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

Top-LOC: Visual Retrieval-Based Topological Localization for Dynamic Indoor Environments

Rafael Flor-Rodríguez-Rabadán[1], Sergio Lafuente-Arroyo[1], Saturnino Maldonado-Bascón[1], Roberto Javier López-Sastre[1], and Carlos Gutiérrez-Alvárez[1]

***Abstract*— Visual place recognition in large-scale, indoor environments often suffers from perceptual aliasing due to structural symmetries and dynamic changes. This work presents a robust hierarchical topological mapping framework designed for long-term robot autonomy. Our system integrates multi-modal data (including 2D LiDAR, odometry, and RGB imagery) into a two-layer architecture. First, a Layout Layer is designed to capture the geometric structure of the environment. Then, a Visual Layer is used to encode image sequences. A key contribution is the dynamic map maintenance mechanism, which monitors the attenuation of edge weights to detect environmental transitions, such as the opening or closing of doors. This allows for seamless lifelong updates without human intervention in large-scale environments. We evaluate our approach using various visual descriptors (*e.g.* SuperGlue, Patch-NetVLAD, and SeqVLAD) within a sequence-based matching pipeline. Experimental results in a 750 m^2 real-world facility demonstrate that the proposed method achieves high discrimination and scalability, even in challenging open areas and symmetric corridors. This framework provides a reliable solution for assistive robotics navigating complex, evolving public spaces.**

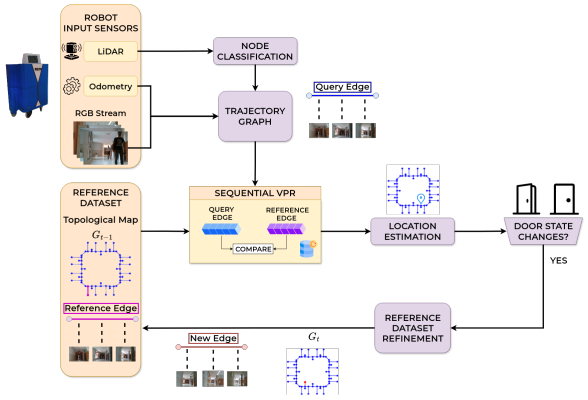

Fig. 1: System architecture. Multi-modal inputs are processed into a two-layer scene graph. The Layout Layer segments the spatial structure into nodes and edges, while the Visual Layer encodes RGB sequences per edge. Sequential VPR estimates location by matching query and reference trajectories. Repeated visits enable dynamic change detection and lifelong map updates.

I. Introduction

Indoor localization remains a fundamental challenge for autonomous mobile robots operating in large-scale public buildings such as hospitals, offices, and residential facilities. These environments present three critical difficulties: (i) *high structural symmetry*, where corridors exhibit nearly identical appearances leading to perceptual aliasing; (ii) *open areas* with few distinctive visual cues; and (iii) *dynamic changes* caused by door openings and closures that alter the traversability of the environment.

Existing visual place recognition (VPR) methods [1]–[3] have shown remarkable performance on standard benchmarks, yet their robustness degrades in large-scale indoor settings characterized by repetitive patterns. Single-image retrieval struggles with perceptual aliasing, while 3D Structure-from-Motion models [4] are storage-intensive and hard to maintain. Moreover, few systems address the dynamic nature of real environments, where topological changes demand continuous map updates [5].

To address these challenges, we propose a hierarchical topological mapping framework that leverages a *two-layer scene graph* combining spatial structure from LiDAR/odometry with visual sequence matching from an egocentric RGB camera. Our main contributions are:

- A topological representation that scales visual retrieval to large environments by segmenting the space into structural sections (edges and nodes), enabling sequence-based comparisons that mitigate perceptual aliasing.
- A modular architecture that supports the integration and benchmarking of diverse VPR methods, from handcrafted descriptors (SIFT [6], ORB [7]) to deep-learning-based approaches (SuperGlue [3], Patch-NetVLAD [2], SeqVLAD [8]).
- A lifelong map maintenance mechanism that detects and incorporates dynamic topological changes, such as door openings and closures, in real time, without human intervention.

II. Method

Two-layer topological graph. The environment is modeled as a graph $G_t = \{V, E\}$. Nodes V represent structural junctions, classified as *End*, *T*, *L*, *Cross*, or *Open-area* via an AI model on 2D LiDAR scans. Edges E represent the traversable corridors in the environment. A two-layer scene graph captures both the *Layout Layer* (node types and edge lengths derived from LiDAR and odometry) and the *Visual Layer* (sequential RGB descriptors associated with each edge), as illustrated in Fig. 1.

[1]Signal Theory and Communications Department, University of Alcalá, Alcalá de Henares, Madrid, Spain. rafael.flor@uah.es

TABLE I: Dynamic route: localization and change detection. $\alpha > \epsilon$=1.3 triggers a graph update. **Bold**: best; underline: 2nd best.

Method	AEL↓	DL↑	$\alpha\uparrow$ (d_A)	$\alpha\uparrow$ (d_B)
SIFT	**0.000**	0.142	1.16	1.19
ORB	0.076	0.110	1.15	1.12
SuperGlue	**0.000**	**0.248**	**1.62**	1.43
Patch-NetVLAD	0.005	0.229	1.49	1.29
SeqVLAD	0.187	0.274	1.20	1.13

Localization. Given a query trajectory $Y = \{y_i\}_{i=1}^{T}$ of T nodes, we compute a similarity weight w for each compatible reference trajectory Y' as:

$$w = \frac{1}{2^{T-1}} \prod_{i=1}^{T-1} \big(w_{i,\text{layout}} + w_{i,\text{visual}}\big), \tag{1}$$

where $w_{i,\text{layout}}$ encodes node-type agreement and edge-length similarity, and $w_{i,\text{visual}}$ captures visual appearance through frame-level or sequence-level descriptors. Location is estimated in real time via a sliding window over a third-order weight tensor, requiring only the last three traversed edges per iteration.

Dynamic map maintenance. A key innovation is the ability to detect and adapt to environmental changes such as door closures that alter the topological structure. During normal navigation, the best-matching weight $w_{i,\max}$ remains relatively stable; however, when the agent traverses an edge that no longer exists in the reference map—because a door has changed the local geometry—no stored reference trajectory can explain the observations, and $w_{i,\max}$ drops sharply. This drives the *reliability ratio* $\alpha = w_{i-1,\max} \,/\, w_{i,\max}$ above a detection threshold $\epsilon > 1$, signalling a topological transition. Upon detection, the system either creates a new direct edge (if the agent re-localises within the K-neighbourhood of a known node) or inserts both a new node and edge, updating the map for lifelong autonomy.

III. Experiments

For evaluation, we collected data in two real-world buildings at the University of Alcalá: a large-scale Polytechnic School ($\sim$10,000 m^2, 64 nodes, 64 edges) and a medium-sized Faculty of Nursing (12 nodes, 12 edges). Data was collected with a robotic platform equipped with an Intel RealSense D435 camera and an RPLIDAR A1 sensor. We define two evaluation metrics: *Accumulative Error Location* (AEL↓), which penalizes localization errors, and *Discrimination Location* (DL↑), which rewards confident correct estimates.

Corridor localization. Across five test sequences covering daytime, nighttime, human presence, and cross-building generalization, SuperGlue achieves error-free localization (AEL = 0) in all of them with the highest average discrimination, while SeqVLAD also attains zero errors consistently.

Dynamic environments. Table I reports performance on a route where two doors (d_A, d_B) are sequentially closed, modifying the node topology. Beyond localization (AEL/DL), we report the observed α at each closure: a value above ϵ=1.3 triggers a map update. CNN-based descriptors yield higher α, confirming greater sensitivity to structural changes.

Open areas & computational cost. In diaphanous spaces ($\sim$750 m^2), all methods reach AEL = 0; SIFT and Patch-NetVLAD achieve the best discrimination (DL up to 2.84), while SuperGlue shows slightly lower performance due to higher pose variability. All methods operate in real time since localization is triggered only at edge transitions; SIFT offers the best cost/accuracy trade-off, while SuperGlue maximises accuracy at moderate overhead.

IV. Conclusions

We have presented a hierarchical topological mapping framework that leverages structural and visual information for scalable indoor localization in large-scale, symmetric, and dynamic environments. By integrating sequence-based VPR into a dynamically maintained two-layer scene graph, our system overcomes the key challenges of perceptual aliasing, open-area ambiguity, and environmental changes. Extensive real-world experiments confirm that integrating SuperGlue into our framework achieves robust, error-free localization across diverse conditions, while SIFT provides a lightweight alternative with competitive performance. The dynamic map maintenance mechanism enables lifelong autonomy by seamlessly adapting the topological representation to environmental transitions. Future work will focus on deploying the framework into a complete assistive navigation system for complex public spaces.

References

[1] R. Arandjelovic, P. Gronat, T. Akihiko, and T. Pajdla, "NetVLAD: CNN architecture for weakly supervised place recognition," in *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, p. 5297–5307.

[2] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[3] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4937–4946.

[4] L. Liu, H. Li, and Y. Dai, "Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] R. R. Wiyatno, A. Xu, and L. Paull, "Lifelong Topological Visual Navigation," *IEEE Robotics and Automation Letters*, vol. 7, pp. 9271–9278, 2021.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, vol. 60, 2004, pp. 91–110.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571.

[8] R. Mereu, G. Trivigno, G. Berton, C. Masone, and B. Caputo, "Learning Sequential Descriptors for Sequence-Based Visual Place Recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 383–10 390, 2022.