Long-term Visual Place Recognition

Farid Alijani[1], Jukka Peltomäki[1], Jussi Puura[2], Heikki Huttunen[3], Joni-Kristian Kämäräinen[1], Esa Rahtu[1]
[1]Tampere University, [2]Sandvik Ltd, [3]Visy Ltd
Finland
{*farid.alijani, jukka.peltomaki, joni.kamarainen, esa.rahtu*}*@tuni.fi, jussi.puura@sandvik.com, heikki.huttunen@visy.fi*

***Abstract*—In this work, we study the long-term performance of visual place recognition in urban outdoor environment. A long-term benchmark is constructed from the Oxford RobotCar dataset. It contains sequences of the same route traversed over a period of approx. 500 days. We carefully selected three gallery sequences, one training sequence and 15 query sequences that cover different seasons, times of day and weather. The RobotCar sequences from the first half year have several problems, for example, only partial routes and inaccurate location data. We circumvent these problems by reversing the time. In the benchmark dataset the gallery and training images are the latest and the query sequences go gradually back in time. Our experiments provide the following findings. 1) the selected gallery sequence has strong impact on performance, and 2) additional training sequences help to mitigate differences between the gallery sequences. In addition, results indicate that 3) there is a long-term trend of performance degradation over time. The degradation can be quantified as about 6 percentage points per 100 days and, therefore, the initial performance of** 40% **eventually drops below** 20% **at the end.**

I. Introduction

Visual place recognition is the process of recognizing a previously visited place using visual information, often under varying appearance conditions and viewpoint changes [1]. Visual place recognition is a crucial part in visual navigation stacks of autonomous robots and vehicles. The recent advances in deep learning have resulted to deep visual place recognition methods that achieve substantially better performance than the previous methods based on engineered features. The "engineered" point feature and deep feature based methods are surveyed in the two recent works [1], [2].

In this work, we focus on the long-term aspect of visual place recognition. This problem is important for practitioners as it helps to estimate for how long a single recorded gallery set can be expected to succeed in a non-static environment such as urban city centers. In place recognition terminology, the gallery set is a set of images for which location tags are added. During test time, new images, queries, are matched against the gallery and their locations retrieved based on the best or multiple best matches. In static conditions, the place recognition performance should not degrade, but in natural environments, such as city centers, there are various short-term and long-term changes. Short-term changes are dynamic components such as other traffic (pedestrians, cars and cyclists) and weather, and long-term changes are new buildings and permanent changes in the existing ones. There are also certain cyclic-changes such as time-of-day (ambient illumination) and seasons from summer to winter and back to summer again. However, there are surprisingly little work on quantitative evaluation of these long-term effects.

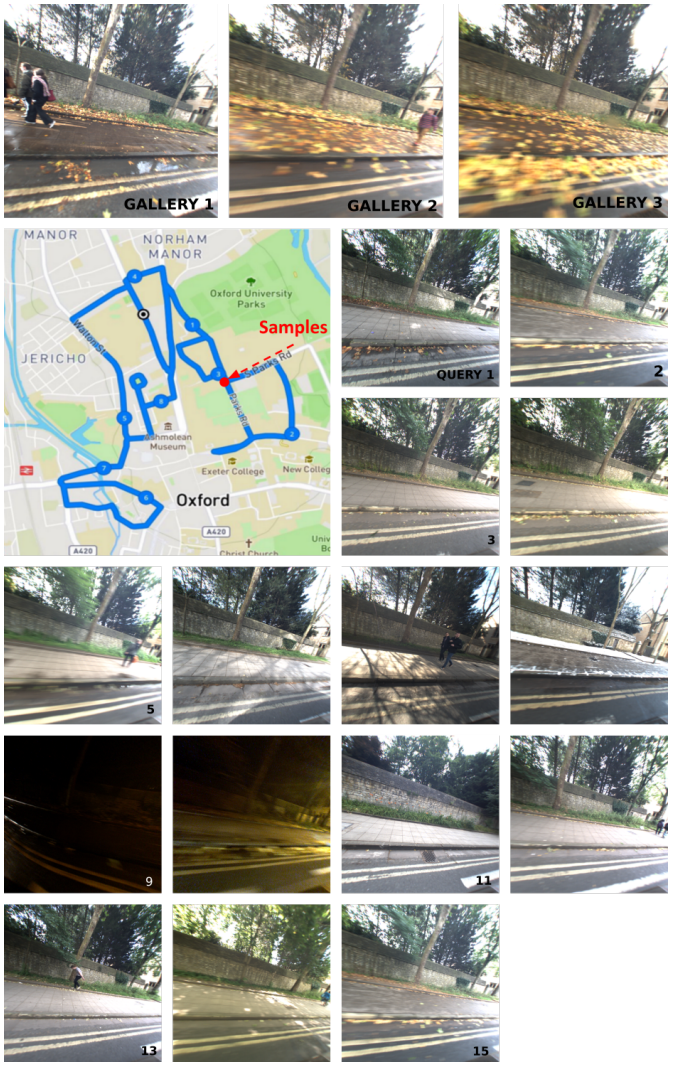


Fig. 1: The Oxford RobotCar based long-term visual place recognition dataset used in this work. The sequences span various seasons, weather and traffic conditions and times of day over the duration of 542 days. Examples (Gallery 1-3 and Query 1-15) are from the same location on the map.

The main contributions of this work are: **1)** we propose a long-term place recognition benchmark that consists of carefully selected sequences from the Oxford RobotCar dataset [3]

and spans more than 540 days between 2014 and 2015 (Figure 1), and **2)** with the help of the benchmark and a state-of-the-art deep visual place recognition method, we are able to show that long-term visual place recognition performance indeed has a degradation trend.

II. Related Work

A significant body of research in visual place recognition has focused on proposing techniques that are invariant to viewpoint, illumination and seasonal variations which are the main sources of variation in outdoor navigation. With the rise of deep convolutional neural network (CNN) architectures in computer vision and robotics, also visual place recognition has leveraged deep learning to outperform conventional approaches which were mainly based on handcrafted point and region features [4], [5], [6], [7].

The main computational paradigm for deep visual place recognition is deep metric learning. The deep methods learn mapping from an input image to a global descriptor. Descriptors from the same location should have small distances and between two different locations large distances. With the help of descriptors, the actual place recognition can be performed as nearest neighbor search. Chen *et al.* [8] is one of the first works to employ off-the-shelf deep features in visual place recognition. In the following work, Chen *et al.* [9], the descriptors were further fine-tuned using place recognition data. They trained two neural networks (AMOSNet and HybridNet) on the SPED dataset. AMOSNet was trained from scratch on SPED, while the weights for HybridNet were initialized from the top-5 convolutional layers of CaffeNet.

Sünderhauf *et al.* [10] also used pre-trained deep features as descriptors and reported good performance with a large-scale outdoor datasets. Subsequently, [11] utilized region proposals to extract ConvNet features separately on each of the regions. Bai *et al.* [12] proposed SeqCNNSLAM to improve the viewpoint invariance and boost the performance of the original SeqSLAM [4]. Bai *et al.* also used pre-trained features. Merrill and Huang [13] proposed a convolutional auto-encoder for visual place recognition, where auto-encoder network was trained in a weakly-supervised manner to recreate HOG-descriptors for viewpoint-variant cropped images of the same place.

The state-of-the-art methods for visual place recognition are NetVLAD and GeM. Arandjelović *et al.* [14] presented NetVLAD as an end-to-end holistic descriptor in which a trainable VLAD layer [15] is integrated into the CNN architecture to achieve excellent place recognition results. Radenović *et al.* [16] introduces a new trainable generalized mean (GeM) pooling layer into the deep image-retrieval architecture which has shown to provide a performance boost. The seminal works of NetVLAD and GeM have been further developed in recent works. For example, Cao *et al.* [17] present a DELG pipeline as an extension for DELF [18] in which GeM pooling layer for global descriptors and attention mechanism for local features were combined into a single deep model.

Advanced and versatile datasets are important for the development of place recognition because they (1) provide benchmark for the evaluation of novel methods and (2) enable training more accurate deep learning models. For long-term place recognition, illumination is a crucial factor since the appearance of the same place may change drastically under different illumination conditions. For several years, researchers have collected and published several datasets for visual localization and place recognition with various sensor modalities, including monocular and stereo cameras, LiDAR, IMU, radar and GNSS/INS sensors. To highlight a few, New College and City Centre [5], KITTI Odometry benchmark [19], Ford campus [20], extended CMU seasons [21], [22], Mapillary Street-level [23], Málaga Urban [24], Alderley [4], Aachen Day-Night [25] and Nordland [26] are the outdoor and InLoc [27], ETH-Microsoft [28], and NAVER LABS localization [29] are indoor datasets. Among these datasets the Oxford RobotCar dataset [3] is popular as it contains sequences of the same route from the time period of more than 500 days.

III. Methodology

For this work, we adopted the GeM method by Radenović et al. [16] which has achieved state-of-the-art performance with various benchmarks and since it was found the best in our previous work [30].

A. GeM network

The working principle of GeM [16] is similar to other deep architectures for place recognition and image retrieval. A deep representation of an input image is obtained by a neural network architecture which produces a global descriptor of visual content of the input image. Describing the whole image using a holistic feature, a global descriptor, is more robust against appearance changes than local descriptors [1].

For training, Radenović et al. [16] adopt the Siamese neural network architecture. This architecture is trained using positive and negative image pairs and the loss function enforces large distances between negative pairs (images from two distant places) and small distances between positive pairs (images from the same place). The architecture is trained with the contrastive loss which acts on matching (positive) and non-matching (negative) pairs and is defined as follows:

$$\mathcal{L} = \begin{cases} l(\vec{f}_a, \vec{f}_b) \text{ for matching images} \\ \max\left(0, M - l(\vec{f}_a, \vec{f}_b)\right) \text{ otherwise} \end{cases} . \quad (1)$$

In (1) $l(\cdot)$ is the pair-wise distance term (Euclidean distance) and M is the enforced minimum margin between the negative pairs. $\vec{f}_a$ and $\vec{f}_b$ denote the deep feature vectors of images I_a and I_b computed using the convolutional head of a backbone feature extraction network such as AlexNet, VGGNet or ResNet. The feature vector lengths K are 256, 512 or 2048, depending on the backbone. Feature vectors are global descriptors of the input images and pooled over the spatial dimensions. The feature responses are computed from C convolutional layers $\mathcal{X}_c$ following with max pooling layers

that select the maximum spatial feature response from each layer (MAC vector):

$$\vec{f} = [f_1 \ f_2 \ \dots \ f_i \ \dots \ f_K]^T, \ f_i = \max_{x \in X_k} x \ . \tag{2}$$

Radenović et al. [31] originally used the MAC vectors, but in [16] they compared MAC vectors with average pooling (SPoC vector) and GeM pooling. Experiments show that GeM consistently outperforms MAC and SPoC, and provides the best average retrieval accuracy. Therefore we named their architecture as "GeM Network".

GeM pooling layer takes χ as an input and produces a vector $\vec{f} = [f_1 \ f_2 \ \dots \ f_i \ \dots \ f_K]^T$ as an output of the pooling process which results in:

$$\vec{f_i} = \left(\frac{1}{|\chi_i|} \sum_{x \in \chi_i} x^{p_i} \right)^{\frac{1}{p_i}} . \tag{3}$$

MAC and SPoC pooling methods are special cases of GeM pooling depending on how pooling parameter p_i is derived. $p_i \to \infty$ and $p_i = 1$ correspond to max-pooling and average pooling, respectively. The GeM feature vector is a single value per feature map and its dimension varies depending on different networks, *i.e.*, $K = [256, 512, 2048]$.

The Radenović et al. [16] pipeline is largely shared by most deep metric learning approaches to image retrieval, but their unique and important components are *supervised whitening* and *positive and negative sample mining*. Despite the fact that these were defined for image and landmark retrieval they seem to be effective for place recognition as well. More details of GeM Network can be found in the previous works of the original authors in [31], [32] and they also provide the code used in the experiments of this work.

B. Training Details

ResNet50 network trained with the ImageNet dataset was used as the feature extraction backbone of GeM Network in our experiments. The experiments were conducted both using backbone features as is and after fine-tuning with dataset specific data. In all cases the network benefits from fine-tuning with training data.

In the contrastive loss the number of positive matches is the same as the number of images in the pool of queries from which they are selected randomly, whereas the number of negative matches is always fixed in the training.

Following [16], we learn whitening through the same training data for two reasons: (1) it mutually complement fine-tuning to boost the performance, (2) applying whitening as a post-processing step expedites training compared to learning it in an end-to-end manner [31]. We also utilize trainable GeM pooling layer which significantly outperforms the retrieval performance while preserving the dimension of the descriptor.

For our experiments, we used a single NVIDIA Quadro RTX 8000 GPU with 32 GB memory using PyTorch 1.7 deep learning framework.

IV. Data

To study and empirically evaluate long-term visual place recognition, a suitable benchmark dataset is needed. The dataset must represent realistic long-term changes in indoor or outdoor environments. In this work, we focus on outdoor evaluation as no suitable indoor dataset is available.

A. Oxford RobotCar Dataset

One of the few suitable outdoor datasets is Oxford RobotCar [3] which contains image sequences from the same approximately 10.0 *km* long route in the downtown Oxford and traversed repeatedly over the period of 543 days. Given traversals over a year, it comprises different seasons and time of the day for short-term, long-term and cyclic changes that occur in real urban environments.

Moreover, short-term weather changes and dynamic components such as pedestrians, parked and/or passing-by vehicles and cyclist make the dataset realistic. Environmental changes, including illumination and appearance changes, roadworks and construction causing detours, and atmospheric conditions impact the route and GPS reception which provide moderate long-term changes suitable for our evaluation pipeline.

1) Overview: Oxford RobotCar dataset [3] enables research for investigating long-term visual place recognition for autonomous vehicles in real-world and dynamic urban environments by frequently traversing the same route over one year. Data acquisition was performed over the period of May 2014 to December 2015 over several traversals in central Oxford with a total route of 1000 *km* urban driving. This dataset addresses a variety of challenging conditions including weather, traffic, and lighting alterations. It also contains several sensor modalities to perceive an environment and localize an agent accurately (Figure 1).

2) Reversed time: Oxford RobotCar, as any new dataset, has problems in the first sequences. For example, only partial route is acquired, sensor data is missing or location ground truth (GPS/INS, RTK) is unreliable or missing. These problems mostly disappear after the first few months, but make it impossible to select gallery and training images from the beginning of the campaign. We wanted to simulate the real situation where gallery and training data are collected and then the autonomous agent should perform localization in changing conditions over time. The problem was solved by reversing the time. In this setting the results are not affected by poor or limited gallery sets and still have a long-term temporal data, now only to the opposite direction, back in time. Therefore, in our dataset the incremental indexes 1, 2, etc. actually index sequences backwards.

3) Sensors: The RobotCar vehicle is equipped with multiple cameras and LiDARs and GPS/INS units for positioning. For position ground truth we use either the RTK, when available, or GPS/INS as suggested by the original authors. Moreover, as we focus on visual localization and the LiDAR data is omitted. Maddern *et al.* [33] post-processed the raw GPS and IMU data with GNSS base station recordings to produce a Real-time Kinematic (RTK) solution with $cm-$accuracy.

TABLE I: The proposed *Long-term Oxford RobotCar Place Recognition Benchmark* that consists of carefully selected and verified sequences from the original Oxford RobotCar [3]. The query sequences are indexed so that 01 is the temporally closest to the gallery sequences and 15 is the furthest.

Seq#	*# of samples*	*Day*	*Date*	*Start [GMT]*	*Condition*	*GT*
	Train					
1	23,893	0	Nov-12-2015	11:22	Sun	RTK
	Gallery					
1	32,030	+1	Nov-13-2015	10:28	Sun-Overcast	INS
2	35,901	-13	Oct-30-2015	13:52	Overcast	RTK
3	26,683	-14	Oct-29-2015	12:18	Rain	RTK
	Query					
01	24,542	-71	Sep-02-2015	10:37	Sun	INS
02	31,043	-91	Aug-13-2015	16:02	Overcast	RTK
03	27,594	-132	Jul-03-2015	15:23	Overcast	INS
04	20,695	-139	Jun-26-2015	08:09	Overcast	RTK
05	20,695	-177	May-19-2015	14:06	Overcast	RTK
06	20,695	-209	Apr-17-2015	09:06	Sun	RTK
07	33,445	-233	Mar-24-2015	13:47	Sun	RTK
08	27,594	-282	Feb-03-2015	08:45	Snow	INS
09	13,797	-330	Dec-17-2014	18:18	Night-Rain	INS
10	20,695	-363	Nov-14-2014	16:34	Night	RTK
11	13,797	-486	Jul-14-2014	15:42	Sun-Cloud	INS
12	13,797	-486	Jul-14-2014	14:49	Overcast	INS
13	6,898	-516	Jun-26-2014	09:53	Overcast	INS
14	5,919	-518	Jun-24-2014	14:47	Sun	INS
15	9,206	-542	May-19-2014	13:05	Sun	INS

RTK data is more accurate than the provided GPS/INS online location, but is not available for all sequences.

The cameras of the vehicle include one PointGrey Bumblebee XB3 trinocular stereo camera and three PointGrey Grasshopper2 monocular cameras that together cover 360° view around the car. The three monocular Grasshopper2 cameras with fisheye lenses are mounted on the back side of the vehicle (left, backward, right).

We conducted experiments using only a single view, raw images from the left view of the Point Grey Grasshopper2. In the left-hand side traffic in the UK it provides a view to the sidewalk and is thus less affected by forthcoming traffic. Example images are shown in Figure 1.

4) Sequences: For place recognition, we need to define three sets of images: gallery, training and query sets. By verifying each sequence and taking as uniform distribution as possible of examples from the whole period of capture we ended up with 3 gallery sequences, 1 training sequence and 15 query sequences. Table I summarizes the sequences and ground truth used to form our long-term Oxford RobotCar place recognition benchmark.

B. Limitations

In Table I, the query sequences 12-15 contain only partial routes (note that this is not directly observable from the number of samples). The reason is the same as why reversed time was used, i.e. sequences from the first months are partial or missing location ground truth. However, we wanted to keep a small number of these sequences to expand the time duration by 60 days. In our experiments, the results for these sequences were good but not comparable to other sequences in which the full route was available. Gallery 1 and Queries 12-15 with INS ground truth are shown in Figure 2.

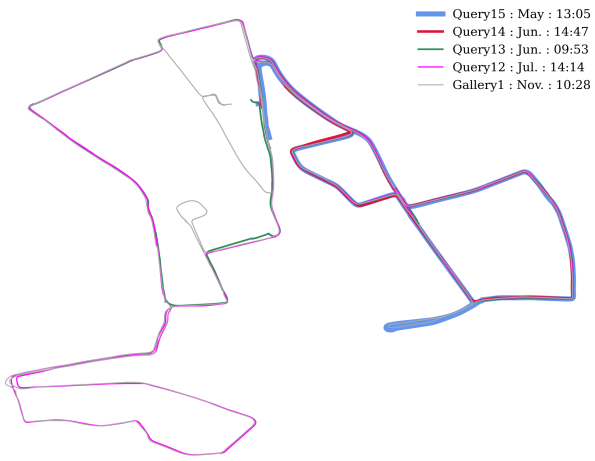


Fig. 2: INS ground-truth trajectories of the Gallery 1 and Queries 12-15 of partial routes.

C. Performance Evaluation

We utilized the method of Section III to compute a feature vector representation for the given query image $f(q)$. Then the feature vector is matched to all gallery image representations of $f(G_i)$ when $\{i = 1,2,...,N\}$ using Euclidean distance $d_{q,G_i} = ||f(q) - f(G_i)||_2$ to obtain proximity matrix of feature vectors and retrieve top-N matches. Given the ground-truth positions of query and gallery sequences, if the best match position is within the given distance threshold (τ), it is marked as *true positive* and otherwise as *false positive*. The ratio of *true positive* to the total number of the query images is identified as recall.

Following the common practices of [14], [34], we measure the place recognition performance by the fraction of correctly matched queries. Similar to [35], we denote the fraction of top-N shortlisted correctly recognized candidates as recall@N which can vary according to the available ground-truth annotations and τ for the dataset. To evaluate the performance of the methods in Section V, we report only the top-1 match recall, *i.e.*, recall@1 for multiple thresholds.

V. Results

In this section, we report the results of experiments with state-of-the-art place recognition method, GeM, described in Section III, using the evaluation dataset and metrics of Section IV. It is noteworthy that the temporal distance measure of days goes backward in our experiments as the Oxford RobotCar dataset has various problems, *e.g.*, lack of precise ground-truth among the first sequences.

A. Overall performance

In the first experiment, we calculate recall@1 separately for all query sequences, *e.g.*, Query 01-15, given each gallery sequence, *e.g.*, Gallery 1-3, with and without training data for feature fine-tuning (Train 1). The overall average performance

TABLE II: Overall average performance results of recall@1, given two recognition distance thresholds, $\tau = 5.0m$ and $\tau = 25.0m$, with and without feature fine-tuning using the training sequence (Train 1).

Gallery #	*Recall@1*			
	$\tau = 5.0m$		$\tau = 25.0m$	
	no training	Train 1	no training	w/ Train 1
1	0.030	**0.294**	0.073	**0.374**
2	**0.050**	0.283	**0.122**	0.371
3	0.034	0.129	0.095	0.186

results of recall@1 for the Oxford RobotCar long-term place recognition dataset are presented in Table II.

These results verify that: **1)** gallery selection has strong impact on the place recognition performance (Gallery 2 performs almost twice better than galleries 1 and 3 without training data), **2)** training data substantially boosts the results ($5-10\times$ better recall@1) and reduces the differences between galleries (Gallery 1 and 2 perform equally well after fine-tuning). The fine-tuned network (w/ Train 1) using GeM method retrieves the best match with correct location for nearly 30% of all query images among all query sequences within 5.0 m of the correct location.

B. Long-term performance

In the second experiment, we evaluated the overall place recognition performance by dividing the temporally distant sequences of Query 01-15 such that Query 01 is the closest to the training and gallery sequences and Query 15 is the furthest. The goal of this experiment was to study the long-term place recognition performance. The results of this experiment are shown in Figure 4.

The long-term results in Figure 4 provide the following findings: **1)** every query sequence remarkably benefits from training data similar to overall results in the previous experiment, **2)** Gallery 1 and Gallery 2 perform equally well and clearly better than Gallery 3 which highlights the rainy weather condition strongly impacting the place recognition performance, **3)** long-term performance degradation becomes clear between the temporally close and distant query sequences.

The results of recall@1 using fine-tuned networks are analogous for both $\tau = 5.0\,m$ and $\tau = 25.0\,m$. This indicates that if the best match is false positive, it evidently is far from the correct location. The best recall rates vary between 0.3 to 0.5 in the first half of the query sets and 0.3 to 0.35 in the second half. This displays a trend of degrading performance when the time from the query and training sequence increases. Query 09 and Query 10 are quite challenging with "night time" conditions. Therefore, their poor performance confirms that gallery sequences with different conditions, *day-time*, cannot perform equally well for *night-time* sequences.

C. Long-term performance trend

In the third experiment, we analyzed the quantified trend of the recall@1 performance, portrayed as a function of the days from the training data. This forms data points $\{x,y\}_i = \{\text{days}, recall@1\}_i$ to which a linear model was fitted and its slope used as the quantitative trend value.

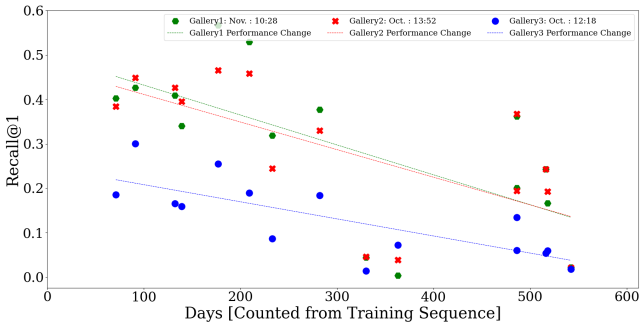


Fig. 3: Recall@1 performance results of each query sequence and for all three galleries (red, green and blue dots) as functions of time in days from the training set (Day 0). Linear models are fitted to data points and the slopes (trend) are $-6.72 \cdot 10^{-4}$, $-6.21 \cdot 10^{-4}$ and $-3.85 \cdot 10^{-4}$ for the galleries G1-3, respectively.

The results of Figure 3 depict a clear performance degradation over time for all three galleries. Gallery 1 and Gallery 2 illustrate an approximately similar long-term performance trend although query sequences with closer dates vary slightly. The slope values correspond to percentage point changes -6.72%, -6.21% and -3.85% per 100 days.

VI. Conclusion

In this work, we reported an experimental study of long-term visual place recognition (VPR) with a SotA deep architecture. For evaluation, we proposed a long-term VPR benchmark using sequences from the Oxford RobotCar dataset.

Our experimental results indicate that the performance of state-of-the-art deep place recognition method steadily degrades over time due to long-term changes of the outdoor environment. This highlights the fact that timely update of the gallery and/or training data can help to maintain the required performance of visual place recognition.

Limitations

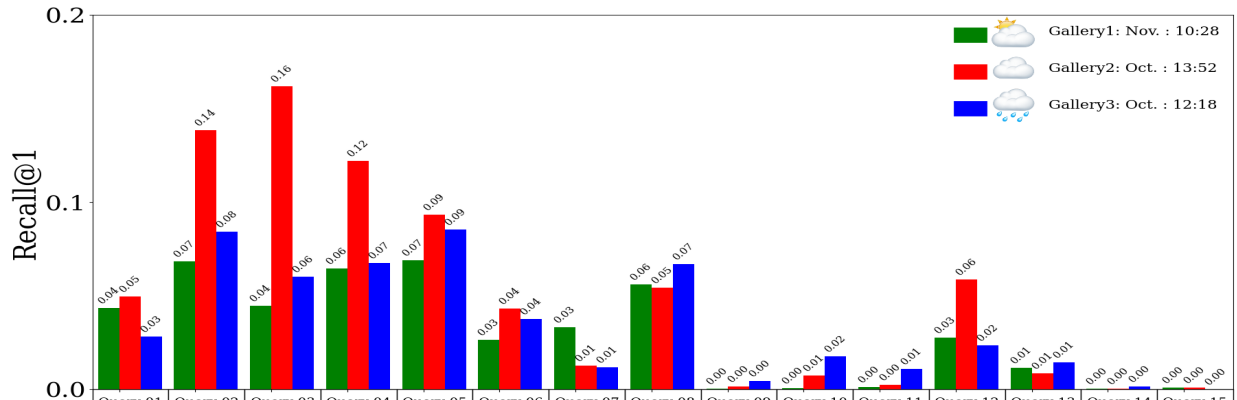
Our results are based on a single dataset with unverified location ground truth and which includes variation over weather and illumination. Therefore, our findings are merely indicative and require verification with another more controllable autonomous driving dataset.

References

[1] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Transactions on Robotics*, vol. 32, no. 1, pp. 1–19, 2016.

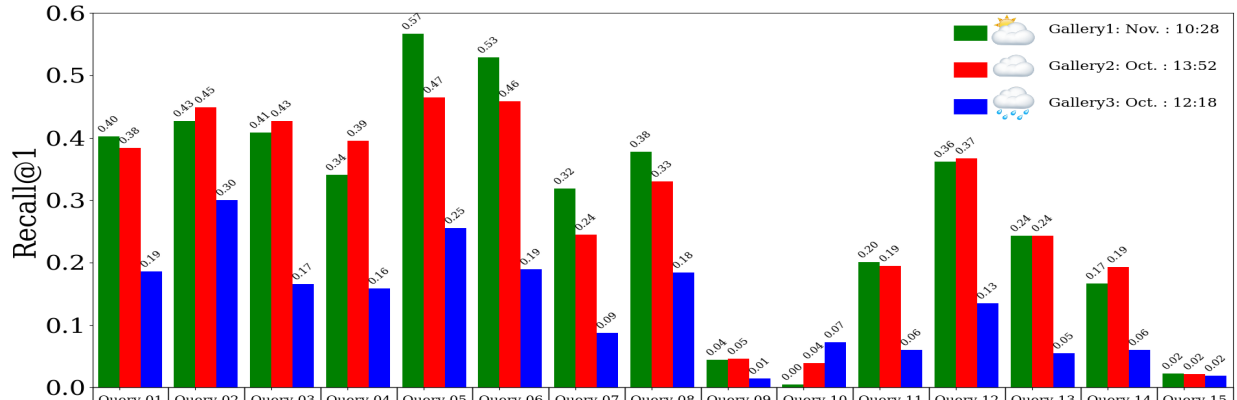
[2] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognition*, vol. 113, p. 107760, 2021.

[3] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The oxford robotcar dataset," *The International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.



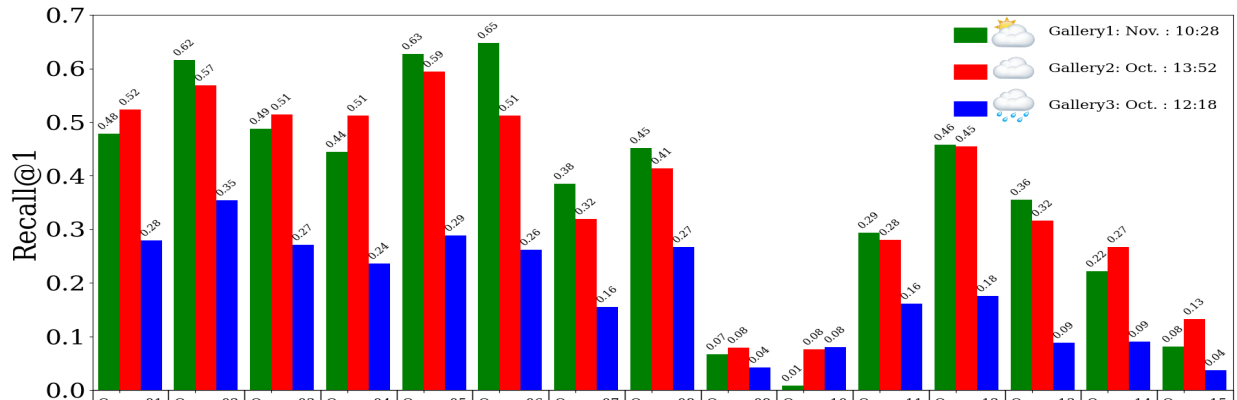
	Query 01	Query 02	Query 03	Query 04	Query 05	Query 06	Query 07	Query 08	Query 09	Query 10	Query 11	Query 12	Query 13	Query 14	Query 15
Month	Sep.	Aug.	Jul.	Jun.	May	Apr.	Mar.	Feb.	Dec.	Nov.	Jul.	Jul.	Jun.	Jun.	May
Start Time	10:37	16:02	15:23	08:09	14:06	09:06	13:47	08:45	18:18	16:34	15:42	14:49	09:53	14:47	13:05
Weather	sun	overcast	overcast	overcast	overcast	sun	sun	snow	night-rain	night	sun-cloud	overcast	overcast	sun	sun

(a) $\tau = 5.0m$, no training



	Query 01	Query 02	Query 03	Query 04	Query 05	Query 06	Query 07	Query 08	Query 09	Query 10	Query 11	Query 12	Query 13	Query 14	Query 15
Month	Sep.	Aug.	Jul.	Jun.	May	Apr.	Mar.	Feb.	Dec.	Nov.	Jul.	Jul.	Jun.	Jun.	May
Start Time	10:37	16:02	15:23	08:09	14:06	09:06	13:47	08:45	18:18	16:34	15:42	14:49	09:53	14:47	13:05
Weather	sun	overcast	overcast	overcast	overcast	sun	sun	snow	night-rain	night	sun-cloud	overcast	overcast	sun	sun

(b) $\tau = 5.0m$, fine-tuned with Train 1



	Query 01	Query 02	Query 03	Query 04	Query 05	Query 06	Query 07	Query 08	Query 09	Query 10	Query 11	Query 12	Query 13	Query 14	Query 15
Month	Sep.	Aug.	Jul.	Jun.	May	Apr.	Mar.	Feb.	Dec.	Nov.	Jul.	Jul.	Jun.	Jun.	May
Start Time	10:37	16:02	15:23	08:09	14:06	09:06	13:47	08:45	18:18	16:34	15:42	14:49	09:53	14:47	13:05
Weather	sun	overcast	overcast	overcast	overcast	sun	sun	snow	night-rain	night	sun-cloud	overcast	overcast	sun	sun

(c) $\tau = 25.0m$, fine-tuned with Train 1

Fig. 4: Long-term performance. Recall of each query sequence is reported separately and in temporal order. Note that the y-axes have different scaling.

[4] 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*, 2012, pp. 1643–1649.

[5] M. Cummins and P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," *The International Journal of Robotics Research*, vol. 27, no. 6, pp. 647–665, 2008.

[6] B. Williams, G. Klein, and I. Reid, "Automatic relocalization and loop closing for real-time monocular slam," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1699–1712, 2011.

[7] D. Galvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1188–1197, 2012.

[8] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *ArXiv*, vol. abs/1411.1509, 2014.

[9] 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)*, 2017, pp. 3223–3230.

[10] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 4297–4304.

[11] N. Sünderhauf, S. A. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, and M. Milford, "Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free," in *Robotics: Science and Systems*, 2015.

[12] D. Bai, C. Wang, B. Zhang, X. Yi, and X. Yang, "Sequence searching with cnn features for robust and fast visual place recognition," *Computers and Graphics*, vol. 70, pp. 270–280, 2018, cAD/Graphics 2017.

[13] N. Merrill and G. Huang, "Lightweight unsupervised deep loop closure," in *Proc. of Robotics: Science and Systems (RSS)*, Pittsburgh, PA, Jun. 26-30, 2018.

[14] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: Cnn architecture for weakly supervised place recognition," *TPAMI*, 2018.

[15] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 3304–3311.

[16] F. Radenović, G. Tolias, 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, 2019.

[17] B. Cao, A. Araujo, and J. Sim, "Unifying Deep Local and Global Features for Image Search," in *Computer Vision – ECCV 2020*. Cham, Switzerland: Springer, Nov 2020, pp. 726–743.

[18] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3476–3485.

[19] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[20] G. Pandey, J. R. McBride, and R. M. Eustice, "Ford campus vision and lidar data set," *The International Journal of Robotics Research*, vol. 30, no. 13, pp. 1543–1552, 2011.

[21] H. Badino, D. Huber, and T. Kanade, "Visual topometric localization," in *2011 IEEE Intelligent Vehicles Symposium (IV)*, 2011, pp. 794–799.

[22] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, F. Kahl, and T. Pajdla, "Benchmarking 6dof outdoor visual localization in changing conditions," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8601–8610.

[23] F. Warburg, S. Hauberg, M. López-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. 2623–2632.

[24] J.-L. Blanco-Claraco, F. Ángel Moreno-Dueñas, and J. González-Jiménez, "The málaga urban dataset: High-rate stereo and lidar in a realistic urban scenario," *The International Journal of Robotics Research*, vol. 33, no. 2, pp. 207–214, 2014.

[25] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt, "Image retrieval for image-based localization revisited," in *Proceedings of the British Machine Vision Conference*. BMVA Press, 2012, pp. 76.1–76.12.

[26] D. Olid, J. M. Fácil, and J. Civera, "Single-view place recognition under seasonal changes," *CoRR*, vol. abs/1808.06516, 2018.

[27] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, M. Pollefeys, J. Sivic, T. Pajdla, and A. Torii, "Inloc: Indoor visual localization with dense matching and view synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 4, pp. 1293–1307, 2021.

[28] ETH Zurich Computer Vision Group and Microsoft Mixed Reality & AI Lab Zurich, "The ETH-Microsoft Localization Dataset," https://github.com/cvg/visloc-iccv2021, 2021.

[29] D. Lee, S. Ryu, S. Yeon, Y. Lee, D. Kim, C. Han, Y. Cabon, P. Weinzaepfel, N. Guérin, G. Csurka, and M. Humenberger, "Large-scale localization datasets in crowded indoor spaces," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3226–3235.

[30] J. Peltomäki, F. Alijani, J. Puura, H. Huttunen, E. Rahtu, and J.-K. Kämäräinen, "Evaluation of long-term LiDAR place recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[31] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *ECCV*, 2016.

[32] F. Radenović, G. Tolias, and O. Chum, "Deep Shape Matching," in *Computer Vision – ECCV 2018*. Cham, Switzerland: Springer, Oct 2018, pp. 774–791.

[33] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, and P. Newman, "Real-time kinematic ground truth for the oxford robotcar dataset," *arXiv preprint arXiv: 2002.10152*, 2020.

[34] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "VPR-Bench: An Open-Source Visual Place Recognition Evaluation Framework with Quantifiable Viewpoint and Appearance Change," *Int. J. Comput. Vision*, vol. 129, no. 7, pp. 2136–2174, Jul 2021.

[35] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *CVPR 2011*, 2011, pp. 737–744.