ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

vS-Graphs: Environment-Aware 3D Scene Graphs for Visual SLAM

Ali Tourani, Saad Ejaz, Miguel Fernandez-Cortizas, Jose Luis Sanchez-Lopez, and Holger Voos
Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg, Luxembourg
{ali.tourani, saad.ejaz, miguel.fernandez, joseluis.sanchezlopez, holger.voos}@uni.lu

I. Introduction & Motivation

Visual Simultaneous Localization and Mapping (VSLAM) is a widely used solution for robotic perception, providing camera pose estimation and environment reconstruction using low-cost vision sensors. Beyond geometric consistency, many robotics applications increasingly require *semantically meaningful* and *structurally interpretable* maps, enabling robots to reason about the environments they operate in. While recent VSLAM works have incorporated object- and scene-level understanding, most VSLAMs still produce maps that are difficult to interpret and offer limited support for explicitly modeling layout-driven semantic information, such as localizing walls and doorways.

In response to this, 3D scene graphs can provide promising structured representations for capturing such **hierarchical, spatial relationships**. However, they are often constructed *offline* and require *complete maps* or *ground-truth semantics* (such as HOV-SG [2]), which means they are not tightly integrated into the SLAM pipeline. Our previous works [3], [4] used fiducial markers to localize environment-driven entities and generate hierarchical 3D scene graphs, enabling richer layout understanding. However, their reliance on pre-placed markers limits deployment in unprepared environments.

To overcome these limitations, we introduce **vS-Graphs** [1]: a publicly available[1] real-time VSLAM framework that tightly couples map reconstruction with online 3D scene graph generation. Inspired by our LiDAR-based S-Graphs [5], [6], vS-Graphs employs visual and depth cues to detect and localize **building components**, such as walls and ground surfaces, from which higher-level **structural elements**, including n-wall-shaped rooms and floors, are inferred. These entities are incorporated into an optimizable hierarchical 3D scene graph, jointly maintained with the SLAM pipeline, enabling richer map semantics and improved localization.

II. Method Overview

vS-Graphs extends ORB-SLAM3 [7] with online *semantic-structural reasoning* to jointly reconstruct the environment and generate an optimizable 3D scene graph. The current implementation operates on RGB-D input, where visual features support camera tracking, while depth data provides geometric cues to validate semantic entities. The framework introduces two newly integrated threads: "**building component recognition (BCR)**" and "**structural element recognition (SER)**." As incoming RGB-D frames are processed, the *tracking* and *local mapping* threads of the baseline estimate camera poses, pick KeyFrames, and maintain the geometric map. For each KeyFrame, the proposed *BCR* thread analyzes visual-spatial observations to detect environment-driven planar entities, including walls and ground surfaces. This is achieved by combining panoptic scene segmentation algorithms (such as YOSO [8] and pFCN [9]) with point cloud filtering and RANSAC-based plane fitting, followed by semantic and geometric validation. Detected *building components* are stored in the active map and used by the *SER* thread, which runs periodically to infer higher-level layout entities. In particular, vS-Graphs groups associated walls and ground surfaces to identify enclosed free-space configurations corresponding to n-walled rooms and further aggregates them into floor-level structures. These inferred entities, together with their spatial relationships, form a hierarchical scene graph that is incrementally built online. Unlike conventional semantic mapping pipelines that remain detached from SLAM optimization, vS-Graphs incorporates building components and structural elements directly into the mapping process. This tight coupling enables the system to maintain semantically enriched and structurally coherent maps while preserving real-time operation. Samples of reconstructed maps using vS-Graphs are shown in Fig. 1.

III. Preliminary Results

We evaluated vS-Graphs on both public RGB-D benchmarks and our in-house *SMapper* dataset [10], which includes diverse multi-room indoor environments with LiDAR-derived ground truth. These evaluations focused on **trajectory estimation**, **map quality**, **semantic structural detection**, and **runtime performance**. Full evaluation results and figures are publicly available[2].

Regarding trajectory estimation, vS-Graphs consistently achieved competitive or state-of-the-art trajectory accuracy, reducing Absolute Trajectory Error (ATE) by an average of **15.22%** (shown in Fig. 2) compared to the baseline. Larger gains were observed in looped and multi-room environments, where structural consistency is especially beneficial. This evaluation proved that the inclusion of higher-level semantic entities further improves localization robustness by reinforcing more spatial constraints. Beyond trajectory estimation, vS-Graphs also produced more coherent reconstructed maps, achieving *lower mapping error* despite generating

[1] https://github.com/snt-arg/visual_sgraphs

[2] https://snt-arg.github.io/vsgraphs-results/

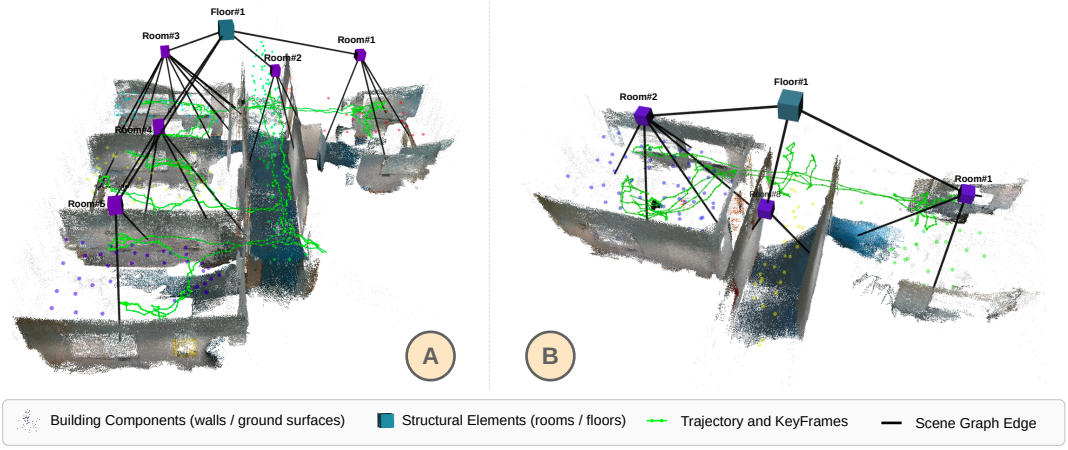

Fig. 1. Reconstructed maps and their corresponding optimizable 3D scene graphs, enriched with environment-driven semantic entities and generated by vS-Graphs [1]. The examples are taken from the in-house dataset and correspond to sequences *MR03* (A) and *MR01* (B).

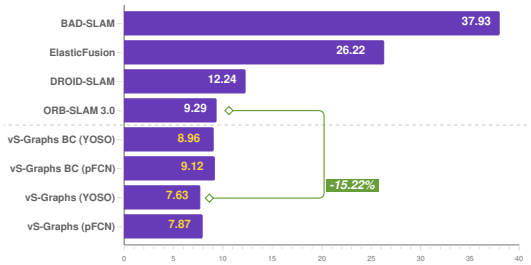

Fig. 2. Mean Absolute Trajectory Error (ATE), reported in *centimeters*, for different RGB-D VSLAM frameworks, with lower values indicating better accuracy. The results are computed across five datasets, including both standard benchmarks and in-house-collected sequences, with a total duration of approximately 44 minutes and a cumulative trajectory length of approximately 620 meters.

fewer points on average than the baseline. In addition, the framework demonstrated environment-driven semantic entity detection performance comparable to the LiDAR-based *S-Graphs* system, while relying only on RGB-D input. This highlights the effectiveness of the proposed visual-semantic structural reasoning pipeline.

IV. Discussions & Future Work

The presented results highlight the potential of tightly coupling VSLAM with online hierarchical scene graph generation for richer, more structurally meaningful environmental understanding. In particular, the ability of vS-Graphs to infer higher-level layout entities from visually detected building components suggests a promising direction for bridging *geometric mapping* and *semantic scene reasoning* within a unified framework.

At the same time, vS-Graphs presents several opportunities for further improvement: As the framework relies on visual semantic segmentation to initialize environment-driven entities, its performance remains partially dependent on the quality and robustness of the underlying perception model, with segmentation errors potentially propagating to 3D reconstruction and structural inference. In addition, maintaining real-time operation motivates continued exploration of lightweight perception backbones and more efficient semantic reasoning strategies. Future work will extend the framework to include **additional building components**, such as ceilings and doorways, and to support more complex or concave room layouts beyond the current structural assumptions. We also plan to investigate graph deep learning-based **structural reasoning** and **semantic loop closure detection**, in which revisited areas can be recognized primarily using previously inferred structural and semantic entities.

References

[1] A. Tourani, S. Ejaz, H. Bavle, M. Fernandez-Cortizas, D. Morilla-Cabello, J. L. Sanchez-Lopez, and H. Voos, "vs-graphs: Tightly coupling visual slam and 3d scene graphs exploiting hierarchical scene understanding," *arXiv e-prints*, pp. arXiv–2503, 2025.

[2] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[3] A. Tourani, H. Bavle, D. I. Avşar, J. L. Sanchez-Lopez, R. Munoz-Salinas, and H. Voos, "Vision-based situational graphs exploiting fiducial markers for the integration of semantic entities," *Robotics*, vol. 13, no. 7, p. 106, 2024.

[4] A. Tourani, H. Bavle, J. L. Sanchez-Lopez, R. M. Salinas, and H. Voos, "Marker-based visual slam leveraging hierarchical representations," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3461–3467.

[5] H. Bavle, J. L. Sanchez-Lopez, M. Shaheer, J. Civera, and H. Voos, "S-graphs+: Real-time localization and mapping leveraging hierarchical representations," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4927–4934, 2023.

[6] ——, "S-graphs 2.0–a hierarchical-semantic optimization and loop closure for slam," *IEEE Robotics and Automation Letters*, 2025.

[7] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[8] J. Hu, L. Huang, T. Ren, S. Zhang, R. Ji, and L. Cao, "You only segment once: Towards real-time panoptic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 819–17 829.

[9] Y. Li, H. Zhao, X. Qi, L. Wang, Z. Li, J. Sun, and J. Jia, "Fully convolutional networks for panoptic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 214–223.

[10] P. M. B. Soares, A. Tourani, M. Fernandez-Cortizas, A. Bikandi-Noya, H. Voos, and J. L. Sanchez-Lopez, "Smapper: A multi-modal data acquisition platform for slam benchmarking," *Journal of Intelligent & Robotic Systems*, vol. 112, no. 20, 2026.