

# PlaceNav: Topological Navigation through Place Recognition

Lauri Suomela<sup>1</sup>, Jussi Kalliola, Harry Edelman and Joni-Kristian Kämäräinen

**Abstract**—Recent results suggest that splitting topological navigation into robot-independent and robot-specific components improves navigation performance by enabling the robot-independent part to be trained with data collected by robots of different types. However, the navigation methods’ performance is still limited by the scarcity of suitable training data and they suffer from poor computational scaling. In this work, we present PlaceNav, subdividing the robot-independent part into navigation-specific and generic computer vision components. We utilize visual place recognition for the subgoal selection of the topological navigation pipeline. This makes subgoal selection more efficient and enables leveraging large-scale datasets from non-robotics sources, increasing training data availability. Bayesian filtering, enabled by place recognition, further improves navigation performance by increasing the temporal consistency of subgoals. Our experimental results verify the design and the new method obtains a 76 % higher success rate in indoor and 23 % higher in outdoor navigation tasks with higher computational efficiency.

## I. INTRODUCTION

Autonomous visual navigation is a well-studied problem in the field of robotics [1, 2]. One line of research frames navigation in known environments as *topological navigation* [3–5], meaning purely vision-based navigation between nodes of a topological map that are represented by images. The advantage of this approach is that it does not require building a geometric reconstruction of the operating environment [6] or training environment-specific control policies [7].

Topological navigation systems have two parts: *subgoal selection* and a *goal-reaching policy*. First, the subgoal selection module chooses the map node to reach next as a subgoal. Then, the goal-reaching policy produces control commands to take the robot to the selected subgoal. A popular approach to subgoal selection utilizes *temporal distance prediction*, which means predicting the number of time steps between the robot’s current camera observation and the subgoal candidates [8–13]. These learning-based models are trained with offline datasets of robot trajectories.

Previous works have demonstrated impressive real-world navigation [10, 12, 13], but the temporal distance prediction approach has two significant shortcomings. First, the distance has to be estimated individually for each subgoal candidate, incurring complexity that scales at  $\mathcal{O}(n)$  with the number of candidate images. This requires heuristics to limit the candidates and constrains the methods available for ensuring subgoal temporal consistency. Second, the fact that temporal distance prediction requires training data that originates from robots, actual or simulated, introduces an unnecessary data

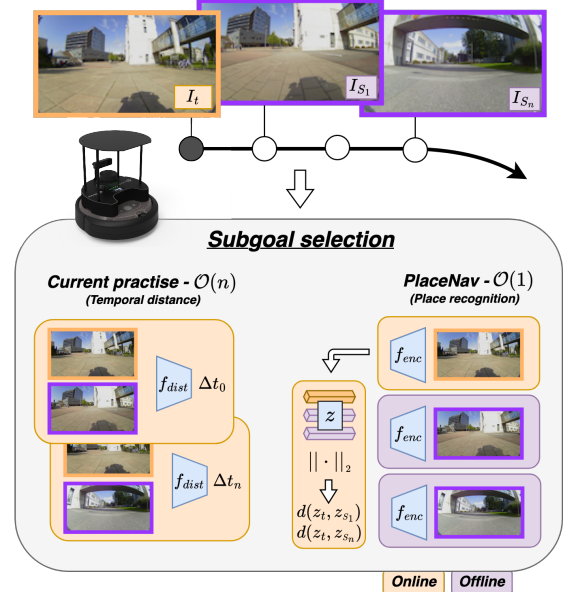


Fig. 1: Visual place recognition finds which map image  $I_s$  was captured closest to the robot’s observation  $I_t$  by efficient matching of image embeddings  $z$ .

bottleneck. High-quality robotics datasets are very scarce compared to general web-scale data, and models trained with simulated data suffer from generalization issues [14].

We claim that subgoal selection is not a unique problem but an instance of the broader concept of image retrieval. To address this, we present **PlaceNav**, which frames the selection as a *place recognition* [15] task. This design provides three advantages. First, the large-scale and high-diversity datasets available for training place recognition models enhance subgoal selection robustness against changes in viewpoint and appearance. Second, subgoal selection is performed by a fast nearest-neighbor search over image embeddings. This, as illustrated in Fig. 1, provides superior scalability and removes the need for heuristics. Finally, place recognition readily integrates with methods for ensuring temporal consistency.

In summary, our contributions are

- A navigation approach that decouples training of subgoal selection models from robotics datasets by treating the selection as a generic place recognition task.
- Integration of learning-based subgoal selection with a Bayesian filter that improves temporal consistency.
- Demonstrations with real robots that experimentally validate our design.

Code and videos are available on the project page<sup>†</sup>.

<sup>1</sup>lauri.a.suomela@tuni.fi. <sup>†</sup>lasuomela.github.io/placenv/

All authors are with Tampere University, Finland. This work was supported by Technology Innovation Institute.

## II. RELATED WORK

**Vision-based topological navigation.** In a recent trend, topological maps have been used to divide long-horizon tasks into short-horizon segments suitable for learned goal-reaching policies [8, 16]. An essential part of such a hierarchical structure is choosing which subgoal to reach next. One approach is to learn to predict the reachability between two images from simulated rollouts [17]. Savinov *et al.* [8] proposed to use the *temporal distance*, or the number of time steps  $\Delta t$ , between the current observation and a subgoal as a proxy for reachability. Its key strength is that it can be learned from offline data. While alternative approaches exist [18], temporal distance is popular and has been adopted in several recent works [9–13]. However, the diversity and size of datasets suitable for temporal distance learning are modest. RECON [19] with 25 h, SACSoN [20] with 75 h, and TartanDrive [21] with 5 h of navigation trajectories from single location each are notable examples. Furthermore, because of the model architectures utilized, the computational complexity of temporal distance prediction scales at  $O(n)$  with the number of subgoal candidates considered.

**Place recognition.** Place recognition involves recognizing places from images, often framed as image retrieval across images captured from different viewpoints and varying appearances [15]. It naturally integrates with topological navigation, as subgoal selection can be viewed as an image retrieval problem. Traditional methods for place recognition rely on aggregating handcrafted local features [22–24], but newer methods utilize deep learning to extract embeddings that can be compared efficiently using nearest-neighbor search [25–27]. The methods are trained to produce embeddings that are similar for images from the same place and dissimilar for images from different places, typically by classification loss [27] or ranking losses such as contrastive [28], triplet [25] or listwise [26] loss.

**Temporal consistency.** While subgoal temporal consistency has been studied in non-learned topological navigation literature [4], it has received limited attention in the context of learning-based methods. A robot moves along a route continually, so the transitions between the subgoals should be smooth. As a heuristic solution, SPTM [8] and GNM [12] adopted the approach of only considering subgoals within a sliding window centered on the previous subgoal. Meng *et al.* [17] utilize a similar approach but resort to global search when the window deviates from the robot’s actual location.

In place recognition literature, the topic has received more attention. Early approaches [29–32] utilized feature similarity matrices to find the best-matching image sequences. A newer line of work [33–35] considers descriptors that represent sequences instead of individual images. As an alternative, Xu *et al.* [36, 37] added a Bayesian filter to the matching process. In this work, we show that learning-based topological navigation also benefits from such methods.

## III. SYSTEM OVERVIEW

In this section, we describe *PlaceNav*, our proposed navigation pipeline. First, we discuss the basic components and

definitions of topological navigation. Then, we elaborate on our contributions related to subgoal selection via place recognition and subgoal temporal consistency.

### A. Topological navigation fundamentals

Autonomous navigation using topological maps generally consists of two stages. Before navigation, an operator has to perform a manual ‘reference run’ to capture the desired route. The robot saves images along the route that compose the topological map  $\mathcal{M}$  for navigation.

During navigation, the robot-agnostic topological navigation algorithm is combined with a robot-specific controller. At each inference step  $t$ , the current robot observation  $I_t$  is compared to the different subgoal candidate images  $I_s \in \mathcal{M}$  at nodes  $s = [0, 1, \dots, S]$  of the topological map. One of the nodes is selected as the next subgoal, and an image-based goal-reaching policy produces the motion plan to reach it. In this work, we experiment with the different subgoal selection methods and adopt the waypoint estimation approach proposed by Shah *et al.* [12] as the goal-reaching policy. This approach defines the motion plan as a sequence of  $\tau$  waypoints  $\{p_i, \psi_i\}_i$ ,  $i = [0, 1, \dots, \tau]$  that guide the robot to the subgoal. The waypoints, defined as metric coordinates  $p_i$  and heading angle  $\psi_i$  in the robot’s local coordinate frame, are tracked by a robot-specific controller.

### B. Subgoal selection via place recognition

PlaceNav introduces the following modifications to the subgoal selection procedure. Instead of computing temporal distances  $\Delta t$  between each observation and subgoal candidate pair, we use a place recognition model  $f_{enc}$  to process the observation and map images separately. The model produces image embeddings  $\mathbf{z}_t$  and  $\mathbf{z}_s$  that can be compared by Euclidean distance, enabling efficient subgoal search. Figure 1 visualizes the concept.

**Training data availability.** The temporal distance prediction models are trained to predict the  $\Delta t$  between two image frames sampled from a robot-driven trajectory. This limits the amount and diversity of potential training data. Place recognition methods can be trained with data from more generic sources. Training utilizes images of different places, preferably captured at various points in time, from different viewpoints, and under different environmental conditions. The images’ rough position and orientation information provide annotations. Google StreetView images, for example, are well-suited as training data. The sizes of place recognition datasets are in the order of millions of images [38–40], the SF-XL [27] alone consisting of 41M images.

**Computational complexity.** The computational complexity of temporal distance prediction scales linearly with the number of subgoal candidates considered at each inference step. Because of this, the number of subgoal candidates must be limited, which is commonly implemented as a *sliding window* over the map nodes. The window is centered on the subgoal from the previous step, and only the nodes within the window are considered potential subgoals.

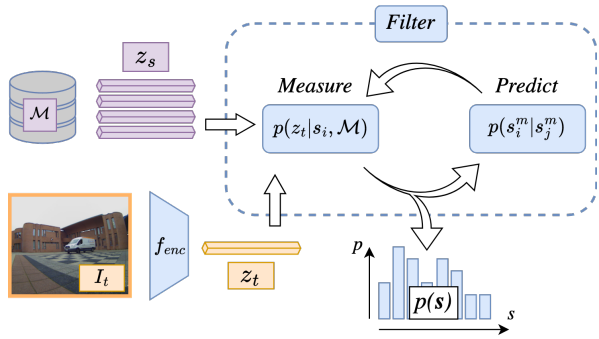


Fig. 2: The discrete Bayesian filter alternates place recognition measurement and motion model prediction.  $p(s)$ , the posterior belief, determines the best matching node.

Place recognition enables computation and storage of the descriptors for the map images offline before robot operation. Thus, the inference computational complexity of subgoal selection is decoupled from the number of subgoal candidates being considered. The descriptors can be matched in milliseconds by nearest neighbor search [41]. Consequently, heuristics to limit the number of subgoal candidates are not needed from the perspective of computational budget.

### C. Temporal consistency

Limiting the number of subgoal candidates also enhances temporal coherence between inference steps, preventing erratic subgoal selection *e.g.* due to visually similar content. A sliding window over the map achieves this to some extent. However, the window may drift away from the robot’s location, making correct subgoal selection impossible. Bayesian filtering, enabled by efficient matching of image embeddings, is an alternative strategy for enforcing temporal consistency.

Xu *et al.* [36] propose one such approach for use with place recognition methods, which we adapt for our problem. The idea, illustrated in Fig. 2, is to formulate place recognition as the measurement step of a discrete Bayesian state estimator. This filter maintains a belief of the robot’s location over the map nodes by recursively updating its posterior distribution. We present the key equations here for the sake of completeness but refer the reader to the original paper for details.

Given an initial posterior belief distribution, a motion model propagates the belief into the future in a prediction step. If the robot’s local movement (*i.e.* odometry) is not being tracked, as is the case with our system, the motion model is simply

$$p(s_i|s_j) \propto \begin{cases} 1 & w_l \leq i - j \leq w_u \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

From each node, the robot has an equal probability of moving up to  $w_u$  steps toward the goal, staying put, or moving  $w_l$  steps back. Other transitions have zero probability.

The prediction step is followed by a measurement step, where a place recognition query between the observation

and the map nodes produces a measurement belief by the measurement function

$$p(\mathbf{z}_t|s, \mathcal{M}) \propto g(\mathbf{z}_t, s, \mathcal{M}) = \exp(-\lambda_1 \|\mathbf{z}_t - \mathbf{z}_s\|_2) , \quad (2)$$

where  $\mathbf{z}_t$  is the observation embedding,  $\mathbf{z}_s$  is a map node embedding at the state  $s$  being considered, and  $\mathcal{M}$  is the map.  $\lambda_1$  scales the effect of each measurement on the posterior. Its value is automatically determined at the beginning of each navigation session as proposed by Xu *et al.* [36].

The measurement belief is multiplied by the belief from the prediction step to acquire the posterior belief distribution  $p(s)$ . The map node with the highest posterior probability is considered the closest to the latest observation. This filter significantly improves the stability of subgoal selection. Unlike the sliding window approach, the filter maintains full posterior belief over all map nodes, so it cannot get lost. It can solve the ‘kidnapped robot’ problem [42], whereas the sliding window requires the start node of the robot to be specified manually.

### D. Implementation details

**Architecture & Training.** The Shah *et al.* [12] waypoint estimation model was chosen as the goal-reaching policy. We do not retrain the model and use the weights provided by the authors, trained with  $85 \times 64$  images.

For the place recognition part of the PlaceNav, we use a CosPlace network [27] because of its high performance and simple training process. The model architecture comprises a convolutional encoder and a generalized mean pooling (GeM) layer. The model is trained via classification loss, enabled by dividing the training data into distinct spatial groups. As the original CosPlace model was trained with high-resolution images ( $512 \times 512$ ), the training checkpoints provided by the authors do not work well with the  $85 \times 64$  images we need to use for comparison with the baseline temporal distance model from Shah *et al.* [12]. For our experiments, we train a CosPlace model from scratch using an EfficientNet-B0 [43] backbone and the 41.2M images of the San Francisco eXtra Large (SF-XL) dataset [27], resized to  $85 \times 85$  during training. The model was configured to extract 512-dimensional descriptors. Otherwise, we followed the training procedure outlined in [27]. We will refer to this low-resolution model as *CosPlace-LR*.

**Deployment.** During inference, the robot uses place recognition to identify the map node that best matches the current observation. The next node along the route after the best-matching node is selected as the subgoal. We implemented two distinct temporal consistency methods in the subgoal selection. The first is the sliding window used in prior works [8, 12, 17], and the second is our implementation of the discrete Bayesian filter proposed by Xu *et al.* [36]. At the beginning of a route, the sliding window is initialized to the first node. The initial belief distribution of the discrete filter is acquired from the first place recognition query, meaning that it operates in a ‘kidnapped robot’ mode. We set the discrete filter motion model transition boundaries to  $w_u = 2$  and  $w_l = -1$  based on calibration experiments.



Fig. 3: The robots: A Turtlebot2 (left) and a Robotnik Summit XL Steel (right)

After choosing the subgoal to reach next, the goal-reaching policy predicts a series of 5 waypoints  $\{p_i, \psi_i\}$ . The robot follows these waypoints in an open-loop fashion until the next prediction using the robot’s low-level controller. The prediction loop of subgoal selection and waypoint prediction runs at 5 Hz. The place recognition and waypoint estimation models receive  $85 \times 64$  resolution images as input.

#### IV. EXPERIMENTAL SETUP

We performed navigation experiments with real robots in diverse environments to enable an informed comparison of PlaceNav and the prior work. We conducted 360 repetitions of different routes with different robots, subgoal selection methods, and temporal consistency approaches, adding up to a total of 19 km of navigation testing. We also evaluated the subgoal selection methods offline using a place recognition benchmark. With our experiments, we aim to answer the following research questions:

- **Q1.** Does subgoal selection require models trained with data originating from robots, or could more efficient place recognition models replace them?
- **Q2.** How robust are temporal distance prediction models to viewpoint and appearance changes?
- **Q3.** How does subgoal temporal consistency affect navigation performance?

##### A. Robot navigation experiments

**The robots.** We experimented with two different robots, a Turtlebot2, and a Robotnik Summit XL Steel, shown in Fig. 3. Turtlebot is a small research platform intended for indoor use. Summit is a 4-wheeled skid-steer ground vehicle with a weight of 90 kg, load capacity of 130 kg, and a maximum velocity of 3 m/s. Both robots carry a front-facing  $175^\circ$  field-of-view fish-eye camera, which is used for navigation, and a laptop computer that runs the navigation algorithms. The laptop has a Nvidia Geforce GTX 1070 GPU and an Intel i7-7700 CPU. The deep learning components of the navigation stack run on the GPU.

**Baseline.** We used the temporal distance prediction from GNM [12] by Shah *et al.* as the baseline. GNM was utilized for waypoint prediction in all experiments and we only experiment with the subgoal selection. We use model weights provided by the authors, obtained by training the model with  $85 \times 64$  images worth 54 hours of navigation. The inference



Fig. 4: Examples along the test routes. The top row is from outdoor tests, bottom row is from indoors.

loop runs at 5 Hz. At each step, the node with the smallest temporal distance  $\Delta t$  above 3 is selected as the subgoal.

As the Bayesian filter requires the distance between the current observation and all the map nodes at each step, using it with the GNM is not computationally feasible. Thus, with GNM, we only utilize the sliding window.

**Indoor navigation.** The indoor experiments were conducted with the Turtlebot2 robot. We tested the methods along 20 different routes, performing 3 repetitions of each route with each method. Fig. 4 shows examples along the routes. Routes where all the methods fail were excluded from the quantitative analysis. The experiments took place at a university campus, with buildings from various eras with diverse appearances.

The lengths of the test routes were uniformly distributed in the range from 15 m to 35 m, and they contained various amounts of difficult content, such as turning maneuvers and passing narrow gaps. We chose routes that do not contain content that would cause navigation failures because of errors in waypoint estimation. Such content, *e.g.* very narrow gaps, and turning maneuvers in open areas with few salient features can cause the robot to veer off course even though the subgoal selection algorithm is not failing.

**Outdoor navigation.** The outdoor experiments were conducted with the Summit XL robot. We experimented on 20 test routes in an urban environment, ranging from 50 m to 150 m in length. Like the indoor experiments, each test route was repeated 3 times with each method.

In indoor calibration tests before the actual experiments, correct subgoal selection led to accurate waypoint estimation. In outdoor tests, we observed more variability. This is likely due to increased environmental noise and larger appearance changes between the reference and test runs. Sometimes changes in ambient illumination led to complete waypoint estimation failure despite a good subgoal choice. In this work we are interested in the performance of subgoal selection, not the goal-reaching policy. Therefore, in the experiments, we maintained the difference between test and reference run illuminances below a threshold of 10 000 lx.

**Evaluation criteria.** We follow the evaluation guidelines for image-goal navigation in [44] and measure the navigation performance by the *success rate* (SR). Success rate describes the ratio of successful and failed repetitions of a test route.

TABLE I: **Indoors experiment** success rates over 60 repetitions, driven with the Turtlebot.

| Method   | Type | Temporal filter | Easy<br>$n = 33$ | Hard<br>27  | Total<br>60 |
|----------|------|-----------------|------------------|-------------|-------------|
| GNM [12] | $T$  | Window          | 0.52             | 0.26        | 0.39        |
| PlaceNav | $P$  | Window          | 0.62             | 0.60        | 0.61        |
|          |      | Bayesian        | <b>0.65</b>      | <b>0.77</b> | <b>0.69</b> |

Repetition is considered successful when the robot reaches a pose where its camera view corresponds to the final node of the topological map. The subgoal selection must also localize to this node, triggering a navigation stop signal. We do not use distance-based success criteria, which are more common but less aligned with the robot’s goal determination. While last-mile metric navigation could enhance goal-reaching precision, as suggested by Wasserman *et al.* [45], it is unnecessary for the scope of this work. Repetition is considered a failure when the robot takes actions that prevent it from reaching the goal, *i.e.* sending the ‘reached goal’ signal without the goal view in sight, risking immediate collisions, or deviating significantly from the test route without recovery prospects.

### B. Offline evaluation

To answer Q2. and enable a reproducible comparison of temporal distance prediction and place recognition, we tested GNM on standard place recognition benchmarks that contain appearance and viewpoint changes. The VPR-Benchmark by Bertoli *et al.* [46] was used to facilitate the experiments. We modified the benchmark to enable the evaluation of GNM that simultaneously takes the query and reference images as inputs. A subset of the VPR-Benchmark datasets where the size of the test database is small enough that GNM can process it in a reasonable time was picked for evaluation. We compared the performance of our low-resolution CosPlace-LR model and GNM temporal distance. The test images were resized to  $85 \times 64$ . For reference, we additionally evaluate the standard CosPlace model with full-resolution images.

Place recognition performance is assessed using the Recall@N score, which measures how often one of the top N images retrieved is within 25 m of the query image location. In the case of temporal distance prediction, the top N images are those with the smallest predicted temporal distances.

## V. RESULTS

### A. Robot navigation experiments

Tables I and II display the results of the navigation experiments with the Turtlebot and Summit XL.

**Indoors.** The GNM baseline has a notably lower success rate than the proposed place recognition based approaches. PlaceNav shows a 56% SR increase, which rises to 77% with the Bayesian filter. This observation is interesting given that GNM training includes indoor images from the GS4 [47] dataset, while CosPlace models are trained on outdoor Google Streetview images. Place recognition models

TABLE II: **Outdoors experiment** success rates over 60 repetitions, driven with the Summit XL.

| Method   | Type | Temporal filter | Easy<br>$n = 24$ | Hard<br>36  | Total<br>60 |
|----------|------|-----------------|------------------|-------------|-------------|
| GNM [12] | $T$  | Window          | <b>0.67</b>      | 0.33        | 0.47        |
| PlaceNav | $P$  | Window          | 0.46             | 0.44        | 0.45        |
|          |      | Bayesian        | <b>0.67</b>      | <b>0.53</b> | <b>0.58</b> |

$T$ : temporal distance,  $P$ : place recognition

exhibit broad generalization capabilities, learning features that generalize across domains.

Similar to Shah *et al.* [12], we split the test routes into ‘Easy’ and ‘Hard’ categories in posterior analysis. The categories are based on the number of narrow turns and tight passages along the routes. We chose a threshold that splits the routes into evenly sized groups. For indoor routes, this threshold was set to 4. ‘Easy’ routes had fewer than 4 such features, while ‘Hard’ routes had 4 or more. The categorization provides further insight into the method performances. While the differences in SR are minimal for ‘Easy’ routes, the advantage of place recognition is evident in the ‘Hard’ category. GNM’s SR decreases by 50% from ‘Easy’ to ‘Hard,’ but PlaceNav maintains the same SR, even improving when the Bayesian filter is employed.

Introducing the Bayesian filter yields clear advantages over the sliding window method. Visual inspection confirms improved stability and performance during turns, especially in narrow gaps. The filter mitigates errors such as mid-turn direction changes that can arise because of the erratic behavior of the sliding window approach.

**Outdoors.** Outdoors, the success rate gap between methods is narrower compared to indoors, despite CosPlace-LR’s outdoor training data. One possible explanation is the higher variation in waypoint estimation performance observed in the calibration tests. Consequently, the waypoint estimation module contributes more significantly to the final SR’s, diminishing the effect of subgoal selection performance. The magnitude of the performance increase brought by the Bayesian filter is consistent with the indoor experiments, the improvement being around 10 percentage points in both cases. The differences in SR’s between the ‘Easy’ and ‘Hard’ categories are similar to the indoor experiments. GNM’s SR drops by half from ‘Easy’ to ‘Hard,’ whereas PlaceNav exhibits minimal to no performance decrease, emphasizing place recognition’s effectiveness in maneuvers where having a correct subgoal is crucial.

In the analysis of outdoor experiments, an even distribution of routes between the ‘Easy’ and ‘Hard’ categories was achieved by a threshold of 3 turns or narrow passages per route. The occurrence of visually bursty content [49], characterized by repetitive geometric patterns that heavily influence image embeddings, poses a challenge for place recognition methods [15, p.12] on certain test routes (see Fig. 5). This issue, causing the relatively low SR of PlaceNav with the sliding window in the ‘Easy’ category can lead

TABLE III: **Offline evaluation** of Recall@1 for the place recognition ( $P$ ) and temporal distance ( $T$ ) models.

| Method        | Type | Resolution     | Multiview   |               | Front-facing |             |               |                  |               |              |             |
|---------------|------|----------------|-------------|---------------|--------------|-------------|---------------|------------------|---------------|--------------|-------------|
|               |      |                | Pitts30k    | Tokyo<br>24/7 | MSLS<br>Val  | St. Lucia   | SVOX<br>Night | SVOX<br>Overcast | SVOX<br>Rainy | SVOX<br>Snow | SVOX<br>Sun |
| CosPlace [48] | $P$  | full*          | <b>89.5</b> | <b>81.9</b>   | <b>80.8</b>  | <b>98.8</b> | <b>32.9</b>   | <b>85.4</b>      | <b>79.6</b>   | <b>85.3</b>  | <b>61.5</b> |
| CosPlace-LR   | $P$  | $85 \times 64$ | 71.8        | 24.1          | 43.8         | 90.4        | 1.2           | 33.9             | 26.1          | 24.5         | 12.4        |
| GNM [12]      | $T$  | $85 \times 64$ | 8.9         | 0.3           | 2.6          | 8.7         | 0.0           | 0.1              | 0.2           | 0.0          | 0.4         |

\*: various resolutions

to the sliding window method becoming trapped within the bursty map region, resulting in navigation failure. In contrast, the Bayesian filter handles bursty content more effectively by maintaining a full posterior belief across all map nodes and avoiding such entrapment. If the robot traverses the bursty segment successfully, the filter accurately localizes and completes the test route without issues.

### B. Offline evaluation

Here, we present the results of evaluating GNM and CosPlace on the VPR-Benchmark. We also discuss the impact of input resolution and domain shift on recall rates.

**Performance Comparison.** Table III shows a comparison of the retrieval performances of the GNM and CosPlace models across several benchmark datasets. Notably, temporal distance prediction performs worse than place recognition across all datasets. GNM’s recall follows a similar trend as CosPlace-LR, with GNM achieving higher recall wherever CosPlace-LR excels. However, GNM’s recall values are consistently an order of magnitude lower. For instance, on Tokyo24/7 and St. Lucia, where CosPlace-LR attains over 70% recall, GNM only reaches approximately 9%. On other datasets, GNM’s performance is significantly lower.

**Impact of Input Resolution.** Decreasing image resolution has a substantial impact on CosPlace’s performance. Reducing the resolution to  $85 \times 64$  pixels decreases recall rates up to 58 percentage points on datasets like Tokyo24/7, MSLS, and SVOX. Interestingly, GNM’s temporal distance prediction performs best on datasets where the performance differences between full and low-resolution CosPlace models are minimal, namely Pitts30k and St. Lucia. This suggests that GNM performance, too, would be improved by training the model with higher-resolution images.

**Viewpoint and Appearance Change.** Pittsburgh30k and Tokyo24/7 datasets capture images from various angles, while the rest feature images from front-facing vehicle cameras. Despite the similarity in viewpoint variation between GNM’s training data and front-facing datasets, this is not reflected in recall rates. GNM performs well on Pittsburgh30k but poorly on SVOX. This discrepancy may stem from other factors contributing to domain shift between query and reference images. Besides viewpoint changes, Pittsburgh30k and St. Lucia exhibit limited variation. The other datasets contain shifts in illumination, weather, and camera which GNM struggles to handle, not having been explicitly trained for such invariance.



Fig. 5: Repetitive patterns can disturb place recognition.

TABLE IV: **Runtimes** for the different methods with sliding window and Bayesian filter. \*: Bayesian filter considers all map nodes

| Method   | Temporal filter | Window size | Runtime (ms) |
|----------|-----------------|-------------|--------------|
| GNM [12] | Window          | 5           | 45           |
|          | Window          | 21          | 174          |
| PlaceNav | Window          | 5           | <b>19</b>    |
|          | Window          | 21          | <b>19</b>    |
|          | Bayesian        | -*          | <u>41</u>    |

### C. Runtime analysis

Table IV shows average runtimes for PlaceNav and the baseline. Replacing temporal distance prediction with place recognition significantly reduces runtime. Place recognition’s runtime does not depend on window size, while temporal distance is computed separately for each subgoal candidate inside the window. The Bayesian filter increases PlaceNav’s runtime slightly. This enables making a resource-performance trade-off based on the navigation requirements.

## VI. CONCLUSION

Our findings show that place recognition enables more accurate subgoal selection than the temporal distance prediction methods at a lower computational cost. The offline evaluation implies that appearance change between the reference run and robot operation conditions would further amplify the difference. These results suggest that the training of learning-based models for robotics should not be unnecessarily limited to robotics-specific data. If some part of the robot’s task can be cast as a more general learning problem, a larger amount of more diverse data can be used, leading to better performance. Our future work will apply this principle to developing appearance-invariant subgoal selection models and goal-reaching policies.

## ACKNOWLEDGMENT

The authors thank Jani Käpylä, Olli Suominen, and Jussi Rantala from CIVIT for access to the Summit XL. We would also like to thank María Andrea Cruz Blandón and German F. Torres for their valuable comments on an earlier version of the manuscript.

## REFERENCES

- [1] R. Brooks, "Visual map making for a mobile robot," in *1985 IEEE International Conference on Robotics and Automation Proceedings*, vol. 2, Mar. 1985, pp. 824–829. 1
- [2] E. Baumgartner and S. Skaar, "An autonomous vision-based mobile robot," *IEEE Transactions on Automatic Control*, vol. 39, no. 3, pp. 493–502, Mar. 1994, conference Name: IEEE Transactions on Automatic Control. 1
- [3] S. Thrun, "Learning metric-topological maps for indoor mobile robot navigation," *Artificial Intelligence*, vol. 99, no. 1, pp. 21–71, Feb. 1998. 1
- [4] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional Vision Based Topological Navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 219–236, Sept. 2007. 2
- [5] D. Filliat, "Interactive learning of visual topological navigation," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept. 2008, pp. 248–254, iSSN: 2153-0866. 1
- [6] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, "Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 1689–1696, iSSN: 2577-087X. 1
- [7] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu, D. Kumaran, and R. Hadsell, "Learning to Navigate in Complex Environments," in *International Conference on Learning Representations (ICLR)*, Nov. 2016. 1
- [8] N. Savinov, A. Dosovitskiy, and V. Koltun, "Semi-parametric topological memory for navigation," in *International Conference on Learning Representations (ICLR)*, 2018. 1, 2, 3
- [9] N. Savinov, A. Raïchuk, D. Vincent, R. Marinier, M. Pollefeys, T. Lillicrap, and S. Gelly, "Episodic Curiosity through Reachability," in *International Conference on Learning Representations (ICLR)*, 2018. 2
- [10] D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, "ViNG: Learning Open-World Navigation with Visual Goals," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE Press, 2021, pp. 13 215–13 222. 1
- [11] L. Mezghan, S. Sukhbaatar, T. Lavril, O. Maksymets, D. Batra, P. Bojanowski, and K. Alahari, "Memory-Augmented Reinforcement Learning for Image-Goal Navigation," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 3316–3323, iSSN: 2153-0866.
- [12] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine, "GNM: A General Navigation Model to Drive Any Robot," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 7226–7233. 1, 2, 3, 4, 5, 6
- [13] D. Shah, A. Sridhar, N. Dashora, K. Stachowicz, K. Black, N. Hirose, and S. Levine, "ViNT: A Large-Scale, Multi-Task Visual Navigation Backbone with Cross-Robot Generalization," in *Proceedings of the 7th Annual Conference on Robot Learning*. PMLR, Aug. 2023. 1, 2
- [14] A. Kar, A. Prakash, M.-Y. Liu, E. Cameracci, J. Yuan, M. Rusiniak, D. Acuna, A. Torralba, and S. Fidler, "Meta-Sim: Learning to Generate Synthetic Datasets," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019, pp. 4551–4560. 1
- [15] C. Masone and B. Caputo, "A Survey on Deep Visual Place Recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021. 1, 2, 5
- [16] D. Singh Chplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural Topological SLAM for Visual Navigation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, June 2020, pp. 12 872–12 881. 2
- [17] X. Meng, N. Ratliff, Y. Xiang, and D. Fox, "Scaling Local Control to Large-Scale Topological Navigation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, May 2020, pp. 672–678, iSSN: 2577-087X. 2, 3
- [18] B. Eysenbach, R. R. Salakhutdinov, and S. Levine, "Search on the Replay Buffer: Bridging Planning and Reinforcement Learning," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. 2
- [19] D. Shah, B. Eysenbach, N. Rhinehart, and S. Levine, "Rapid Exploration for Open-World Navigation with Latent Goal Models," in *Proceedings of the 5th Conference on Robot Learning*. PMLR, Jan. 2022, pp. 674–684, iSSN: 2640-3498. 2
- [20] N. Hirose, D. Shah, A. Sridhar, and S. Levine, "SACSoN: Scalable Autonomous Control for Social Navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 49–56, Jan. 2024. 2
- [21] S. Triest, M. Sivaprakasam, S. J. Wang, W. Wang, A. M. Johnson, and S. Scherer, "TartanDrive: A Large-Scale Dataset for Learning Off-Road Dynamics Models," in *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 2546–2552. 2
- [22] Sivic and Zisserman, "Video Google: a text retrieval approach to object matching in videos," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2003, pp. 1470–1477 vol.2. 2
- [23] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, June 2010, pp. 3304–3311, iSSN: 1063-6919.
- [24] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and

- T. Pajdla, “24/7 place recognition by view synthesis,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 1808–1817, iSSN: 1063-6919. 2
- [25] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “NetVLAD: CNN Architecture for Weakly Supervised Place Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1437–1451, June 2018. 2
- [26] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, “Learning With Average Precision: Training Image Retrieval With a Listwise Loss,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 5106–5115. 2
- [27] G. Berton, C. Masone, and B. Caputo, “Rethinking Visual Geo-localization for Large-Scale Applications,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 4868–4878. 2, 3
- [28] F. Radenović, G. Toliás, and O. Chum, “Fine-Tuning CNN Image Retrieval with No Human Annotation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, July 2019. 2
- [29] M. J. Milford and G. F. Wyeth, “SeqSLAM: Visual route-based navigation for sunny summer days and stormy winter nights,” in *2012 IEEE International Conference on Robotics and Automation (ICRA)*, May 2012, pp. 1643–1649, iSSN: 1050-4729. 2
- [30] P. Hansen and B. Browning, “Visual place recognition using HMM sequence matching,” in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sept. 2014, pp. 4549–4555, iSSN: 2153-0866.
- [31] E. Pepperell, P. I. Corke, and M. J. Milford, “All-environment visual place recognition with SMART,” in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, May 2014, pp. 1612–1618, iSSN: 1050-4729.
- [32] T. Naseer, W. Burgard, and C. Stachniss, “Robust Visual Localization Across Seasons,” *IEEE Transactions on Robotics*, vol. 34, no. 2, pp. 289–302, Apr. 2018. 2
- [33] A. Gawel, C. D. Don, R. Siegwart, J. Nieto, and C. Cadena, “X-View: Graph-Based Semantic Multi-View Localization,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1687–1694, July 2018. 2
- [34] Y. Latif, R. Garg, M. Milford, and I. Reid, “Addressing Challenging Place Recognition Tasks Using Generative Adversarial Networks,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018, pp. 2349–2355, iSSN: 2577-087X.
- [35] S. Garg and M. Milford, “SeqNet: Learning Descriptors for Sequence-Based Hierarchical Place Recognition,” *IEEE Robotics and Automation Letters*, vol. PP, pp. 1–1, 2021. 2
- [36] M. Xu, N. Snderhauf, and M. Milford, “Probabilistic Visual Place Recognition for Hierarchical Localization,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 311–318, Apr. 2021. 2, 3
- [37] M. Xu, T. Fischer, N. Sünderhauf, and M. Milford, “Probabilistic Appearance-Invariant Topometric Localization With New Place Awareness,” *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6985–6992, Oct. 2021. 2
- [38] N. Carlevaris-Bianco, A. Ushani, and R. Eustice, “University of Michigan North Campus long-term vision and lidar dataset,” *The International Journal of Robotics Research*, vol. 35, Dec. 2015. 2
- [39] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, “Deep learning features at scale for visual place recognition,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 3223–3230.
- [40] F. Warburg, S. Hauberg, M. Lopez-Antequera, P. Gargallo, Y. Kuang, and J. Civera, “Mapillary Street-Level Sequences: A Dataset for Lifelong Place Recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2626–2635. 2
- [41] J. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, July 2021, conference Name: IEEE Transactions on Big Data. 3
- [42] S. Thrun, D. Fox, W. Burgard, and F. Dellaert, “Robust Monte Carlo localization for mobile robots,” *Artificial Intelligence*, vol. 128, no. 1, pp. 99–141, May 2001. 3
- [43] M. Tan and Q. Le, “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, May 2019, pp. 6105–6114, iSSN: 2640-3498. 3
- [44] P. Anderson, A. Chang, D. S. Chaplot, A. Dosovitskiy, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva, and A. R. Zamir, “On Evaluation of Embodied Navigation Agents,” *arXiv:1807.06757 [cs]*, July 2018. 4
- [45] J. Wasserman, K. Yadav, G. Chowdhary, A. Gupta, and U. Jain, “Last-Mile Embodied Visual Navigation,” in *Proceedings of The 6th Conference on Robot Learning*. PMLR, Mar. 2023, pp. 666–678, iSSN: 2640-3498. 5
- [46] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, “Deep Visual Geo-Localization Benchmark,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5396–5407. 5
- [47] N. Hirose, F. Xia, R. Martín-Martín, A. Sadeghian, and S. Savarese, “Deep Visual MPC-Policy Learning for Navigation,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3184–3191, Oct. 2019. 5
- [48] G. Berton, C. Masone, and B. Caputo, “Rethinking Visual Geo-localization for Large-Scale Applications,” in *2022 IEEE/CVF Conference on Computer Vision*

*and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 4868–4878. 6

- [49] H. Jegou, M. Douze, and C. Schmid, “On the burstiness of visual elements,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, June 2009, pp. 1169–1176, iSSN: 1063-6919. 5