CrossMaps: Confidence-Aware Open-Vocabulary Semantic Mapping for Rover Navigation

Jan-Niklas Klein, Sona Ghahremani*, Christian Medeiros Adriano* and Holger Giese

Hasso Plattner Institute for Digital Engineering, Potsdam, Germany.

Abstract

Rovers rely on perception to maintain spatial maps that encode both objects and sensor quality (e.g., range reliability, lighting artifacts, data density), guiding data fusion, embedding updates, and navigation under partial observability. To study these coupled perception–navigation processes, we present CrossMaps, a real-time confidence-aware open-vocabulary semantic mapping pipeline that constructs language-queryable maps from RGB-D data. Building on VLMaps-style approaches, CrossMaps integrates multi-scale CLIP embeddings with confidence-aware fusion and a dual-memory architecture consisting of Short-Term Memory (STM) and Long-Term Memory (LTM). The STM aggregates noisy visual observations using geometric, semantic, and temporal confidence cues, while confident and coherent cells are promoted to the LTM as persistent semantic landmarks. Designed for deployment with a Jetson Orin-powered UGV alongside SLAM, CrossMaps runs in real time and produces semantic heatmaps that can be queried with natural language to guide rover navigation.

Keywords

Semantic mapping, Open-vocabulary perception, Dual-memory maps, SLAM, Visual–language maps

1. Introduction

Context and Limitations – Autonomous rovers in unstructured environments must continuously build representations of their surroundings to navigate safely and achieve mission goals, a task that is especially challenging for planetary rovers due to partial observability, sensor noise, and dynamic conditions. Traditional navigation systems rely mainly on SLAM-based geometric maps [3, 4], which capture spatial structure but lack semantic understanding of the environment. To address this semantic gap, recent advances in VLMs (Vision Language Models) have enabled open-vocabulary semantic mapping, allowing rovers to associate spatial locations with natural language concepts using methods such as VLMaps [7], which augment geometric maps with CLIP-based embeddings [5] for text-queryable navigation beyond predefined object classes. However, deploying such systems on mobile rovers remains challenging: observations are noisy and viewpoint-dependent, embeddings vary across viewpoints and lighting, objects may appear or disappear, and the map must remain stable over long horizons while still adapting to new evidence, requiring mechanisms that carefully balance plasticity with robustness.

Research Problems – We investigate the transfer of perception knowledge along three main axes:

RP.1 - *How to capture and represent perception knowledge?* We focus on perception knowledge with both qualitative (e.g., semantic information) and quantitative (e.g., distance) properties.

RP.2 - *How to pre-process, clean, and rank perceptual hypotheses, given their spatial and temporal information as well as their inherent uncertainty?* The primary source of perception knowledge is sensor data, which can be noisy, incomplete, and biased due to imperfections that may depend on the rover's actions (e.g., viewpoint, occlusions) or not (e.g., lighting, weather).

RP.3 - *How can we maintain, merge, and share perception knowledge across multiple agents with different downstream tasks while preserving semantic consistency?* We study update and fusion rules that keep shared perception maps semantically consistent while exposing task-relevant subsets for agents with different intentions (e.g., exploration, inspection).

ICRA'26: ROSE International Workshop on Robotics Software Engineering, June 01, 2026, Vienna, Austria

*Corresponding author.

jan-niklas.klein@student.hpi.uni-potsdam.de (J. Klein); sona.ghahremani@hpi.de (S. Ghahremani); christian.adriano@hpi.de (C. M. Adriano); holger.giese@hpi.de (H. Giese)

0000-0003-0697-9195 (S. Ghahremani); 0000-0003-2588-9937 (C. M. Adriano); 0000-0002-4723-730X (H. Giese)

Approach – We introduce *CrossMaps*, a confidence-aware open-vocabulary semantic mapping pipeline designed for mobile rover navigation. CrossMaps extends the VLMaps paradigm by introducing a dual-memory representation and confidence-aware data fusion mechanisms that improve robustness during online mapping. The pipeline maintains a *Short-Term Memory (STM)* that aggregates recent observations while filtering noise through temporal decay and consistency checks, and a *Long-Term Memory (LTM)* that stores reliable semantic landmarks derived from confident observations.
Contributions – (i) a confidence-aware open-vocabulary semantic mapping framework that assigns point- and cell-level confidence to CLIP-based observations using geometric, semantic, and temporal cues during map fusion, (ii) a dual-memory semantic map representation that aggregates observations into cell-wise embeddings, confidence, and coherence in a short-term grid map, and promotes only sufficiently confident, coherent, and multi-view–supported cells into a persistent long-term landmark map, (iii) a real-time semantic mapping pipeline that fuses multi-scale, tile-level CLIP embeddings with SLAM-based spatial alignment to build language-queryable 3D maps under rover motion, and finally, (iv) a confidence- and coherence-based promotion and querying mechanism that uses coherence (intra-cell agreement in embedding space) to shape heatmaps and cell selection, while using cell-wise confidence to gate long-term promotion and top-1 navigation targets.

2. Related Work

Research on rover navigation mapping has progressed from geometric SLAM toward semantically enriched, language-aware representations. Classical feature-based and graph-based SLAM systems such as ORB-SLAM3 [4] and RTAB-Map [3] focus on accurate localization and geometric reconstruction using feature tracking, bundle adjustment, loop closure, and multi-modal inputs, but they do not provide language-aligned semantics suitable for open-vocabulary interaction. More recent dense and neural SLAM approaches like NICE-SLAM [6] improve geometric fidelity using neural implicit representations, but lack open-vocabulary semantic reasoning for navigation.

Semantic SLAM augments geometric maps with semantic categories or object instances, typically relying on closed-set detectors such as YOLO [1] or Mask R-CNN [2] sometimes combined with general segmentation models like SAM [9]. While effective in structured settings, these systems are constrained by predefined semantic vocabularies and struggle to generalize in open-world environments, limiting their applicability to open-vocabulary rover missions [1, 2, 9].

Open-vocabulary 3D scene understanding leverages vision–language models to overcome fixed label sets. Methods such as OpenScene [10] and OpenOcc [12] project CLIP-based embeddings into 3D reconstructions or volumetric occupancy grids, enabling zero-shot recognition beyond predefined object categories. However, these approaches are designed for offline scene processing and do not address dynamic, resource-constrained navigation with continuous map updates.

Recent work integrates vision–language models directly into online mapping. VLMaps [7], LERF [8], Open-Fusion [13] and subsequent systems construct spatially indexed or object-centric maps that can be queried with natural language, and more recent frameworks like FindAnything [18] and DualMap [15] begin to address online operation and dynamic environments. These methods demonstrate the feasibility of open-vocabulary mapping but generally fuse all evidence into a single representation with limited treatment of uncertainty.

3. Approach - The CrossMaps Pipeline

Overview – We extended DualMaps [15] and VLMaps [7] by constructing an open-vocabulary semantic map that integrates vision-language embeddings with spatial mapping and confidence-aware memory management. The pipeline processes a stream of RGB-D observations during rover navigation and, given an RGB image I_t, depth image D_t, and rover (camera) viewpoint T_t estimated by SLAM at time t, extracts CLIP-based semantic embeddings, back-projects them into a 3D spatial grid map, and fuses them into short-term memory using confidence-aware updates. It maintains both an STM that aggregates recent

Algorithm 1 CrossMaps Semantic Mapping Pipeline

Require: RGB image I_t, depth map D_t, rover viewpoint T_t, confidence τ_c and coherence τ_h thresholds

Compute CLIP embeddings e_i for all coarse and fine image tiles in Image I_t
for each sampled depth point p **do**
    Back-project p using D_t and transform using T_t; retrieve corresponding coarse- and fine-tile embeddings; fuse them into a point embedding e_p; estimate geometric confidence c_p^{geom}; assign point observation to grid cell $\mathscr{C}$
end for
for each affected grid cell $\mathscr{C}$ **do**
    Compute semantic consistency + temporal decay; update embedding, $c_{\mathscr{C}}$, $coh_{\mathscr{C}}$
end for
for each STM cell $\mathscr{C}$ **do**
    if $C_{\mathscr{C}} > \tau_c$ **and** $H_{\mathscr{C}} > \tau_h$ **and** viewpoint.diversity **then**
        Promote $\mathscr{C}$ to LTM
    end if
end for
Query: encode + execute q; generate heatmap; select top navigation targets

observations and an LTM that stores persistent semantic landmarks. The pipeline (Figure 1) consists of embedding extraction, spatial projection into spatial map cells using depth and viewpoint information, confidence-aware fusion of observations in STM, and STM-to-LTM promotion (see Algorithm 1).

Real-Time Open-Vocabulary Semantic Mapping – CrossMaps performs online semantic mapping by combining multi-scale vision–language embeddings with SLAM-based spatial alignment. Each incoming RGB frame is divided into tiles at multiple spatial scales (Figure 2a), and each tile is encoded with a pretrained CLIP visual transformer (ViT-L/14, OpenAI weights) to obtain a semantic embedding aligned with natural language. These tile-level embeddings support open-vocabulary queries (e.g., "duck") while avoiding the cost of dense, pixel-wise embeddings that require pre-segmentation. To reduce computation and increase robustness, we adopt tile-level aggregation instead of pixel-wise embeddings: this trades some spatial precision for higher throughput and resilience to camera motion. Using the corresponding depth image D_t and viewpoint T_t, depth pixels are sampled and back-projected into 3D, transformed into the global frame, and associated with the fused embeddings of their corresponding coarse and fine image tiles. The resulting point-wise semantic observations are then accumulated in a top-down spatial grid map whose cells store semantic embeddings and statistics. Multi-view observations are fused via a top-down scheme that supports real-time mapping.

Confidence-Aware STM Semantic Fusion – After associating sampled 3D observations with fused tile-level semantic embeddings and assigning them to spatial grid cells, CrossMaps explicitly models perception uncertainty by assigning confidence values to observations during map fusion—see Figure 2b. Confidence is computed at both the *point-level* and *cell-level* using *geometric*, *semantic*, and *temporal* cues. **Geometric confidence** captures viewing conditions such as sensor distance and visibility; distant observations are down-weighted via distance-based decay, while repeated observations from multiple viewpoints increase confidence. **Semantic confidence** measures agreement between new embeddings and the current cell evidence via cosine similarity in CLIP space, reinforcing consistent observations and down-weighting inconsistent ones, while a coherence measure assesses agreement among fused observations. **Temporal confidence** applies time-based decay so that confidence in rarely observed regions decreases, enabling the STM to adapt to changes while the LTM remains persistent.

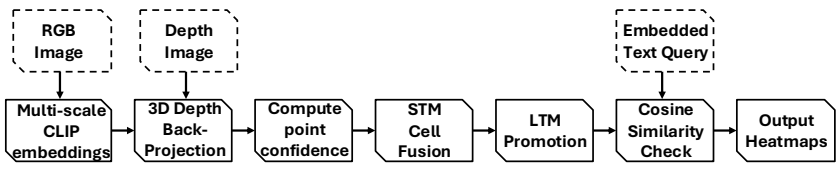
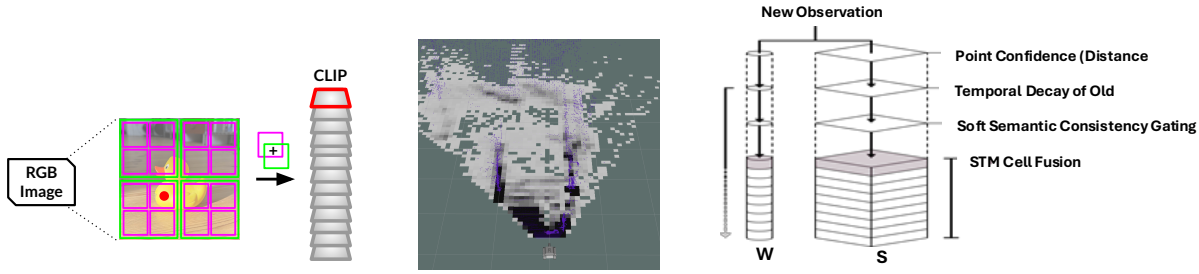


Figure 1: CrossMaps pipeline



(a) Tile-level coarse (green) and fine (pink) embeddings

(b) Cell-confidence semantic map

(c) STM cell fusion

Figure 2: Visualization of transformations across data representations

These confidence components are combined into confidence-weighted updates during semantic cell fusion so that reliable observations influence the map more strongly. To fuse multiple observations in the same grid cell, the STM performs confidence-weighted semantic fusion: each cell $\mathscr{C}$ maintains a semantic accumulator $S_{\mathscr{C}}$ and a confidence accumulator $W_{\mathscr{C}}$, where $S_{\mathscr{C}}$ stores the sum of confidence-weighted embeddings and $W_{\mathscr{C}}$ stores the accumulated confidence used for normalization—see Figure 2c.

Before fusion, the confidence of a new observation i is computed from the geometric, semantic, and temporal factors, and the stored cell confidence is decayed to reduce the influence of outdated observations. The incoming observation is then checked for semantic consistency with the current cell embedding using cosine similarity in CLIP space and softly down-weighted through a gating function if it deviates. The fusion step performs confidence-weighted aggregation of embeddings $\mathbf{e}_i$ with confidence c_i and is given by: $S_{\mathscr{C}} \leftarrow S_{\mathscr{C}} + c_i \mathbf{e}_i, \quad W_{\mathscr{C}} \leftarrow \lambda W_{\mathscr{C}} + c_i, \quad \mathbf{e}_{\mathscr{C}} = \frac{S_{\mathscr{C}}}{\|S_{\mathscr{C}}\|}$, where λ denotes the temporal decay factor and $\mathbf{e}_{\mathscr{C}}$ is the normalized semantic cell embedding. Afterwards, viewpoint coverage and semantic coherence statistics are updated to assess the reliability of the fused representation.

STM to LTM Promotion – To maintain both adaptability and stability, CrossMaps generates a dual-memory semantic map representation consisting of STM and LTM. While the STM acts as a dynamic grid map that aggregates recent observations, the LTM represents semantic landmarks and other stable structures in the environment (e.g., furniture). When sufficient evidence accumulates for a cell in the STM, the pipeline evaluates whether the observation is reliable enough to be promoted to the LTM. Only cells that exhibit high confidence, strong semantic coherence, and support from multiple viewpoints are transferred to the LTM. CrossMaps therefore selects STM cells for promotion using three criteria—confidence, semantic coherence, and viewpoint diversity.

Confidence – A cell is eligible for promotion when its accumulated confidence ($W_{\mathscr{C}}$) exceeds a threshold (τ_c); high-confidence observations are moved to LTM, while low-confidence, noisy, or unstable detections remain in STM and decay over time. If a new high-confidence observation contradicts an existing LTM cell, it can replace the older entry, allowing the long-term map to adapt.

Coherence – Beyond confidence, we measure how consistently observations within a cell agree in embedding space. Let $\mathbf{e}_i$ be the normalized embeddings fused into a cell; coherence is the norm of their mean $\text{coh}_{\mathscr{C}} = \left\| \frac{1}{n} \sum_{i=1}^{n} \mathbf{e}_i \right\|$, which estimates angular variance in embedding space. Well-aligned embeddings yield a large norm and coherence near 1, while conflicting embeddings cancel out and drive coherence toward 0, so only cells with coherence above a threshold (τ_h) are promoted.

Viewpoint Diversity – Finally, CrossMaps requires confirmation from multiple viewpoints. For each cell, the relative rover–cell angle is discretized into viewpoint bins and stored as a bitmask (e.g., 00011011) that records viewing directions. Cells observed from several viewpoints are treated as more reliable, are more likely to be promoted to LTM, and can receive initial boosted confidence in STM.

Natural-Language Semantic Querying – The CrossMaps semantic map supports open-vocabulary queries by embedding natural language into the same space as map cells. Given a query q, we compute

its CLIP ViT-L/14 text embedding $\mathbf{e}_q = \text{CLIP}_{\text{text}(q)}$ and for each cell with embedding $\mathbf{e}_{\mathscr{C}}$, we evaluate the *cosine similarity* (as semantic matching score) $cos(\mathbf{e}_q, \mathbf{e}_{\mathscr{C}}) = \frac{\mathbf{e}_q \cdot \mathbf{e}_{\mathscr{C}}}{\|\mathbf{e}_q\| \|\mathbf{e}_{\mathscr{C}}\|}$. To obtain a spatial heatmap H, we weight this similarity by the cell coherence $coh_{\mathscr{C}}$ as $H_{\mathscr{C}} = s_{\mathscr{C}} \cdot coh_{\mathscr{C}}$, so that semantically consistent regions remain visible even at lower confidence. For the top-ranked cell, its raw confidence $c_{\mathscr{C}}$ influences ranking, yielding a natural-language-driven spatial index where high-confidence cells act as landmarks while potentially useful low-confidence regions are still highlighted.

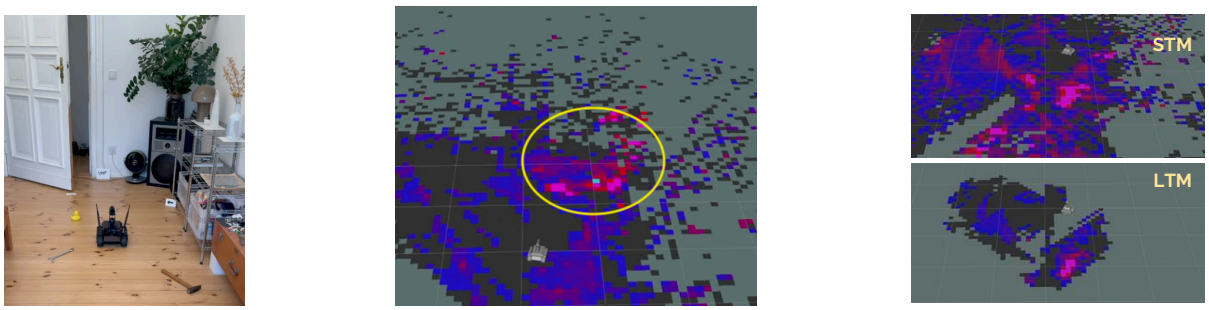
4. Implementation and Proof-of-Concept Demonstration

To validate the framework, we implemented CrossMaps as a real-time perception and semantic mapping pipeline on a Waveshare UGV Jetson Orin rover (a compact GPU-equipped platform).

Pipeline Implementation – CrossMaps is implemented as a modular ROS 2 node operating alongside a SLAM-based localization pipeline. The node subscribes to RGB and depth images, camera calibration data, and pose estimates via standard ROS 2 topics. Semantic maps are maintained in the odometry frame and transformed to the global map frame through ROS 2 TF, ensuring stability under SLAM updates such as loop closures. Semantic features are extracted using a pretrained CLIP ViT-L/14 model, balancing expressiveness and computational efficiency for onboard inference. RGB frames are divided into multi-scale tiles whose embeddings are projected into 3D using depth measurements and pose estimates. These observations are integrated into the semantic grid map described in Section 3. Configurable parameters control mapping behavior, including grid resolution, STM decay rate, confidence thresholds, and STM-to-LTM promotion rules. The code and replication package are publicly available [17].

Platform Deployment – Our prototype was deployed on a Waveshare UGV rover equipped with an RGB-D camera and an NVIDIA Jetson Orin GPU. The rover streams RGB-D observations and pose estimates via ROS 2 to an external computer running CrossMaps, while RTAB-Map provides localization for global alignment. CrossMaps processes these inputs to build semantic maps and publishes embeddings, confidence grids, and heatmaps via ROS 2 topics, enabling visualization in RViz.

Semantic Query Demonstration – CrossMaps supports open-vocabulary semantic queries at runtime. A natural-language query is encoded using the CLIP text encoder and compared with stored cell embeddings using cosine similarity. The resulting similarity scores are visualized as spatial heatmaps over the semantic grid map. Figure 3 illustrates representative heatmaps generated for example queries such as ”plant” and ”hammer”. These heatmaps highlight map regions that match the query while suppressing inconsistent observations through the STM update and confidence mechanisms. As in Figure 3, for the query ”hammer”, the STM tends to be noisy, while LTM shows more sparse and reliable suggestions. A demonstration of the rover exploration is available at https://youtu.be/vMQndtoBYTU.



(a) Rover in the room (b) STM for query ”plant” (c) Query ”hammer”

Figure 3: STM and LTM heatmaps for different natural language queries in CrossMaps

5. Conclusion, Future Work and Declaration on Generative AI

CrossMaps shows that combining open-vocabulary perception with confidence-aware fusion and a dual-memory representation enables rovers to build language-queryable semantic maps while mitigating

perceptual noise. The STM absorbs uncertain observations, while the LTM retains only stable semantic structures supported by sufficient confidence and coherence. Our proof-of-concept demonstrates real-time operation with natural-language queries, and future work will integrate segmentation-based feature extraction [9] and extend CrossMaps toward multi-agent sharing of causal knowledge [16] and neuro-symbolic reasoning [14, 11, 19] to tackle safety-critical rover navigation scenarios. Generative AI tools were used solely for language editing. The authors are responsible for the content.

References

[1] Joseph Redmon et al. "You only look once: Unified, real-time object detection". In: *Proc. of the IEEE CVPR*. 2016, pp. 779–788.

[2] Kaiming He et al. "Mask r-cnn". In: *Proc. of the IEEE ICCV*. 2017, pp. 2961–2969.

[3] Mathieu Labbé and François Michaud. "RTAB-Map as an Open-Source Lidar and Visual SLAM Library for Large-Scale and Long-Term Online Operation". In: *Journal of Field Robotics* 36.2 (Mar. 2019), pp. 416–446.

[4] Carlos Campos et al. "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM". In: *IEEE Transactions on Robotics* 37.6 (2021), pp. 1874–1890.

[5] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proc. of the ICML*. Vol. 139. PMLR, 2021, pp. 8748–8763.

[6] Zihan Zhu et al. "NICE-SLAM: Neural Implicit Scalable Encoding for SLAM". In: *2022 IEEE/CVF CVPR*. New Orleans, LA, USA: IEEE, June 2022, pp. 12776–12786.

[7] Chenguang Huang et al. "Visual language maps for robot navigation". In: *Proc. of the IEEE ICRA*. IEEE. 2023, pp. 10608–10615.

[8] Justin Kerr et al. "Lerf: Language embedded radiance fields". In: *Proc. of the IEEE/CVF ICCV*. 2023, pp. 19729–19739.

[9] Alexander Kirillov et al. "Segment anything". In: *Proc. of the IEEE/CVF ICCV*. 2023, pp. 4015–4026.

[10] Songyou Peng et al. "Openscene: 3d scene understanding with open vocabularies". In: *Proc. of the IEEE/CVF CVPR*. 2023, pp. 815–824.

[11] Christian Medeiros Adriano, Sona Ghahremani, and Holger Giese. "Principled Transfer Learning for Autonomic Systems: A Neuro-Symbolic Vision". In: *ACSOS-C*. Sept. 2024, pp. 79–84.

[12] Haochen Jiang et al. "Openocc: Open vocabulary 3d scene reconstruction via occupancy representation". In: *Proc. of the IEEE/RSJ IROS*. IEEE. 2024, pp. 1145–1152.

[13] Kashu Yamazaki et al. "Open-fusion: Real-time open-vocabulary 3d mapping and queryable scene representation". In: *Proc. of the ICRA*. IEEE. 2024, pp. 9411–9417.

[14] Christian Medeiros Adriano et al. "Neuro-symbolic causal reasoning for cautious self-adaptation under distribution shifts". In: *In Proc. of ACSOS*. IEEE. 2025, pp. 88–99.

[15] Jiajun Jiang et al. "DualMap: Online Open-Vocabulary Semantic Mapping for Natural Language Navigation in Dynamic Changing Scenes". In: *IEEE Robotics and Automation Letters* (2025).

[16] Kathrin Korte et al. "Causal knowledge transfer for multi-agent reinforcement learning in dynamic environments". In: *In Proc. of ACSOS*. IEEE. 2025, pp. 154–159.

[17] Niklas Klein. *CrossMaps: Confidence-Aware Open-Vocabulary Semantic Mapping for Rover Navigation*. https://github.com/jniklasklein/crossmaps.git. Accessed: 2026-03-15. 2026.

[18] Iro-Elizabeth Laina et al. "FindAnything: Open-Vocabulary and Object-Centric Mapping for Robot Exploration in Any Environment". In: *arXiv preprint arXiv:2504.08603* (2026).

[19] Zainab Rehan et al. "Towards Neuro-symbolic Causal Rule Synthesis, Verification, and Evaluation". In: *NEXUS, Companion workshop of AAMAS*. Cyprus, 2026. URL: https://nexus.telecom-paris.fr/.