2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

GPD-AP: A Grasp Pose-Driven Active Perception Framework for Occlusion-Robust Robotic Manipulation

Yancong Wei[1], Yunyi Pang[2], Sicheng Liu[3], Kangkang Dong[1], and Houde Liu[1,*]

***Abstract*— Humans instinctively adjust their viewpoints to resolve occlusions and infer spatial relationships, enabling effective perception and navigation in cluttered environments. This capability, however, remains a significant challenge for robotic systems. To address this, we propose GPD-AP, a novel active perception framework that leverages grasp pose estimation and associated scoring to systematically tackle grasping tasks in occluded and cluttered settings. The core innovation lies in an end-to-end system where a computationally efficient grasp pose estimation module directly informs a Next-Best-View (NBV) planner. This integration shifts the focus from generic scene exploration to a grasp-oriented visual search, guiding the robot to viewpoints that minimize uncertainty about potential grasps. To train and validate GPD-AP, we introduce a simulation reset method capable of generating highly challenging scenes with partially or fully occluded target objects. Experimental results demonstrate that GPD-AP improves grasping success rates by 30% in dense obstacle environments, effectively enabling the transition of target objects from invisible to visible and graspable states. This work marks a significant step towards autonomous and intelligent robotic manipulation in unstructured real-world scenarios.**

I. INTRODUCTION

Robotic grasping and manipulation in unstructured, cluttered environments remains a critical challenge[1], particularly when performing target-oriented grasping in scenarios where the target object is partially or fully occluded. Effective execution of such tasks requires the robot to accurately detect and localize the target object while precisely maneuvering the end effector to perform grasping or other manipulation tasks as efficiently as possible [2].

In complex occluded scenes, many existing approaches rely on removing obstacles to gradually expose the target object [3]. While effective in certain controlled environments, this strategy often proves impractical in real-world settings where occlusions are difficult to remove or obstacles cannot be conveniently repositioned. These limitations necessitate the adoption of active vision strategies to address occlusions more intelligently, enabling the robot to dynamically adjust its viewpoint to reveal the target object and facilitate subsequent manipulation tasks.

We propose GDP-AP, a grasp pose-driven grasping framework designed to address occlusion-rich environments with a focus on the target object. This framework integrates state-of-the-art pose estimation models [4] and leverages the estimated poses and their associated confidence scores to guide reinforcement learning (RL)-based action generation. Unlike prior work that primarily applies grasp poses to simple grasping tasks [5], GDP-AP extends their application to more complex, cluttered scenes. This is achieved by incorporating grasp poses as a key feature in RL policy networks, enabling richer environmental understanding and more adaptive decision-making. Moreover, while traditional methods rely on inverse kinematics to directly plan trajectories, our approach deepens the integration of pose estimation into RL policies for more dynamic and responsive action generation.

The main contributions of this work are as follows:

1) We propose an efficient framework for constructing occluded scenes, enabling streamlined training and evaluation of reinforcement learning algorithms in environments with varying obstacle configurations. The adaptability of this framework to diverse cluttered scenes is also demonstrated.
2) We explore the integration of pose estimation into RL policies, using object poses and their confidence scores as input features for the policy network. This enables an end-to-end optimization pipeline that seamlessly links environmental perception with action generation.
3) We design a reward-based decision-making mechanism that dynamically balances the trade-off between acquiring new perspectives for environmental understanding and executing grasping actions efficiently. This mechanism effectively resolves the inherent conflict between information acquisition and task efficiency.

II. RELATED WORKS

A. Multi-View Strategies for Occlusion-Robust Grasping

Vision-driven robotic grasping systems often face significant challenges in cluttered environments, where sensor noise and visual occlusions are prevalent. Several studies have attempted to mitigate these effects, such as Ten et al. [6] fused point clouds from multiple views to compute grasp poses, achieving a 9% improvement in grasping success rates. While effective, this approach risks incorporating irrelevant viewpoints from a pre-defined set, which may degrade overall performance by adding unnecessary noise to the fused point cloud. For instance, Yang et al. [7] further explored the use of information gain metrics to guide multi-view detection and utilization, optimizing grasping in cluttered scenes.[8]

[1]Tsinghua University Shenzhen International Graduate School, Shenzhen, China
[2]Huazhong University of Science and Technology, Wuhan, China
[3]Actibot Intelligence, Shenzhen, China
Corresponding author: Houde Liu, liu.hd@sz.tsinghua.edu.cn
This work was supported by the Shenzhen Science and Technology Program (Grant No. RCJC20210706091946001) and the Shenzhen Science and Technology Program (Grant No. ZDCY20250901104207008).

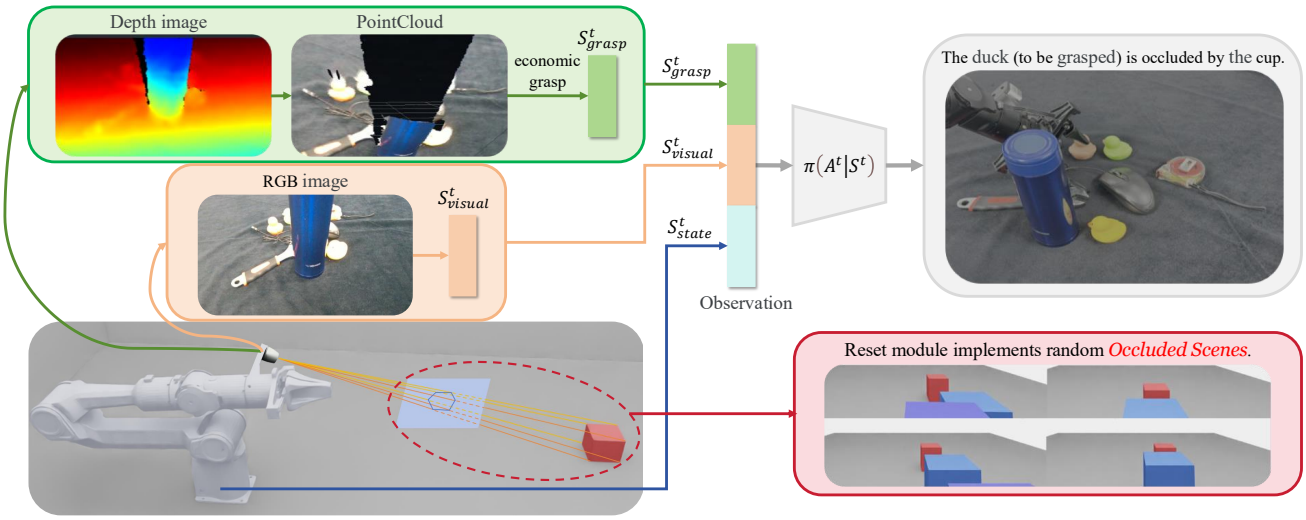

Fig. 1: Method overview. Our method processes the gripper depth image to point cloud, which then be sent to economic grasp for predicting grasping poses S^t_{grasp}, along with S^t_{visual} and S^t_{state} for learning $\pi(A^t|S^t)$. The reset module automatically generates randomized occluded scenes during environment resets.

proposed a deep learning architecture with an augmented memory capacity to achieve zero-shot grasp.

Despite these advancements, most existing methods are constrained by discrete viewpoint sampling and a focus on generic scene reconstruction rather than grasping efficiency. In this work, we argue that for grasping tasks in occluded scenes, the most effective viewpoint is not necessarily the one that provides the most general information, but rather the one that provides the most actionable information for a successful grasp. In contrast, our approach prioritizes object-centric active vision exploration, directly targeting occlusion resolution to enhance grasping success in cluttered environments.

B. Grasp pose estimation and reinforcement learning

Building on the advancements in active perception and multi-view strategies, a parallel research area has focused on improving grasp pose estimation itself through learning-based methods. For example, Fang et al. [9] developed a general-purpose 6D grasp pose estimation system trained on a large-scale dataset, while Fang et al. [10, 11] extended this work by directly predicting 6D grasp poses from visual scenes through large-scale data-driven training. Building on these efforts, Back et al. [12, 13] focused on robust pose estimation in occlusion-heavy environments, achieving promising results using an even larger dataset. Additionally, Li et al. [3] introduced a Neural Graspness Field (NGF) network that effectively links grasp pose estimation with next-best-view (NBV) planning, demonstrating improved performance in complex scenarios.

The integration of these learning-based models has significantly enhanced grasping systems' adaptability across diverse environments. However, these models often produce a dense set of potential grasp poses, many of which are redundant or noisy due to occlusions and sensor limitations. To address this issue, researchers have proposed fusion strategies that combine nearby grasp poses in the world coordinate system, refining the selection of optimal candidates. While previous studies have established the importance of high-quality grasp pose estimation, they often treat it as a static perception step. In contrast, our work fundamentally integrates these estimated poses and their associated confidence scores into a dynamic, active perception framework, where the pose distribution itself becomes a direct input for intelligent viewpoint planning and grasping execution.

C. Reinforcement learning and active perception

Reinforcement Learning and Active Perception Recent active grasping frameworks [14, 15, 16] have begun integrating next-best-view (NBV) planning into grasping policies. These frameworks predominantly focus on tasks like 3D reconstruction [17, 18], rather than goal-oriented view exploration for grasping. When faced with limited or insufficiently diverse camera perspectives, robots often need to perform exploratory actions, such as shifting viewpoints or focusing on specific regions, to gather a more comprehensive understanding of the scene. This enables them to complete tasks with greater accuracy.

Initially, active view planning relied on rule-based planners [14, 15], which manually designed strategies for viewpoint selection. While effective in certain scenarios, these methods often struggle with adaptability to dynamic or cluttered environments. With the advent of reinforcement learning (RL), learning-based approaches for active view planning have emerged, offering increased flexibility and adaptability. Some studies have even explored imitation learning methods [19, 20, 21] to model exploratory behavior.

Unlike rule-based methods, RL inherently incorporates active exploration, enabling it to achieve high generalization and robustness in complex scenarios. This makes RL particularly well-suited for tasks requiring goal-driven view selection. In this paper, we leverage RL to concentrates grasp

pose distributions on the target object, effectively mitigating occlusion and improving the overall success rate of grasping tasks in cluttered environments.

III. Method

A. Overview

We address the problem of robotic grasping in cluttered environments, where the target object is partially or fully occluded by surrounding objects. We propose a comprehensive framework consisting of three key components: (1) a multi-modal observation system that integrates fused 6D grasp poses, visually encoded RGB images, and the robot's joint state for effective perception and motion planning; (2) a tailored reward system that incentivizes viewpoint exploration and occlusion resolution; and (3) an innovative reset module that automatically generates randomized occluded scenes during environment resets, as shown in Fig. 1. By diversifying object arrangements and occlusion relationships, the reset module provides a robust and scalable training setup, enhancing the system's adaptability to cluttered environments.

Together, these components enable the robot to balance viewpoint optimization and grasping success, effectively addressing the unique challenges posed by complex occluded scenes.

B. Pose Estimation

To address the challenges of maintaining efficient inference during reinforcement learning training, we propose a pose estimation and fusion strategy tailored for cluttered scenes. Specifically, we adopt a lightweight, cost-effective pose estimation model [4] that takes point clouds as input and outputs grasp poses along with their corresponding scores. During training, pose estimation is performed every N steps, and the resulting grasp poses are used as part of the robot's observations. However, due to the large number of grasp poses generated in cluttered environments, there is often a significant degree of redundancy and overlap. To address this, we apply a filtering and fusion approach based on voxelized clustering to maintain only the most representative grasp poses.

Unlike traditional methods that directly select the top-scoring grasp poses, we employ a pose integration strategy that balances pose quality and spatial distribution. For each voxel, we maintain an average grasp pose, which is iteratively updated when new grasp poses are introduced. Specifically, if a newly introduced grasp pose has an angular deviation greater than a predefined threshold θ from the average grasp pose, it is discarded. Otherwise, the new pose is fused with the existing average pose using weighted averaging for position and spherical linear interpolation (SLERP) for orientation. For example, in Fig. 2, the yellow grasp pose should be fused and the red one should be discarded. The fusion rules are as follows:

The fused position $\text{pos}_{\text{fused}}$ is computed as:

$$\text{pos}_{\text{fused}} = \frac{\text{score}_i \cdot \text{pos}_i + \text{score}_j \cdot \text{pos}_j}{\text{score}_i + \text{score}_j}, \tag{1}$$

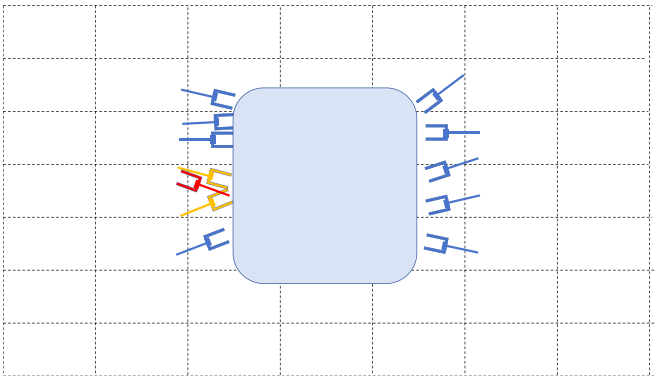

Fig. 2: Grasp pose filtering and fusion process.

where pos_i and pos_j are the positions of the average pose and the new pose, respectively, and score_i and score_j are their corresponding scores.

The fused orientation $\mathbf{q}_{\text{fused}}$ is computed using spherical linear interpolation (SLERP):

$$\mathbf{q}_{\text{fused}} = \frac{\sin\left(\left(1 - \frac{\text{score}_j}{\text{score}_i + \text{score}_j}\right)\theta\right)}{\sin(\theta)}\mathbf{q}_i + \frac{\sin\left(\frac{\text{score}_j}{\text{score}_i + \text{score}_j}\right)\theta}{\sin(\theta)}\mathbf{q}_j, \tag{2}$$

where θ is the angular difference between orientations $\mathbf{q}_i$ and $\mathbf{q}_j$, computed as $\cos(\theta) = \mathbf{q}_i \cdot \mathbf{q}_j$.

After filtering and fusion, only the refined grasp poses are retained to represent the scene. Notably, while the grasp pose scores are used for filtering and fusion, we do not include them in the observation vector during reinforcement learning. This design choice is motivated by the fact that scores primarily represent the confidence of the pose estimation model, which can introduce bias into the learning process. Instead, we rely on the fused spatial and orientation information, which provides a more robust representation of the scene for downstream tasks.

By applying this filtering and fusion strategy, our method reduces redundancy in grasp poses while preserving the essential spatial features of the scene. This approach ensures that the robot receives high-quality grasp pose information without being overburdened by excessive or redundant data.

C. Reset Module

To enable efficient training in occluded scenes, we designed a Reset Module that systematically generates randomized yet meaningful training environments. This module employs a domain randomization strategy tailored for creating occluded scenes, which our experiments show significantly enhances the model's learning efficiency. By dynamically resetting the scene, we ensure sufficient diversity and challenging configurations for the robot to learn robust grasping strategies.

Most desktop scenarios involve objects that can be approximated as cubes of varying sizes. When only a few cubes are placed randomly, occlusion rarely occurs, leading to underutilization of training data, as illustrated in Fig. 3. To address this, the Reset Module introduces purposeful scene

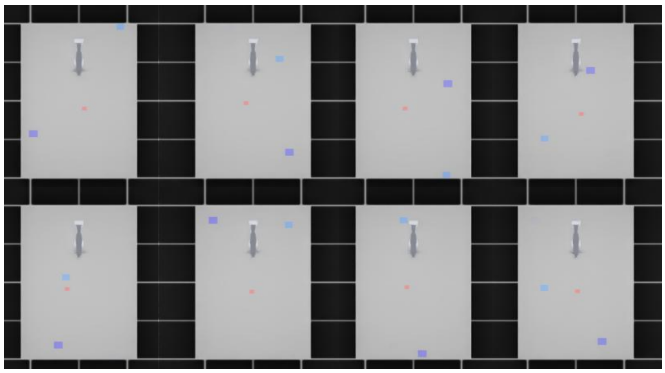

Fig. 3: Random placement of obstacles hardly generates occlusion scenes (where the red block is the target object and other blocks are obstacles).

configurations that maximize occlusion while maintaining variability.

Farthest Target Criterion. The Reset Module first defines the position (x, y, z) of the target object. To maximize occlusion, the target object is placed at the farthest point within the robot's reachable workspace. This configuration ensures that obstacles positioned between the target object and the robot create occluded scenes. Objects placed farther than the obstacles are deprioritized, as they contribute little to occlusion. Obstacles are strategically positioned closer to the robot, while the target object remains farther away, facilitating the formation of meaningful occlusion.

View Cone-Based Placement. The generation of occluded scenes is closely tied to the wrist camera's field of view (FOV), which is represented as a cone-like volume called the "view cone". Two primary types of occlusion are considered: **(1)** The target object lies outside the view cone, necessitating camera adjustment; **(2)** An obstacle blocks the view cone, preventing the camera from directly perceiving the target object.

To model these scenarios, the Reset Module focuses on obstacles within the view cone that occlude the target object. A plane is defined at the obstacle's height, intersecting the view cone. This plane determines the placement surface for the first obstacle, ensuring that occlusion occurs within the relevant visual field. This targeted placement strategy reduces randomness and increases the likelihood of creating occlusion directly relevant to the task.

Size-Based Obstacle Placement. After the first obstacle is placed using the view cone-based method, remaining obstacles are placed systematically to prevent overlap with the robot or other objects. To ensure effective occlusion, obstacles are prioritized by size, with larger obstacles placed first. Sorting obstacles in descending order of size prevents scenarios where a large rear obstacle overshadows a smaller front obstacle, as depicted in Fig. 4. This strategy ensures that occluded scenes are both challenging and diverse.

By dynamically resetting environments with the Reset Module, we achieve a balance between scene variability and task difficulty. This systematic approach enables the robot to encounter a wide range of occluded scenes, promoting the development of robust grasping strategies.

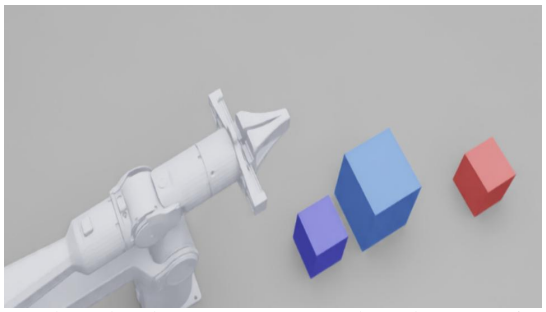

Fig. 4: Obstacle placement: Larger obstacles are prioritized to avoid smaller obstacles being overshadowed.

D. Reward Design

In grasping tasks, there exists a natural trade-off between observing the target object to gather meaningful information and approaching the target efficiently for grasp execution. To address this, we design a reward mechanism that dynamically balances these two objectives over time. Specifically, in the early stages of the task, the reward prioritizes observation, encouraging the robot to explore the target object from diverse perspectives. As the task progresses, the reward shifts its focus toward approaching the target object and completing the grasp. This balance is achieved through a temporal weighting mechanism.

Dynamic Weighting Mechanism. We introduce a time-dependent weight w_t that governs the balance between the observation reward $R_{\text{grasp\_score}}$ and the distance reward R_{distance}. The total reward is given by:

$$R_{\text{total}} = w_t \cdot R_{\text{grasp\_score}} + (1 - w_t) \cdot R_{\text{distance}}, \tag{3}$$

where w_t is a temporal weight that decreases smoothly from 1 to 0 as time progresses.

We define w_t using a cosine-based interpolation function:

$$w_t = \frac{1 + \cos\left(\pi \cdot \frac{t}{T}\right)}{2}, \tag{4}$$

where t is the current timestep, and T is the total number of timesteps in the task. At the beginning of the task ($t = 0$), $w_t = 1$, meaning the reward prioritizes observation. At the end of the task ($t = T$), $w_t = 0$, meaning the reward fully shifts to approaching the target.

Observation and Proximity Rewards. The observation reward $R_{\text{grasp\_score}}$ encourages the robot to gather meaningful information by evaluating the quality of grasp poses around the target object:

$$R_{\text{grasp\_score}} = \frac{\sum_{i \in \text{ROI}} s_i}{S_{\text{total}}}, \tag{5}$$

where s_i represents the score of the i-th grasp pose, and S_{total} is the sum of scores of all grasp poses within the region of interest (ROI).

The proximity reward R_{distance} incentivizes the robot to approach the target object efficiently:

$$R_{\text{distance}} = -\|p_{\text{ee}} - p_{\text{target}}\|, \tag{6}$$

where p_{ee} and p_{target} are the positions of the robot's end-effector and the target object, respectively.

Balancing Efficiency and Accuracy. This temporal weighting mechanism enables the robot to balance efficiency and accuracy dynamically. Early in the task, the observation reward dominates, guiding the robot to collect meaningful information and identify high-quality grasp poses. As the task progresses, the proximity reward becomes dominant, encouraging the robot to approach the target object quickly and execute the grasp. This design ensures that the robot achieves both robust grasping accuracy and operational efficiency in complex environments.

IV. Experiments

A. Experiment Description

We conducted algorithm training and evaluation in the Isaac Lab simulation environment, utilizing the Piper robotic arm to set up a desktop environment. Obstacles were placed based on our proposed placement strategy. To ensure the target object remained within the robotic arm's workspace and avoided collisions with obstacles, the target object was randomly sampled within a predefined square area.

Baselines: To benchmark the performance of our proposed method, GPD-AP, we selected two representative baselines:

- **AnyGrasp** ([10]): A state-of-the-art grasping model.
- **GAMMA** ([5]): A reinforcement learning method with grasp pose estimation, using 9D grasp features to represent a single grasp.

Evaluation Metrics: The algorithm's performance was evaluated based on three key metrics that reflect both the perception and execution capabilities of the robotic arm:

- **Visibility**: The proportion of the target object visible within the camera's field of view, with at least 80% visibility considered valid.
- **Proximity**: The robotic arm's end-effector approaches the target object within 10 cm.
- **Collision**: A trial is considered a failure if the robotic arm collides with any obstacles during the process.

The experiments were performed across scenarios with increasing complexity:

- **No Obstacles**: A simple, unobstructed environment.
- **Simple**: Two obstacles placed in the workspace.
- **Medium**: Three obstacles added.
- **Hard**: Five obstacles to simulate highly cluttered environments.

Fig. 5 illustrates an example of the obstacle configurations used. The results across all scenarios are summarized in Table I.

TABLE I: Experiment Results

	No Obstacle	Simple	Medium	Hard
AnyGrasp	78%	52%	33%	29%
GAMMA	73%	66%	45%	37%
GPD-AP (Ours)	**76%**	**67%**	**62%**	**55%**

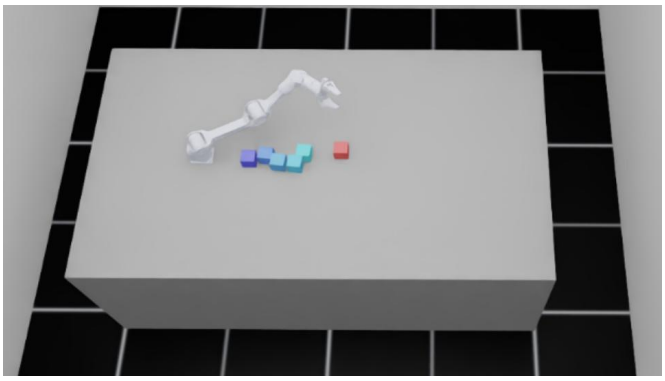

Fig. 5: Random obstacle placement with varying numbers of obstacles.

The results demonstrate the following trends:

- In non-occluded scenes (*No Obstacles*), GPD-AP performs comparably to GAMMA and Anygrasp, reflecting the simplicity of the task.
- As the number of obstacles increases (*Simple*, *Medium*, *Hard*), GPD-AP consistently outperforms both baselines, achieving a **55% success rate in the hardest scenario**, which demonstrates its robustness in highly cluttered environments.

B. Ablation Studies

To further investigate the contributions of individual components in GPD-AP, we performed ablation studies by disabling specific modules.

Grasp Pose Estimation Module: The grasp pose estimation module was removed by excluding the observation vector and the corresponding reward. Without this module, the success rate dropped significantly to 44%, as shown in Table II. This result highlights the module's critical role in enabling the robot to identify high-quality grasp poses, especially in occluded scenes.

Reset module: We compared the effectiveness of training with and without the proposed reset module. Without this module, obstacles and target objects were randomly placed without consideration of occlusion or collision constraints. The success rate dropped drastically to 28%, emphasizing the importance of high-quality training samples and realistic scene distribution.

TABLE II: Ablation Study for Our Modules

Model Variant	Success Rate
GPD-AP (without grasp pose)	44%
GPD-AP (without reset module)	28%
GPD-AP (Full)	**76%**

C. Real-World Experiments

To validate the effectiveness of GPD-AP in real-world scenarios, we deployed the algorithm on a physical robotic system comprising the Piper robotic arm and a Realsense D405 camera, as shown in Fig. 6. The evaluation focused

Fig. 6: Real-world experimental setup: Piper robotic arm paired with a Realsense camera.

on grasping performance in cluttered environments with randomly placed objects, where occlusion and object proximity pose significant challenges.

We conducted 50 grasp trials across 5 different clutter configurations. GPD-AP achieved a success rate of **62%** (31/50), with failures primarily occurring when the target object was heavily occluded or when grasp poses were kinematically unreachable due to joint limits.

These results demonstrate that GPD-AP successfully generalizes from simulation to real-world deployment, maintaining robust grasping performance despite challenging conditions.

V. CONCLUSIONS

In this paper, we propose a novel grasp pose-driven active perception (GPD-AP) framework designed for robotic grasping in occluded scenes. Unlike traditional methods that often rely on static or pre-defined viewpoints, GPD-AP enables the robot to autonomously adjust its viewing perspective to optimize grasping performance. By seamlessly integrating grasp pose estimation with active perception, the framework achieves significant improvements in grasping success rates, particularly under challenging occluded scenarios.

Through extensive simulation experiments and real-world validations, GPD-AP demonstrated its ability to adapt to cluttered environments, navigate occlusions, and execute efficient grasping operations. The proposed approach leverages environmental understanding via pose estimation, enabling the robot to make informed decisions about both perception and action. Ablation studies further highlight the essential contributions of key modules, including the grasp pose estimation module and scenario distribution strategy, to overall system performance.

This work represents a step forward in addressing the challenges of robotic manipulation in occluded and cluttered environments. Future research directions include extending the framework's applicability to dynamic or semi-structured environments, improving computational efficiency, and exploring its integration into large-scale robotic systems for tasks such as warehouse automation and industrial assembly.

REFERENCES

[1] Yaoyao Qian et al. "ThinkGrasp: A vision-language system for strategic part grasping in clutter". In: *arXiv preprint arXiv:2407.11298* (2024).

[2] Jue Wang, Xiaoxiang Sun, and Wei Wang. "Trajectory planning for manipulator grasping with obstacle avoidance and visual occlusion in a complex environment". In: *Mechanical Sciences* 16.2 (2025), pp. 445–455.

[3] Yitong Li et al. "Broadcasting support relations recursively from local dynamics for object retrieval in clutters". In: *arXiv preprint arXiv:2406.02283* (2024).

[4] Xiao-Ming Wu et al. "An economic framework for 6-dof grasp detection". In: *European Conference on Computer Vision*. Springer. 2024, pp. 357–375.

[5] Jiazhao Zhang et al. "Gamma: Graspability-aware mobile manipulation policy learning based on online grasping pose fusion". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 1399–1405.

[6] Andreas Ten Pas et al. "Grasp pose detection in point clouds". In: *The International Journal of Robotics Research* 36.13-14 (2017), pp. 1455–1473.

[7] Jun Yang, Dong Li, and Steven L Waslander. "Probabilistic multi-view fusion of active stereo depth maps for robotic bin-picking". In: *IEEE Robotics and Automation Letters* 6.3 (2021), pp. 4472–4479.

[8] Hamidreza Kasaei et al. "Simultaneous multi-view object recognition and grasping in open-ended domains". In: *Journal of Intelligent & Robotic Systems* 110.2 (2024), p. 62.

[9] Hao-Shu Fang et al. "Graspnet-1billion: A large-scale benchmark for general object grasping". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 11444–11453.

[10] Hao-Shu Fang et al. "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains". In: *IEEE Transactions on Robotics* 39.5 (2023), pp. 3929–3945.

[11] An Dinh Vuong et al. "Grasp-anything: Large-scale grasp dataset from foundation models". In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, pp. 14030–14037.

[12] Seunghyeok Back et al. "GraspClutter6D: A Large-scale Real-world Dataset for Robust Perception and Grasping in Cluttered Scenes". In: *arXiv preprint arXiv:2504.06866* (2025).

[13] Yi-Lin Wei et al. "Grasp as you say: Language-guided dexterous grasp generation". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 46881–46907.

[14] Michel Breyer et al. "Closed-loop next-best-view planning for target-driven grasping". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2022, pp. 1411–1416.

[15] Xuechao Zhang et al. “Affordance-driven next-best-view planning for robotic grasping”. In: *arXiv preprint arXiv:2309.09556* (2023).

[16] Haoxiang Ma et al. “Active perception for grasp detection via neural graspness field”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 38122–38141.

[17] Jacopo Aleotti, Dario Lodi Rizzini, and Stefano Caselli. “Perception and grasping of object parts from active robot exploration”. In: *Journal of Intelligent & Robotic Systems* 76.3 (2014), pp. 401–425.

[18] Gregory Kahn et al. “Active exploration using trajectory optimization for robotic grasping in the presence of occlusions”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2015, pp. 4783–4790.

[19] Pooya Abolghasemi and Ladislau Bölöni. “Accept synthetic objects as real: End-to-end training of attentive deep visuomotor policies for manipulation in clutter”. In: *2020 IEEE International conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 6506–6512.

[20] Haoyu Xiong et al. “Vision in Action: Learning Active Perception from Human Demonstrations”. In: *arXiv preprint arXiv:2506.15666* (2025).

[21] Xuxin Cheng et al. “Open-television: Teleoperation with immersive active visual feedback”. In: *arXiv preprint arXiv:2407.01512* (2024).