A Scene Graph Backed Approach to Open Set Semantic Mapping

Martin Günther[1] Felix Igelbrink[1] Oscar Lima[1]
Lennart Niecksch[1,2] Marian Renz[1,2] Martin Atzmueller[2,1]

[1]German Research Center for Artificial Intelligence (DFKI), Osnabrück, Germany
[2]Osnabrück University, Semantic Information Systems Group, Osnabrück, Germany

AAAI Spring Symposium on Machine Learning and Knowledge Engineering
April 2026

Motivation: Beyond Geometric Mapping

- **The Goal:** Autonomous mobile robots in large-scale, real-world environments require high-level semantic understanding to perform complex tasks.
- **The Limitation of Implicit Representations:**
  - Volumetric maps (e.g., TSDFs) and implicit representations (e.g., NeRFs) provide high-fidelity geometry.
  - However, they lack an *explicit model* of the scene's structure.
  - The environment is treated as a monolithic volume, decoupling perception from high-level reasoning.
- **The Solution: 3D Semantic Scene Graphs (3DSSGs)**
  - Bridge the gap between sub-symbolic sensor data and symbolic reasoning.
  - Transform raw perception into a queryable knowledge base.
  - Align the robot's internal representation with human-understandable concepts.

Problem Statement

Current Status Quo

Many existing frameworks treat the scene graph as a **derivative layer generated post-hoc** (after the geometric map is fully built).

Why is this an issue?

- Limits the system's ability to maintain topological consistency over extended operations.
- Prevents online, real-time interactions and reasoning.
- Fails to seamlessly integrate top-down expectations (knowledge priors) with bottom-up sensory data.

Our Contributions

We propose a mapping architecture where the **3DSSG serves as the foundational backend** for the entire mapping process.

1. **Modular Mapping Architecture:** A multi-layered 3DSSG acts as the primary map representation, supporting both flat and hierarchical topologies.
2. **Incremental Predicate Prediction (IPP):** Inferring and updating structural relationships (e.g., *on top of*, *part of*) dynamically via a Graph Neural Network.
3. **Open-Set Grounding:** Incorporating Vision-Language Foundation Models (CLIP) to facilitate natural language queries.
4. **Real-World Deployment:** Online processing (7–8 Hz) deployed on a mobile robot (TIAGo) in large-scale environments.

The 3DSSG Backend: Layers of Abstraction

Our approach adopts an instance-based, object-centric mapping paradigm. The backend is a multi-layered 3DSSG:

- **Frames Layer:** Tracks 6D pose information and raw 2D segmentation masks for all keyframes.
- **Segments Layer:** Intermediate representation tracking all active 3D object fragments, raw geometry, and feature vectors. Edges describe co-visibility.
- **Objects Layer:** Persistent, consolidated object instances maintaining spatio-semantic relations and accumulated features.

Strictly incremental: the 3DSSG backend transitions from one valid state to the next at each step.

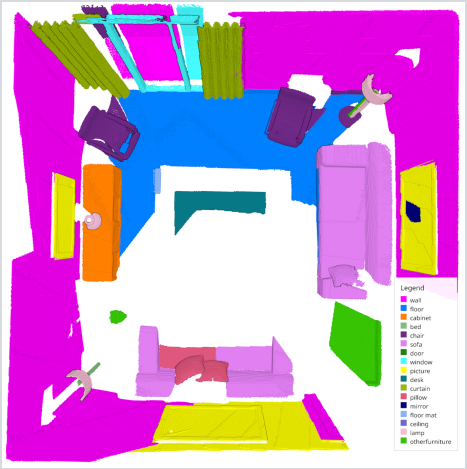
Mapping Pipeline: Segmentation & Local Graphs

1. Segmentation & Extraction

- Input: RGB-D frames.
- Segmentation: FastSAM (refined via depth discontinuities).
- Features: DINOv2 (per-patch features to bypass heavy computation on individual crops).

2. Local Graph Generation

- Depth projected to 3D.
- Custom CUDA DBSCAN for fast, batched noise filtering.
- Populates a *local* 3DSSG mirroring the current frame.



Resulting objects after semantic segmentation

Mapping Pipeline: Global Integration

How do we merge the local frame into the global knowledge map safely?

Stage 1: Greedy Association

- Matches *local* segments to *global* segments based on 3D Bounding Box IoU and DINOv2 cosine similarity.
- Highly conservative: only merges unambiguous matches.

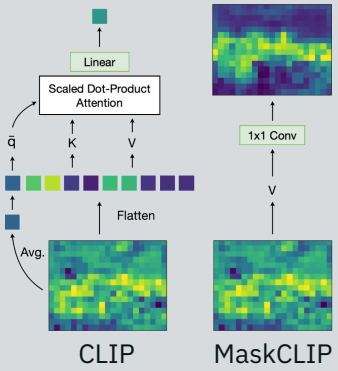
Stage 2: Active Refinement

- Resolves temporal inconsistencies and over-segmentation.
- Runs immediately after Stage 1, but **only evaluates the active subset** of the graph (nodes just modified/created).
- Merges fragments into coherent instances based on feature stability and voxel-grid overlap.

Vision-Language Feature Integration

To allow humans (or agentic models) to query the map:

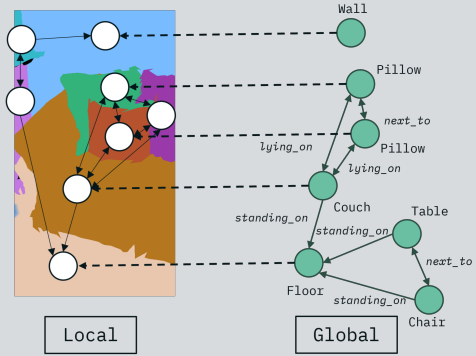
- **MaskCLIP Paradigm:** We extract per-patch features from the penultimate layer of a standard CLIP model. (Dong et al. 2023)
- **Gated Integration:** Local patch features are strongly biased to local textures. We modulate them using the *global* CLIP embedding of the entire frame (cosine similarity).
- **View Quality Score:** Features are weighted by viewing conditions (segment size, distance to image center, surface structure) to ensure robust embeddings against noisy or fleeting observations.



Incremental Predicate Prediction (IPP)

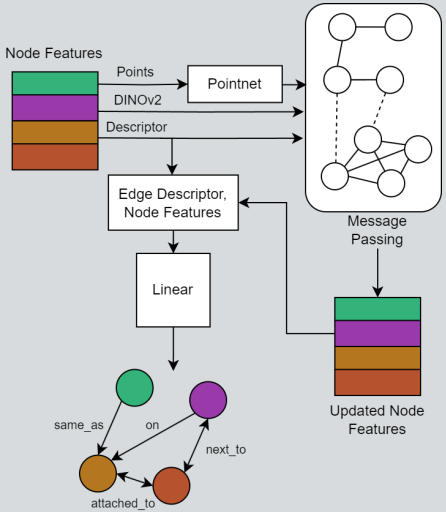
Moving beyond mere spatial proximity:

- Edges in standard maps only show hierarchy or adjacency.
- We predict semantic predicates (e.g., *attached to*, *on top of*) to assist task planners and LLMs.
- Uses a heterogeneous **Graph Neural Network (GNN)**.
- Operates on both the local and global layers of the graph simultaneously.



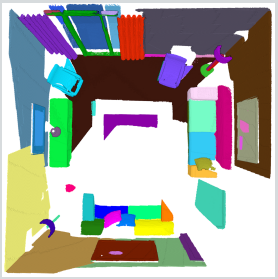
Heterogeneous Graph Neural Network Architecture

- **Node Features:** Points (via PointNet), DINOv2, and geometric descriptors.
- **Message Passing:** Heterogeneous GraphSage (Hamilton et al. 2017).
- **Semantic Priors:** Global layer includes text embeddings (GloVe (Pennington et al. 2014)) of previously predicted edges.



Preliminary Results: Simulation (ICL Dataset)

RGB



Instances (Segment Layer)

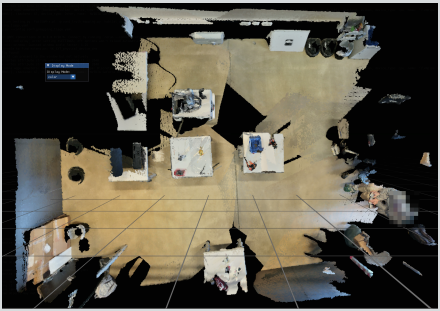
Semantic Segmentation (Object Layer)

- wall
- floor
- cabinet
- bed
- chair
- sofa
- door
- window
- picture
- desk
- curtain
- pillow
- mirror
- floor mat
- ceiling
- lamp
- otherfurnit.

- System successfully isolates major structural elements.
- Semantic labels generated purely by querying open-set CLIP features against NYU-40 label embeddings.
- Conservative merging strategy ensures objects are not irreversibly fused by accident.

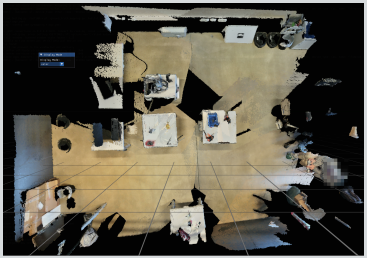
Real-World Deployment: TIAGo Robot

- **Hardware:** TIAGo mobile robot, Ouster OS0-128 LiDAR, Femto Bolt RGB-D camera.
- **Localization:** MICP-L (Mock et al. 2024) aligning LiDAR against a static polyhedral reference mesh of the building.
- **Processing:** 440 frames processed incrementally.
- Successfully registers objects despite real-world sensor noise and minor pose inaccuracies.

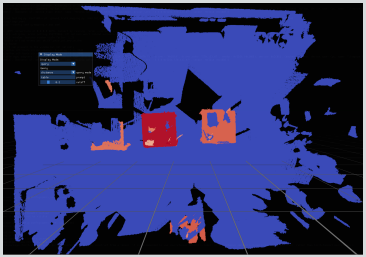


Reconstructed RGB Point Cloud

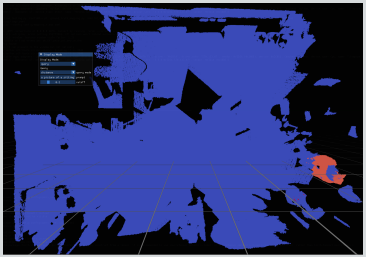
Real-World Results: Open-Set Queries



RGB Reconstruction



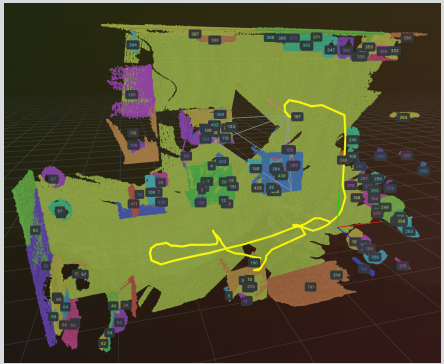
Query: “Table”



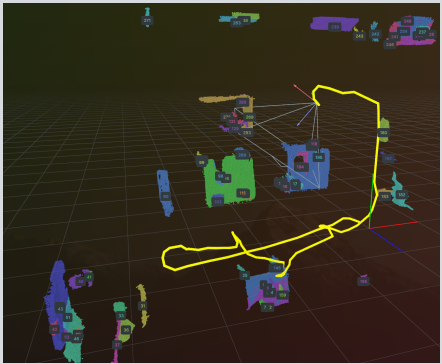
Query: “Sitting man”

- Visualizing cosine similarity between instance CLIP embeddings and text queries.
- Accurately isolates targeted objects, ranging from large furniture to detailed descriptions.

Handling Clutter: Geometric Background Filtering



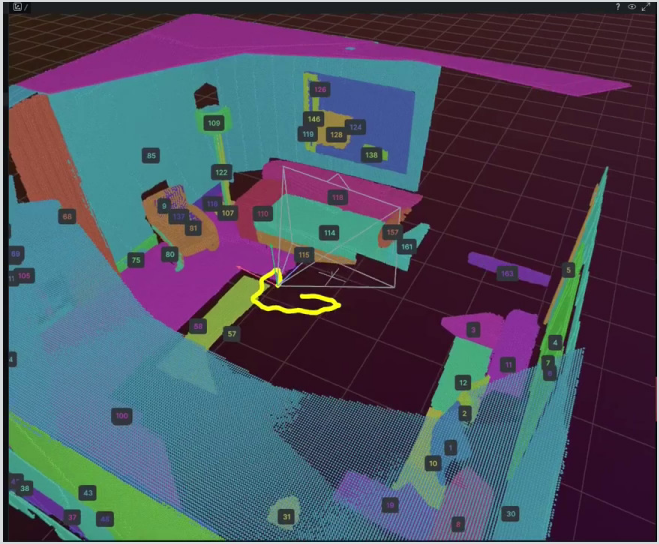
Unfiltered Depth Data



Filtered Data (Simulated Panoptic)

- **Challenge:** Structural *stuff* classes (floors, walls) create spurious instances.
- **Solution:** Used Rmagine (Mock et al. 2023) to match depth against the building model, filtering out static structures prior to mapping.

Video Demonstration



Discussion & Future Outlook

The Hybrid Intelligence Advantage

Establishing the 3DSSG as the **foundational backend** successfully bridges statistical learning (perception/VLMs) with symbolic reasoning (knowledge graphs).

Current Limitations:

- *Resolution Constraint:* MaskCLIP patch size currently limits the granularity for very small or highly cluttered objects.
- *Supervised Dependency:* The GNN relies on supervised learning (3DSSG dataset labels), limiting generalization to novel environments.

Future Work:

- Inferring spatial relations in an **unsupervised manner** (using geometric heuristics like SEMAP).
- Tightening the integration of relation prediction directly within the real-time mapping loop.

Thank You!

Code and Video available at:
https://dfki-ni.github.io/SSG-MAKE-2026/

Questions?