Grant agreement No: 101017008

Harmony

Assistive robots for healthcare

Enhancing Healthcare with Assistive Robotic Mobile Manipulation

(HARMONY) | H2020-ICT-2018-20| RIA

Start of the project: 01.01.2021
Duration: 42 months

Deliverable Number	D4.3
Deliverable Name	Autonomous Model Acquisition
WP Number	4
Lead Beneficiary	BONN
Dissemination Level	Public
Internal Reviewer	ETH
Due Date	31.12.2023
Date of Submission	22.12.2023
Version	1.0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017008

Revision History

Version	Date	Author(s)	Comments
0.1	10-12-2023	Haofei Kuang	Initial draft
0.2	18-12-2023	Cyrill Stachniss, Haofei Kuang, Jens Behley	Revised version
1.0	22-12-2023	Haofei Kuang, Jens Behley	Incoporated feedback from ETH

Table of Contents

Revision History **2**
Table of Contents **3**
Summary **4**
Acronyms **5**
Introduction **6**
Autonomous Exploration **7**
Object-centric Model Completion **12**
Conclusion **19**
References **20**

Summary

In this deliverable, we describe the autonomous model acquisition system, a component of our mobile robot designed to acquire additional knowledge about the world to enhance its operations. This deliverable includes two components developed by BONN for exploration and model completion using the combination of an occupancy grid map and a visual-language-based 3D semantic map that can be used for object-centric exploration in indoor environments.

In the first part, we deploy a frontier-based exploration algorithm in a completely unknown environment to acquire a map representation. Based on a 2D laser SLAM algorithm, we collect the prior observations of the novel environment and construct a 2D occupancy grid map. Model acquisition is achieved by identifying frontiers between explored and unexplored areas on the map and guiding the robot to explore these areas. Throughout this process, we collect partial data of the scene, which assists in constructing a map that serves as a prior for the subsequent stage of model completion. Moreover, the approach yields a comprehensive 2D occupancy grid map, encompassing the entire indoor scene, which proves invaluable for future navigation tasks. Through testing in both simulated environments and real-world settings, we have demonstrated the effectiveness of our exploration algorithm in autonomously exploring unknown environments and acquiring complete geometric representations of the scenes.

In the second part, we introduce our object-centric model completion method, which navigates the robot to collect additional observations around specific objects of interest to enhance their 3D semantic representations. We construct the 3D map using RGB-D data obtained from the first stage and employ a comprehensive visual-language model for scene understanding, which is zero-shot and not constrained by specific categories of objects. By extracting semantic information from the scene, we identify specific objects and navigate the robot to collect more data around it. The novel observations are then integrated into the 3D map, resulting in a more complete representation of the objects in the environment. Our experiments validate that this method enables the collection of more detailed object data within the maps constructed in the first stage, thus enriching the representation of objects within the map with more completeness.

Through the integration of our exploration and object-centric model completion, we have successfully developed a comprehensive system for autonomously collecting data and constructing detailed metric-semantic maps of hospital environments.

Acronyms

LiDAR	Light Detection and Ranging
ROS	Robot Operating System
SLAM	Simultaneous Localization and Mapping
TSDF	Truncated Signed Distance Function

BONN	Rheinische Friedrich-Wilhelms-Universität Bonn
ETH	Eidgenössische Technische Hochschule Zürich

Introduction

In order to enable the Harmony robot to autonomously and efficiently explore a novel environment that contains multiple objects, it is key to completely explore the novel environment and establish a comprehensive object-centric representation of the environment to support further localization and navigation tasks. The Harmony robot needs to the capable of independently understanding and exploring complex and dynamic environments.

To address the problem, we develop a system that allows robots to autonomously explore new environments and build an object-based understanding of these settings using geometric and semantic perception. This involves tackling two main challenges: first, creating an accurate geometric representation of unfamiliar environments, and second, enriching this representation with information about objects. To achieve these goals, the project is structured into two main stages: autonomous exploration and object-centric model completion.

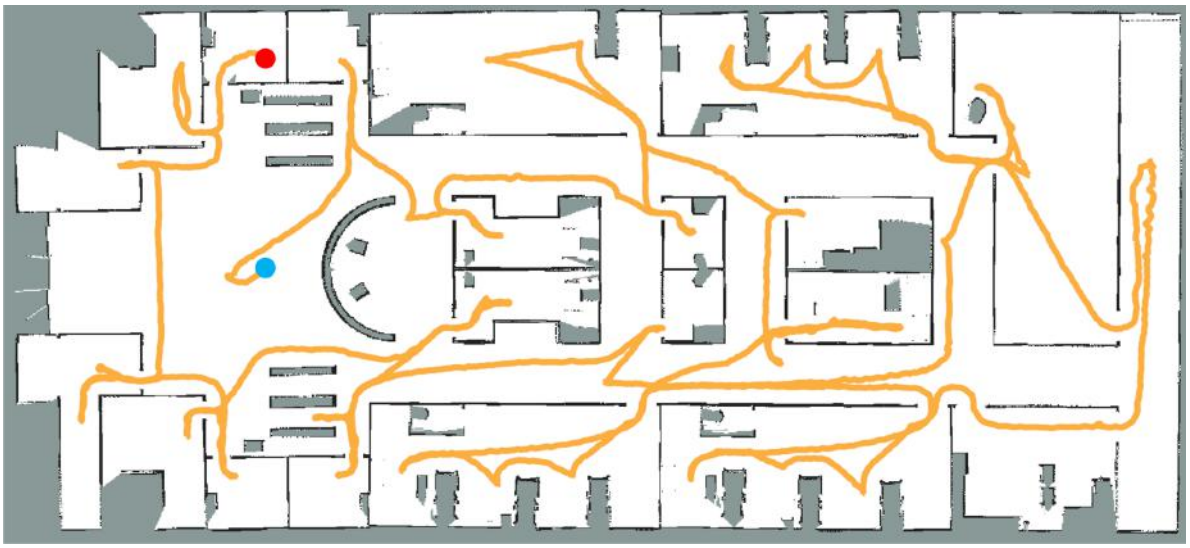
In the autonomous exploration stage, the focus is on deploying the robust and efficient 2D LiDAR SLAM system previously developed by BONN for mapping and exploration. This stage is crucial for enabling Harmony robots to navigate unknown environments safely and efficiently, ensuring thorough coverage and detailed geometric mapping.

Following this, in the object-centric model completion stage, the emphasis changes to enhancing the semantic understanding of these environments. Utilizing advanced visual-language models with RGB-D-based TSDF fusion and open-vocabulary semantic querying, this stage aims to identify and categorize key objects within the scene. This accurate semantic representation enables the robot to understand the distribution of objects in the scene and autonomously collect more observations of key objects. This not only improves the depth of environmental understanding but also enables object-centric interaction and navigation within these spaces.

Both parts of the model acquisition system have been built to utilize the sensor suite described in Deliverable D3.2, i.e., our framework leverages the forward-looking RGB-D sensor, one SICK 2-D LiDAR, and the wheel odometry.

Together, these two stages represent a comprehensive approach to autonomous model acquisition, tackling the geometric and semantic aspects of environmental understanding. This approach is designed to be robust, adaptable, and applicable across a wide range of scenarios, improving the efficiency of autonomous exploration in hospital environments.

(a)



(b)

Figure 1: An example of our autonomous exploration in a simulated hospital environment. (a) The simulation environment in Gazebo. (b) Result of the exploration, the orange line is the trajectory of exploration, and the blue and red points are the start and end points of the trajectory used for exploration, respectively.

Autonomous Exploration

Robots need the ability to operate in new environments. Autonomous exploration is crucial in enabling robots to deploy quickly and accurately in these novel settings, especially in the context of mapping and data collection. Leveraging autonomous exploration and advanced sensors, robots can easily adapt to and navigate unknown environments. While necessary in some scenarios, traditional manual operation is costly and limited in its ability to capture

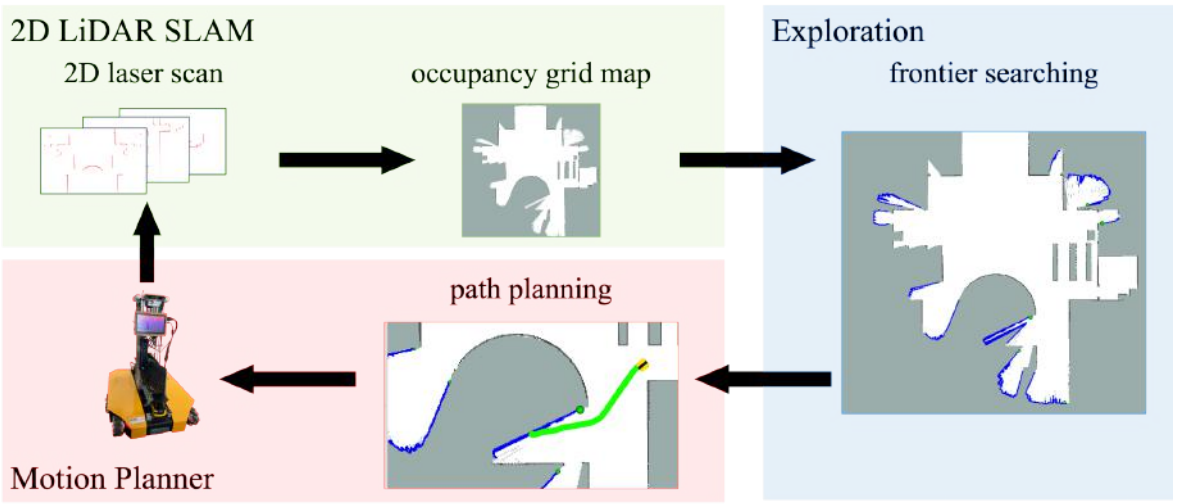


Figure 2: Overview of our autonomous exploration systems. It includes three modules. 2D LiDAR SLAM (shown in green) uses the 2D laser scan as input to estimate the robot pose and builds the occupancy grid map. The exploration module (shown in blue) finds unexplored areas in the scene. The motion planner (shown in red) will plan a path (green trajectory) to navigate the robot to reach the next best exploration area.

every detail of an environment. In contrast, autonomous exploration reduces human operations in potentially complex environments and increases operational efficiency. Using our developments, a robot can comprehensively search and access every reachable area, ensuring a complete and detailed geometric understanding of the environment.

For this reason, our approach enables robots to autonomously navigate and explore every robot-accessible area within a novel environment. Our task is to ensure that the robot acquires a complete understanding of the environment's geometry in its initial exploration stage. This level of understanding is essential for safe and efficient navigation in subsequent missions. To this end, we employ 2D LiDAR SLAM with a frontier-based exploration on our robots. This implementation is crucial in ensuring that the robot can effectively construct a map of the environment and navigate to any specified location with precision and reliability. Figure 1 shows an example of our autonomous exploration. The successful outcome of this exploration is two aspects: a 2D occupancy grid map of the novel environment and the extensive collection of environmental data for building the semantic understanding of the scene.

Figure 2 shows an overview of our autonomous exploration. Our approach to autonomous exploration is achieved through a combination of BONN's 2D LiDAR SLAM [1], frontier-based exploration, and a motion planner. In this task, we deploy our mapping algorithm for 2D SLAM. It creates accurate maps by aligning LiDAR scans using an occupancy grid map representation. It provides accurate pose estimation and maps for autonomous exploration.

(a) (b)

Figure 3: Example of the frontier-based exploration. (a) Frontier searching, which finds all areas between the free area (white) and the unexplored area (gray) in the occupancy grid map, shown by red circles. (b) Next best exploration area (shown by a red circle), which is decided by the size of the frontier and its distance to the robot.

For geometry-aware exploration, we use frontier-based exploration [2], which is an efficient and fundamental algorithm in robotics. It operates by identifying and navigating robots to frontiers, i.e. regions on the edge of the explored and unexplored areas, in a 2D occupancy grid map. The algorithm works by first identifying these frontiers using the map data. Once identified, the robot then plans a path to these frontiers, ensuring that unexplored areas are systematically covered. The process of frontier searching is shown in Figure 3.

In our implementation, we build upon the frontier-based exploration module from the ROS framework. This method is efficient and effective for searching the frontier to explore unknown environments. Our decoupled approach allows for easy adaptation and integration with various robotic systems, ensuring that our robots can reliably and autonomously explore new environments. Navigation commands are sent and executed by the robot's navigation module.

The Navigation Stack [3] uses the A* algorithm as a path planner for finding the shortest path efficiently to a goal. For motion control and obstacle avoidance, the basic control model of the robot is utilized.

In our experiments, we tested our autonomous exploration system in both, simulated and real-world environments. The simulation employed a setup similar to that in Deliverable 4.2, using an omnidirectional Dingo robot. However, we use only a subset of the employed sensors here, i.e., a 2D LiDAR and a forward-facing RGB-D camera given the compute resources on our robot.

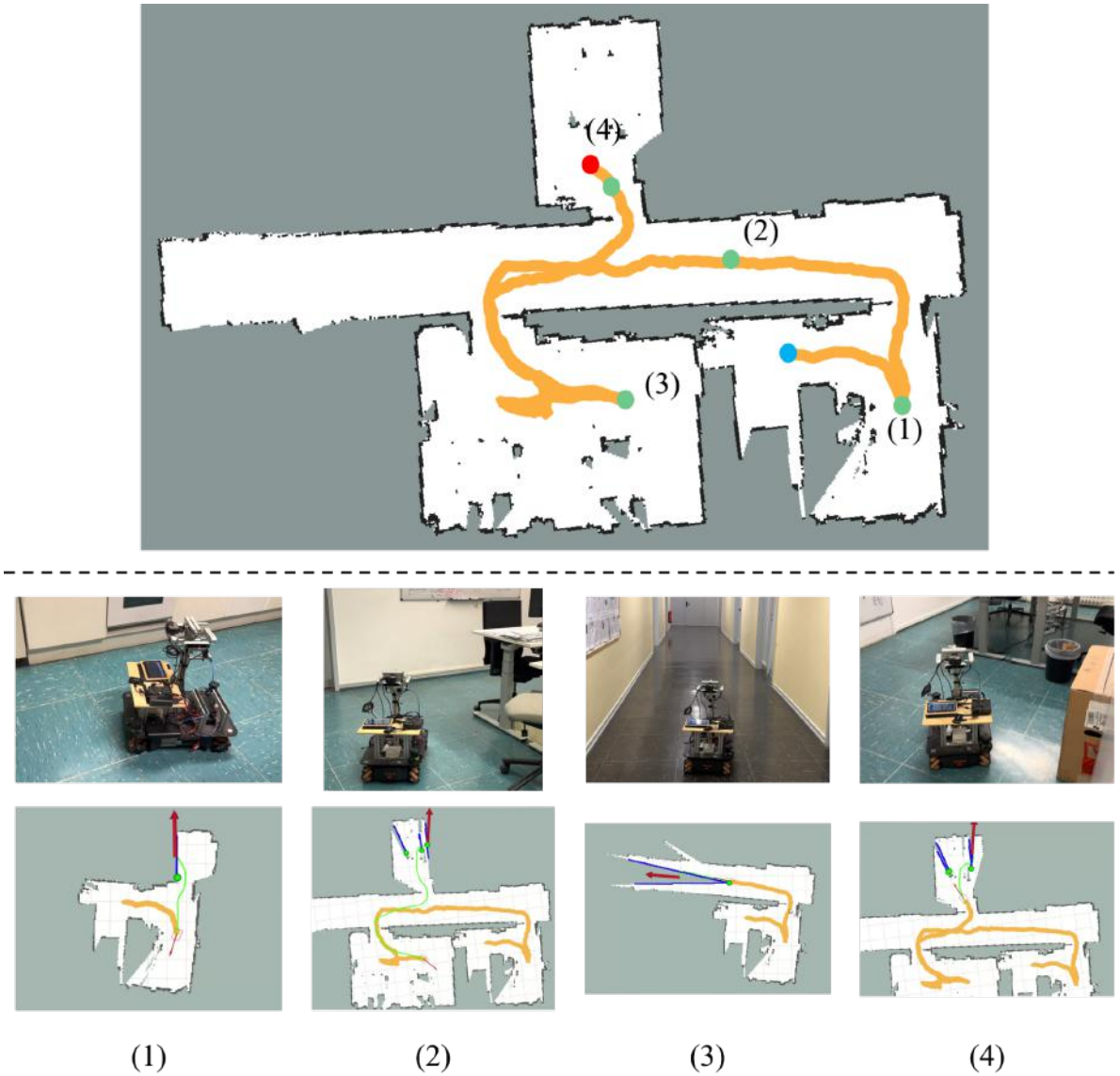


Figure 4: The process of autonomous exploration in our lab in BONN. The upper figure shows the exploration trajectory and the map built during the exploration, the blue and red points are the start and end points, and the green points are positions of the robot during exploration shown below. In the bottom figures, we show snapshots of the exploration of the 4 intermediate positions of the upper figure. The orange line is the past trajectory. The blue line is the frontier, the red arrow indicates the current exploration target, and the green line depicts the global path from the current robot position to the exploration target.

In the real-world scenario, we additionally utilized a YouBot equipped with the same sensors as Deliverable 4.1 to highlight the generalizability of the approach. Figure 4 shows the process of the autonomous exploration of our lab in BONN. The results of our experiments can be found in Figure 5. The results from these experiments demonstrated that our system could efficiently and comprehensively capture the geometric representation of an entirely

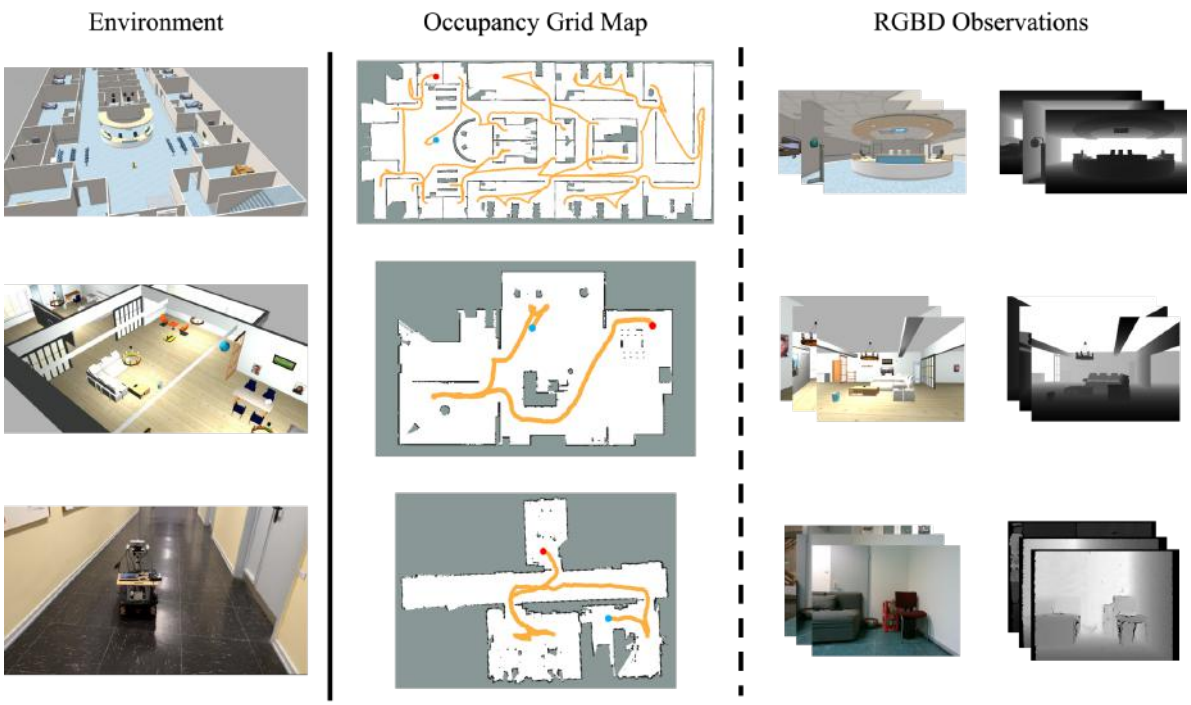


Figure 5: The experimental results (top to bottom) of the autonomous exploration on two simulated environments and one real-world environment recorded in BONN. We can get an occupancy grid map (middle) and a set of RGB-D observations (right) from our autonomous exploration. The orange lines in the middle part are the exploration trajectories.

novel environment, i.e. we can get a complete 2D occupancy grid map for further localization and safe navigation from this stage. Furthermore, the robot successfully collected an initial set of observations for constructing the semantic information of the scene as prior knowledge for the second stage of object-centric model completion.

In summary, our autonomous exploration framework has successfully demonstrated the ability to autonomously navigate and construct a 2D occupancy grid map of novel environments using a combination of 2D LiDAR SLAM and frontier-based exploration. Tested in both simulated and real-world settings, our system efficiently captured comprehensive geometric representations and collected vital data for semantic scene understanding. This foundational work provides the prior knowledge for further developments in autonomous exploration and object-centric model completion.

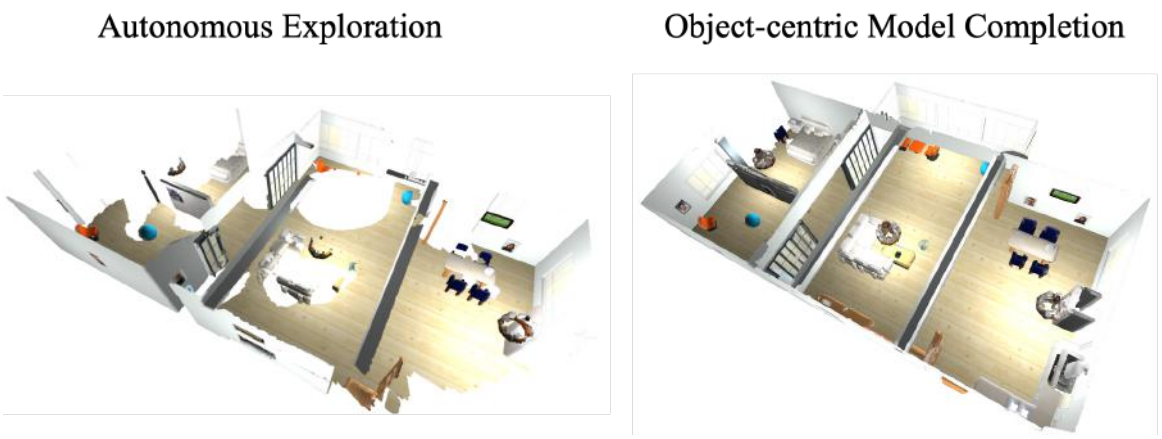


Figure 6: An example showing the output 3D map in the two stages of the autonomous model acquisition system. The object-centric model completion acquires an more complete 3D structure of the scene.

Object-centric Model Completion

In the dynamic and complex healthcare environments where Harmony robots operate, a comprehensive understanding of the semantic information of the novel environment is desirable. This scene understanding with the completed objects is the foundation for tasks such as navigation, human-robot interaction, and manipulation.

However, in the first autonomous exploration stage, the robot primarily focuses on acquiring a complete geometric representation of the scene for safe and efficient navigation, which often leads to an incomplete semantic understanding of the environment, i.e. it lacks enough observation of the object in the scene for building a complete metric-semantic representation of the novel environment. Addressing this gap is our motivation for developing an advanced object-centric model completion system, aiming to enrich the initial data with deeper semantic insights. Figure 6 shows an example of scene completion after deploying our object-centric model completion.

Building on the initial stage of autonomous exploration, our objective in the next stage is to enhance the semantic understanding of the novel environment through a second round of exploration, known as object-centric model completion. Unlike the first stage, which emphasizes the geometric representation of the scene, this stage concentrates on key objects within the scene. The Harmony robot which is equipped with sensors such as RGB-D cameras, will collect additional observations focused on key objects. The target is to utilize this newly collected data to update the map of the scene, thereby enriching its semantic representation and creating a more comprehensive understanding of the environment. This targeted exploration is crucial for developing a detailed and complete metric-semantic representation of the novel environment.

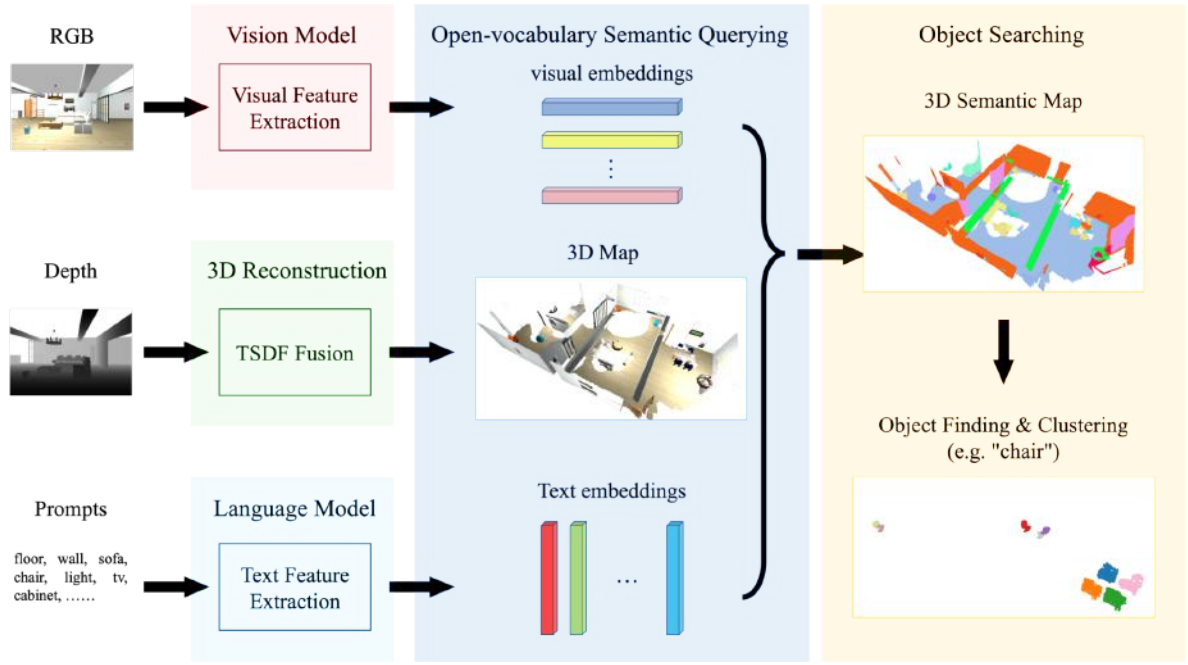


Figure 7: Overview of the object-centric scene understanding. The collected RGB-D data in the first stage will be used to reconstruct the 3D structure and extract visual features. The scene understanding is achieved by querying the visual feature with a set of prompts. During the model completion stage, a specific object classes will be extracted from the 3D semantic map and clustered into individual instances.

Defining the problem for object-centric model completion, we focus on two key tasks for the Harmony robot. Firstly, the robot must utilize data collected from the initial autonomous exploration stage to construct a preliminary semantic understanding of the novel scene described in the previous section. This involves identifying the categories of objects likely present in the current environment. Secondly, the robot needs to localize and navigate to specific key objects to collect more detailed observations about these objects. These tasks are crucial for developing a deeper, more complete semantic representation of the environment, building upon the foundational geometric information obtained in the first exploration stage. Figure 7 shows the overview of our object-centric scene understanding algorithm and the different stages to determine task-relevant objects in the map.

First of all, our method for object-centric model completion involves a key process: 3D reconstruction using TSDF fusion with the collected RGB-D data from the first autonomous exploration stage, which accumulates information from multiple RGB-D frames to produce a cohesive 3D representation. We build upon Open3D's voxel hashing structure to efficiently represent the 3D scene. This allows us to reconstruct a detailed 3D metric map of the environment, which is essential for identifying specific key objects in the subsequent stage.

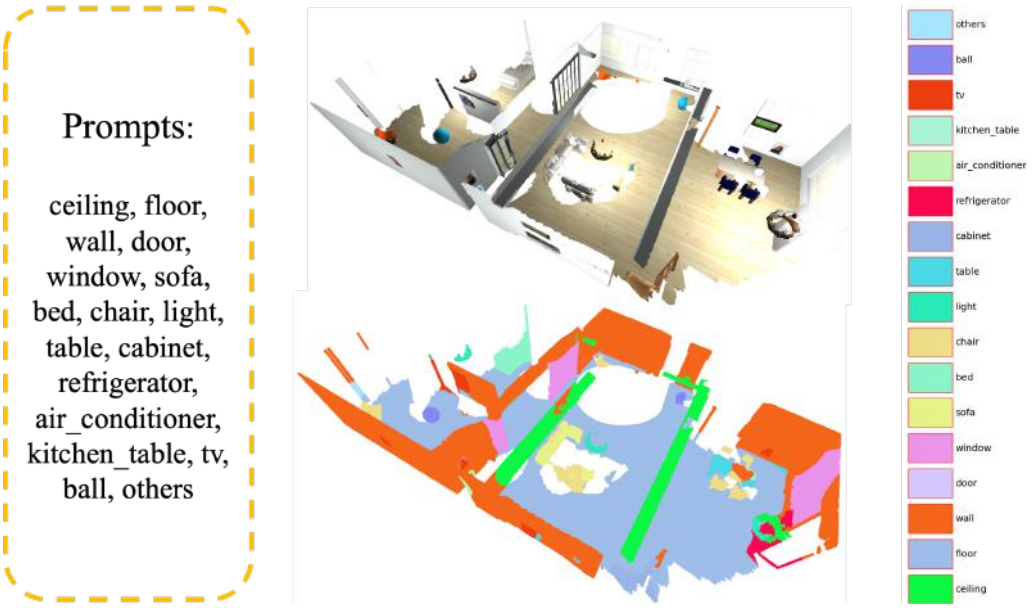
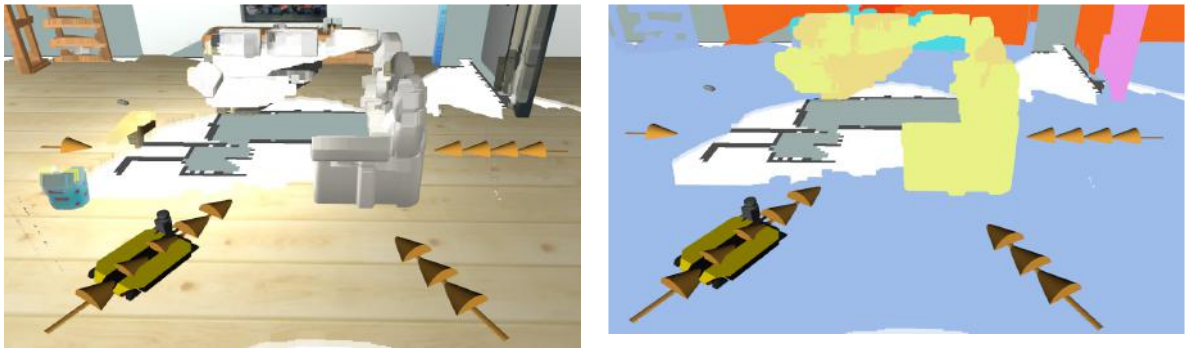


Figure 8: The 3D semantic map building using the open-vocabulary scene understanding.

Regarding the scene understanding model, we utilize the large-scale visual-language model to achieve efficiency and generalization. Specifically, we use a pre-trained SEEM [4] model as the foundation for scene understanding. SEEM learns a joint visual-semantic space in which visual prompts are naturally aligned with textual prompts, as the model continuously learns a common visual-semantic space. It uses a text encoder to encode text queries and mask labels into the same semantic space for open-vocabulary segmentation. This setup enables SEEM to effectively generalize to novel prompts or combinations thereof, exhibiting a remarkable capacity for handling various types of segmentation tasks without supervision.

In our implementation of the metric-semantic map construction with the visual-language model to extract and fuse the semantic information of objects to the voxel representation. For each RGB-D frame, the visual feature will be extracted by SEEM from the RGB image, and fused to the corresponding voxel according to the projection by the depth image. We simply average all visual features in the same voxel as the final visual feature. After that, the object categories will be identified by open-vocabulary semantic querying. We use natural language descriptions to categorize objects within the scene. Each category of object is then associated with a unique textual feature which is extracted using the text encoder of SEEM. By matching these visual features with the extracted textual features, we can accurately determine the semantic category of each voxel. This method of open-vocabulary semantic querying is crucial for ensuring that our metric-semantic map is not only geometrically accurate but also rich in semantic information. By leveraging the flexibility of natural language, this approach is highly adaptable and can be generalized to different scenes by varying the natural language descriptions used for categorization. Figure 8 shows an example of extracting semantic information of the scene by querying the map with a set of prompts.



(a) (b)

Figure 9: An example of the robot navigation to collect additional observations of the sofa. The orange arrows depict the target view of the object to complete the object model.

Exploiting the information of the metric-semantic map, we will search for key objects within the scene. In detail, we filter the pointcloud with specific categories and cluster the pointcloud into different instances. Once we have identified and localized these object instances, we engage in object-centric exploration. For each identified object instance, we compute a bounding box that covers the entire object. Then, we plan a coverage path along the edges of this bounding box. The goal of this coverage path is to comprehensively perceive the entire object, allowing the Harmony robot to collect detailed data about the target. Figure 9 shows an example of the coverage views around objects. The 2D occupancy grid map which is obtained from the first stage of exploration ensures that the path is not only optimal for data collection but also navigable and safe for the robot. Through this meticulous process, we ensure that our object-centric exploration is both thorough and efficient, capturing the essential details needed for a complete understanding of each object in the novel environment.

In our experiments, we maintained the same settings as used in the autonomous exploration stage, and tested our algorithm in both simulated and real-world environments. The results from these experiments demonstrate the effectiveness of our scene understanding results based on the visual-language model. Furthermore, the entire process is zero-shot and open-vocabulary, meaning that the model is not limited to a specific domain of the scene. This versatility allows it to be easily generalized to any new scene. By utilizing the collected target data to update the map, we were able to achieve a more complete and consistent 3D metric-semantic map which is enriched with accurate semantic information. These results indicate that our approach not only enhances the geometric details obtained from the first exploration stage but also significantly improves the overall semantic understanding of the environment. This improvement in semantic mapping is crucial for various applications where detailed environmental understanding is paramount. Figure 10 shows an example of the object-centric model completion with the prompt “bookshelf” at our lab in BONN. Figure 11 shows the results of our autonomous model acquisition system in at our lab in BONN, and Figure 12 shows the results of our system deployed in different simulated environments.

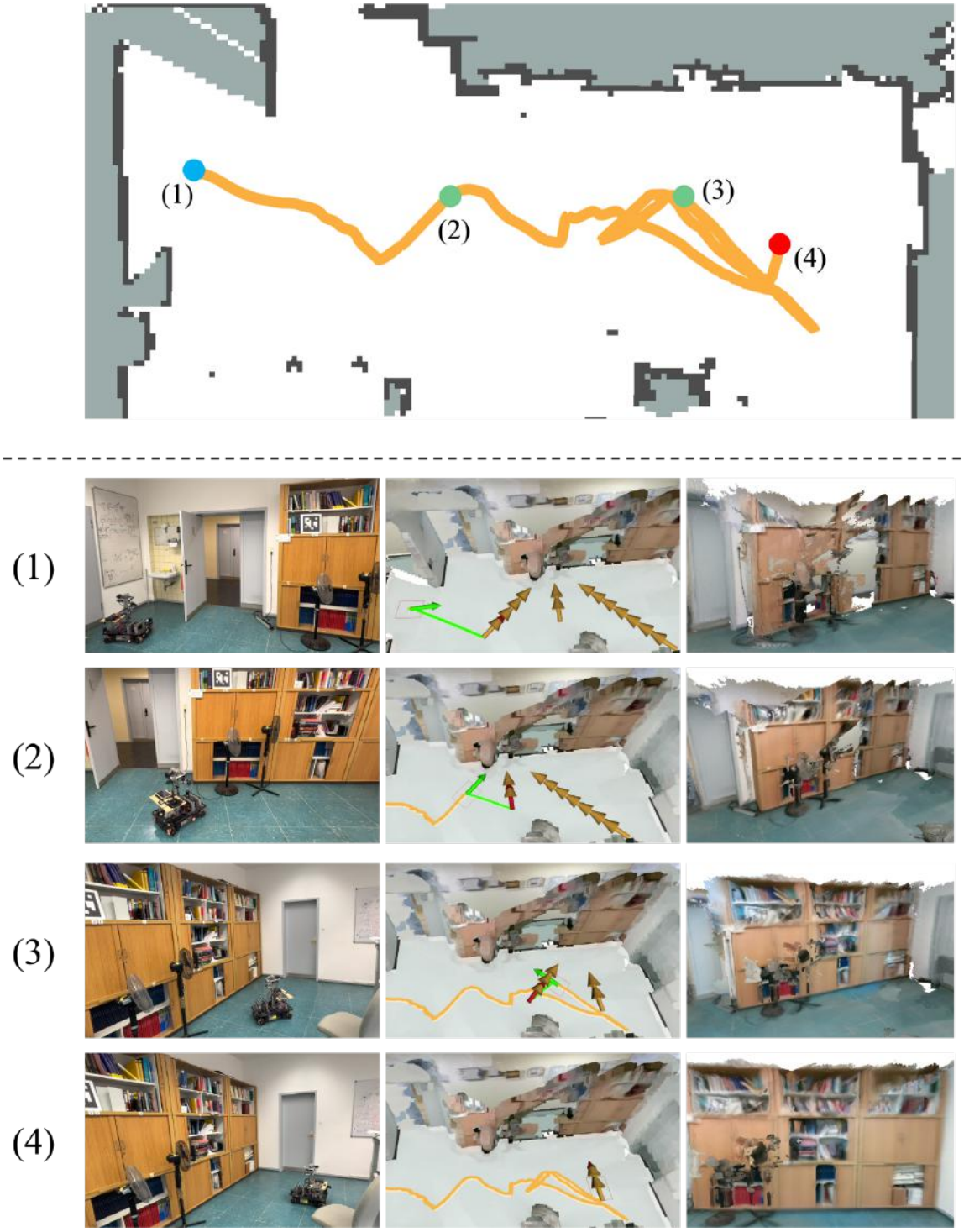


Figure 10: The process of object-centric model completion in our lab. The robot tries to collect more observations of the bookshelf. In the upper figure, the blue and red points are the start and end points, and the green points are intermediate positions. In the bottom figure, we show the intermediate stages of the model completion. In the middle images, the orange line is the past trajectory and the orange arrows are the rest of the views that need to be collected. The green arrow is the current robot pose and the green line is the global path from the current robot position to the next view. The images on the right show the progress of the model completion.

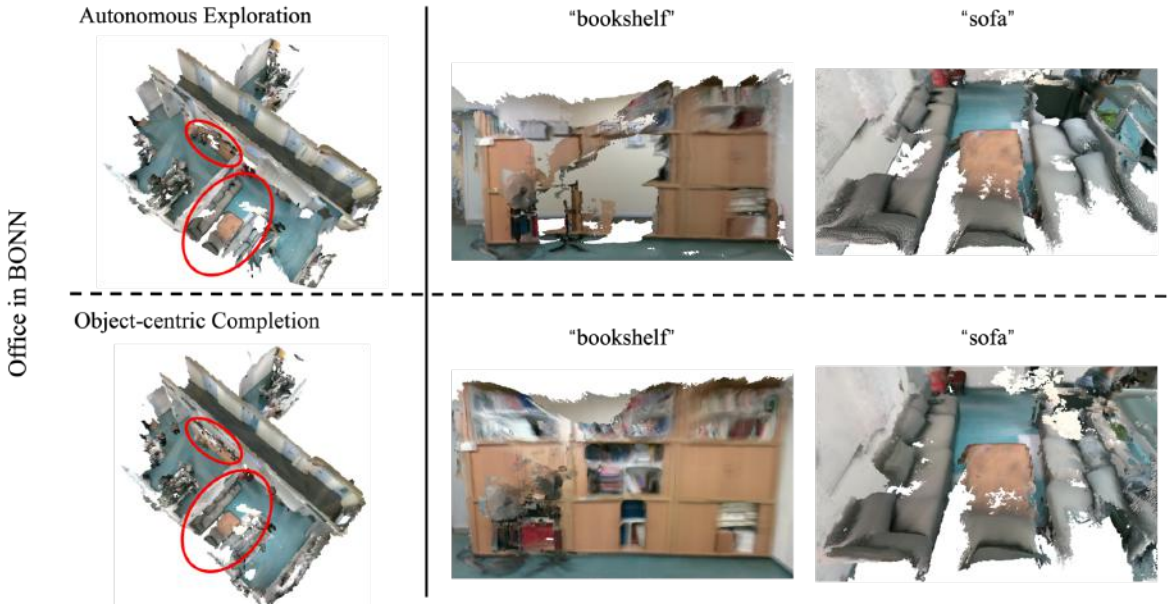


Figure 11: The experimental results on a real-world office scene in BONN. The red circle highlights the locations of the example objects (shown on the right) in the map. The upper part shows the model before the model completion right after the exploration. The bottom part of the images show the completed model after the object-centric completion.

We can observe that more observations from the model completion stage can supplement the limited view points acquired in the exploration stage and generate a more complete 3D map. Besides, more details of specific objects are represented in the final 3D map.

In summary, our object-centric model completion effectively enriches the semantic understanding of novel environments covering the scene and its relevant objects and enhances the 3D metric-semantic maps with detailed and accurate information. This improvement supports more robust navigation or localization systems for Harmony robots, aligning well with the map construction tasks in Deliverables 4.1 and 4.2, and the WP5. Moreover, the model's zero-shot, open-vocabulary capabilities ensure easy adaptability to new scenes, showcasing its versatility for various applications.

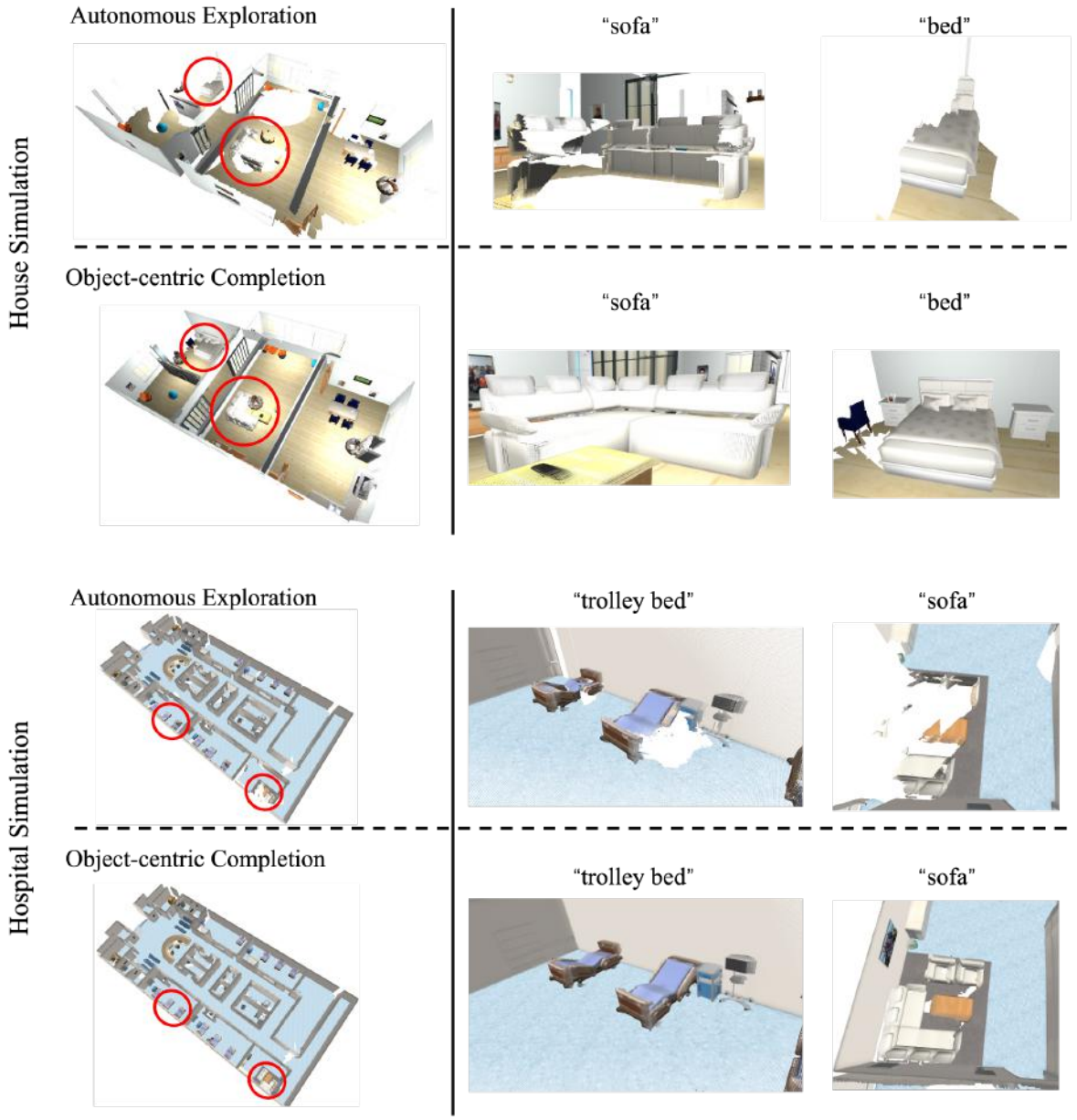


Figure 12: Experimental results on simulated house and hospital environments. The red circle highlightes the locations of the example objects on the map shown on the right. As s

Conclusion

In conclusion, our work in autonomous exploration and object-centric model completion has enhanced the capabilities of model acquisition for Harmony robots in complex indoor environments. By integrating advanced techniques like 2D LiDAR SLAM, frontier-based exploration, and the use of large-scale visual-language models, we have successfully addressed the challenges of mapping and understanding novel environments. The project's robustness is highlighted by extensive evaluations in both simulated and real-world settings, proving the system's effectiveness in comprehensive environment mapping and semantic understanding. This progress aligns with our previous Deliverables 4.1 and 4.2 and yields more complete data for mapping and localization. It also provides a completed semantic map representation for navigation.

References

[1] G. Grisetti, C. Stachniss, and W. Burgard. *Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters.* IEEE Transactions on Robotics, vol. 23(1), p. 34-46, 2007.

[2] B. Yamauchi. *Frontier-based exploration using multiple robots.* Proc. of the Intl. Conf. on Autonomous agents. 1998.

[3] E. Marder-Eppstein, E. Berger, T. Foote, B. Gerkey, and K. Konolige. *The office marathon: Robust navigation in an indoor office environment*. Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA), 2010.

[4] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Gao, and Y. J. Lee. *Segment everything everywhere all at once*. Proc. of the Conf. on Neural Information Processing Systems (NeurIPS) , 2023.