A Manipulation Pipeline for Grasping Unknown Objects in Heavy Clutter for Decontamination

Engjell Hyseni*,[1], Sebastian Nutto*,[1], Janna Nefzer[1], Miguel de Diego Pérez[2], Antonio Morales[2], Tamim Asfour[1]

***Abstract*— The decontamination of nuclear waste remains a challenging and largely manual process, exposing human workers to physical strain and potential health risks due to radiation. In this work, we present a manipulation pipeline for grasping unknown objects in heavily cluttered environments, motivated by real-world decontamination scenarios addressed in the ROBDEKON project. The proposed system integrates a robust perception pipeline for scene segmentation, a manipulation framework for grasp generation and selection and a failure detection and recovery mechanism. Our approach enables autonomous grasping of previously unseen objects in a cluttered scene from containers and prepares them for subsequent decontamination steps, thereby improving worker safety and increasing overall process efficiency.**

I. Introduction

Nuclear decontamination still relies largely on manual labor, with workers removing contaminated objects from cluttered containers for processing. This task is physically demanding, requires heavy protective equipment, and exposes workers to ionizing radiation, making the duration of human work strictly regulated. These limitations reduce productivity and motivate the development of robotic solutions that can take over hazardous and repetitive tasks. The ROBDEKON project seeks to automate key decontamination steps to improve safety, working conditions, and efficiency in nuclear decommissioning.

A crucial prerequisite for autonomous manipulation is robust scene understanding in complex, cluttered environments, as objects are unknown, unordered, and often partially occluded. Previous work in decontamination scenarios relied on local surface analysis for scene understanding and grasp generation [1]. While effective in cluttered scenes, such approaches are limited by hand-crafted geometric features and local heuristics, particularly under occlusions, sensor noise, or challenging surfaces. Recent advances in foundation models for perception offer new opportunities to address these limitations. In particular, segmentation models trained on large-scale datasets enable robust decomposition of complex scenes into individual object instances, improving scene understanding over purely geometry-based approaches.

*These authors contributed equally to the manuscript

The research leading to these results has received funding from the German Federal Ministry of Research, Technology and Space (BMFTR) under the competence center ROBDEKON and the Robotics Institute Germany (RIG).

[1]The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. Correspondence to: {engjell.hyseni, sebastian.nutto, asfour}@kit.edu

[2]The authors are with the Robotic Intelligence Laboratory, Jaume I, Castellon, Spain.

For grasp generation and execution, our work builds on the *MAkEable* mobile manipulation framework [2], which structures manipulation tasks into distinct stages: Discovery, Parameterization, Validation, Selection, and Execution. This enables systematic reasoning over multiple manipulation hypotheses and flexible task execution, facilitating the transfer of capabilities and knowledge. However, in its current form, *MAkEable* primarily treats grasp actions as independent unimanual actions and does not explicitly reason about handedness or paired grasps for grasping unknown objects. Such reasoning was previously performed manually via a teleoperation, limiting autonomous manipulation of heavy and large objects. Furthermore, the execution stage operates in an open-loop manner without explicit grasp success verification. These limitations are addressed in this work.

In summary, this work contributes (i) a foundation-model-based perception pipeline for cluttered decontamination scenes, (ii) an extension of the *MAkEable* framework towards bimanual grasp selection for unknown objects, and (iii) an execution-level failure detection and recovery mechanism.

II. Approach

Our manipulation pipeline follows the stages of the *MAkEable* framework and extends them with improved perception, bimanual grasp handling of unknown objects, and execution level feedback. An overview of the complete pipeline is shown in Fig. 1 and is discussed in the following sections.

1) Perception: The perception module forms the basis for robust grasp generation in challenging environments. We employ *FoundationStereo* [3], a neural network–based stereo vision model, to acquire dense and accurate point cloud representations, including transparent, reflective, or absorbing objects. For scene decomposition, we use the *Segment Anything Model 2* (*SAM2*) [4] to segment the scene into individual object instances. To obtain more uniform scene sampling, the initial prompt grid is generated using a Sobol sequence [5] instead of a Cartesian grid. Building on this, we implemented a self-prompting configuration of *SAM2* similar to [6], in which mask centroids from the initial prompts are reused for subsequent inference. This implicitly encodes scene structure and object-level priors, resulting in more consistent masks and reduced oversegmentation at the cost of increased runtime.

2) Discovery: Based on the segmented point cloud provided by the perception module, object oriented bounding boxes (OOBBs) are fitted for each object instance [7]. These

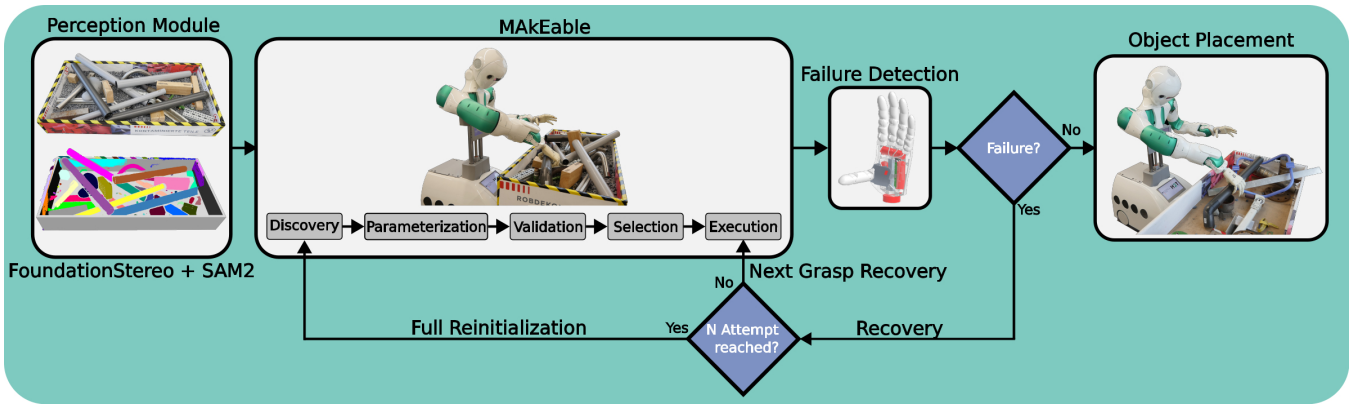

Fig. 1: The pipeline begins with the perception of the scene and its segmentation into individual objects instances. Based on this representation, grasp hypotheses are generated, parameterized, validated, and ranked. The best ranked grasp is then executed, after which in-hand sensing is used to detect failures. If needed, the system triggers appropriate recovery strategies to continue the manipulation task.

OOBBs serve as geometric abstractions for subsequent grasp generation. Grasp hypotheses are then generated on the OOBBs along their first principle axis. At this stage, the grasp hypotheses are robot-agnostic and represent abstract manipulation intents rather than executable action.

3) Parameterization: During the parameterization stage, abstract grasp hypotheses are instantiated as robot-specific actions, containing all information required for execution (e.g., robot base poses, and end-effector trajectories).

4) Validation: The generated actions are evaluated in a validation stage to ensure feasibility and success. This includes reachability checks, collision checking at the execution pose, and hand orientation validation. Optionally, for a bimanual grasp it is checked, if the grasps handedness matches the corresponding side of the object. Actions that violate any validation criterion are filtered out.

5) Selection: In the selection stage, the remaining valid actions are ranked according to multiple heuristics, such as the height of the grasp point and the distance the robot must move to reach the grasp [8]. For bimanual manipulation, the best ranked candidate grasp for the left and right hand for each object are combined into coordinated, symmetric grasp actions [9], effectively treating the object as two unimanually graspable subregions subject to handedness constraints. The result is a ranked list of executable actions.

6) Execution, Failure Detection and Recovery: During first time execution, the robot performs the highest ranked grasp action. After execution, grasp success is evaluated using multiple sensor modalities, including finger joint encoders, a wrist-mounted force-torque sensor, and a hand integrated time-of-flight (ToF) depth sensor [10]. Grasp success is inferred from threshold-based cues on joint closure, force-torque signatures, and depth measured by the in-hand ToF sensor. If a grasp is detected as unsuccessful, the method automatically applies one of two recovery strategies:

- **Next-Grasp Recovery:** The next best grasp from the ranked list is executed.
- **Full Reinitialization:** The perception and grasp generation pipeline is restarted, updating the scene representation to account for any changes caused by prior interactions.

The system first attempts *Next-Grasp Recovery*, and only if repeated attempts fail, *Full Reinitialization* is triggered. This structured recovery mechanism allows the system to recover from failures in grasp execution and achieve robust emptying of the cluttered box.

III. Preliminary Evaluation

We evaluate the proposed pipeline for unimanual grasps on the humanoid robot ARMAR-DE in a heavily cluttered scenario with unknown objects. The robot autonomously grasps objects from a cluttered box and places them into a target box. Five experiments were conducted, each transferring five objects. In total, 42 grasp attempts were executed, 25 of which were successful, yielding a success rate of $\approx 60\%$. *Next-Grasp Recovery* was applied in 82% of cases.

All experiments were run using the pipeline's fastest configuration. Even though grasp failures occurred, every experiment was ultimately completed successfully, with all five objects transferred. These results demonstrate that the method reliably completes the task autonomously, achieving a throughput of approximately one object every 90 seconds.

IV. Conclusion and Future Work

In this work, we presented a manipulation pipeline for grasping unknown objects in heavily cluttered environments, developed in the context of the ROBDEKON project. The proposed system integrates robust perception, manipulation planning, and failure recovery to enable autonomous grasping in challenging real-world scenarios.

Current limitations include naive grasp generation based on box-shaped primitives. Future work will focus on data-driven grasp generation directly from point cloud data, more intelligent grasp execution under environmental constraints, improved integration of bimanual grasping, reactive grasping using the in-hand ToF sensor, and multi-robot execution via a ROS bridge.

REFERENCES

[1] C. Pohl and T. Asfour, "Probabilistic spatio-temporal fusion of affordances for grasping and manipulation," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 2, pp. 3226–3233, 2022.

[2] C. Pohl, F. Reister, F. Peller-Konrad, and T. Asfour, "Makeable: Memory-centered and affordance-based task execution framework for transferable mobile manipulation skills," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 3674–3681.

[3] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "Foundationstereo: Zero-shot stereo matching," *CVPR*, 2025.

[4] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[5] I. Sobol', "On the distribution of points in a cube and the approximate evaluation of integrals," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, no. 4, pp. 86–112, 1967.

[6] A. S. Wahd, J. Küpper, J. L. Jaremko, and A. R. Hareendranathan, "Semantic autosam: Self-prompting segment anything model for semantic segmentation of medical images," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[7] C. Pohl, K. Hitzler, R. Grimm, A. Zea, U. D. Hanebeck, and T. Asfour, "Affordance-based grasping and manipulation in real world applications," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, USA, October 2020, pp. 9569–9576.

[8] W.-J. Baek, C. Pohl, P. Pelcz, T. Kröger, and T. Asfour, "Improving Humanoid Grasp Success Rate based on Uncertainty-aware Metrics and Sensitivity Optimization," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, Ginowan, Okinawa, Japan, 2022.

[9] F. Krebs and T. Asfour, "A bimanual manipulation taxonomy," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 031–11 038, 2022.

[10] J. Starke, F. Hundhausen, P. Weiner, S. Rader, E. Hyseni, and T. Asfour, "The KIT Robotic Hands – A Scalable Humanoid Hand Platform With Multi-Modal Sensing and In-Hand Embedded Processing," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, accepted.