controller is used for this process. The three processes are coordinated under the Global Scheduling process.

By decoupling the manipulation process into three different parts, our approach has several advantages over existing methods. First, it allows the robot to capture high-resolution visual information while also estimating the 6D pose of the object, which is crucial for accurate manipulation. Second, it enables the robot to adapt to different tasks and objects, enhancing its versatility and effectiveness. Finally, our approach provides an option to balance accuracy and efficiency, solving the trade-off between exploration and exploitation.

II. Related Work

A. Vision-based Robotic Manipulation

Human decision-making and locomotion heavily rely on visual perception. Similarly, visual perception plays a crucial role for robots to adapt to and interact with the real world. Recent years have witnessed significant advancements in vision-based robotic manipulation, where robots utilize visual information to perceive and understand their environment.

Various visual modalities have been explored for perceiving the environment in robotic manipulation. Some studies, such as Where2Act [21], SAGCI-System [19], RLAfford [10] and Flowbot3D [8], have employed point clouds as observations, leveraging the 3D information they provide. On the other hand, approaches from Geng *et al.* [9], Xu *et al.* [32] and Wu *et al.* [29] have utilized both RGB images and depth maps for tasks like articulated object manipulation and object grasping. However, there has been limited exploration of using only RGB images as input, mainly due to the belief that depth information is essential for determining the actual 3D coordinates of pixels in an image, thus RGB-only input may result in spatial ambiguity. For instance, in the work Where2Act [21], authors compared performance of policies with RGB image input and RGBD image input, and from the experimental result, we can see a large performance drop due to the removal of depth information. This highlights the trade-off between using RGB-only observations and the depth-rich RGBD input. Nonetheless, the use of depth information poses challenges when dealing with specular and transparent textures, as they can interfere with the depth capturing process and result in noisy depth maps [24]. To address this issue, some works, such as [13], [6], have employed Neural Radiance Field (NeRF) [20] to recover depth information from multi-view RGB images.

Despite the advancements in RGBD-based methods, our approach focuses on utilizing RGB images as the sole input modality for robotic manipulation. Unlike NeRF-based approaches [13], [6], our method does not rely on depth recovery. Instead, we directly estimate the 6D poses of objects using a multi-view pose estimator. This unique approach allows us to circumvent the trade-off between RGB and depth observations. Through our RGB-only approach with pose estimation, we contribute a novel perspective to the field of vision-based robotic manipulation, highlighting the potential of leveraging only RGB images for perception and control.

B. Object Pose Estimation

Object pose estimation provides position and orientation information for the target object, which are important for robotic manipulation. In this work, we focus on category-level object pose estimation [23], which aims to predict the pose for unseen objects belonging to a specific object category. Once the pose estimation model is trained, it can be directly applied to novel objects for robotic manipulation. Most existing methods adopt a prior-based pose estimation paradigm [27], [28]. Typically, SPD [26] optimizes a shape deformation field to reconstruct the 3D object model. Then, it densely matches the reconstructed model and the observed instance point cloud for object pose estimation. SGPA [4] developes a prior adaptation module, which dynamically adjusts the prior feature to handle intra-class variation and achieves a higher category-level pose accuracy. RBP-Pose [35] further leverages a residual-vector-based representation to enhance the 3D spatial cues in the object point cloud for robust object pose estimation. Recently, Liu *et al.* proposed IST-Net [18], a prior-free framework. It resorted to an implicit space transformation module, which associates the camera-space feature with the object-space feature in an implicit way without relying on any shape prior point cloud. However, both prior-based and prior-free methods highly rely on object point clouds, which are not applicable when high-quality point cloud observation is not available. To tackle this limitation, StereoPose [5] proposed a pure image-based framework. It leverages a parallax-aware module to fuse stereo image features and model intra-class object shape variation. Stereo coordinate maps are further regressed from stereo images for accurate object pose estimation. Inspired by StereoPose, in this work, we propose a novel multi-view image-based method for category-level object pose estimation. Different from StereoPose, our method will recover object pose from images captured at multiple viewpoints along the robot trajectory. To reduce the pose ambiguity of monocular images, we will leverage robot kinematics data to effectively fuse multi-view image features. In addition, by actively adjusting the robot trajectory, we manage to utilize the most informative views to recover the object pose accurately.

III. Method

As shown in Fig. 2, our method mainly consists of 3 modules, the Global Scheduling Policy S, the Active Per-

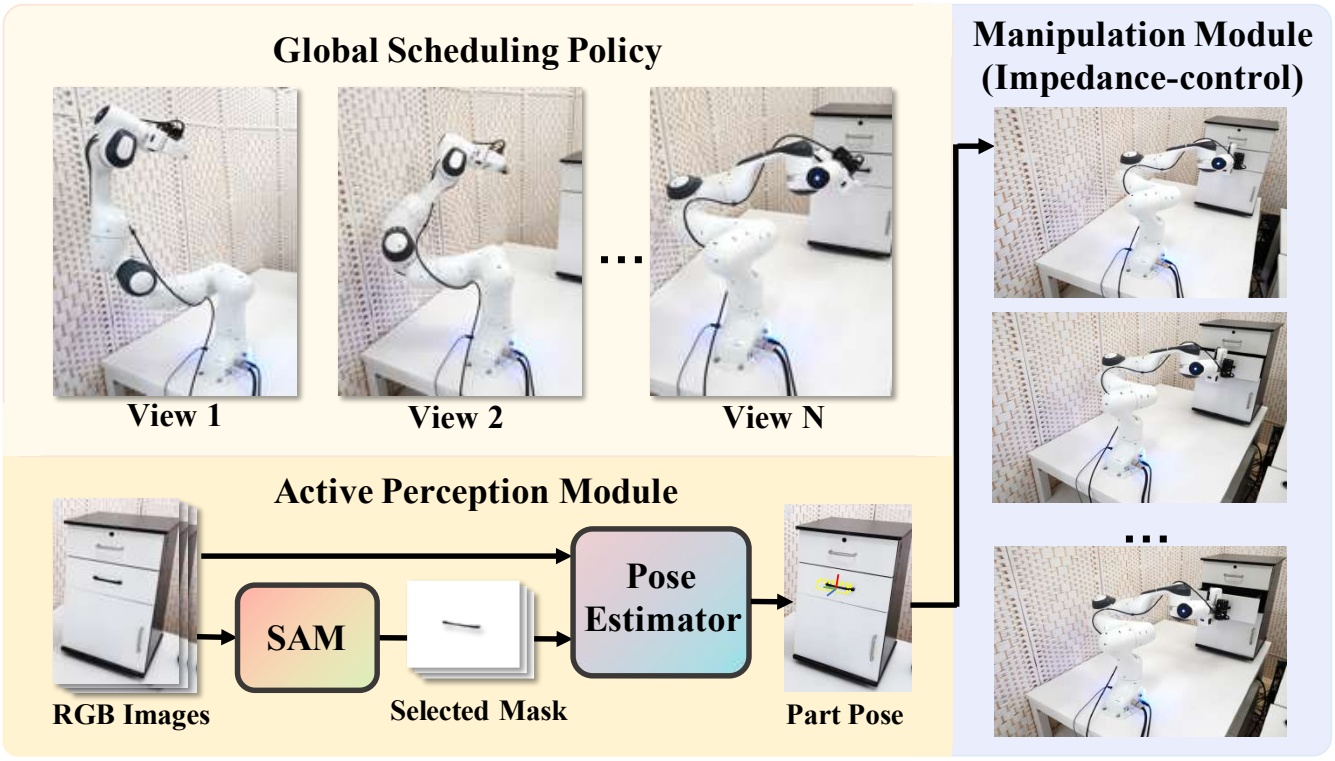

Fig. 2. In our pipeline, the Global Scheduling Policy serves as a high-level decision making policy to schedule Active Perception Module and Manipulation Module. The Active Perception Module learns to perceive the environment to predict pose information with the help of a pre-trained segmentation model (SAM [16]). The Manipulation Module is used to complete the manipulation task through impedance control.

ception Module P and the Manipulation Module M. The three modules are coordinated under the control of the Global Scheduling Policy S.

Once the robot was deployed in the environment, it will explore the environment while completing the assigned manipulation task. In other word, the robot will actively perceive the surrounding environment through a camera mounted on its end-effector through the procedures described below.

A. Exploration via Global Scheduling

The Global Scheduling Policy, denoted as S, serves as a high-level decision-making mechanism powered by reinforcement learning algorithms. Its primary role is to decide whether to further explore the environment from a novel perspective or to initiate manipulation, taking into account the accumulated information and the feedback from the Active Perception Module. When opting for exploration, the policy specifies the 6D extrinsic parameters for the robotic arm to relocate to capture an image. This captured image is then utilized by the Active Perception Module to produce pose estimations. Conversely, if the policy concludes that it's appropriate to stop the exploration, control will be transferred to the Manipulation Module.

More precisely, at time step t, the Global Scheduling Policy takes all previous views $\mathrm{V}_1, \cdots \mathrm{V}_{t-1}$ and the current prediction from Active Perception Module as input, and outputs two values: p_t and f_t . The second output f_t decides whether to terminate the view-point planning process and try to finish the manipulation task based on current information. If $f_t = 0$, then the robot will continue exploration process and go to way-point p_t to obtain view V_t, otherwise, the Manipulation Module will take over the control of the robot. This modeling allows us to train the Global Scheduling Policy with Proximal Policy Optimization [25].

B. Kinematics-Guided Multi-view Object Pose Estimation

The core role of the Active Perception Module is to estimate the pose for the object of interest, given all information gathered during the exploration. At time step t, the camera mounted on the robot arm will capture an RGB image I_t for the target object. We exploited a segmentation model [16] to crop the object region, as shown in Fig. 2. In order to handle the intra-class variation for category-level object pose estimation, similar to [4], [35], [18], we first resorted to a canonical object representation [27] and estimated a normalized coordinate map based on F_t, which is the deep feature of I_t. The predicted coordinate map M_t encodes dense 2D-3D correspondences between the camera and object frame, which are essential for object pose estimation. However, the category-level pose cannot be fully recovered with monocular RGB image, due to the depth ambiguity. In this regard, we further proposed a kinematics-guided depth-aware module to fuse multi-view image features. It aims to leverage the robot kinematics data to reduce the pose estimation ambiguity. For adjacent two RGB images I_t and I_{t+1}, their relative extrinsic $(\mathbf{R}_t^{t+1}, \mathbf{t}_t^{t+1})$ can be derived from the kinematics data between t and $t+1$. Then, we fused adjacent image features by warping F_t to F_{t+1} with a multi-homography mapping. Specifically, we uniformly sample a set of hypothetical depth planes $\{d_i\}_{i=1}^{N}$ between d_{min} and d_{max}. At each hypothetical

depth plane, we warp F_t to F_{t+1} based on the corresponding homography, which is computed as:

$$H(d_i) = \mathbf{K} \cdot \mathbf{R}_t^{t+1} (\mathbf{I} + \frac{\mathbf{t}_t^{t+1} \cdot \mathbf{n}^\top}{d_i}) \cdot \mathbf{K}^{-1}, \quad (1)$$

where $\mathbf{K}$ denotes the camera intrinsics and $\mathbf{n}$ denotes the principle axis of the camera at time step $t+1$. The warped features would exhibit different similarities on different depth planes. Therefore, by concatenating features at different depth, we can construct a 4D depth-aware feature volume. This volume is then regularized with a volume regularization layer similar to [33], [3] to derive the fused image feature $\hat{F}_{t+1}$. $\hat{F}_{t+1}$ is further concatenated with the features extracted from M_{t+1} and passed through MLP-based networks to predict object size, rotation and translation, respectively.

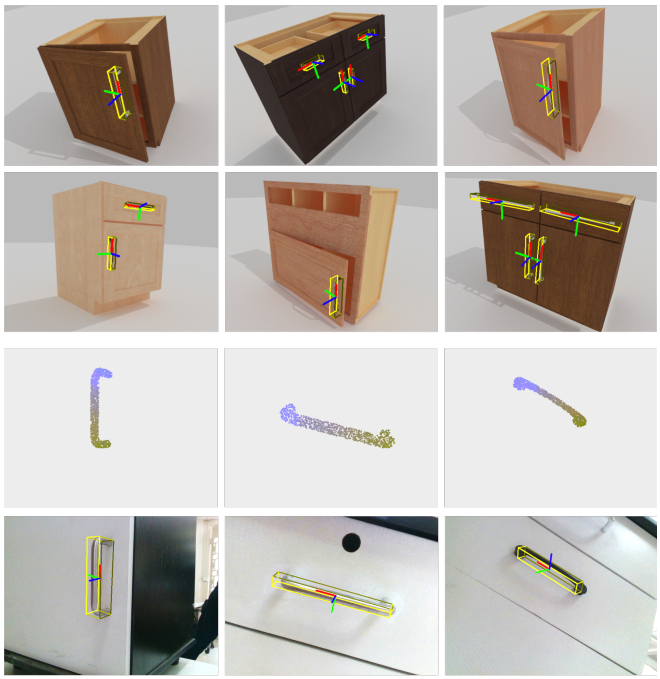

Fig. 3. Category-level object pose estimation results for handles of different cabinets in the simulator (the first two rows) and the real world (the bottom two rows).

C. Domain Randomization

To better adapt to real-world scenarios, we added domain randomization to the training environment. While training the active perception module, the texture of objects (including transparent, specular and diffuse material, which presents challenges for point-cloud cameras but not for RGB cameras), the position and intensity of light source and the initial pose of the object will change.

D. Balancing Accuracy and Efficiency (A & E)

While the balance of exploration and exploitation is a fundamental problem in policy learning [17], [1], there is a similar problem which we name it the balance of accuracy and efficiency. Higher accuracy of pose estimation may enhance the success rate of the entire task, but usually requires more views and increases the total distance between way-points, harming the efficiency of the method. On the other hand, increasing the efficiency necessitates fewer way-points and less variety of view points, negatively impacts the estimation accuracy and success rate.

To consider both the precision and efficiency in this balance, we introduced a parameter α in the reward computation of the Active Perception Module. This parameter is defined as $\alpha = \frac{r_{pen}}{r_{prec}}$, where r_{prec} is the reward for the precision of pose-estimation and r_{pen} is the penalty for moving distance. A smaller α biases the system towards better precision, while larger α tends to minimize the time required.

E. Impedance-control Manipulator

Using visual-aware closed-loop control in the context of active perception is less favored because it often results in inadequate visual information to estimate crucial states. For instance, when our monocular robot opens a door using the door handle, the camera is positioned too closely, making it challenging to observe the door rotation.

To overcome this limitation, we turned our attention to harnessing more information from robotic kinematics. The force exerted on the robotic arm can be used as a signal to adjust the manipulation. Our manipulator employs an impedance controller. Here, the end-effector has the freedom of movement but tends to return to its target pose, ensures its tolerance to minor errors. Given $\mathbf{X}$ and $\mathbf{R}$ as the translational and rotational error of the end-effector from its target pose, the torque τ for each robot joint is computed as:

$$\tau = \mathbf{J}^T \left(-k \begin{pmatrix} \mathbf{X} \\ \mathbf{R} \end{pmatrix} - b(\mathbf{J}\dot{\mathbf{q}}) \right) + \mathbf{N}, \quad (2)$$

Where $\mathbf{J}$ denotes the Jacobian matrix of the robot, k, b represent the stiffness and damping terms, respetively. The variable $\mathbf{q}$ is the current robot joint state, and $\mathbf{N}$ is the additional term account for handle Jacobian nullspace and Coriolis force. Viewing the manipulation trajectory as a time-dependent function, we can dynamically predict the subsequent point on the trajectory, leading to a reliable manipulation policy that remains resilient to disturbances and can effectively manage both revolute and prismatic articulated objects. Let $\mathbf{p}$ denote the pose of the end-effector over time. Then we determine the current target pose as:

$$\mathbf{p}^* = \mathbf{p} + k_1 \dot{\mathbf{p}} + k_2 \ddot{\mathbf{p}}, \quad (3)$$

Here k_1 and k_2 are coefficients for correcting direction and curvature of the trajectory.

IV. Experiment

A. Task Settings

We designed six challenging tasks to evaluate our method. In all tasks, a robotic arm is required to accomplish a specific manipulation goal with different objects.

Open Door: A door is initially closed, the agent needs to open the door larger than 0.15 rad (8.6 degrees). The position and rotation of the door is randomized within a range to make the task more challenging.

Open Door 45°: The harder version of Open Door. The agent needs to open the door to more than 45 degrees.

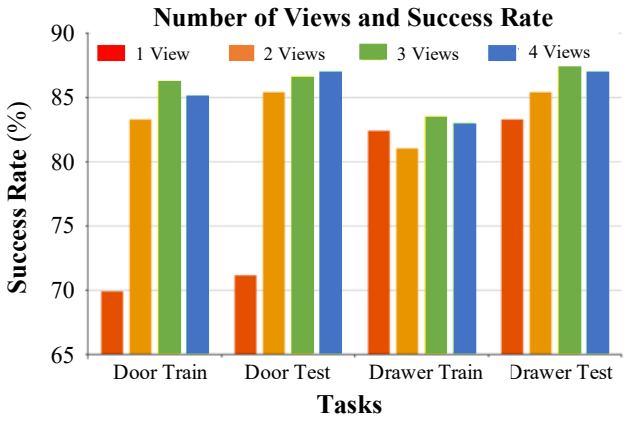

Fig. 4. Performance of our method under different number of views.

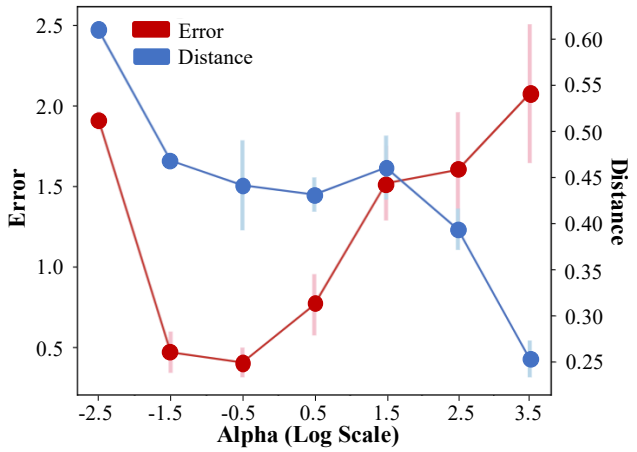

Fig. 5. Our method under different α for balancing A&E. The x-axis indicates the value of α, the red curve corresponds the error of pose estimation and the blue curve is the average moving distance during manipulation. The dot and bar are the mean and standard deviation (respectively) over 5 differently evaluated policies.

Open Drawer: A drawer is initially closed, the agent needs to open the drawer larger than a specific distance (15 centimeters). The position and rotation of the drawer is also randomized.

Open Drawer 30cm: The harder version of Open Drawer. The agent needs to open the drawer to more than 30 centimeters, which is fully open.

Open Pot: A kitchen-pot is initially on the floor with its lids on, the agent needs to lift the lid to a specific height. The position and rotation of the pot is also randomized.

Pick Mug: A mug is initially on the floor, the agent needs to pick up it to a specific height. The position and rotation of the mug is also randomized.

To evaluate our method, we trained our models using two datasets: PartNetMobility, a 3D articulated object dataset [22], and ShapeNet, a comprehensive rigid 3D shape dataset [2]. All training was conducted within the SAPIEN simulator [30], [11]. Within the simulator, our experiments spanned 184 shapes from 4 distinct object categories. Additionally, we selected real-world objects for testing.

For each task, we divided the objects into a training set and a testing set, trained our method, baselines and ablations fully on the training set and saved checkpoints every 25 time-steps within 2000 total time-steps. Then, we selected the checkpoint with the highest reward for comparison. We use average success rate to evaluate our method.

B. Baselines and Ablation

We benchmarked our method against six other algorithms, categorizing them into two groups based on their input type: four point-cloud-based and two that exclusively use RGB. The results are summarized in Table. I. The following is a brief description of each:

Where2Act [21]: Operates on point-cloud inputs, estimating per-point action scores. To execute the task, we integrated it with our manipulation policy, selecting the point with the highest score for interaction.

Flowbot3D [8]: Predicts the point-wise motion direction on the point cloud, denoting it as 'flow'. The point with the largest flow magnitude serves as our interaction point. Subsequent manipulations utilize our policy. Notably, we replaced the original suction gripper with a parallel one to ensure a fair comparison.

UMPNet [32]: Accepts RGBD images, predicting an action point on the image which is then projected into 3D space based on the depth data of the predicted pixel. Similar to the above methods, we paired it with our manipulation policy. The original suction gripper in this method was also replaced with a parallel gripper for comparison.

GAPartNet [9]: A pose-centric approach that predicts the pose of an object's part from point-cloud inputs. Manipulations are executed based on the predicted pose of the relevant task's component using our policy.

DrQ-v2 [34]: Represents the cutting-edge in pure RL methodologies. Here, reinforcement learning directly trains the manipulation policy. Inputs for this policy encompass both the robot's state and an RGB image, culminating in an output specifying the desired 6D pose of the robot's end-effector.

LookCloser [14]: A multi-view RL model combining third-person and egocentric viewpoints. While DrQ-v2 is confined to a single eye-on-hand camera's image input, LookCloser's used of multi-view input and visual transformers [7] enables the fusion of data from varied angles.

To elucidate the contribution and effectiveness of individual modules within our approach, we conducted an extensive ablation study. Six experiments were carried out, each omitting or adjusting specific components:

Ours w/o Global Scheduling: Rather than leveraging the observation perspective determined by the Global Scheduling Policy (Sec III-A), this experiment uses two manually set fixed perspectives for perception.

Ours w/o Impedance Control: This variant employs an open-loop manipulation policy. In the absence of the impedance control manipulator (Sec III-E), the policy operates by moving directly to the desired position.

Ours w/o Domain Randomization: We trained our method without employing the domain randomization process outlined in Sec III-C.

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND BASELINES

Methods	Modality	Open Door		Open Door 45°		Open Drawer		Open Drawer 30cm		Open Pot		Lift Mug	
		Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Where2Act	Point-clouds	8.0	7.0	1.8	2.0	5.9	7.5	1.1	0.6	**30.0**	55.3	20.9	19.6
Flowbot3D	Point-clouds	19.5	20.4	6.8	6.4	27.3	25.8	16.9	11.3	2.5	7.4	4.9	4.3
UMPNet	Point-clouds	27.1	28.1	11.0	10.9	16.6	18.8	4.4	5.6	19.1	36.9	26.6	22.9
GAPartNet	Point-clouds	69.5	74.5	39.4	43.6	50.6	59.3	44.6	48.6	5.3	10.8	0.0	0.0
DrQ-v2	RGB	1.8	2.5	0.8	0.8	1.9	1.0	1.4	0.5	0.1	0.0	0.0	0.0
LookCloser	RGB	1.5	1.25	0.8	0.8	0.8	0.0	0.0	0.0	0.3	0.0	4.8	6.5
Ours	RGB	**89.3**	**88.9**	**47.8**	**52.9**	**83.0**	**87.0**	**63.5**	**61.9**	22.8	**55.6**	**48.4**	**41.9**

TABLE II

ABLATION STUDY OF OUR METHOD

Methods	Open Door		Open Door 45°		Open Drawer		Open Drawer 30cm	
	Train	Test	Train	Test	Train	Test	Train	Test
Ours w/o Global Scheduling	78.6	78.4	49.8	39.1	82.4	83.3	**64.8**	56.6
Ours w/o Impedance Control	74.5	74.0	24.6	29.3	68.9	68.4	24.6	29.3
Ours w/o Domain Randomization	66.6	73.0	32.5	36.4	77.9	77.6	40.6	30.1
Ours w/o Pose Estimation Tricks	65.5	36.0	48.3	22.8	35.9	42.0	43.0	20.3
Ours	**89.3**	**88.9**	**51.1**	**52.9**	**83.0**	**87.0**	63.5	**61.9**

Ours w/o Pose Estimation Tricks: This experiment omits the kinematics-guided depth-aware fusion module from the object pose estimator, as detailed in Sec III-B.

Ours w/ Different Number of Views: This set of tests alters the approach by varying the number of views. The results underscore the diminishing returns of adding extra viewpoints, emphasizing the importance of striking a balance between accuracy and time-efficiency.

Ours w/ Balancing A&E: For this group, the number of views is held constant at four. By adjusting the parameter α in the reward computation of the Global Scheduling Policy, this experiment showcases the interplay between accuracy and efficiency.

C. Quantitative Results in Simulator

Table. I shows our large-scale evaluation in simulator over different tasks. The results indicate that our method consistently surpasses all baselines across nearly all tasks. Notably, in the Open Pot task, the majority of methods exhibit improved performance on the test set. This observation can be attributed to the train-test split used by all methods, with the test set encompassing a greater number of simpler instances. It's also worth highlighting the significant performance decline of GAPartNet in the Pot and Mug tasks. This is likely a consequence of substantial pose estimation errors, especially considering the heightened precision required to pinpoint critical parts of the objects in these tasks.

As depicted in Table II, each module within our proposed methodology plays a pivotal role. The effects of omitting the impedance control become increasingly pronounced in more intricate tasks (Open Door 45° and Open Drawer 30cm). This underscores the indispensability of the closed-loop impedance controller, particularly in long-horizon manipulative tasks.

Fig. 4 reveals an intriguing trend: the augmentation of way-points directly correlates with an elevation in the average success rate of manipulations. However, the addition of more way-points inevitably leads to diminishing returns. With up to 4 way-points, no significant improvement is observable. These results reinforce the importance of finding a harmony between accuracy and efficiency.

Lastly, Fig. 5 illustrates that the pose estimation accuracy diminishes when α is excessively large. Concurrently, the average moving distance experiences a decline as α increases. The unintuitive U-shaped error curve is possibly due to imperfect reward design, which the terms other than error and distance penalty dominates the overall reward when α is too small, leading to a sub-optimal policy.

D. Real-world Experiment

We employed a Franka Panda robot arm as our manipulator and fixed a Realsense camera with RGB-only output onto the robot's end-effector. The observations from the camera are directly used by the agent without refinement. Videos can be found on https://rgbmanip.github.io/.

V. CONCLUSION

In this study, we presented a pioneering approach to active pose estimation for monocular robotic manipulation. Our method uniquely equips robots with the ability to handle different tasks with monocular RGB inputs. This is achieved through a three-pronged process. 1) The robot explores the environment actively. 2) Pose information of interested objects is derived from the exploration. 3) Manipulation is achieved with a closed-loop impedance control policy.

A notable implication of our work is the attainment of robust manipulation control without the necessity for point-cloud sensors. Furthermore, experimental evaluations solidify the superiority of our method, as it consistently surpassed all baseline approaches.

ACKNOWLEDGEMENT

This project was supported by The National Youth Talent Support Program (8200800081) and National Natural Science Foundation of China (No. 62136001). This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Number: 24209223).

REFERENCES

[1] Susan Amin, Maziar Gomrokchi, Harsh Satija, Herke van Hoof, and Doina Precup. A survey of exploration methods in reinforcement learning. *arXiv preprint arXiv:2109.00157*, 2021.

[2] 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] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021.

[4] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782, 2021.

[5] Kai Chen, Stephen James, Congying Sui, Yun-Hui Liu, Pieter Abbeel, and Qi Dou. Stereopose: Category-level 6d transparent object pose estimation from stereo images via back-view nocs. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2855–2861. IEEE, 2023.

[6] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: Multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. *arXiv preprint arXiv:2210.06575*, 2022.

[7] A Dosovitskiy, L Beyer, A Kolesnikov, D Weissenborn, X Zhai, and T Unterthiner. Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Ben Eisner, Harry Zhang, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. In *Robotics: Science and Systems (RSS)*, 2022.

[9] 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.

[10] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. End-to-end affordance learning for robotic manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2023.

[11] 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.

[12] Radu Horaud, Miles Hansard, Georgios Evangelidis, and Clément Ménier. An overview of depth cameras and range scanners based on time-of-flight technologies. *Machine vision and applications*, 27(7):1005–1020, 2016.

[13] Jeffrey Ichnowski, Yahav Avigal, Justin Kerr, and Ken Goldberg. Dexnerf: Using a neural radiance field to grasp transparent objects. *arXiv preprint arXiv:2110.14217*, 2021.

[14] Rishabh Jangir, Nicklas Hansen, Sambaran Ghosal, Mohit Jain, and Xiaolong Wang. Look closer: Bridging egocentric and third-person views with transformers for robotic manipulation. *IEEE Robotics and Automation Letters*, 7(2):3046–3053, 2022.

[15] Achuta Kadambi, Ayush Bhandari, and Ramesh Raskar. 3d depth cameras in vision: Benefits and limitations of the hardware: With an emphasis on the first-and second-generation kinect models. *Computer vision and machine learning with RGB-D sensors*, pages 3–26, 2014.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[17] Pawel Ladosz, Lilian Weng, Minwoo Kim, and Hyondong Oh. Exploration in deep reinforcement learning: A survey. *Information Fusion*, 2022.

[18] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Prior-free category-level pose estimation with implicit space transformation. *arXiv preprint arXiv:2303.13479*, 2023.

[19] Jun Lv, Qiaojun Yu, Lin Shao, Wenhai Liu, Wenqiang Xu, and Cewu Lu. Sagci-system: Towards sample-efficient, generalizable, compositional, and incremental robot learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 98–105. IEEE, 2022.

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[21] 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.

[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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[23] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[24] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642. IEEE, 2020.

[25] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[26] 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.

[27] 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.

[28] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021.

[29] Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, and Kui Jia. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. *Advances in Neural Information Processing Systems*, 33:13174–13184, 2020.

[30] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[31] Hantong Xu, Jiamin Xu, and Weiwei Xu. Survey of 3d modeling using depth cameras. *Virtual Reality & Intelligent Hardware*, 1(5):483–499, 2019.

[32] Zhenjia Xu, Zhanpeng He, and Shuran Song. Universal manipulation policy network for articulated objects. *IEEE Robotics and Automation Letters*, 7(2):2447–2454, 2022.

[33] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018.

[34] Denis Yarats, Rob Fergus, Alessandro Lazaric, and Lerrel Pinto. Mastering visual continuous control: Improved data-augmented reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

[35] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022.

[36] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

VI. APPENDIX

A. Dataset

We used 3D models from PartNetMobility [22] and ShapeNet [2]. Details can be found on `https://`

rgbmanip.github.io.

B. Training Details

1) Global Scheduling Policy: The Global Scheduling Policy solves the scheduling problem as a Markov Decision Process. We used Proximal Policy Optimization (PPO) to train our Global Scheduling Policy. The reward function is the weighted sum of the following terms:

- Move-target difference reward: $\|p_{cam} - p_{tar}\|_2$, where p_{cam} is the current position of the camera. p_{tar} is current target position of the camera.
- Move success reward: $\mathbb{I}$(Move success)
- Move period penalty: the number of steps used to move from previous position to the new one.
- Distance penalty: $\left\|p_{cam} - p_{prop}\right\|$, where p_{prop} is a point $0.9m$ above the base of robot.
- Orientation reward: $q_{cam} \cdot q_{prop}$, q_{cam} is the quaternion which the camera can face directly to the object.
- Look-at regularization penalty: $(\|p_{look\ at} - p_{tar}\|_2 - 1)^2$, where $p_{look\ at}$ is the target position where the camera should be facing at.
- Mask bounding-box penalty: $\left\|mid - [0.5\ 0.5]^T\right\|_2$, mid is the coordinate of the central pixel of the bounding box of the object in the current view.
- Mask bounding-box boundary penalty: $\mathbb{I}(l \leq 0.1) + \mathbb{I}(r \geq 0.9) + \mathbb{I}(d \leq 0.1) + \mathbb{I}(u \geq 0.9)$, where l, r, d, u are the boundaries of the bounding box.
- Object-in-view reward: $\mathbb{I}$(object in current view)
- Center reward: $\frac{1}{1+\left\|p_{pred} - p_{gt}\right\|_2^2}$, here p_{pred} and p_{gt} are the predicted and ground-truth center of the object.
- Orientation reward: $\frac{1}{1+\left\|o_{pred} - o_{gt}\right\|_2^2}$, here o_{pred} and o_{gt} are the predicted and ground-truth orientation of the object.
- View diversity reward: $\mathbb{I}(\langle p_{cam} - p_{obj}, p'_{cam} - p_{obj}\rangle > 0.3)$.

The model itself consists of a policy network and a value network. Each network uses a separate MLP with hidden layer size [96, 96, 32]. We used adaptive learning rate varying from 2×10^{-4} to 5×10^{-3} to train this model.

2) Multi-view Object Pose Estimator: The object pose estimation model was trained ahead of the Global Scheduling Policy based on pre-collected synthetic data. We captured multi-view images in the simulator at viewpoints that were uniformly sampled from the hemisphere around the target object. Then, random image pairs were fed into the pose estimator for model training. For the homography-based feature fusion, we sampled the hypothetical depth plane between 0.1 and 2.4 with an interval of 0.1. For each object part, the model predicted its normalized coordinate map, depth map, and object pose and size parameters. The object rotation was parameterized with a continuous 6D representation [36]. The training loss is the weighted sum of the following terms:

- Pose loss: $\|\mathbf{R}_{pred} - \mathbf{R}_{gt}\|_2 + \|\mathbf{t}_{pred} - \mathbf{t}_{gt}\|_2 + \|\mathbf{s}_{pred} - \mathbf{s}_{gt}\|_2$, where $(\mathbf{R}, \mathbf{t}, \mathbf{s})_{pred}$ and $(\mathbf{R}, \mathbf{t}, \mathbf{s})_{gt}$ are the predicted and ground-truth pose parameters.
- Coordinate map loss: $\|C_{pred} - C_{gt}\|_1$, where C_{pred} and C_{gt} are predicted and ground-truth coordinate maps respectively.
- Depth loss: $\|D_{pred} - D_{gt}\|_1$, where C_{pred} and C_{gt} are predicted and ground-truth depth maps respectively.

For different object categories, the object pose estimation model was trained separately.

3) Impedance-control Manipulator: This model does not require training.

C. Environment Settings

In preparation for the setup of our real-world experimental environment, we assembled a collection of fifteen test objects, distributed across three categories: mugs, cabinets, and pots, with each category comprising five distinct varieties.

1) Simulation: The Franka robotic arm was designated as our agent, onto which an RGB camera was mounted at the gripper's location. An object is located in front of the robotic arm as the manipulation target. For the observation of our agent, a mask for the object is captured along with the RGB image from the mounted camera.

2) Real-world: Our operational pipeline in the real-world setting closely paralleled that of the simulator, with a singular distinction being our approach to mask selection. This process was made more flexible and observable by allowing mask selection to be guided either by prompts from the SAM model or through manual annotation.

3) Domain-randomization for pose estimation: We employed the Sapien [30] rendering engine to advance texture and lighting randomization, aiming to improve synthetic dataset realism for our tasks. The randomization includes different materials (transparent, specular and diffuse) of the same object, the intensity (strong or weak) and location (sampled on vertices and edge-centers on a surrounding cube) of light source will also change. This randomization process enriches our dataset for the pose estimator with varied appearances and lighting.

The initial pose of the target object is selected randomly. More precisely, the pose can be designated as a tuple of four values (α, β, d, h). α is the rotation along the z-axis of the object, β is the azimuth relative to the robotic arm, d is the distance from the robotic arm and h is the height of the object. For different tasks, those values has different distributions. The unit of angles are radians in Table.III.

TABLE III
DISTRIBUTION OF PARAMETERS

Tasks	α		β		d		h	
	Low	High	Low	High	Low	High	Low	High
Open Door	-0.20	0.20	-0.40	0.40	0.50	0.85	0.01	0.05
Open Drawer	-0.20	0.20	-0.40	0.40	0.50	0.80	0.01	0.05
Open Pot	-0.20	0.20	-0.40	0.40	0.20	0.38	0.01	0.30
Lift Mug	1.57	4.71	-0.40	0.40	0.44	0.50	0.10	0.15