MASTER THESIS

Thesis submitted in fulfillment of the requirements for the degree of Master of Science in Engineering at the University of Applied Sciences Technikum Wien - Degree Program Mechatronics/Robotics

SAGE: Multi object semantic aware guided exploration with persistent memory

By: Kevin Eppacher, BSc

Student Number: 2310331013

Supervisor: Simon Schwaiger, MSc

Vienna, July 16, 2025

This work was conducted in the context of the project “Stadt Wien Kompetenzteam für Drohnentechnik in der Fachhochschulausbildung” (project number MA23 35-02, financed by the Department MA23 for Economic Affairs, Labour and Statistics of the City of Vienna).

Funded by

City of Vienna | Economic Affairs, Labour and Statistics

Declaration

“As author and creator of this work to hand, I confirm with my signature knowledge of the relevant copyright regulations governed by higher education acts (see Urheberrechtsgesetz /Austrian copyright law as amended as well as the Statute on Studies Act Provisions / Examination Regulations of the UAS Technikum Wien as amended).

I hereby declare that I completed the present work independently and that any ideas, whether written by others or by myself, have been fully sourced and referenced. I am aware of any consequences I may face on the part of the degree program director if there should be evidence of missing autonomy and independence or evidence of any intent to fraudulently achieve a pass mark for this work (see Statute on Studies Act Provisions / Examination Regulations of the UAS Technikum Wien as amended).

I further declare that up to this date I have not published the work to hand nor have I presented it to another examination board in the same or similar form. I affirm that the version submitted matches the version in the upload tool.“

Vienna, July 16, 2025 Signature

Kurzfassung

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Schlagworte: Keyword1, Keyword2, Keyword3, Keyword4

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Keywords: Keyword1, Keyword2, Keyword3, Keyword4

Acknowledgements

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

1 Introduction

1.1 Problem Statement

1. Motivation & Relevance
2. Domain and Motivation: e.g. Search and Rescue (SAR), semantic Navigation, exploration in unknown environments, etc.
3. Which solutions exist? Comparison of general properties and performance metrics:
   a) One Map to Find Them All (OneMap) [1]
   b) Visual Language Maps for Robot Navigation (VLMaps) [2]
   c) Vision-Language Frontier Maps for Zero-Shot Semantic Navigation (VLFM) [3]
   d) Open-Vocabulary 3D Scene Graphs for Perception and Planning (ConceptGraphs) [4]
   e) Object Goal Navigation using Goal-Oriented Semantic Exploration (SemExp) [5]
   f) Learning Generalizable Feature Fields for Mobile Manipulation (GeFF) [6]

Methods	Persistent Memory	Real-Time Cabability	Representation (2D/3D)	Open-Vocab Cabability	Explicit Frontiers
VLFM	x	✓	2D	✓	✓
VLMaps	✓	X (offline)	2.5D	✓	x
OneMap	✓	2 Hz (drops heavily proportional to map size)	2.5D	✓	✓
ConceptGraphs	✓	X (offline)	3D	✓	x
SemExp	✓	✓	2D	x	✓
GeFF	✓	✓	3D	✓	x

Figure 1: Comparison of state-of-the-art methods regarding common properties.

4. What is the technical problem? A combination of the following:
   a) No persistent memory
   b) Not Real-Time
   c) 3D-Representation
   d) Sensibility to false positives for zero shot object detection
5. Scientific Contribution:

Methode	Success Rate (SR)	Success Rate (Multi-Obj.)	SPL (Pfadlänge)	Semantic Precision	Robustheit kleine Objekte
VLFM	~81%	~60%	~60%	mittel	niedrig
VLMaps	~63%	~46%	~41%	mittel (~65%)	mittel
OneMap	89%	70%	70%	hoch (~85%)	mittel (~65-70%)
ConceptGraphs	hoch (geschätzt ~85-90%)	mittel (~65-70%, geschätzt)	mittel (~60%)	hoch	hoch
SemExp	64%	32%	35%	mittel	niedrig
GeFF	hoch (~85%, geschätzt)	mittel (~70%, geschätzt)	mittel/hoch (~65%)	hoch	mittel/hoch

Figure 2: Comparison of state-of-the-art methods regarding performance metrics.

- Development of a hybrid semantic exploration framework that combines VLFM-based [3] frontier scoring with Open-Fusion's [7] global 3D scene representation, enabling multi-object search with open-vocabulary queries (text~~, image, audio~~) during autonomous exploration. The approach is evaluated for improvements in overall and multi-object success rate (Success Rate (SR)) as well as path efficiency measured by success weighted by path length (Success weighted by Path Length (SPL)).
- Implementation of a lightweight VLFM [3] component leveraging Segment Everything Everywhere All at Once (SEEM) [8] as a unified vision-language model, substantially reducing GPU memory requirements compared to traditional VLFM [3] pipelines using multiple separate models.
- Design of a sensor-based fusion strategy that integrates semantic detections from SEEM [8] with clustered relevance fields from Open-Fusion [7], applying spatial confidence weighting to enhance robustness against false positives in zero-shot object detection.
- Experimental validation of the proposed system on a real mobile robot, focusing on practical aspects such as real-time performance and robustness to sensor noise including depth inaccuracies and changing lighting conditions.

While existing approaches either rely on short-term detection or isolated 3D mapping, this thesis introduces a hybrid architecture that tightly integrates real-time semantic frontier scoring with persistent 3D relevance fields.

1.2 Thesis Structure

2 State of the Art

In this chapter, the current state of research in semantic multi-object search, map reconstruction, and object detection is reviewed. The goal is to identify strengths and limitations of existing methods and establish the technological context for the proposed hybrid approach. The chapter is divided into three key areas: approaches for searching multiple objects semantically, techniques for building and maintaining persistent semantic maps, and recent advances in object detection and promptable models for open-vocabulary tasks.

2.1 Semantic Multi-Object Search Approaches

- Review of methods targeting simultaneous or sequential search for multiple objects in unknown environments.
- Analysis of VLFM, VLMaps, OneMap, GeFF,... regarding:
  - Success Rate (SR) and Success weighted by Path Length (SPL) as key metrics.
  - Real-time capabilities and computational requirements.
  - Handling of open-vocabulary queries.
- Discussion of semantic exploration frameworks combining language models with spatial reasoning.
- Challenges of maintaining semantic context across multiple targets.

2.2 Map Reconstruction and Persistent Semantic Mapping

- Overview of approaches to build and update semantic maps during exploration.
- Techniques for fusing sensor data into persistent 2D/3D representations.
- Comparison of representations (Octomaps, point clouds, voxel grids) in terms of:
  - Memory efficiency.
  - Ability to store semantic labels persistently.
- Discussion of ...
  - ConceptGraphs
  - ConceptGraph-Online
  - OpenFusion
  - Clio
  - OpenScene
  - GeFF

- CLIP-Fields
  - ConceptFusion
  - VLMaps
  - LERF

  as examples of global 3D semantic maps.

- Limitations in updating or correcting the map after wrong detections.

2.3 Object Detection and Promptable Models

- Review of traditional and open-vocabulary object detection methods.
- Analysis of grounding-capable detectors and segmentation models for zero-shot tasks.
- Specific evaluation of the following models for their suitability in semantic multi-object search:
  - YOLOv7
  - GroundingDINO
  - MobileSAM
  - GroundedSAM
  - SEEM
  - OWL-ViT
  - MaskDINO
- Discussion of promptable vision-language models supporting multi-modal queries (text, image, audio).
- Challenges with false positives in zero-shot settings and their implications for reliable multi-object detection.

3 Methods

This chapter details the methods developed for semantic exploration, persistent 3D mapping, promptable object detection, and robust fusion strategies for multi-object search.

3.1 System Overview

- Presentation of the overall architecture of the exploration, detection, mapping, and fusion pipeline.
- Description of data flow between exploration (frontier evaluation), detection (promptable models), and exploitation (persistent semantic mapping).
- Explanation of how exploration and mapping components interact to progressively build a semantic understanding of the environment.

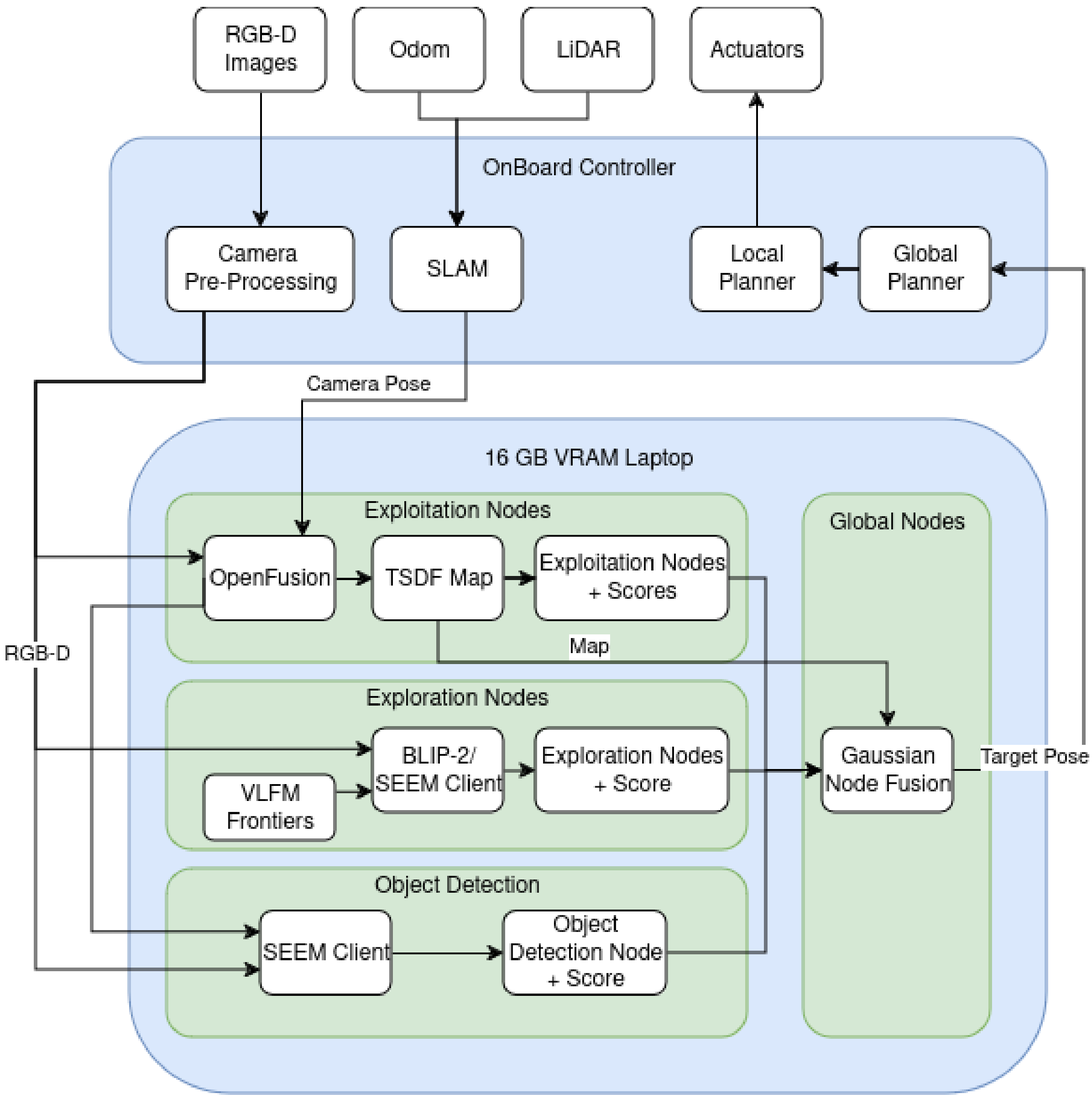

Figure 3: System architecture

3.2 Semantic Frontier Mapping

3.2.1 Frontier Detection and Calculation

- Detection of frontiers on a 2D occupancy grid to identify candidate regions for exploration.
- Application of classical frontier-based exploration algorithms extended with semantic information.

3.2.2 Value Map Generation using Vision-Language Models

- Computation of value maps by evaluating cosine similarity between multi-modal queries (text, image, audio) and scene observations.
- Use of promptable vision-language models (e.g., SEEM) to assign semantic relevance scores to each region.
- Dynamic update of value maps as new observations are integrated.

3.2.3 Navigation to High-Value Frontiers

- Selection of the frontier with the highest semantic relevance score.
- Continuous re-evaluation of frontiers during exploration to adapt to changing scene semantics.
- Strategy for balancing exploration efficiency and semantic search goals.

3.3 Persistent 3D Semantic Mapping

3.3.1 Global Map Construction with Open-Fusion

- Incremental creation of a global semantic point cloud map integrating RGB-D observations over time.
- Registration of observations using robot poses to maintain a consistent world representation.
- Association of semantic labels with 3D points based on query relevance scores.

3.3.2 Semantic Clustering and Graph Node Generation

- Clustering of points with similar semantic labels to form object-level hypotheses.
- Construction of semantic graph nodes representing detected object instances with aggregated confidence scores.
- Maintenance of the semantic graph as a persistent memory for multi-object search tasks.

3.4 Promptable Zero-Shot Detection

- tbd

3.5 Fusion Strategy

- Formulation of a fusion strategy combining frontier-based semantic relevance with persistent 3D semantic mapping.
- Design of decision-making algorithms leveraging combined semantic information to robustly answer multi-object search queries.

4 Implementation

This section details the practical implementation of the proposed approach, covering the simulation and real-world setup, datasets, software stack, and hardware configuration.

4.1 Simulation Environment

- Evaluation of simulation frameworks for indoor semantic navigation:
  - HabitatSim: Realistic Matterport3D-based environments with semantic annotations.
  - Isaac Sim / Isaac Lab: GPU-accelerated simulation, advanced physics, support for RTX ray tracing.
  - MuJoCo: High-speed physics engine, limited support for complex indoor scenes.
  - Ignition Gazebo: Modular simulator, ROS2 integration, good for real-robot transfer.
- ...

4.2 Dataset

- Use of **Matterport3D** scenes for realistic indoor environments with ground truth 3D reconstruction and semantic annotations.
- Incorporation of the **Habitat Navigation Challenge 2023** tasks to benchmark exploration and navigation performance (SR, SPL).

4.3 Used Software

- ROS2-based implementation (Humble Hawksbill) as middleware.
- Navigation stack: Navigation2 (Nav2) for frontier-based exploration and path planning.
- DDS communication layer for distributed communication between detection, mapping, and control nodes.
- Custom RobotDriver for interfacing with real robot hardware.
- Integration of promptable models (SEEM, GroundingDINO, etc.) for real-time zero-shot detection during exploration and exploitation.
- ...

4.4 Used Hardware

- **PC:**
  - CPU: AMD Ryzen 9 5950X 16-Core Processor
  - Motherboard: B550 Gaming X V2
  - GPU: ASUS TUF Gaming RTX 4090 24GB OC Edition
  - RAM: 64GB Corsair Vengeance LPX DDR4
- **Real Robot:** Configuration and components to be determined (tbd).

4.5 Evaluation Metrics

This section defines the evaluation metrics used throughout the experiments and assigns them to each corresponding experiment.

4.5.1 Experiment 1: Success Rate (SR)

- Metric: Ratio of successfully reached single goal objects.
- Evaluation against: VLFM, VLMaps, OneMap, GeFF

4.5.2 Experiment 2: Path Efficiency (SPL)

- Metric: Success weighted by path length (SPL).
- Used to assess navigational efficiency of successful trajectories.

4.5.3 Experiment 3: Multi-Object Success Rate (MSR)

- Metric: Percentage of tasks where all queried objects are found.
- Important for evaluating real-world multi-object search performance.

4.5.4 Experiment 4: Ablation – Exploitation Component (OpenFusion)

- Metric: MSR change with vs. without OpenFusion (i.e. only VLFM).
- Purpose: Understand OpenFusion's contribution in the hybrid setup.

4.5.5 Experiment 5: Robustness to False Positives (Fusion Strategy)

- Metric: **Semantic Precision** and **False Positive Rate**.
- Description: Evaluates how fusion between SEEM and OpenFusion mitigates false detections under occlusion and ambiguous inputs.

4.5.6 Experiment 6: Real-World System Performance

- Metrics:
  - **SR, MSR, SPL**: for search performance under real-world conditions.
  - **System Metrics:** CPU/GPU usage, FPS, inference latency.
- Objective: Assess robustness, efficiency, and deployability in physical environments.

4.5.7 Experiment 7: Vision-Language Model Comparison (SEEM vs. BLIP2)

- Metrics: Cosine Similarity Consistency, SR, SPL, Memory Footprint.
- Purpose: Compare SEEM and BLIP2 in the context of frontier scoring and value map generation.

5 Discussion and Results

This chapter presents the experimental evaluation of the proposed hybrid semantic exploration system. Each experiment targets a specific research question and is evaluated using quantitative performance metrics.

5.1 Results on Semantic Multi-Object Search Tasks

5.1.1 Experiment 1: Single-Object Success Rate (SR)

5.1.2 Experiment 2: Navigation Efficiency (SPL)

5.1.3 Experiment 3: Multi-Object Success Rate (MSR)

5.1.4 Experiment 4: Ablation of Exploitation (OpenFusion)

5.2 Experiment 5: Improving Detection Robustness via Semantic Fusion

5.3 Experiment 6: Real-World Deployment

5.4 Experiment 7: Comparison of VL-Models for Frontier Scoring

The goal is to assess whether SEEM, as a lightweight unified model, can maintain comparable exploration performance to BLIP2 while reducing computational overhead.

6 Summary and Outlook

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Bibliography

[1] F. L. Busch, T. Homberger, J. Ortega-Peimbert, Q. Yang, and O. Andersson, One map to find them all: Real-time open-vocabulary mapping for zero-shot multi-object navigation, 2025. arXiv: 2409.11764 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2409.11764.

[2] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 2023.

[3] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in International Conference on Robotics and Automation (ICRA), 2024.

[4] Q. Gu et al., Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning, 2023. arXiv: 2309.16650 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2309.16650.

[5] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in In Neural Information Processing Systems, 2020.

[6] R.-Z. Qiu et al., "Learning generalizable feature fields for mobile manipulation," arXiv preprint arXiv:2403.07563, 2024.

[7] K. Yamazaki et al., "Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation," arXiv preprint arXiv:2310.03923, 2023.

[8] X. Zou et al., Segment everything everywhere all at once, 2023. arXiv: 2304.06718 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2304.06718.

List of Figures

List of Tables

List of source codes

A Appendix A

B Appendix B