Unordered Landmark Visual Navigation

Hao Ren[1], Junzhe Zhu[1], Yihan Li[1], Zetong Bi[1], Le Zheng[1], Zhi Li[1], Yiqing Yuan[1], Zhaoliang Wan[2], Dizhe Zhang[2], Lu Qi[2], Hui Cheng[1⋆]

[1] Sun Yat-sen University, Guangzhou, China
[2] Insta360 Research, Shenzhen, China

https://hren20.github.io/ulvn-website

Abstract. Image-goal navigation is a fundamental capability for embodied AI, yet its practical deployment is strained by strong prior assumptions. Existing methods predominantly rely on temporally ordered video streams or auxiliary sensors (e.g., depth, LiDAR) to maintain spatial consistency. These sequential and multimodal dependencies severely restrict scalability, especially when deploying robots using crowd-sourced or pre-recorded unordered image collections. When temporal priors are removed, current methods struggle with severe perceptual aliasing, noisy associations, and catastrophic mapping failures. To address this underexplored challenge, we propose Unordered Landmark Visual Navigation (ULVN), a unified RGB-only framework free from temporal and odometric priors. ULVN systematically mitigates error accumulation by integrating mapping, localization, and planning. Specifically, it constructs a robust 2D topological map directly from unstructured images via calibrated geometric verification and maximum spanning forest refinement. For closed-loop execution, ULVN abandons sequential heuristics, utilizing a graph-based belief propagation filter with entropy-adaptive fusion for global localization and dynamic subgoal planning. Extensive experiments in simulation and real-world deployments demonstrate that ULVN significantly outperforms state-of-the-art methods.

Keywords: Visual Navigation · Topological Mapping · Visual Localization

1 Introduction

Visual navigation [6,16,21,77] is a fundamental capability for autonomous agents, supporting applications from motion planning [17] to domestic robotics [34]. From the perspective of cognitive science, human navigation relies primarily on visual observations to form a topological understanding of the environment, enabling route finding without requiring precise metric coordinates or expensive active sensors such as LiDAR [51,56]. Emulating this capability in embodied AI has therefore become a highly desirable objective [1,56,81]. Ideally, an embodied

⋆ Corresponding author: Hui Cheng (chengh9@mail.sysu.edu.cn).

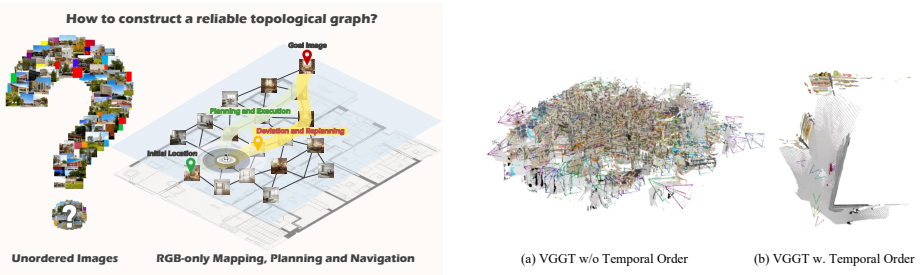

Fig. 1: From unordered images to navigation. Our system, ULVN, automatically constructs a topological representation purely from an unstructured set of RGB images. This enables closed-loop planning, execution, and replanning. In contrast, foundation models such as VGGT struggle to handle unordered images.

agent should operate predominantly on RGB observations within a pre-built topological map [7,10,11,51,58,64,66], developing holistic spatial awareness and reliable goal-directed behavior.

In contrast to this ideal, many existing visual navigation systems implicitly rely on simplifying conditions that rarely hold in the wild. They often require temporally ordered video streams, which provide strong sequential priors for graph construction and policy learning [64,78]. Absent this temporal structure, such as when using an unordered photo collection from real estate listings, learned transitions often fail. Without sequential priors, disambiguating similar places is challenging, making topological connections noisy and action selection unstable, which frequently causes oscillatory behavior. Furthermore, many systems depend on auxiliary sensors like depth or LiDAR to stabilize localization [19,25]. Relying solely on appearance complicates geometric verification, and even advanced reconstruction models struggle with pose estimation from sparse, disconnected observations [60,65,70]. Consequently, current success heavily depends on temporal continuity, auxiliary sensing, or both.

Eliminating these assumptions presents two progressive challenges. First, removing temporal priors forces topological mapping to infer connectivity from appearance alone. This vulnerability to perceptual aliasing and spurious edges makes planning and localization fragile, as errors accumulate without sequential updates to suppress ambiguity [7,11,38,72]. Second, removing multimodal sensors alongside temporal priors compounds the difficulty. Without depth, LiDAR, or odometry, resolving scale and viewpoint ambiguities becomes significantly harder, increasing cumulative drift and complicating closed-loop correction using only RGB evidence [6,8,73,77] . These challenges motivate our central question: *Can we build an RGB-only navigation framework that achieves reliable closed-loop navigation using only unordered images, without relying on odometry, depth, or temporal priors?*

To address this question, we introduce ULVN, an unordered landmark navigation framework designed from a unified system optimization perspective. This framework ensures that mapping, localization, and planning are co-designed for consistency and reliability. For topological mapping, ULVN employs a vi-

sual place recognition and geometric verification scheme equipped with a one-shot, data-driven threshold calibration to enhance edge verification robustness. We further refine the graph by extracting a maximum spanning forest skeleton and reinserting strong loop closures to yield a reliable topological structure from unordered images. For localization, we propose a global belief propagation scheme that updates spatial probabilities through Bayesian estimation and entropy-based adaptive fusion, effectively mitigating localization drift under visual ambiguity. Finally, ULVN utilizes this global belief state to plan paths across the graph and dynamically replans when deviations occur. We comprehensively validate ULVN in both simulated and real-world environments. The main contributions of this work are as follows:

- **A unified framework for unordered landmark visual navigation.** We consider RGB-only navigation without temporal or multimodal priors, an important yet still underexplored setting, and provide a unified framework that integrates mapping, localization, and navigation to help control error accumulation.
- **A more reliable topological mapping from unordered RGB images.** To obtain a high-quality topological map, ULVN adopts a two-stage scheme that calibrates VPR-based candidates retrieval and geometric verification using a one-shot threshold estimation and then refines the graph through an MSF skeleton with loop reinsertion.
- **Odometry-free global localization and navigation on 2D graphs.** ULVN updates a global belief state through Bayesian estimation with entropy-based adaptive fusion to limit drift under visual ambiguity, and uses this belief to plan and re-plan paths on the topological graph.
- **Benchmark and comprehensive experimental validation.** We publicly release the unordered landmark visual navigation dataset in simulation and real-world environments, together with data-collection pipelines, and evaluation metrics to support reproducible research. We provide ablation studies on error accumulation across these components and show that ULVN achieves stronger performance than existing methods.

2 Related Work

Topological Map Construction. Early appearance-based visual navigation and SLAM used topological maps for efficiency and robustness, coupling visual place recognition (VPR) with probabilistic inference to mitigate perceptual aliasing [11, 38, 79]. This line of work also indicated that image-only representations can serve as a scalable alternative to dense metric reconstructions [15, 26, 50] in many settings. A large fraction of recent RGB-only systems assumes temporally ordered capture [46, 57, 62, 63], effectively yielding a one-dimensional chain of landmarks [20, 48, 61, 64] and focusing on local planners [36, 47, 49, 68, 77]. While practical, these designs rely on sequential structure and are incompatible with the *bag-of-images* setting where no temporal signal is available. When

sequences are absent, a 2D topological graph is typically built via VPR-based candidate retrieval followed by geometric verification [18,37]. Classical pipelines retrieve neighbors with local features and verify edges via epipolar or homography tests [53, 54], but are sensitive to hyperparameters, such as RANSAC inlier thresholds, which trade off edge precision and recall, ultimately affecting connectivity. Learned global descriptors improve retrieval invariance (viewpoint, illumination, season), yet descriptor similarity does not imply traversability; verification remains essential. Recent robust estimators [2, 45, 59, 71] mitigate, but do not remove, the reliance on thresholds and upper bounds across environments [14].

Localization and Planning on Topological Maps. Given a map, the agent must localize and plan a path. Many planners assume precise global information (e.g., depth or metric maps) [5, 40], which is often unavailable in RGB-only scenarios. A popular alternative is to estimate a "temporal distance" between the current and goal views [27, 35, 41, 51, 56, 58]. However, this proxy may correlate poorly with geodesic distance [4, 31, 42], often presumes near-constant velocity [29, 39, 80], and is expensive to evaluate over many candidates [64], while requiring robot-specific trajectory data for training [56, 64, 75]. Recent work reframes subgoal selection as VPR-based localization [9, 64], improving efficiency and robustness, yet many models implicitly retain 1D-topology assumptions [25, 64], which fail on general 2D graphs with intersections, branches, and loops. In contrast to such sequential assumptions, our approach localizes on 2D graphs without odometry by using a topological transition model defined by graph adjacency.

3 Methodology

To achieve image-goal navigation strictly from unordered RGB images, our framework, ULVN, adopts a purely topological approach, abandoning error-prone metric reconstruction and temporal heuristics. Assuming the provided image library contains sufficient visual overlap, ULVN systematically mitigates error accumulation through three tightly integrated modules: robust topological graph construction (RAVEL, Sec. 3.1), global state estimation (BPL, Sec. 3.2), and closed-loop visual planning (BASS, Sec. 3.3).

Problem Definition and Notation. Let $\mathcal{L} = \{I_i\}_{i=1}^{N}$ be an unordered set of collected images, I_g be the goal image, and I_t be the live observation at time t, with no odometry or temporal priors available. A global image encoder $f(\cdot)$ produces L2-normalized embeddings $\mathbf{z}_i^r = f(I_i)$, $\mathbf{z}_g = f(I_g)$, and $\mathbf{z}_t = f(I_t)$. ULVN constructs a directed, weighted graph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W}), \quad \mathcal{V} \subseteq \{1, \ldots, N\}, \quad \mathbf{W} = \{W_{ij}\}_{(i,j)\in\mathcal{E}} \tag{1}$$

Because metric distance is unavailable, we use visual overlap as a proxy for spatial proximity. Each edge $(i, j) \in \mathcal{E}$ has an integer weight $W_{ij} \in \mathbb{Z}_{>0}$ representing the number of verified local feature inliers.

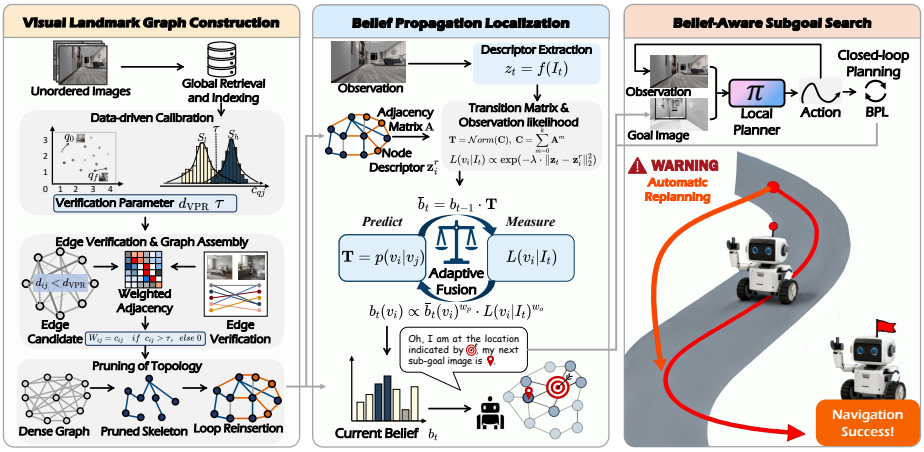

Fig. 2: Overview of Unordered Landmark Visual Navigation. (left) build a calibrated, pruned topology from feature retrieval over unordered images. (middle) embed the observation and update node belief with observation likelihoods and transition $\mathbf{T}$ via adaptive fusion. (right) pick the next landmark from the belief and plan in closed loop; belief shifts trigger automatic replanning to the goal.

3.1 Visual Landmark Graph Construction

We introduce RAVEL (Robust Augmentation and VErification of Landmarks), a pipeline that transforms $\mathcal{L}$ into a weighted, directed visual graph. Unordered collections naturally produce dense, redundant candidate pairs that can cause memory bloat and degrade downstream planning. RAVEL mitigates this by applying global retrieval for fast recall, followed by geometric verification as a high-precision bottleneck, and ultimately pruning the graph into a compact, structurally meaningful skeleton.

Global Retrieval and Indexing. To avoid exhaustive pairwise matching, we first extract an L2-normalized global descriptor $\mathbf{z}_i^r = f(I_i)$ for each image using an image encoder $f(\cdot)$. While $f(\cdot)$ can be a general-purpose backbone (e.g., ResNet, DINOv2), our experiments show that representations trained explicitly for Visual Place Recognition (VPR) yield the most robust retrieval candidates. All global descriptors are indexed using FAISS [13, 28].

To automatically calibrate verification thresholds, we select the first image q_0 and retrieve its farthest valid neighbor $q_f = \arg\max_{j \in \mathcal{N}_{\text{FAISS}}(q_0)} \|\mathbf{z}_{q_0} - \mathbf{z}_j\|_2^2$, explicitly excluding invalid index outputs. For each anchor $q \in \{q_0, q_f\}$, we perform full feature matching against all retrieved images, yielding an inlier count c_{qj} and anchor descriptor distance d_{qj} for each pair. Applying k mean clustering to pooled inlier counts yields a cluster of highest-confidence S_h and a cluster of second-highest-confidence S_l. The inlier threshold τ is defined as:

$$\tau = \frac{1}{2}\big(\min(S_h) + \max(S_l)\big). \tag{2}$$

The global distance threshold is derived from S_h: $d_{\text{VPR}} = \max_{(q,j)\in S_h} \|\mathbf{z}_q - \mathbf{z}_j\|_2$.

Local Geometric Verification and Graph Pruning. For every candidate pair where the global descriptor distance $s_{ij} \leq d_{\text{VPR}}$, we perform local geometric verification using a deep matcher (e.g., LightGlue [33]) to obtain the precise inlier match count c_{ij}. The weighted adjacency matrix is updated as:

$$W_{ij} = W_{ji} = \begin{cases} c_{ij}, & c_{ij} > \tau \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Raw graphs built from unordered collections are inherently dense and contain weak, redundant edges that not only cause memory bloat but also degrade the subsequent belief propagation by diffusing probability mass into noisy pathways. To extract a compact, reliable skeleton, we compute the Maximum Spanning Forest (MSF) $\mathcal{F} = (\mathcal{V}, \mathcal{E}_{\text{MSF}})$. By maximizing the total edge weight $\sum_{(i,j)\in\mathcal{E}_{\text{MSF}}} W_{ij}$, the MSF ensures the navigational backbone consists strictly of the strongest geometric transitions, minimizing local execution failures.

While the MSF effectively eliminates visual aliases, it destroys cyclic structures essential for alternate routing and topological error correction. To recover these, we introduce data-driven *Strong-Loop Reinsertion*. Instead of brittle heuristics, we cluster all edge weights in $\mathcal{G}$ into $k = 10$ clusters via k-means. We define a dynamic threshold τ_{add} as the minimum weight within the two highest-centroid clusters to ensure adaptability to different scenes. Reinserting excluded edges where $W_{ij} > \tau_{\text{add}}$ yields the final robust graph $\mathcal{G}_{\text{pruned}}$, seamlessly balancing structural reliability with critical topological redundancy.

3.2 Belief Propagation Localization

Building upon the foundations of classical Bayesian filtering [67], we extend the 1D sequential localization filters utilized in prior works [64, 72] to operate on arbitrary 2D spatial graphs with complex branching and loops. We introduce a Belief Propagation Localization (BPL) filter that maintains a belief distribution b_t over all nodes $v_i \in \mathcal{V}$ via a recursive prediction-correction cycle.

Topological Prediction Step. Let $\mathbf{A}$ be the binary adjacency matrix derived from $\mathcal{G}_{\text{pruned}}$. To model the robot's possible motion over multiple steps without odometry, we define a cumulative reachability matrix aggregating paths up to length $K = 3$:

$$\mathbf{C} = \sum_{m=0}^{K} \mathbf{A}^m, \tag{4}$$

where $\mathbf{A}^0 = \mathbf{I}$. Row-normalizing $\mathbf{C}$ yields a stochastic transition matrix $\mathbf{T}_{ij} = \mathbf{C}_{ij} / \sum_{j'} \mathbf{C}_{ij'}$. The predicted belief is $\bar{b}_t = b_{t-1} \cdot \mathbf{T}$.

Adaptive Observation Fusion. The observation likelihood is defined as $L(v_i|I_t) = \exp(-\lambda \cdot \|\mathbf{z}_t - \mathbf{z}_i^r\|_2^2)$ with $\lambda = 5$. We introduce an entropy-adaptive fusion mechanism. We quantify uncertainty using the normalized Shannon entropy of the prior belief:

$$H_n(b_{t-1}) = -\frac{1}{\log|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} b_{t-1}(v_i) \log b_{t-1}(v_i). \tag{5}$$

To dynamically balance prediction and observation, we set the fusion weights directly from the entropy such that $w_p = 1 - H_n(b_{t-1})$ and $w_o = H_n(b_{t-1})$, ensuring they strictly sum to 1. The posterior belief is computed via a normalized weighted geometric mean:

$$b_t(v_i) \propto \bar{b}_t(v_i)^{w_p} \cdot L(v_i|I_t)^{w_o}. \tag{6}$$

The estimated location is the Maximum A Posteriori (MAP) node: $\hat{v}_t = \arg\max_{v_i} b_t(v_i)$.

3.3 Belief-Aware Subgoal Search for Navigation

With the graph constructed and the localization filter initialized, Belief-Aware Subgoal Search (BASS) plans and executes paths entirely from images.

Task Initialization. A navigation task begins with a goal image I_g. Before planning, we must anchor the start and goal states on the graph $\mathcal{G}$. We establish the start node v_s using the maximum-a-posteriori estimate of the initial global belief b_0: $v_s = \arg\max_{v_i \in \mathcal{V}} b_0(v_i)$. The goal node v_g maximizes descriptor similarity: $v_g = \arg\max_{v_i \in \mathcal{V}} (\mathbf{z}_g^\top \mathbf{z}_i^r)$.

Topological Planning. We treat $\mathcal{G}$ as a weighted graph where edge weights W_{ij} reflect visual overlap. Because a visual navigation path is only as robust as its weakest visual link, we define the confidence of a candidate path $\mathcal{P}$ on the graph $\mathcal{G}$ from v_s to v_g as its minimum edge weight:

$$\text{conf}(\mathcal{P}) = \min_{(v_i, v_j) \in \mathcal{P}} W_{ij}. \tag{7}$$

BASS extracts the widest path $\mathcal{P}^* = \arg\max_{\mathcal{P}} \text{conf}(\mathcal{P})$ with the Dijkstra's algorithm . This max-min objective guarantees that the chosen sequence of visual subgoals $(v_s, v_1, \ldots, v_g)$ has the highest probability of successful image-to-image local control.

Closed-Loop Vision-Only Execution. Our framework operates strictly without odometry. The robot follows $\mathcal{P}^*$ by sequentially passing the planned subgoals to an image-based local planner (e.g., a visual foundation model like ViNT [58] or a generic visual servoing policy). At each step, the local planner takes the observation I_t and the current topological subgoal image I_{s_n} and directly outputs low-level velocity commands to drive the robot toward the target.

During execution, BPL continuously updates the belief b_t. A deviation is detected if the MAP node v_t falls off the planned path $\mathcal{P}^*$ or if the graph distance from v_t to the current subgoal exceeds $D_{\text{thres}} = 3$. Upon deviation, BASS dynamically replans a new widest path from v_t to v_g. Navigation succeeds when the belief converges on the goal node with a probability exceeding 0.5.

4 Experiments

4.1 Experimental Settings

Evaluation Scenes. We evaluate our method, ULVN, within the NVIDIA Isaac Sim simulator using the GRScenes dataset [69], a high-fidelity benchmark for

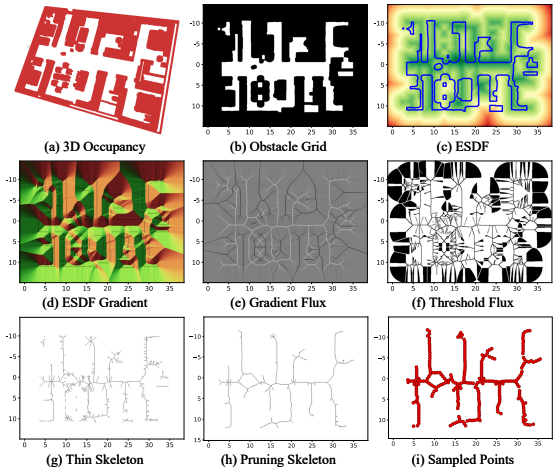

Fig. 3: The topological skeleton extraction in data collection pipeline. It starts with (a) a 3D occupancy map, which is projected into (b) a 2D binary obstacle grid. (c) A Euclidean Signed Distance Function (ESDF) and (d) its gradient. (e) The ESDF gradient flux is then calculated, (f) thresholded to isolate the medial axis, (g) thinned, (h) pruned to produce the clean topological skeleton, and finally (i) indicates the node of acquired images.

general-purpose robotics. We utilize 10 distinct scenes, equally split between *home* and *commercial* environments. All scenes feature realistic physics and materials. We also employ the CARLA [12] simulator to specifically isolate and validate the RAVEL and BPL modules under varying perceptual conditions.

Tasks and Metrics. We evaluate ULVN across three core capabilities: 1) *Topomap Construction:* We report Precision (P), Recall (R), F1-score ($F1$), and Accuracy ($Acc.$) of the generated graph edges evaluated against the ground-truth spatial connectivity (derived via Sec. 4.2). We also analyze the Pearson correlation (r) between global descriptor similarity and true geometric inliers to justify our retrieval design. 2) *Localization:* We report the localization Accuracy (success rate of the MAP estimate correctly identifying the nearest topological node) when localizing a continuous stream of noisy images against the pre-built, non-sequential graph. 3) *Navigation:* We adopt standard Habitat ImageNav metrics [52]: Success Rate (SR), Success weighted by Path Length (SPL), and average number of collisions.

Baselines. We compare against strong state-of-the-art representations and systems. For *Topomap Construction*, we evaluate ResNet-50 [22], DINOv2 [44], MegaLoc [3], ViNT [58], and PlaceNav [64]. Because ViNT and PlaceNav traditionally rely on temporally ordered videos, we adapt them for unordered mapping by using their temporal distance predictions or VPR retrieval scores as direct edge-weight proxies. For *Localization*, we compare against single-frame VPR (MegaLoc, PlaceNav), temporal distance (ViNT), and sequence smooth-

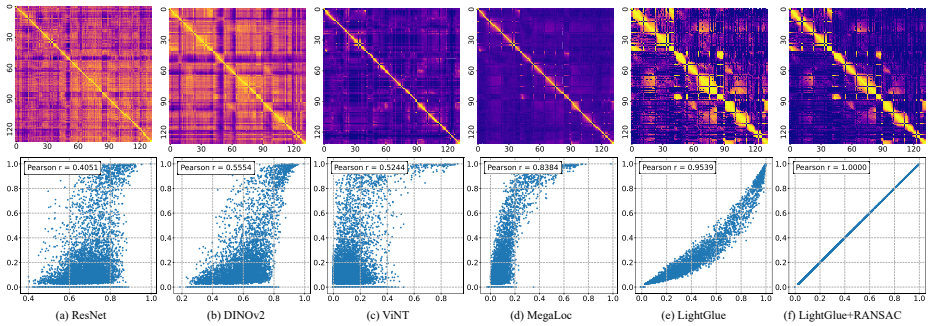

Fig. 4: Global Descriptors as Efficient Proxies for Geometric Verification. We compare pairwise similarity scores from various global descriptors: (a) ResNet, (b) DINOv2, (c) ViNT (temporal distance), and (d) MegaLoc (VPR-trained), against a ground-truth geometric consistency score from (f) LightGlue+RANSAC. Top row: Pairwise affinity matrices for an unordered image set. Bottom row: Correlation plots (with Pearson's r) for each descriptor's score (Y-axis) versus the ground truth (X-axis). The VPR-trained MegaLoc (d) demonstrates a strong linear correlation, validating it as an effective and efficient proxy for geometric verification.

Table 1: Comparison of different retrieval baselines and our method on GRScenes.

Method	P	R	F1	Acc.
VGGT [70]	0.1599	0.1659	0.1629	0.9931
PlaceNav top 2 [64]	0.5262	0.7113	0.6043	0.9948
PlaceNav top 5 [64]	0.2699	0.7655	0.3731	0.9853
ViNT top 2 [58]	0.5201	0.6775	0.5815	0.9946
ViNT top 5 [58]	0.2660	**0.8246**	0.3806	0.9816
RAVEL (Ours)	**0.7104**	0.7656	**0.7365**	**0.9970**

ing (JIST [4]). For *Navigation*, we evaluate the integration of our framework with visual local planners (ViNT, NoMaD) and compare with other SOTA methods.

Implementation Details. Global descriptors are extracted using MegaLoc [3] with L2-normalization, and geometric verification is performed via LightGlue [33]. During RAVEL calibration, if the high-confidence cluster S_h is valid ($k = 2$), τ and d_{VPR} are dynamically set from its statistics; otherwise, we default to $\tau_{\text{default}} = 15$ and $d_{\text{VPR,default}} = 1.7$. The MSF is computed using Kruskal's algorithm, with strong-loop reinsertion bounded by $\tau_{\text{add}} = 1.5\tau$. For BPL, the observation likelihood scaling factor is $\lambda = 10$.

4.2 Data Collection and Ground Truth Pipeline

To rigorously evaluate Topomap metrics, ground-truth spatial connectivity is required. As shown in Fig. 3, we developed an automated in-simulation pipeline that exports 3D scene occupancy and converts it into a 2D traversability grid based on the robot's dimensions. Following [32], we use the Euclidean Signed Distance Field (ESDF) [43] to extract the topological skeleton of the traversable area. By applying non-maximum suppression along this skeleton, we sample sparse and spatially uniform nodes. The underlying skeletal structure provides

Table 2: Quantitative evaluation of topological graph construction on GRScenes. Ablations compare our RAVEL with a Top-k ANN retrieval baseline and the variant ablation of our method.

Method	P	R	F1	Acc.
Top-k ANN	0.1245	**0.9121**	0.2156	0.9584
RAVEL w/o (MSF, AVP τ)	0.3676	0.7177	0.4496	0.9898
RAVEL w/o MSF	0.6910	0.4220	0.5157	0.9956
RAVEL (Ours)	**0.7104**	0.7656	**0.7365**	**0.9970**

Table 3: Robustness of RAVEL

Method	Precision	Recall	F1-Score	Accuracy
RAVEL+noise	0.6703 (-5.64%)	0.7072 (-7.63%)	**0.6882 (-6.56%)**	**0.9963 (-0.07%)**
CosPlace top 2+noise	0.4592 (-12.73%)	0.6615 (-7.00%)	0.5420 (-10.29%)	0.9936 (-0.12%)
CosPlace top 5+noise	0.2192 (-18.78%)	0.7739 (+1.10%)	0.3416 (-8.44%)	0.9826 (-0.27%)
ViNT top 2+noise	0.4197 (-19.30%)	0.6064 (-10.49%)	0.4961 (-15.70%)	0.9913 (-0.33%)
ViNT top 5+noise	0.2065 (-22.36%)	0.7976 (-3.27%)	0.3281 (-18.43%)	0.9769 (-0.48%)

the absolute ground-truth connectivity matrix used to evaluate the precision and recall of RAVEL's generated graphs. Further details are provided in the Appendix.

4.3 Evaluation of RAVEL

Global Descriptors as Proxies for Geometric Verification. We first assess the viability of using global descriptors as a fast-recall proxy for costly geometric verification. Using CARLA data, we analyze the Pearson correlation (r) between descriptor similarity and the ground-truth geometric score (LightGlue + RANSAC inliers). As shown in Figure 4, general-purpose features like ResNet (a) and DINOv2 (b) exhibit weak, non-linear correlations, leading to severe perceptual aliasing and false-positive retrievals. Similarly, ViNT's temporal distance (c) fails to reliably map to geometric overlap in unordered settings. Conversely, the VPR-trained MegaLoc (d) demonstrates a strong, tightly-clustered linear correlation with geometric reality. This confirms that VPR is strictly necessary for the initial retrieval stage to filter candidates effectively without discarding valid topological neighbors.

Mapping Performance Evaluation. Table 1 and Table 2 detail the graph construction performance across 10 GRScenes. Retrieval-only baselines (CosPlace, ViNT) highlight a fundamental tension: increasing the Top-k threshold boosts recall but decimates precision, yielding poor F1 scores. This indicates that raw appearance similarity cannot substitute for structural verification.

Our ablation (Table 2) reveals the mechanics of our solution. The adaptive-threshold variant (*Ours w/o MSF*) achieves high precision by pruning weak visual matches, but suffers from low recall because purely local edge rejection fragments the graph. Applying the MSF resolves this. By shifting the objective from isolated edge evaluation to global structural coherence, the MSF reconstructs a unified skeleton. This allows RAVEL to aggressively filter false pos-

Table 4: Quantitative comparison of global localization under Global and Difficult Path scenarios.

Scenario	Metric	Ours	MegaLoc [3]	ViNT [58]	JIST [4]
All	Nodes	976	976	976	976
	Success	**932**	889	845	829
	Fail	**44**	87	131	147
	Acc.(%)	**95.49**	91.09	86.58	84.94
Difficult Path	Nodes	383	383	383	383
	Success	**360**	341	270	315
	Fail	**23**	42	113	68
	Acc.(%)	**93.99**	89.03	70.50	82.25

Table 5: Comparison of localization accuracy of different methods under different conditions. Results are reported as mean ± standard deviation across four datasets.

Condition	Localization Accuracy			
	MegaLoc [3]	**JIST** [4]	**ViNT** [58]	**BPL (Ours)**
Rotation	0.913 ± 0.104	0.722 ± 0.066	0.895 ± 0.044	**0.966 ± 0.020**
Rot + Gauss	0.743 ± 0.088	0.452 ± 0.177	0.885 ± 0.045	**0.914 ± 0.038**
Rot + Poisson	0.717 ± 0.114	0.445 ± 0.176	0.888 ± 0.037	**0.896 ± 0.036**
Rot + Crop	0.855 ± 0.151	0.525 ± 0.076	0.640 ± 0.096	**0.945 ± 0.027**
Average	0.807 ± 0.133	0.536 ± 0.167	0.827 ± 0.124	**0.930 ± 0.040**

itives while maintaining high recall, ultimately achieving the highest F1-score and demonstrating robust generalization across diverse scenes.

Robustness to Visual Perturbations. To evaluate the robustness of our graph construction against real-world sensor imperfections, we introduce a *+noise* setting (Table 3). We apply a suite of random visual perturbations to the images prior to mapping. This includes lighting variations (random brightness scaling $\in [0.75, 1.25]$, RGB color shifting $\pm 5\%$, contrast adjustments $\in [0.85, 1.15]$, and low-light Gaussian noise with $\sigma \in [3, 10]$ triggered when brightness drops below 0.9) and directional motion blur (random kernel lengths $\in \{3, 5, 7, 9, 11\}$ at angles $\in \{0°, 45°, 90°, 135°\}$).

As shown in Table 3, purely retrieval-based methods with VPR (CosPlace) and temporal distance (ViNT) are highly sensitive to these pixel-level perturbations, suffering massive drops in precision and F1-score. This confirms that appearance-based distances alone degrade rapidly under environmental noise. In contrast, RAVEL demonstrates remarkable resilience, experiencing only a 6.56% drop in F1-score. While the noise inevitably reduces the raw number of local feature inliers, RAVEL's geometric verification and the MSF's global structural constraints successfully prevent the corrupted images from forming spurious edges, thereby preserving a high accuracy and structural integrity.

4.4 Evaluation of BPL

To evaluate localization, we simulate robot motion by collecting continuous image streams along random trajectories in GRScenes. Crucially, these test images

Table 6: Ablation Study for Belief Propagation Localization

k	1	2	3	4	5	1
Adaptive Fusion	✓	✓	✓	✓	✓	✗
Acc. (%)	95.49	94.47	93.65	92.42	92.73	90.57

are captured at novel poses distinct from the reference nodes in the topological graph. We introduce a *Difficult Path* subset, characterized by long routes, large-angle turns, and low start-to-goal visual overlap.

Comparison with Other Paradigms. As shown in Table 4, BPL delivers the highest accuracy. While baselines degrade noticeably on complex routes, BPL's performance remains stable. 1D temporal-distance methods like ViNT fail significantly on challenging paths with intersections or loops. BPL's robustness stems from its graph-derived transition matrix, which seamlessly diffuses belief over multi-hop neighborhoods, allowing probability mass to branch at junctions and reconverge during loop closures.

Ablation for BPL. Table 6 analyzes BPL's internal mechanisms. Modifying the propagation depth (k) shows that performance is relatively stable, peaking around 2–3 hops; excessive diffusion over-smooths the belief and slightly increases errors. More importantly, disabling the entropy-adaptive fusion causes a severe accuracy drop from 0.9549 to 0.9057. This proves that while structural diffusion is helpful, dynamically weighting the prediction and observation based on current uncertainty is the critical factor for resolving visual ambiguity.

Robustness to Perceptual Degradation. To evaluate the resilience of BPL against real-world sensor noise and viewpoint shifts during live execution, we test localization accuracy under severe image perturbations (Table 5). We use continuous trajectories from each of four diverse datasets: RECON [55], SCAND [30], GoStanford [24], and SACSoN [23]. The live query images are subjected to four degradation conditions: pure angular viewpoint shift (Rotation), and rotation compounded with Gaussian noise, Poisson noise, or random structural cropping.

As shown in Table 5, BPL achieves the highest and most stable localization accuracy (averaging 0.930 ± 0.040) across all degradation conditions. While pure VPR (MegaLoc), sequence smoothing (JIST), and temporal models (ViNT) exhibit severe vulnerability to compounded pixel noise or structural occlusions, BPL consistently maintains an accuracy above 0.89. This resilience validates our entropy-adaptive fusion design: when severe perceptual degradation flattens the observation likelihood, the filter dynamically shifts its reliance to the topological prediction, leveraging the multi-hop spatial prior to coast through temporary visual disruptions without losing track.

4.5 Evaluation of BASS

We evaluate BASS on random trajectories across GRScenes. Figures 5 and 7(a-d) demonstrate our closed-loop execution: the robot smoothly tracks subgoals in simple layouts, and dynamically replans recovery routes when BPL detects physical deviations exceeding D_{thres} in complex topologies. Furthermore, we compare

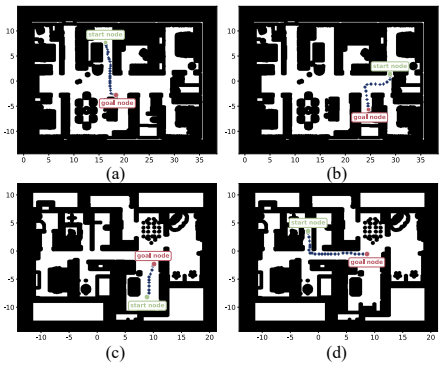

Fig. 5: BASS navigation in easy (a,c) and hard (b,d) tasks. Green: start; red: goal; blue: executed trajectory.

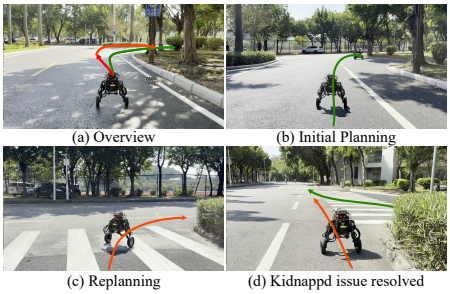

Fig. 6: Real-world case. (a) Overview. (b) Initial plan. (c) Replanning after deviation. (d) Kidnapped-robot case resolved via relocalization and replanning. Green: planned path; red: observed deviation; orange: replanned path.

Table 7: Comparison of Different Algorithms on GRScenes. "**-A**" means retraining with only the action head loss function. "**w.** d_{temp}" denotes that localization with temporal distance instead of BPL.

Method	**SR** (%)	**Avg. Colli.**	**Avg. SPL**
Uni-Navid [76]	32.0	1.96	0.2391
UniGoal [74]	61.6	0.88	0.3176
ULVN+VINT-A [58]	31.0	1.13	0.812
ULVN+NoMaD-A [61]	54.3	0.94	0.7458
ULVN+ViNT w. d_{temp} [58]	59.6	0.88	0.7752
ULVN+ViNT [58]	68.1	0.76	**0.8398**
ULVN+NoMaD [61]	**71.9**	**0.42**	0.7978

ULVN against Uni-Navid, a state-of-the-art end-to-end model (Fig. 7e). While Uni-Navid's purely reactive, egocentric navigation suffers severe trajectory oscillations, ULVN executes a smooth, efficient trajectory guided by its topological graph. This confirms that lightweight structural memory is crucial for mitigating the myopic planning inherent in purely reactive methods.

Table 7 quantitatively compares our framework against end-to-end baselines and ablates local planner integrations. ULVN significantly outpaces purely reactive models like Uni-Navid and UniGoal, achieving a peak Success Rate (SR) of 71.9% with NoMaD and maintaining a substantially higher SPL. We also ablate the temporal distance loss during the local planner's training (denoted as *-A*). Removing this loss severely degrades performance, particularly for the regression-based ViNT, whereas the generative NoMaD is comparatively more robust. This reveals that temporal distance acts as a crucial auxiliary training signal. However, replacing our BPL filter with temporal-distance-based localization during execution (*ULVN+ViNT w.* d_{temp}) reduces SR to 59.6%, proving that while temporal distance benefits local control training, our VPR-based BPL is strictly necessary for reliable global localization.

We observe a clear performance gap between localization accuracy (∼95%) and navigation Success Rate (∼71%). This gap is primarily due to the physical

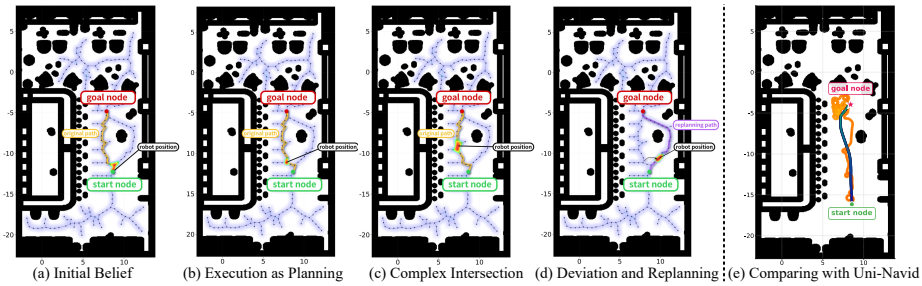

Fig. 7: Visualization of replanning and comparison with Uni-Navid.

limitations of the local planners. Specifically, limited obstacle perception and oscillatory trajectory generation, which can occasionally cause the robot to fail despite possessing the correct topological subgoal.

4.6 Real-world Evaluation

To validate sim-to-real transfer, we deployed ULVN on a Diablo wheeled robot equipped with an Azure Kinect camera and NVIDIA Jetson Orin. Figure 6 illustrates a representative trial. The initial planned route required a right turn, but physical disturbances caused the robot to drift off-path. Because BPL continuously tracks the observation against the graph, the entropy-adaptive filter concentrates belief on an off-route node. Once the topological distance between the MAP node and the target subgoal exceeded our deviation threshold, BASS successfully triggered a real-time replanning event, issuing new subgoals that safely guided the robot to the destination. This confirms the framework's practical viability for robust, odometry-free physical deployment.

5 Conclusion

We present ULVN, an image-goal navigation framework operating entirely from unordered RGB image collections, eliminating the need for odometry or temporal priors. To resolve severe perceptual aliasing, our RAVEL pipeline constructs a robust topological skeleton by reconciling local geometric verification with global structural coherence. On this graph, our Belief Propagation Localization (BPL) filter employs entropy-adaptive fusion to maintain reliable tracking despite severe visual noise, structural occlusions, and complex loops. Supported by this state estimation, our Belief-Aware Subgoal Search (BASS) enables consistent closed-loop execution with dynamic drift recovery. Extensive sim-to-real evaluations demonstrate that our mapping-then-navigating paradigm is highly resilient to perceptual degradation and overcomes the myopic trajectory oscillations inherent in purely reactive end-to-end models. Ultimately, ULVN validates lightweight topological memory as a highly scalable, robust solution for generalizable navigation in unstructured environments.

Acknowledgement.

This work was supported by the National Natural Science Foundation of China (U22A2095).

References

1. Anderson, P., Chang, A., Chaplot, D.S., Dosovitskiy, A., Gupta, S., Koltun, V., Kosecka, J., Malik, J., Mottaghi, R., Savva, M., et al.: On evaluation of embodied navigation agents. arXiv preprint arXiv:1807.06757 (2018)
2. Barath, D., Mishkin, D., Cavalli, L., Sarlin, P.E., Hruby, P., Pollefeys, M.: Stereoglue: Robust estimation with single-point solvers. In: European Conference on Computer Vision. pp. 421–441. Springer (2024)
3. Berton, G., Masone, C.: Megaloc: One retrieval to place them all. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2861–2867 (2025)
4. Berton, G., Trivigno, G., Caputo, B., Masone, C.: Jist: Joint image and sequence training for sequential visual place recognition. IEEE Robotics and Automation Letters **9**(2), 1310–1317 (2023)
5. Blochliger, F., Fehr, M., Dymczyk, M., Schneider, T., Siegwart, R.: Topomap: Topological mapping and navigation based on visual slam maps. In: 2018 IEEE International Conference on Robotics and Automation (ICRA). pp. 3818–3825. IEEE (2018)
6. Bonin-Font, F., Ortiz, A., Oliver, G.: Visual navigation for mobile robots: A survey. Journal of intelligent and robotic systems **53**, 263–296 (2008)
7. Booij, O., Terwijn, B., Zivkovic, Z., Krose, B.: Navigation using an appearance based topological map. In: Proceedings 2007 IEEE International Conference on Robotics and Automation. pp. 3927–3932. IEEE (2007)
8. Cadena, C., Carlone, L., Carrillo, H., Latif, Y., Scaramuzza, D., Neira, J., Reid, I., Leonard, J.J.: Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on robotics **32**(6), 1309–1332 (2016)
9. Claxton, O., Malone, C., Carson, H., Ford, J.J., Bolton, G., Shames, I., Milford, M.: Improving visual place recognition based robot navigation by verifying localization estimates. IEEE Robotics and Automation Letters (2024)
10. Cui, X., Liu, Q., Liu, Z., Wang, H.: Frontier-enhanced topological memory with improved exploration awareness for embodied visual navigation. In: European Conference on Computer Vision. pp. 296–313. Springer (2024)
11. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. The International journal of robotics research **27**(6), 647–665 (2008)
12. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
13. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)
14. Edstedt, J.: Less biased noise scale estimation for threshold-robust ransac. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2811–2820 (2025)

15. Fan, S., Liu, R., Wang, W., Yang, Y.: Navigation instruction generation with bev perception and large language models. In: European Conference on Computer Vision. pp. 368–387. Springer (2024)
16. Furgale, P., Barfoot, T.D.: Visual teach and repeat for long-range rover autonomy. Journal of field robotics **27**(5), 534–560 (2010)
17. Gao, C., Liu, S., Chen, J., Wang, L., Wu, Q., Li, B., Tian, Q.: Room-object entity prompting and reasoning for embodied referring expression. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(2), 994–1010 (2023)
18. Garg, S., Fischer, T., Milford, M.: Where is your place, visual place recognition? In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21). pp. 4416–4425. International Joint Conferences on Artificial Intelligence (2021)
19. Garg, S., Rana, K., Hosseinzadeh, M., Mares, L., Sünderhauf, N., Dayoub, F., Reid, I.: Robohop: Segment-based topological map representation for open-world visual navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 4090–4097. IEEE (2024)
20. Gode, S., et al.: Flownav: Learning efficient navigation policies via conditional flow matching. arXiv (2024)
21. Gornet, J., Thomson, M.: Automated construction of cognitive maps with visual predictive coding. Nature Machine Intelligence **6**(7), 820–833 (2024)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Hirose, N., Shah, D., Sridhar, A., Levine, S.: Sacson: Scalable autonomous control for social navigation. IEEE Robotics and Automation Letters (2023)
24. Hirose, N., Xia, F., Martín-Martín, R., Sadeghian, A., Savarese, S.: Deep visual mpc-policy learning for navigation. IEEE Robotics and Automation Letters **4**(4), 3184–3191 (2019)
25. Hossain, J., Faridee, A.Z., Roy, N., Freeman, J., Gregory, T., Trout, T.: Toponav: Topological navigation for efficient exploration in sparse reward environments. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 693–700. IEEE (2024)
26. Irschara, A., Zach, C., Frahm, J.M., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2599–2606. IEEE (2009)
27. Jiang, Y., Liu, Q., Yang, Y., Ma, X., Zhong, D., Hu, H., Yang, J., Liang, B., XU, B., Zhang, C., et al.: Episodic novelty through temporal distance. In: The Thirteenth International Conference on Learning Representations (2025)
28. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with gpus. IEEE Transactions on Big Data **7**(3), 535–547 (2019)
29. Kamila, S., Hasanuzzaman, M., Ekbal, A., Bhattacharyya, P.: Measuring temporal distance focus from tweets and investigating its association with psychodemographic attributes. IEEE Transactions on Affective Computing **13**(2), 1086–1097 (2020)
30. Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., Stone, P.: Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. IEEE Robotics and Automation Letters (2022)
31. Li, H., Peng, G., Zhang, J., Wen, M., Ma, Y., Wang, D.: Casevpr: Correlation-aware sequential embedding for sequence-to-frame visual place recognition. IEEE Robotics and Automation Letters (2025)

32. Li, Z., Zheng, K., Yuan, Y., Huang, J., Zhang, X., Wu, J., Cheng, H.: Learning to explore efficiently: Heterogeneous topological graphs and lightweight global reasoning for robotic exploration. IEEE Robotics and Automation Letters (2025)
33. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 17627–17638 (2023)
34. Liu, P., Orru, Y., Paxton, C., Shafiullah, N.M.M., Pinto, L.: Ok-robot: What really matters in integrating open-knowledge models for robotics. Proceedings of Robotics: Science and Systems (RSS) (2024)
35. Ma, Y.J., Sodhani, S., Jayaraman, D., Bastani, O., Kumar, V., Zhang, A.: Vip: Towards universal visual reward and representation via value-implicit pre-training. arXiv preprint arXiv:2210.00030 (2022)
36. Majumdar, A., Aggarwal, G., Devnani, B., Hoffman, J., Batra, D.: Zson: Zero-shot object-goal navigation using multimodal goal embeddings. Advances in Neural Information Processing Systems **35**, 32340–32352 (2022)
37. Miao, J., Jiang, K., Wen, T., Wang, Y., Jia, P., Wijaya, B., Zhao, X., Cheng, Q., Xiao, Z., Huang, J., Zhong, Z., Yang, D.: A survey on monocular re-localization: From the perspective of scene map representation. IEEE Transactions on Intelligent Vehicles **10**(4), 2519–2550 (2025)
38. Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE international conference on robotics and automation. pp. 1643–1649. IEEE (2012)
39. Montano-Oliván, L., Placed, J.A., Montano, L., Lázaro, M.T.: G-loc: Tightly-coupled graph localization with prior topo-metric information. IEEE Robotics and Automation Letters (2024)
40. Muravyev, K., Melekhin, A., Yudin, D., Yakovlev, K.: Prism-topomap: online topological mapping with place recognition and scan matching. IEEE Robotics and Automation Letters (2025)
41. Myers, V., Zheng, C., Dragan, A., Levine, S., Eysenbach, B.: Learning temporal distances: Contrastive successor features can provide a metric structure for decision-making. In: International Conference on Machine Learning. pp. 37076–37096. PMLR (2024)
42. Neubert, P., Schubert, S., Protzel, P.: A neurologically inspired sequence processing model for mobile robot place recognition. IEEE Robotics and Automation Letters **4**(4), 3200–3207 (2019)
43. Noël, T., Lehuger, A., Marchand, E., Chaumette, F.: Skeleton disk-graph roadmap: A sparse deterministic roadmap for safe 2d navigation and exploration. IEEE Robotics and Automation Letters **9**(1), 555–562 (2023)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
45. Piedade, V., Miraldo, P.: Bansac: A dynamic bayesian network for adaptive sample consensus. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3738–3747 (2023)
46. Ren, H., Bi, Z., Zeng, Y., Wan, Z., Qi, L., Cheng, H.: Strnet: Visual navigation with spatio-temporal representation through dynamic graph aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42464–42473 (2026)
47. Ren, H., Bi, Z., Zeng, Y., Zheng, L., Li, Z., Wan, Z., Qi, L., Cheng, H.: Fisher-preserving guidance: Training-free manifold constraints for safe diffusion control. arXiv preprint arXiv:2605.29937 (2026)

48. Ren, H., Zeng, Y., Bi, Z., Wan, Z., Huang, J., Cheng, H.: Prior does matter: Visual navigation via denoising diffusion bridge models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12100–12110 (2025)
49. Roth, P., Nubert, J., Yang, F., Mittal, M., Hutter, M.: Viplanner: Visual semantic imperative learning for local navigation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5243–5249. IEEE (2024)
50. Sattler, T., Leibe, B., Kobbelt, L.: Efficient & effective prioritized matching for large-scale image-based localization. IEEE transactions on pattern analysis and machine intelligence **39**(9), 1744–1756 (2016)
51. Savinov, N., Dosovitskiy, A., Koltun, V.: Semi-parametric topological memory for navigation. In: International Conference on Learning Representations (2018)
52. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9339–9347 (2019)
53. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
54. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European conference on computer vision. pp. 501–518. Springer (2016)
55. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Rapid exploration for open-world navigation with latent goal models. arXiv preprint arXiv:2104.05859 (2021)
56. Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., Levine, S.: Ving: Learning open-world navigation with visual goals. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13215–13222. IEEE (2021)
57. Shah, D., Sridhar, A., Bhorkar, A., Hirose, N., Levine, S.: Gnm: A general navigation model to drive any robot. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 7226–7233. IEEE (2023)
58. Shah, D., Sridhar, A., Dashora, N., Stachowicz, K., Black, K., Hirose, N., Levine, S.: Vint: A foundation model for visual navigation. arXiv preprint arXiv:2306.14846 (2023)
59. Shi, P., Yan, S., Xiao, Y., Liu, X., Zhang, Y., Li, J.: Ransac back to sota: A two-stage consensus filtering for real-time 3d registration. IEEE Robotics and Automation Letters **9**(12), 11881–11888 (2024)
60. Song, M., Zhang, D., Ren, H., Zhang, R., Du, B., Yang, M.H., Qi, L.: Unisharp: Universal sharp monocular view synthesis. arXiv preprint arXiv:2606.07514 (2026)
61. Sridhar, A., Shah, D., Glossop, C., Levine, S.: Nomad: Goal masked diffusion policies for navigation and exploration. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 63–70. IEEE (2024)
62. Sun, X., Liu, L., Zhi, H., Qiu, R., Liang, J.: Prioritized semantic learning for zero-shot instance navigation. In: European Conference on Computer Vision. pp. 161–178. Springer (2024)
63. Suomela, L., Arachchige, S.K., Torres, G.F., Edelman, H., Kämäräinen, J.K.: Synthetic vs. real training data for visual navigation. arXiv preprint arXiv:2509.11791 (2025)
64. Suomela, L., Kalliola, J., Edelman, H., Kämäräinen, J.K.: Placenav: Topological navigation through place recognition. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5205–5213. IEEE (2024)

65. Thoma, J., Paudel, D.P., Chhatkuli, A., Probst, T., Gool, L.V.: Mapping, localization and path planning for image-based navigation using visual features and map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7383–7391 (2019)
66. Thrun, S.: Learning metric-topological maps for indoor mobile robot navigation. Artificial intelligence **99**(1), 21–71 (1998)
67. Thrun, S.: Probabilistic robotics. Communications of the ACM **45**(3), 52–57 (2002)
68. Wan, Z., Bi, Z., Zhou, Z., Ren, H., Zeng, Y., Li, Y., Qi, L., Yang, X., Yang, M.H., Cheng, H.: Rapid hand: A robust, affordable, perception-integrated, dexterous manipulation platform for generalist robot autonomy. arXiv preprint arXiv:2506.07490 (2025)
69. Wang, H., Chen, J., Huang, W., Ben, Q., Wang, T., Mi, B., Huang, T., Zhao, S., Chen, Y., Yang, S., et al.: Grutopia: Dream general robots in a city at scale. arXiv preprint arXiv:2407.10943 (2024)
70. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025)
71. Wei, T., Patel, Y., Shekhovtsov, A., Matas, J., Barath, D.: Generalized differentiable ransac. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17649–17660 (2023)
72. Xu, M., Fischer, T., Sünderhauf, N., Milford, M.: Probabilistic appearance-invariant topometric localization with new place awareness. IEEE Robotics and Automation Letters **6**(4), 6985–6992 (2021)
73. Yasuda, Y.D., Martins, L.E.G., Cappabianco, F.A.: Autonomous visual navigation for mobile robots: A systematic literature review. ACM Computing Surveys (CSUR) **53**(1), 1–34 (2020)
74. Yin, H., Xu, X., Zhao, L., Wang, Z., Zhou, J., Lu, J.: Unigoal: Towards universal zero-shot goal-oriented navigation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19057–19066 (2025)
75. Zeng, Y., Ren, H., Wang, S., Huang, J., Cheng, H.: Navidiffusor: Cost-guided diffusion model for visual navigation. arXiv preprint (2025)
76. Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 (2024)
77. Zhang, T., Hu, X., Xiao, J., Zhang, G.: A survey of visual navigation: From geometry to embodied ai. Engineering Applications of Artificial Intelligence **114**, 105036 (2022)
78. Zhang, W., Hara, Y., Nakamura, S.: Topological mapping and navigation using a monocular camera based on anyloc. In: 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE). pp. 3077–3083. IEEE (2025)
79. Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. Pattern Recognition **113**, 107760 (2021)
80. Zhao, J., Zhang, F., Cai, Y., Tian, G., Mu, W., Ye, C., Feng, T.: Learning sequence descriptor based on spatio-temporal attention for visual place recognition. IEEE Robotics and Automation Letters **9**(3), 2351–2358 (2024)
81. Zhu, Y., Mottaghi, R., Kolve, E., Lim, J.J., Gupta, A., Fei-Fei, L., Farhadi, A.: Target-driven visual navigation in indoor scenes using deep reinforcement learning. In: 2017 IEEE international conference on robotics and automation (ICRA). pp. 3357–3364. IEEE (2017)