Evaluation of Range Sensing-Based Place Recognition for Long-Term Urban Localization

Weixin Ma, *Student Member, IEEE*, Huan Yin, *Member, IEEE*, Lei Yao, *Student Member, IEEE*, Yuxiang Sun, *Member, IEEE*, and Zhongqing Su

***Abstract*—Place recognition is a critical capability for autonomous vehicles. It matches current sensor data with a pre-built database to provide coarse localization results. However, the effectiveness of long-term place recognition may be degraded by environment changes, such as seasonal or weather changes. To have a deep understanding of this issue, we conduct a comprehensive evaluation study on several state-of-the-art range sensing-based (i.e., LiDAR and radar) place recognition methods on the Borease dataset, which encapsulates long-term localization scenarios with stark seasonal variations and adverse weather conditions. In addition, we design a novel metric to evaluate the influence of matching thresholds on place recognition performance for long-term localization. Our results and findings provide fresh insights to the community and potential directions for future study.**

***Index Terms*—Autonomous vehicles, long-term localization, place recognition, range sensing, urban environments.**

I. Introduction

PLACE recognition (PR) refers to determining whether the current place has been visited previously against a pre-built keyframe-based database [1]. Such capability is critical for applications in mobile robots and autonomous vehicles, such as global localization and loop closure in Simultaneous Localization and Mapping (SLAM) [2], [3], [4], [5]. Significant progress has been made in the last two decades for both vision-based and range sensing-based methods [6], [7], [8]. However, reliable long-term PR still remains a challenge in complex road environments, where long-term variations frequently occur in both geometry and visual appearance.

Vision-based methods could be easily degraded due to dramatic changes in viewpoint, illumination, or weather conditions [9], [10]. Recently, range sensors, such as LiDAR and Frequency-Modulated Continuous Wave (FMCW) Radar have shown reasonable robustness to diverse weather conditions. Some methods have used them in SLAM and localization tasks under diverse weather conditions [11], [12], [13], [14], [15], [16], [17]. Radars are more robust to extreme weather conditions (e.g., rain or snow) than LiDARs. This is because radars work in GHz, which is lower than that of LiDARs (THz).

However, *to what extent do weather conditions influence radar- and LiDAR-based PR methods?* We find that this question has scarcely been investigated in existing literature, so we attempt to answer it by comparing state-of-the-art (SOTA) range sensing-based methods in this study. In addition to experiments, we also design a novel evaluation metric to assess influences caused by the matching thresholds [7] in long-term PR. The motivation for us to design the new metric is that existing works usually use the retrieval precision and recall [18], however, in long-term PR, the performance might be degraded when using the same matching threshold to determine whether a place has been visited. For example, a threshold that can achieve high precision and recall in summer might not provide satisfactory performance in winter. We argue that a robust long-term PR method should be able to achieve performance with acceptable variations by using a general threshold under seasonal changes.

To the best of our knowledge, this is the first comprehensive evaluation that explores the impact of seasonal and weather variations on range sensing-based place recognition methods in long-term scenarios, considering both performance metrics and matching thresholds. Our contributions are as follows:

1) We design a novel metric to evaluate the influences of matching thresholds on long-term place recognition performance.
2) We conduct a comprehensive evaluation of SOTA range sensing-based place recognition methods on a dataset with long-term localization scenarios to explore the impact of season and weather conditions.
3) We open-source our evaluation code and make the experimental results publicly available, which could inspire further works in this area from the research community[1].

This paper is organized as follows: Section II reviews related works; Section III presents related preliminaries on place

Manuscript received 22 February 2024; accepted 6 March 2024. Date of publication 21 March 2024; date of current version 26 July 2024. This work was supported in part by the Hong Kong Research Grants Council under Grant 15222523, in part by the Hong Kong Innovation and Technology Commission via the project "Smart Railway Technology and Applications" under Grant K-BBY1, and in part by the City University of Hong Kong under Grant 9610675 *(Corresponding authors: Yuxiang Sun; Zhongqing Su.)*

Weixin Ma and Zhongqing Su are with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: weixin.ma@connect.polyu.hk; zhongqing.su@polyu.edu.hk).

Huan Yin is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong (e-mail: eehyin@ust.hk).

Lei Yao is with the Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong (e-mail: rayyoh.yao@connect.polyu.hk).

Yuxiang Sun is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong (e-mail: yx.sun@cityu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIV.2024.3380083.

Digital Object Identifier 10.1109/TIV.2024.3380083

[1] [Online]. Available: https://github.com/Weixin-Ma/PR_Evaluation_Project

recognition; Section IV describes our proposed metric; Section V discusses the evaluation experiments; Conclusions and future work are drawn in the last section.

II. Related Work

This paper mainly focuses on single-shot PR methods using 360° LiDARs and radars. For filtering- or aggregating-based approaches, as well as PR methods using Non-repetitive LiDAR, readers may refer to the survey papers [1], [8].

A. Range Sensing-Based Place Recognition

LiDAR-based PR methods usually design global descriptors to compare the similarity between different LiDAR scans to retrieve places. They can be generally divided into handcrafted feature-based methods and data-driven methods. Early works usually rely on handcrafted local features, such as mean surface curvature [19] and local Normal Distribution Transform (NDT) [20]. Differently, some researchers use semantic objects to build their global descriptors instead of using low-level geometric information. Fan et al. [21] and Zhu et al. [22] both used topological information of objects in environments to build global descriptors. Instead of building global descriptors on the extracted features, projection-based methods generate global descriptors based on the projection results of raw point clouds. M2DP [23] projects raw point cloud into multiple 2D planes to extract Histogram descriptors. Scan Context [24] is a typical bird-eye-view (BEV) projection-based method, where a 2D matrix descriptor is used to embed the geometirc information. Similarly, Scan Context++ [25], SSC [26], Intensity Scan Context [27], DiSCO [28], RING++[29], and LiDAR-Iris [30] all project point clouds into BEV, followed by different global descriptor extraction methods. Instead of using handcrafted features, data-driven methods extract features from point clouds using deep neural networks. Uy et al. proposed PointNetVLAD [31], which uses NetVLAD [32] to aggregate features extracted from PointNet [33] into a global descriptor. MinkLoc3D [34] and its extension [35] use generalized-mean pooling layer [36] to aggregate local features extracted from sparse voxelized point could into global descriptors.

Radar-based PR remains a challenge due to its low spatial resolution and noise. Gadd et al. [37] used sequence matching to reduce influences of noise clusters in a single radar scan, achieving a 30% boost in performance. The authors further introduced a temporal data augmentation method to obtain a more robust descriptor [38]. Suaftescu et al. [39] combined cylindrical convolutions, anti-aliasing blurring, and azimuth-wise max-pooling to extract more reliable features from polar radar scans. Different from all the data-driven methods above, Hong et al. [13] used M2DP [23] to extract global descriptors from filtered 2D radar point clouds.

For long-term range sensing-based PR, there are few works. Alijani et al. [40] evaluated a SOTA visual PR method GEM [36] on the Oxford RobotCar dataset [41]. Their results show a performance degradation of approximately 6% every 100 days. Peltomaki et al. [42] fed LiDAR depth images into an image retrieval method CNNRetr [43] to assess the performance of long-term LiDAR PR. Instead of using depth images, Zywanowski et al. [44] combined camera images and LiDAR intensity images, which benefits the PR performance across weather conditions. However, only one data-driven-based method and one metric are evaluated in [42], [44]. Cao et al. [45] developed a global descriptor from a cylindrical image representation of a 3-D point cloud, which enhances robustness with a sequence-based check. They later proposed a two-head classification network for end-to-end long-term localization [46]. However, all these methods [45], [46] require a sequence of LiDAR scans and odometry to build a submap.

B. Performance Evaluation for Place Recognition

The evaluation of PR methods typically focuses on their place retrieval performance. Machine learning metrics for classification tasks are widely adopted in PR. Popular metrics include Precision-recall curves [24], maximum F_1 score [26], Recall@100% [31], Extended Precision (EP) [18], AUC-PR [28], and Recall@N [7]. Given similarity values between every query frame and their retrieved frame, different values of Precision and Recall can be computed by varying the matching threshold. The Precision-recall curve can be obtained by plotting Precision against the Recall, which summarizes the trade-off between the true positive rate and the positive predictive value using different matching thresholds. Maximum F_1 score, EP, and AUC-PR are all computed from the Precision-recall curve, indicating the performance of a PR method with a single value between 0 and 1. Recall@100% represents the Recall value at which Precision drops from 100%, which shows the highest Recall that can be reached before the first false positive occurs. Recall@N is computed by dividing the number of query frames with correct matches among the top-N retrieved frames by the total number of query frames.

C. Research Gap

Existing literature scarcely investigate and evaluate influences of seasonal changes on the range sensing-based (i.e., both LiDAR and radar) PR methods. Meanwhile, existing evaluation metrics almost focus on the performance from the perspective of Precision and Recall. Influences of matching thresholds on the performance in long-term conditions have also not been studied.

III. Preliminaries

A. Problem Formulation of Place Recognition

Given a query LiDAR frame Q and a prior database D_{ref} as shown in Fig. 1, place recognition aims to determine whether the frame Q has been visited previously in D_{ref} or not. This is determined by the similarity between Q and its most similar frame in D_{ref}. Once the similarity exceeds a predefined matching threshold, the frame Q is determined as a positive match, indicating its corresponding place has been visited previously in D_{ref}. Otherwise, the PR method considers the frame Q as a negative match.

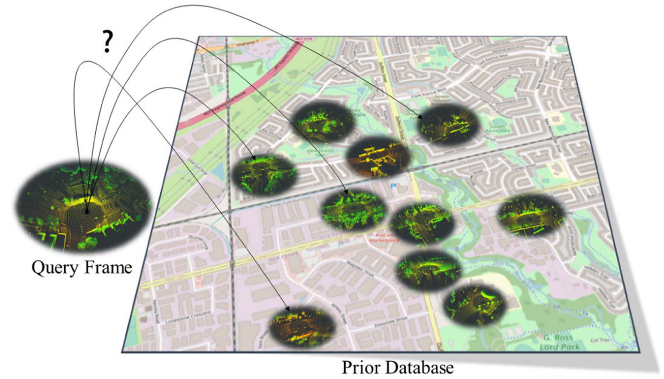

Fig. 1. Problem formulation of place recognition. Given a query frame from the current sensor data, place recognition methods determine whether it has been previously visited by matching it with a pre-built database.

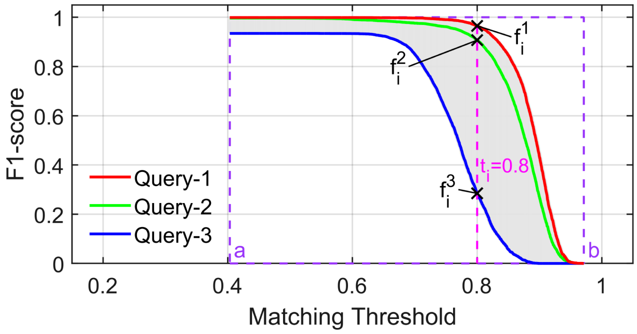

Fig. 2. Example for how to compute the proposed $AwC\text{-}FT$ (i.e., the area of gray area). The normalized $AwC\text{-}FT$, $\overline{AwC\text{-}FT}$ is the ratio of the original $AwC\text{-}FT$ to the area of the purple dashed box.

B. Precision, Recall, and F-Score

In PR, positive matches are fewer than negative matches. Precision-recall curve has been widely used to evaluate this imbalanced matching problem. Based on the matching results and the ground-truth information, correct positive matches are regarded as True-Positives (TP) whereas incorrect positive matches are regarded as False-Positives (FP). Similarly, True-Negatives (TN) and False-Negatives (FN) represent correct negative matches and incorrect negative matches, respectively. Precision and Recall are computed by Precision $= \frac{\text{TP}}{\text{TP+FP}}$ and Recall $= \frac{\text{TP}}{\text{TP+FN}}$, respectively.

Precision is the ratio of correctly identified positive matches, while Recall is the ratio of TP to actual positives. By adjusting the matching threshold, we can compute corresponding precision and recall values. The threshold typically ranges from the lowest to the highest similarity (or distance). A Precision-recall curve, plotting precision against recall, illustrates the trade-off between them under different thresholds.

F-score, F_β, is another widely used evaluation metric for PR, especially F_1 score. F_β considers both precision and recall. It is calculated by: $F_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{ Precision })+ \text{Recall}}$, where $\beta \in R^+$ is chosen to make Recall β times as important as Precision. When $\beta = 1$, we have F_1 score, which is the harmonic mean of Precision and Recall.

IV. The Proposed Evaluation Metric

The Precision-recall curve retains no information about the matching threshold. Other common evaluation metrics, like maximum F_1 score, EP, and AUC-PR, all summarize a Precision-recall curve into a single value between 0 and 1 to indicate the performance of a PR method. Therefore, the matching threshold information is missed. This information is important for long-term PR since a reliable long-term PR method is expected to achieve similar performance with a general matching threshold.

To assess such ability of a PR method, an intuitive idea is to keep the matching threshold unchanged and measure the performance variations for a PR method. For example, given a PR method, we can set a constant matching threshold and calculate metrics like Precision, Recall, or F-score for different sequences. The variations of the values of each metric can show the performance variations of a PR method. However, only the performance variation at a single point (i.e., the matching threshold) is evaluated. Meanwhile, it is not easy to select a specific value for the matching threshold. First, a higher/lower matching threshold will result in lower/higher Recall and higher/lower Precision across all the testing sequences, creating an illusion that the matching threshold has little influence on the performance variations. Second, the range of similarity values of the query results for different PR methods might be very different even using the same testing sequence. So it may be very difficult to find a specific threshold to fairly compare the influences of the matching threshold on the performance of different PR methods.

To solve these problems, we use the statistical result of the performance variations of a PR method instead of the performance variations at a single point. Considering the performance metrics, we choose to use the F-score since it provides a more comprehensive evaluation of a PR method by considering both Precision and Recall. Note that it is also acceptable to use other metrics like Precision and Recall here. Specifically, we introduce the metric $AwC\text{-}FT$, Area within Curves for FT-curves. For a given reference sequence $Seq\text{-}j$ and a query sequence $Seq\text{-}k$, where $k \in \{1, 2, \ldots, M\}$ is the index of the query sequence, a F-score Threshold curve (FT-curve) can be obtained by plotting F-score values against the matching thresholds. This is illustrated by the red, green, or blue curves in Fig. 2. In long-term PR, the reference sequence and the query sequence are usually different in terms of collection dates and environmental conditions, i.e., $j \neq k$. We denote $\{FT^k\}_j$ as the set of FT-curves that use different query sequences $Seq\text{-}k$ with the same reference sequence $Seq\text{-}j$. $AwC\text{-}FT$ can be computed based on the set $\{FT^k\}_j$:

$$AwC\text{-}FT \approx \sum_{i=1}^{N-1} \frac{\Delta_i + \Delta_{i+1}}{2} \times (t_{i+1} - t_i)\,,$$
$$\Delta_i = \text{Max}\,(S_i) - \text{Min}\,(S_i)\,,\ \ S_i = \left\{f_i^k\right\}_j. \tag{1}$$

where N is the number of matching thresholds. t_i is the i-th matching threshold. We use matching similarity (i.e., probability) instead of descriptor distance to determine the value for t_i, which can ensure the ideal maximum value and minimum value for t_i are 1 and 0, respectively. $\{f_i^k\}_j$ is the set of F-score values from the set $\{FT^k\}_j$ when matching threshold equals to t_i. Max$(\cdot)$ and Min$(\cdot)$ are respectively the maximum and minimum values of the given set. Δ_i is the maximum difference of the F-score values when the matching threshold equals t_i. An example for AwC-FT ($\beta = 1$) calculation is shown in Fig. 2. Three different sequences (i.e., the red, green, and blue curve) are used as query sequences to form FT-curves $\{FT^k\}_j$. $\{f_i^1, f_i^2, f_i^3\}$ are F_1 score values from the set $\{FT^k\}_j$ when $t_i = 0.8$. We have $\Delta_i = f_i^1 - f_i^3$. The value of AwC-FT equals to the area of the gray area shown in Fig. 2.

To simplify the calculation, we make the matching threshold t_i as a discrete uniform distribution $U\{a, b\}$. Let $s_l^k \in [0, 1]$ be the similarity between l-th frame of query sequence Seq-k and the retrieved frame from the given reference sequence, then $\{s_l^k\}_j$ is the set of matching similarities for all frames from query sequence Seq-k when using Seq-j as the reference. Values of a and b can be calculated as following:

$$a = \text{Min}\left(\text{Min}\left(\left\{s_l^1\right\}_j\right), \ldots, \text{Min}\left(\left\{s_l^M\right\}_j\right)\right). \tag{2}$$

$$b = \text{Max}\left(\text{Max}\left(\left\{s_l^1\right\}_j\right), \ldots, \text{Max}\left(\left\{s_l^M\right\}_j\right)\right). \tag{3}$$

Therefore, $t_{i+1} - t_i$ can be simplified as to $\frac{b-a}{N}$. Equation (1) can be rewritten as:

$$AwC\text{-}FT \approx \frac{b-a}{N} \times \left(\frac{\Delta_1}{2} + \sum_{i=2}^{N-1} \Delta_i + \frac{\Delta_N}{2}\right). \tag{4}$$

The calculated AwC-FT is then normalized by dividing $((b-a) \times (1-0))$. The normalized AwC-FT is defined as $\overline{AwC\text{-}FT} \in [0, 1]$, representing the ratio of the original AwC-FT and the area of the purple dashed box, as shown in Fig. 2.

$$\overline{AwC\text{-}FT} \approx \frac{1}{N} \times \left(\frac{\Delta_1}{2} + \sum_{i=2}^{N-1} \Delta_i + \frac{\Delta_N}{2}\right). \tag{5}$$

The $\overline{AwC\text{-}FT}$ metric approximately represents the average performance variations of a PR method when conducting place recognition using different query sequences (i.e., query sequences collected on different date and weather conditions) and the same reference sequence, with the same matching thresholds. Theoretically, a larger value of $\overline{AwC\text{-}FT}$ indicates that the performance of the PR method is more sensitive to the matching thresholds under seasonal changes in long-term scenarios. Since $\overline{AwC\text{-}FT}$ only represents the performance variation, it is better to use the metric and the other metrics (e.g., Maximum F_1 score) at the same time for more comprehensive evaluation results. In this paper, we use the F_1 score for $\overline{AwC\text{-}FT}$. Unless specifically stated, otherwise all experimental results related to $\overline{AwC\text{-}FT}$ are calculated based on the F_1 score.

TABLE I
DETAILS OF THE EVALUATION SEQUENCES

Seq-ID	Weather Condition	Frame Number
2020-11-26 (Seq-01)	overcast, snow	3305
2020-12-01 (Seq-02)	overcast, snow, snowing	3291
2021-01-26 (Seq-03)	overcast, snow, snowing (heavy)	4047
2021-02-02 (Seq-04)	overcast, snow (severe)	3324
2021-04-08 (Seq-05)	sun	3104
2021-04-29 (Seq-06)	overcast, rain	3181

Seq-ID refers to the collection date. Notation in a blanket is used to represent the corresponding sequence for simplicity. The frame number here is the number after down-sampling and filtering.

V. The Evaluations

A. Dataset and Experimental Setup

KITTI [48], Oxford Radar [49], and MulRan [50] are urban environment datasets that have been widely used in range sensing-based PR. However, none of these datasets contain both season and weather variations, since their collection dates span less than 3 months. Alternatively, we use Boreas Dataset [47], which includes more than 350 km of data collected by driving a repeated route over one year. Seasonal variations and adverse weather conditions, such as rain and snowstorms, can be found in the dataset. As for sensor configurations, the data-collection vehicle has a camera, a 360° radar, a 128-beam LiDAR, and GPS/IMU. Similar to [12], we select 6 sequences with stark weather variations for evaluation. Sample weather variations can be found in Fig. 3. We down-sample LiDAR frames to the scan frequency of the 360° radar (i.e., 4 Hz). The down-sampled sequences are further filtered to keep the distance between two consecutive frames not less than 1 m. Details of the evaluation sequences can be found in Table I.

We use several SOTA open-sourced PR methods: Scan Context [24], LiDAR-Iris [30], MinkLoc3Dv2 [51], and Overlap-Transformer [52]. Scan Context and LiDAR-Iris are handcrafted feature-based methods. They both extract descriptors from the projection results of point clouds, while LiDAR-Iris compares the similarity between two LiDAR frames in the frequency domain. MinkLoc3Dv2 and OverlapTransformer are data-driven methods, while their loss functions are based on localization and overlap, respectively. We use the default parameters for Scan Context and LiDAR-Iris provided by the authors. Following the existing long-term PR works [40], [42], we fine tune MinkLoc3Dv2 and OverlapTransformer on another sequence 2020-12-18 which is different from all the other evaluation sequences.

During evaluation, we alternately use each sequence as the reference sequence, and the remaining five sequences as the query sequence. For example, if Seq-01 is the reference, Seq-02 to 06 are queries. Alternatively, if Seq-02 is the reference, Seq-01, 03, 04, 05, and 06 are queries. This allows testing of long-term PR performance under various seasonal conditions, like recognizing places on rainy days against a snowy day database, or vice versa. Specifically, we denote a sequence pair as $\langle k, j\rangle$, which contains a query sequence Seq-k and a reference sequence Seq-j. $\{\langle k, j\rangle\}_j$ is the set of sequence pairs that have the same reference sequence Seq-j. Here we

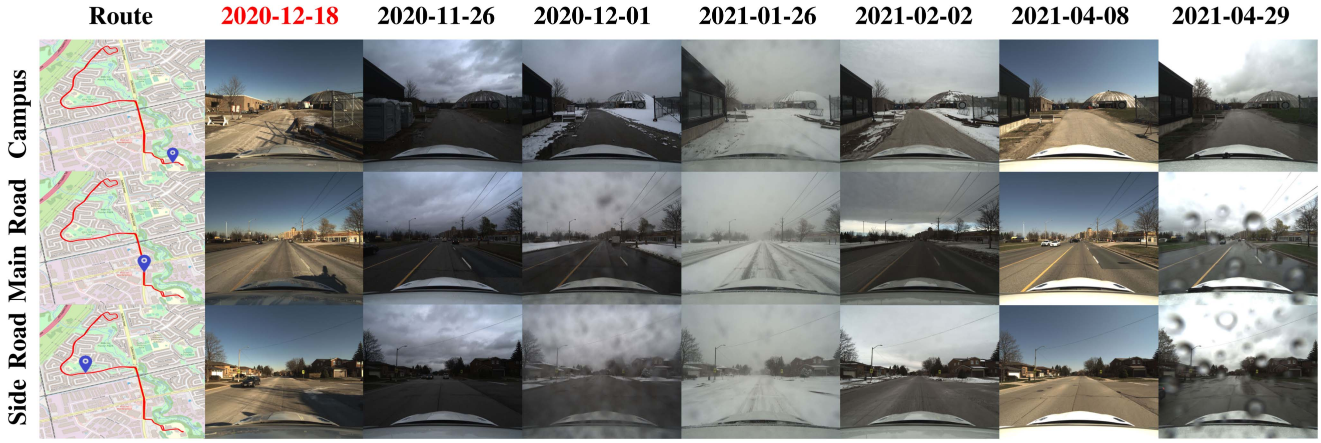

Fig. 3. Examples of seasonal variations across sequences from Boreas Dataset [47]. All the sequences are collected along the same route on different dates. Pictures in each row are captured in the same place. Sequence 2020-12-18 is used to fine tune the data-driven place recognition methods. All the other sequences are used for evaluation.

have $j, k \in \{01, 02, 03, 04, 05, 06\}$ *and* $j \neq k$, representing the 6 evaluation sequences by the order of collection date. Following [40] and [42], we conduct all experiments using the top-1 retrieval results, i.e., only the retrieved frame with the highest similarity is used. In accordance with prior research [24] and [30], a matching detection is classified as a true positive if the ground-truth pose distance between the query frame and the matched frame is less than 4 m.

B. Evaluation With Widely-Used Metrics

For a given PR method and $\langle k, j\rangle$, we evaluate its performance with several widely-used metrics: maximum F_1 score, EP, and Recall@1. The experimental results are displayed in Table II. We also calculate the average values and standard deviations of the above metrics for each $\{\langle k, j\rangle\}_j$ to show the variations of PR performance.

LiDAR-Iris achieves the best performance in terms of all the metrics under all the seasonal conditions. The average values of the maximum F_1 score, EP, and Recall@1 are all higher than 99.5% with a small standard deviation (less than 0.5%) except the average of EP when using sequence Seq-03 as the reference (i.e., 97.83%). Scan Context also shows a competitive performance. There is only a minor decline observed across all the evaluation metrics. We can find prominent degradation for both OverlapTransformer and MinKLoc3Dv2.

According to Recall@1 values, there are respectively around 60% and 50% of the matching results that are correct for OverlapTransformer and MinkLock3Dv2, while more than 96% and 97% of the matching results are positive for Scan Context and LiDAR-Iris under all seasonal conditions. EP provides a good summary on both P_{R0} (i.e., the Precision at the minimum Recall value) and R_{P100} (i.e., the Recall value where the Precision drops from 100%) [18]. When EP is less than 0.5, P_{R0} is less than 1.0, meaning there exist false positives even using the highest matching threshold and the PR method can never provide a matching result at 100% Precision. We can see that the averages of EP for both MickLoc3Dv2 and OverlapTransformer in almost $\{\langle k, j\rangle\}_j$ are less than 0.5, while the average values of EP for both Scan Context and LiDAR-Iris are both larger than 0.94. This shows that Scan Context and LiDAR-Iris can reach a higher Recall without any false positives.

Intuitively, when the seasonal and weather conditions change between query and reference sequences, the performance of a PR method should vary. Interestingly, such performance variations are almost negligible for LiDAR-Iris. For instance, when using a snowstorm sequence Seq-03 as the reference, there was only a slight decrease in Recall@1 compared to using the other sequences as references. Regarding Scan Context, there is an average performance degradation of about 8% in EP when using Seq-03 as the reference. Meanwhile, unlike LiDAR-Iris, the values of average EP and Recall@1 for Scan Context both decrease when using the snowstorm Seq-03 as reference sequence. Similar trends of performance variations can also be found in the results for OverlapTransformer and MinkLoc3Dv2. However, such degradation is more dramatic than that for Scan Context and LiDAR-Iris. For example, for OverlapTransformer, the maximum F_1 score and Recall@1 decrease by about 40% on average when using Seq-03 as a reference sequence. Such degradation is even more prominent for MinkLoc3Dv2, which is more than 50%.

Our results show that the SOTA handcrafted feature-based LiDAR methods can be robust to seasonal variations in long-term PR. Two SOTA data-driven methods (i.e., OverlapTransformer and MinkLoc3Dv2) demonstrate a heightened sensitivity to seasonal variations. We guess the main reason is that these two networks were initially designed and trained using other datasets that vary from Boreas Dataset in terms of both scene layouts and sensor configuration (i.e., resolution and installation position). Limited by the network generalizability, the overall performance degrades than those reported in [51] and [52]. In addition, insufficient training data collected under different weather conditions may result in a significant performance variation during evaluations.

TABLE II
PR PERFORMANCE FOR DIFFERENT $\{\langle k, j\rangle\}_j$

Ref	Que	SC			Iris			OT			Mink		
		F_1	EP	R@1	F_1	EP	R@1	F_1	EP	R@1	F_1	EP	R@1
Seq-01	Seq-02	99.68	96.45	99.36	99.91	99.66	99.64	85.91	51.05	75.30	62.96	51.20	45.94
	Seq-03	93.46	82.42	87.72	100.00	100.00	100.00	27.53	50.02	15.96	1.23	0.00	0.62
	Seq-04	99.95	99.83	99.91	99.98	99.98	99.94	86.49	53.64	76.20	88.10	56.35	78.73
	Seq-05	99.98	99.81	99.97	99.98	99.98	99.97	86.30	51.80	75.90	86.49	56.68	76.19
	Seq-06	99.89	99.73	99.78	99.87	99.69	99.75	80.32	51.51	67.12	84.73	54.92	73.50
	ave	98.59↑	95.65↑	97.35↑	**99.95**↑	99.86↑	99.86↑	73.31↗	51.61→	62.10↗	64.70↗	43.83→	55.00→
	std	2.57	6.74	4.82	0.05	0.15	0.14	23.01	1.18	23.31	33.03	22.00	29.64
Seq-02	Seq-01	99.62	96.14	99.24	99.91	99.29	99.73	86.31	52.26	75.92	62.86	51.09	45.84
	Seq-03	97.91	90.97	95.90	99.99	99.94	99.75	31.30	0.00	18.56	1.37	0.00	0.69
	Seq-04	99.44	95.66	98.89	99.95	99.50	99.82	85.56	50.95	74.76	60.42	51.21	43.29
	Seq-05	99.29	97.92	99.72	99.95	98.76	99.90	83.59	50.40	71.81	68.61	50.48	52.22
	Seq-06	99.49	96.19	98.99	99.95	99.94	99.81	80.03	51.26	66.71	65.98	51.16	49.23
	ave	99.15↑	95.38↑	98.55↑	**99.95**↑	99.49↑	99.80↑	73.36↗	40.97→	61.55↗	51.85→	40.79→	38.26↘
	std	0.63	2.33	1.36	0.02	0.44	0.06	21.14	20.50	21.73	25.39	20.40	19.02
Seq-03	Seq-01	98.36	88.46	96.67	99.91	99.77	98.03	36.22	50.03	22.12	2.39	0.00	1.21
	Seq-02	98.82	89.75	97.11	99.95	99.95	97.54	41.05	0.00	25.83	2.99	0.00	1.52
	Seq-04	98.62	83.82	96.84	99.86	99.83	97.92	40.88	50.02	25.69	1.73	0.00	0.87
	Seq-05	98.04	89.50	96.04	99.77	99.74	97.58	34.54	50.02	20.88	2.10	0.00	1.06
	Seq-06	98.27	84.52	96.32	99.84	99.63	98.05	31.11	0.00	18.42	1.75	0.00	0.88
	ave	98.42↑	87.21↑	96.60↑	**99.87**↑	99.78↑	97.83↑	36.76↘	30.01↘	22.59↘	2.19↓	0.00↓	1.11↓
	std	0.27	2.53	0.38	0.06	0.11	0.22	3.81	24.51	2.85	0.47	0.00	0.24
Seq-04	Seq-01	99.91	99.55	99.82	100.00	100.00	99.85	85.84	51.23	75.19	88.17	54.17	78.85
	Seq-02	99.50	96.89	99.00	99.92	99.82	99.73	85.21	50.98	74.23	59.51	51.17	42.36
	Seq-03	93.35	82.59	87.52	99.98	99.90	99.83	28.23	50.01	16.43	0.84	0.00	0.42
	Seq-05	99.98	99.32	99.98	99.98	99.73	99.97	86.36	51.08	76.00	85.63	54.24	74.87
	Seq-06	99.94	99.87	99.87	99.97	99.84	99.94	80.10	50.30	66.80	83.03	52.00	70.98
	ave	98.53↑	95.64↑	97.24↑	**99.97**↑	99.86↑	99.86↑	73.15↗	50.72→	61.73↗	63.44↗	42.31→	53.50→
	std	2.60	6.61	4.87	0.03	0.09	0.09	22.57	0.48	22.89	32.93	21.19	29.48
Seq-05	Seq-01	99.94	99.94	99.88	99.98	99.88	99.97	85.98	51.47	75.40	86.39	56.35	76.04
	Seq-02	99.47	95.96	98.94	99.97	99.85	99.94	82.28	50.36	69.89	67.11	50.44	50.50
	Seq-03	91.43	79.75	84.21	99.88	99.70	99.75	25.49	50.01	14.60	0.64	0.00	0.32
	Seq-04	99.94	99.68	99.88	100.00	100.00	100.00	86.82	50.77	76.71	84.72	52.90	73.50
	Seq-06	99.94	99.34	99.87	99.98	99.98	99.98	81.59	50.68	68.91	84.95	53.13	73.84
	ave	98.14↑	94.93↑	96.56↑	**99.96**↑	99.88↑	99.93↑	72.43↗	50.66→	61.10↗	64.76↗	42.57→	54.84→
	std	3.36	7.73	6.18	0.04	0.11	0.09	23.56	0.48	23.45	32.84	21.37	28.81
Seq-06	Seq-01	99.82	99.06	99.61	99.92	99.48	99.73	79.95	50.86	66.60	84.54	54.83	73.22
	Seq-02	99.59	96.57	99.18	99.97	99.94	99.70	81.30	50.62	68.49	64.54	51.10	47.65
	Seq-03	92.30	76.38	85.69	99.90	99.76	99.70	23.47	0.00	13.29	1.08	0.00	0.54
	Seq-04	99.95	99.88	99.91	99.98	99.82	99.97	81.98	50.32	69.46	81.83	51.84	69.25
	Seq-05	99.92	99.47	99.84	99.94	99.39	99.87	83.13	50.73	71.13	85.10	56.29	74.07
	ave	98.32↑	94.27↑	96.85↑	**99.94**↑	99.68↑	99.79↑	69.97↗	40.51→	57.80→	63.42↗	42.81→	52.95→
	std	3.01	9.02	5.58	0.03	0.21	0.11	23.27	20.25	22.30	32.07	21.49	27.92

"F_1" is short for maximum F_1 score (unit:%). "EP" is short for Extended Precision (unit:%). "R@1" is short for Recall@1 (unit:%). "ave" is short for average. "std" is short for standard deviation. "Ref" is short for reference. "Que" is short for query. "SC" is short for Scan Context. "Iris" is short for LiDAR-Iris. "OT" is short for overlapTransformer. "Mink" is short for MinkLoc3Dv2. The best result for each evaluation metric is emphasized with different formats (i.e., bold face for F_1, wavy line for EP, and underline for R@1). Arrows with different directions indicate different value ranges. ↓ refers to [0, 20]. ↘ refers to [20, 40]. → refers to [40, 60]. ↗ refers to [60, 80]. ↑ refers to [80, 100].

C. Influences by Matching Thresholds

In the last part, we evaluate the performance of several SOTA LiDAR PR methods by using three widely-used metrics. Such evaluation provides a general statement about how good the performance can be. In this section, we evaluate a long-term PR method from a new perspective by using the proposed metric. For every $\{\langle k, j\rangle\}_j$, we compute its $\overline{AwC\text{-}FT}$, as shown in Table III. The results are also visualized in Fig. 4.

We summarize the main findings as follows: Firstly, using the same matching thresholds, the average F_1 score variations for handcrafted feature-based methods is smaller than those for data-driven methods. As shown in Table III, the average $\overline{AwC\text{-}FT}$ values for LiDAR-Iris and Scan Context are

TABLE III
RESULTS OF $\overline{AwC\text{-}FT}$ FOR DIFFERENT $\{\langle k, j\rangle\}_j$

Ref	Input: LiDAR				Input: Radar	
	SC	Iris	OT	Mink	SC	Iris
Seq-01	25.54	23.06	56.79	85.78	**4.78**	5.84
Seq-02	13.72	17.77	52.78	66.16	**3.32**	5.32
Seq-03	8.49	3.29	9.17	**0.90**	4.65	4.49
Seq-04	25.94	21.33	55.53	67.88	**3.45**	3.86
Seq-05	25.98	20.90	58.32	84.60	**3.14**	4.42
Seq-06	25.31	19.48	55.31	82.89	**5.10**	7.66
ave	20.83	17.64	47.98	64.70	**4.07**	5.26

The best result for each $\{\langle k, j\rangle\}_j$ (i.e., each row) is highlighted in bold font. "Ref" is short for reference.

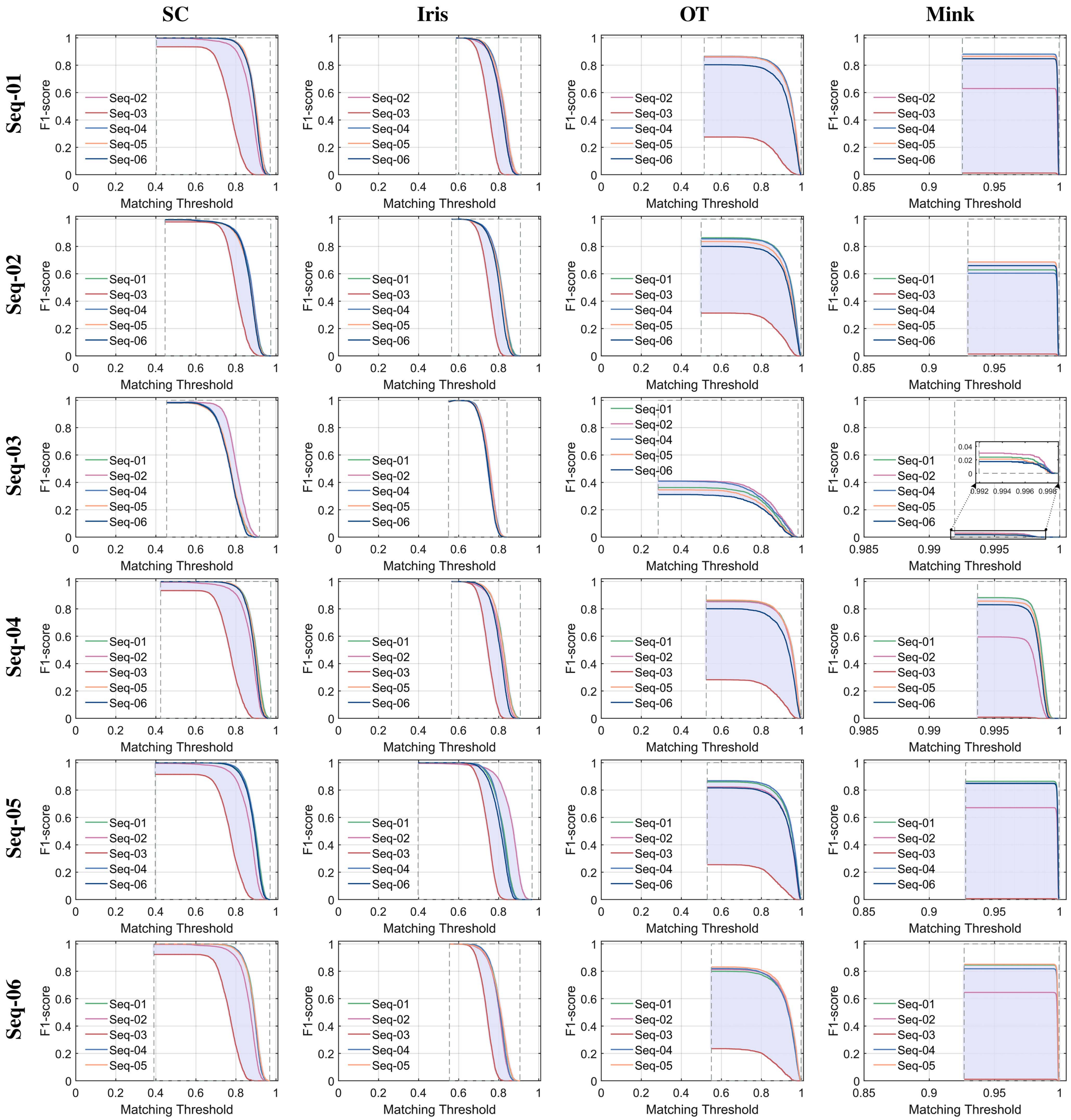

Fig. 4. FT-curves $\{FT^k\}$ for different $\{\langle k, j\rangle\}_j$ using several SOAT LiDAR-based PR methods. Sequences ID on the left refers to the used reference sequence. Figures in each column are based on the same PR method.

much smaller than those for OverlapTransformer and MinkLoc3Dv2. Such variations can be also observed in Fig. 4. Secondly, we interestingly find that $\overline{AwC\text{-}FT}$ decreases when the reference sequence Seq-j for $\{\langle k, j\rangle\}_j$ is significantly different from all the other query sequence Seq-k. Specifically, the snowstorm sequence Seq-03 differs from all the other sequences in terms of weather conditions. When using Seq-03 as the reference sequence, the variation in performance is much smaller than those results when using the other sequences as the reference, as shown in the third row in both Table III and Fig. 4.

Considering practical applications, $\overline{AwC\text{-}FT}$ not only can evaluate the influence of matching thresholds on the performance of different PR methods in long-term localization but also can directly benefit the choice of reference sequences. For example, from Table III, we can find that the performance variation is the smallest when using Seq-03 as a reference. However, the average values for maximum F_1 score, EP, and R@1 are smaller than those when using the other sequences as reference (see Table II). In other words, using Seq-03 as the reference sequence leads to a poor yet stable result in long-term PR. In addition, the absolute performances for all the methods are very close when

TABLE IV
AVERAGE OF RECALL@1 (%) FOR DIFFERENT REGIONS FOR DIFFERENT $\{\langle k, j\rangle\}_j$

Ref	Region-01				Region-02				Region-03			
	SC	Iris	OT	Mink	SC	Iris	OT	Mink	SC	Iris	OT	Mink
Seq-01	94.68↑	**99.84**↑	73.26↗	52.12→	93.90↑	99.52↑	34.55↘	49.83→	99.68↑	$\underline{100.00}$↑	68.82↗	57.88→
Seq-02	98.28↑	**99.94**↑	75.34↗	41.10→	94.41↑	99.17↑	31.25↘	22.08↘	99.85↑	$\underline{100.00}$↑	68.50↗	43.48→
Seq-03	85.66↑	**89.60**↑	40.26→	1.44↓	97.65↑	99.42↑	8.87↓	0.25↓	99.89↑	$\underline{99.99}$↑	21.86↘	1.33↓
Seq-04	94.54↑	**99.67**↑	75.60↗	54.24→	93.50↑	99.67↑	34.56↘	47.88→	99.69↑	$\underline{100.00}$↑	67.62↗	55.37→
Seq-05	93.71↑	**99.83**↑	76.44↗	54.66→	91.50↑	98.92↑	33.77↘	47.38→	99.59↑	$\underline{100.00}$↑	66.73↗	57.79→
Seq-06	93.72↑	**99.60**↑	71.45↗	49.59→	93.64↑	99.47↑	33.94↘	50.08→	99.23↑	$\underline{99.98}$↑	62.11↗	55.19→

Results in each row are computed using the same $\{\langle k, j\rangle\}_j$. The best result for each row is highlighted. Different formats are used to represent the best results from different regions (i.e., bold face for Region-01, wavy line for Region-02, and underline for Region-03).

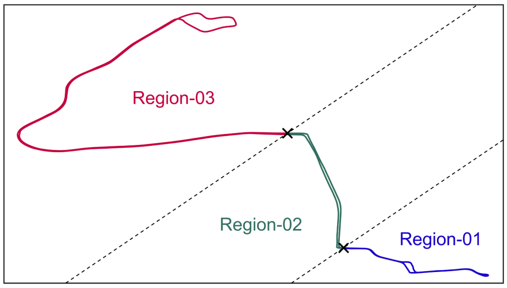

Fig. 5. Example of the segmented regions along the route. × refers to the locations used to segment each sequence into three different regions.

TABLE V
FRAME NUMBER OF EACH SEGMENTED REGION FOR DIFFERENT SEQUENCES

Seq-ID	Frame Number			Total
	Region-01	Region-02	Region-03	
Seq-01	632	774	1899	3305
Seq-02	656	731	1904	3291
Seq-03	818	927	2302	4047
Seq-04	654	707	1963	3324
Seq-05	636	639	1829	3104
Seq-06	604	774	1803	3181

using Seq-01, 02, 04, 05, and 06 as references. It is reasonable to select any of them as the reference. Nevertheless, when using Seq-02 as the reference sequence, we find that the $\overline{AwC\text{-}FT}$ values for all the methods are the smallest. So, choosing Seq-02 as the reference achieves a robust and competitive long-term performance.

Overall, the experimental results show that the performance influenced by different matching thresholds vary across handcrafted feature-based and data-driven methods. The proposed metric can effectively evaluate such influence to show the robustness of the methods to the matching thresholds. Moreover, it can also benefit the choice of reference sequence.

D. Matching Similarity Distributions

In large-scale environments, there may be significant variations in geometry across different regions, such as building styles and road layouts. These variations can directly influence the performance of PR methods. To assess influences of these variations on PR performance, we divide each query sequence into distinct regions based on ground-truth poses. Specifically, there are three different regions, denoted as Region-01 (campus region), Region-02 (main-road region), and Region-03 (side-road region), as shown in Fig. 5. Region-01 and 03 exhibit fewer dynamic objects, such as vehicles and pedestrians, whereas Region-02 is highly dynamic with many moving vehicles. Details about the frames of each region on different sequences can be found in Table V.

For each sequence pair $\langle k, j\rangle$ from the set $\{\langle k, j\rangle\}_j$, we first compute the Recall@1 for each region, i.e., the ratio of positive matches within a specific region to the number of query frames within the same region. We then compute the average of Recall@1 for each region. Results are displayed in Table IV. Values in each row are computed using the same $\{\langle k, j\rangle\}_j$. To better present the differences in matching results in different regions, we visualize some examples of the matching similarity distributions along the trajectory of query sequence as heat maps, as shown in Fig. 6.

In general, LiDAR-Iris demonstrates superior performance across all three regions, particularly in Region-03 where the average of Recall@1 approaches nearly 100% for all $\{\langle k, j\rangle\}_j$. As expected, for all the methods, the average Recall@1 of Region-02 is generally lower than those of the other two regions. This can be also supported by the matching similarity distributions in different regions as shown in Fig. 6. The heat map colors indicating the similarity of matching results in Region-02 are lighter than those in Region-01 and Region-03. This is particularly noticeable in the results for LiDAR-Iris (i.e., second row) and the radar-based Scan Context (i.e., fifth row). Compared to the handcrafted feature-based methods, the average Recall@1 for Region-02 significantly decreases when using the data-driven methods. This suggests, to some extent, that the influences of dynamic objects on Recall@1 is more pronounced for data-driven methods.

Based on our experimental results, the handcrafted feature-based methods are more robust to the variations in geometry across different regions in large-scale environments.

E. Comparisons With Radar-Based Methods

Due to the robustness to diverse weather conditions, radar has recently gained a lot of attention, showing significant potential in long-term localization. However, radar suffers from multiple sources of artifacts and clutters, for example, speckle noise, receiver saturation, and multi-path reflections. So, we first use the filtering method [53] to reduce noise. The filtered points are

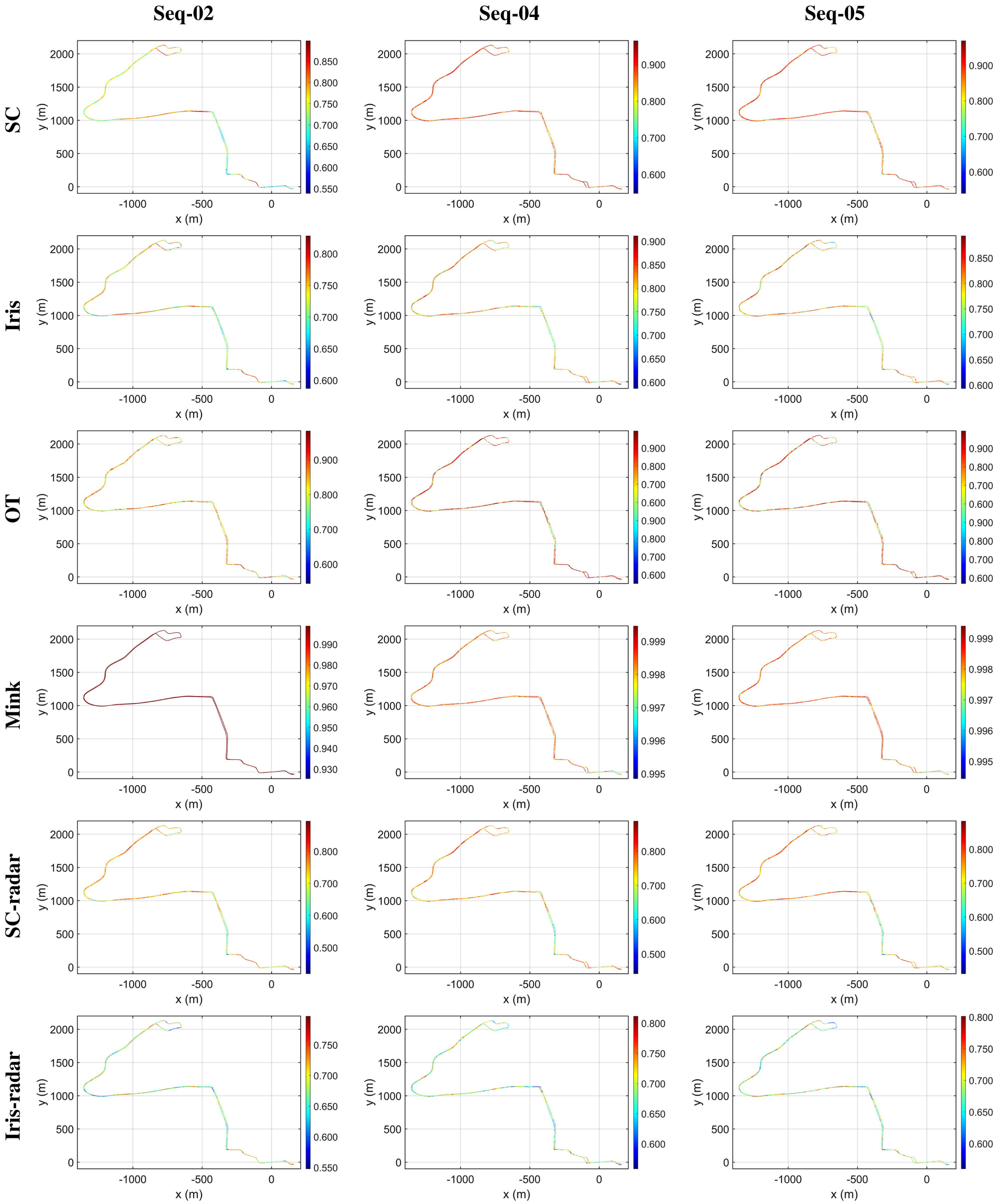

Fig. 6. Examples of top-1 matching similarity distributions. All the results are based on the same reference sequence, Seq-01. The sequence ID on the top refers to the query sequence used in each column.

then transferred from 2D radar images into 3D point clouds while the z-coordinate is set as 1 for all the filtered points. We then use the generated point clouds as inputs to run Scan Context and LiDAR-Iris. Results for maximum F_1 score, EP, and Recall@1 are displayed in Table VI.

Different from LiDAR-based methods, radar-based Scan Context outperforms the radar-based LiDAR-Iris. The radar-based Scan Context exhibits improvements in both maximum F_1 score and Recall@1 compared to the LiDAR-based Scan Context. The average values of maximum F_1 score and average values of Recall@1 for different $\{\langle k, j\rangle\}_j$ are all larger than 99.6% and 99.0%, respectively. The value of average EP increases in almost $\{\langle k, j\rangle\}_j$. Regarding radar-based LiDAR-Iris, the decrease of the maximum F_1 score

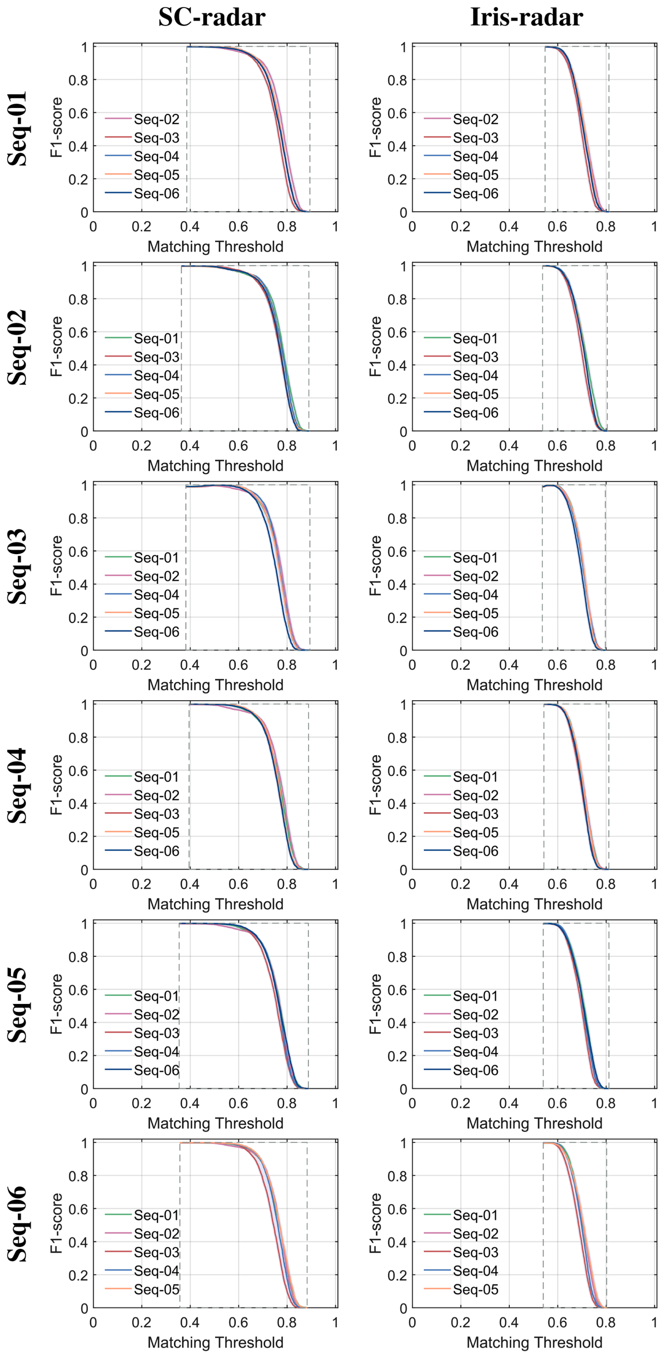

Fig. 7. FT-curves $\{FT^k\}$ for different $\{\langle k, j\rangle\}_j$ using different radar-based PR methods. Sequence ID on the left refers to the used reference sequence in each row. Figures in each column are based on the same method.

and Recall@1 can be deemed insignificant. From the perspective of EP, there is about a 10% drop. We conjecture the reason could be the loss of height information in the radar-based point cloud, which is important for LiDAR-Iris to generate its global descriptor. Surprisingly, these two handcrafted feature-based LiDAR methods show good generalizability on radar sensors.

TABLE VI
PR PERFORMANCE FOR DIFFERENT $\{\langle k, j\rangle\}_j$ USING RADAR-BASED METHODS

Ref	Que	SC-radar			Iris-radar		
		F_1	EP	R@1	F_1	EP	R@1
Seq-01	Seq-02	99.59	94.42	99.18	99.68	95.68	99.36
	Seq-03	99.80	99.23	98.00	99.80	98.35	97.97
	Seq-04	99.89	99.12	99.79	99.92	99.86	99.79
	Seq-05	99.91	92.84	99.82	99.95	99.30	99.91
	Seq-06	99.68	96.08	99.36	99.54	77.20	99.03
	ave	99.77↑	96.34↑	99.23↑	**99.78**↑	94.08↑	99.21↑
	std	0.12	2.53	0.66	0.15	8.56	0.69
Seq-02	Seq-01	99.76	93.69	99.51	99.63	80.38	99.27
	Seq-03	99.49	95.61	97.42	99.49	94.67	97.36
	Seq-04	99.65	94.86	99.30	99.68	95.81	99.33
	Seq-05	99.66	93.66	99.33	99.80	79.02	99.60
	Seq-06	99.73	97.23	99.45	99.71	98.67	99.42
	ave	**99.66**↑	95.01↑	99.00↑	**99.66**↑	89.71↑	99.00↑
	std	0.09	1.33	0.80	0.10	8.29	0.83
Seq-03	Seq-01	99.95	99.60	99.90	99.93	84.40	99.85
	Seq-02	99.79	96.93	99.58	99.79	74.17	99.58
	Seq-04	99.85	97.33	99.70	99.88	99.46	99.68
	Seq-05	99.80	94.53	99.60	99.84	98.80	99.68
	Seq-06	99.84	98.37	99.60	99.65	90.21	99.28
	ave	**99.85**↑	97.35↑	99.68↑	99.82↑	89.41↑	99.61↑
	std	0.06	1.69	0.12	0.09	9.46	0.19
Seq-04	Seq-01	99.92	82.24	99.85	99.95	83.12	99.91
	Seq-02	99.80	95.92	99.61	99.85	88.81	99.70
	Seq-03	99.82	97.02	97.89	99.79	99.52	97.86
	Seq-05	99.92	95.32	99.85	99.94	93.80	99.88
	Seq-06	99.86	91.58	99.73	99.83	95.05	99.67
	ave	**99.87**↑	92.42↑	99.39↑	**99.87**↑	92.06↑	99.40↑
	std	0.05	5.41	0.75	0.06	5.63	0.78
Seq-05	Seq-01	99.87	91.38	99.74	99.89	86.47	99.77
	Seq-02	99.56	94.41	99.13	99.73	81.12	99.45
	Seq-03	99.74	96.27	97.49	99.69	99.06	97.42
	Seq-04	99.90	99.02	99.81	99.89	99.68	99.77
	Seq-06	99.87	99.19	99.74	99.61	85.33	99.23
	ave	**99.79**↑	96.06↑	99.18↑	99.76↑	90.33↑	99.13↑
	std	0.13	2.94	0.88	0.11	7.59	0.88
Seq-06	Seq-01	99.76	94.26	99.53	99.70	77.85	99.40
	Seq-02	99.76	96.13	99.53	99.91	99.65	99.81
	Seq-03	99.81	98.61	97.89	99.68	97.61	97.80
	Seq-04	99.91	99.39	99.81	99.84	72.05	99.69
	Seq-05	99.92	92.94	99.84	99.75	77.03	99.50
	ave	**99.83**↑	96.26↑	99.32↑	99.78↑	84.84↑	99.24↑
	std	0.07	2.46	0.73	0.09	11.45	0.73

The best result for each evaluation metric is emphasized with different formats (i.e., bold face for F_1, wavy line for EP, and underline for R@1).

As expected, radar-based methods show small performance variations using the same matching thresholds in long-term scenarios. As shown in Table III, the average $\overline{AwC\text{-}FT}$ values for radar-based Scan Context and radar-based LiDAR-Iris are 4.07% and 5.26%, respectively, which are greatly less than the other LiDAR-based methods. These results indicate that in radar-based methods, influences of matching thresholds on long-term performance are typically smaller compared to those of the LiDAR-based method.

As for influences of geometry variation across different regions on PR performance in large-scale environments, we find that the average Recall@1 values for different regions are very close. For almost $\{\langle k, j\rangle\}_j$, the values of average Recall@1 for different regions are all more than 99% (see Table VII). We guess the primary reason is that radar possesses specific penetration capabilities, enabling it to reliably observe through various obstacles, such as moving vehicles. However, the heat map colors indicating the similarity of matching results in Region-02

TABLE VII
AVERAGE OF RECALL@1 (%) FOR DIFFERENT REGIONS FOR DIFFERENT $\{\langle k, j\rangle\}_j$ USING DIFFERENT RADAR-BASED PR METHODS

Reference	Region-01		Region-02		Region-03	
	SC-radar	Iris-radar	SC-radar	Iris-radar	SC-radar	Iris-radar
Seq-01	**99.77**↑	99.72↑	99.13↑	98.69↑	99.91↑	99.99↑
Seq-02	99.84↑	**99.91**↑	97.56↑	98.28↑	99.97↑	99.97↑
Seq-03	**89.50**↑	89.48↑	99.34↑	99.17↑	99.90↑	99.88↑
Seq-04	99.54↑	**99.57**↑	99.10↑	99.06↑	99.95↑	99.91↑
Seq-05	99.52↑	**99.81**↑	99.38↑	99.00↑	99.87↑	99.97↑
Seq-06	**99.54**↑	99.43↑	98.79↑	98.13↑	99.90↑	99.74↑

tend to be lighter than those in Region-01 and Region-03, as shown in the last two rows of Fig. 6. So, geometry difference, such as building layout and traffic conditions, still makes place recognition more difficult.

The above comparisons show the potential of radar-based methods in long-term scenarios, which is expected to achieve robust performance using general matching thresholds under diverse conditions.

VI. Conclusion and Future Work

In this study, we conduct a comprehensive evaluation of range sensing-based long-term place recognition in large-scale urban environments. We propose a novel metric to evaluate the influences of matching thresholds on long-term performance, which provides a new perspective of evaluation. Our experimental results provide the following important findings: i) current SOTA handcrafted feature-based LiDAR PR methods are more robust to season and weather variations in long-term and large-scale scenarios; ii) with a general matching threshold, current SOTA data-driven LiDAR-based PR methods tend to provide results with larger variations in long-term scenarios; iii) the variation in geometry information across different regions in large-scale environments, such as building layouts and traffic conditions, can lead to performance degradation. Such degradation is much smaller in handcrafted feature-based LiDAR methods; iv) radar-based PR methods show potential in long-term and large-scale scenarios, which achieve superior robustness under diverse weather and traffic conditions. However, only urban environments are evaluated in this work. We believe more long-term datasets with diverse scenarios, such as wild and rural environments, can provide more insights into the community. Meanwhile, it will be interesting to investigate the performance of current SOTA range sensing-based and visual PR methods under more extreme weather conditions (e.g., hails and hurricanes) and longer time span (e.g., 3 years and 5 years).

References

[1] H. Yin et al., "A survey on global LiDAR localization: Challenges, advances and open problems," *Int. J. Comput. Vis.*, pp. 1–33, 2024.

[2] A. Chalvatzaras, I. Pratikakis, and A. A. Amanatiadis, "A survey on map-based localization techniques for autonomous vehicles," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1574–1596, Feb. 2023.

[3] K. Nielsen and G. Hendeby, "Multi-Hypothesis SLAM for non-static environments with reoccurring landmarks," *IEEE Trans. Intell. Veh.*, vol. 8, no. 4, pp. 3191–3203, Apr. 2023.

[4] W. Ma, S. Huang, and Y. Sun, "Triplet-Graph: Global metric localization based on semantic triplet graph for autonomous vehicles," *IEEE Robot. Automat. Lett.*, vol. 9, no. 4, pp. 3155–3162, Apr. 2024.

[5] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, pp. 194–220, Sep. 2017.

[6] L. Chen et al., "HVP-Net: A hybrid voxel- and point-wise network for place recognition," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 395–406, Jan. 2024.

[7] M. Zaffar et al., "VPR-Bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *Int. J. Comput. Vis.*, vol. 129, no. 7, pp. 2136–2174, 2021.

[8] P. Yin et al., "General place recognition survey: Towards the real-world autonomy age," 2022, *arXiv:2209.04497*.

[9] D. Xiao, S. Li, and Z. Xuanyuan, "Semantic loop closure detection for intelligent vehicles using panoramas," *IEEE Trans. Intell. Veh.*, vol. 8, no. 10, pp. 4395–4405, Oct. 2023.

[10] W. Zhao, H. Sun, X. Zhang, and Y. Xiong, "Visual SLAM combining lines and structural regularities: Towards robust localization," *IEEE Trans. Intell. Veh.*, early access, Sep. 04, 2023, doi: 10.1109/TIV.2023.3311511.

[11] C. D. Monaco and S. N. Brennan, "RADARODO: Ego-motion estimation from doppler and spatial data in radar images," *IEEE Trans. Intell. Veh.*, vol. 5, no. 3, pp. 475–484, Sep. 2020.

[12] K. Burnett, Y. Wu, D. J. Yoon, A. P. Schoellig, and T. D. Barfoot, "Are we ready for radar to replace LiDAR in all-weather mapping and localization?," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10328–10335, Oct. 2022.

[13] Z. Hong, Y. Petillot, A. Wallace, and S. Wang, "RadarSLAM: A robust simultaneous localization and mapping system for all weather conditions," *Int. J. Robot. Res.*, vol. 41, no. 5, pp. 519–542, 2022.

[14] A. Venon, Y. Dupuis, P. Vasseur, and P. Merriaux, "Millimeter wave FMCW RADARs for perception, recognition and localization in automotive applications: A survey," *IEEE Trans. Intell. Veh.*, vol. 7, no. 3, pp. 533–555, Sep. 2022.

[15] N. J. Abu-Alrub and N. A. Rawashdeh, "Radar odometry for autonomous ground vehicles: A survey of methods and datasets," *IEEE Trans. Intell. Veh.*, early access, Dec. 07, 2023, doi: 10.1109/TIV.2023.3340513.

[16] S. Lu et al., "Efficient deep-learning 4D automotive radar odometry method," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 879–892, Jan. 2024.

[17] J. Chang, R. Hu, F. Huang, D. Xu, and L.-T. Hsu, "LiDAR-based NDT matching performance evaluation for positioning in adverse weather conditions," *IEEE Sensors J.*, vol. 23, no. 20, pp. 25346–25355, Oct. 2023.

[18] B. Ferrarini, M. Waheed, S. Waheed, S. Ehsan, M. J. Milford, and K. D. McDonald-Maier, "Exploring performance bounds of visual place recognition using extended precision," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1688–1695, Apr. 2020.

[19] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (FPFH) for 3D registration," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2009, pp. 3212–3217.

[20] M. Magnusson, H. Andreasson, A. Nüchter, and A. J. Lilienthal, "Automatic appearance-based loop detection from three-dimensional laser data using the normal distributions transform," *J. Field Robot.*, vol. 26, no. 11/12, pp. 892–914, 2009.

[21] Y. Fan, Y. He, and U.-X. Tan, "Seed: A segmentation-based egocentric 3D point cloud descriptor for loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5158–5163.

[22] Y. Zhu, Y. Ma, L. Chen, C. Liu, M. Ye, and L. Li, "GOSMatch: Graph-of-semantics matching for detecting loop closures in 3D LiDAR data," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5151–5157.

[23] L. He, X. Wang, and H. Zhang, "M2DP: A novel 3D point cloud descriptor and its application in loop closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 231–237.

[24] G. Kim and A. Kim, "Scan Context: Egocentric spatial descriptor for place recognition within 3D point cloud map," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 4802–4809.

[25] G. Kim, S. Choi, and A. Kim, "Scan Context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1856–1874, Jun. 2022.

[26] L. Li et al., "SSC: Semantic scan context for large-scale place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2092–2099.

[27] H. Wang, C. Wang, and L. Xie, "Intensity Scan Context: Coding intensity and geometry relations for loop closure detection," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 2095–2101.

[28] X. Xu, H. Yin, Z. Chen, Y. Li, Y. Wang, and R. Xiong, "DiSCO: Differentiable scan context with orientation," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 2791–2798, Apr. 2021.

[29] X. Xu et al., "RING++: Roto-translation invariant gram for global localization on a sparse scan map," *IEEE Trans. Robot.*, vol. 39, no. 6, pp. 4616–4635, Dec. 2023.

[30] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, "LiDAR Iris for loop-closure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 5769–5775.

[31] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4470–4479.

[32] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5297–5307.

[33] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[34] J. Komorowski, "MinkLoc3D: Point cloud based large-scale place recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1790–1799.

[35] K. Żywanowski, A. Banaszczyk, M. R. Nowicki, and J. Komorowski, "MinkLoc3D-SI: 3D LiDAR place recognition with sparse convolutions, spherical coordinates, and intensity," *IEEE Robot. Automat. Lett.*, vol. 7, no. 2, pp. 1079–1086, Apr. 2021.

[36] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jul. 2018.

[37] M. Gadd, D. D. Martini, and P. Newman, "Look Around You: Sequence-based radar place recognition with learned rotational invariance," in *Proc. IEEE/ION Position Locat. Navig. Symp.*, 2020, pp. 270–276.

[38] M. Gadd, D. D. Martini, and P. Newman, "Contrastive learning for unsupervised radar place recognition," in *Proc. IEEE Int. Conf. Advanc. Robot.*, 2021, pp. 344–349.

[39] Ş. Săftescu, M. Gadd, D. D. Martini, D. Barnes, and P. Newman, "Kidnapped Radar: Topological radar localisation using rotationally-invariant metric learning," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 4358–4364.

[40] F. Alijani, J. Peltomäki, J. Puura, H. Huttunen, J.-K. Kämäräinen, and E. Rahtu, "Long-term visual place recognition," in *Proc. 26th Int. Conf. Pattern Recognit.*, 2022, pp. 3422–3428.

[41] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford robotcar dataset," *Int. J. Robot. Res.*, vol. 36, no. 1, pp. 3–15, 2017.

[42] J. Peltomäki, F. Alijani, J. Puura, H. Huttunen, E. Rahtu, and J.-K. Kämäräinen, "Evaluation of long-term LiDAR place recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 4487–4492.

[43] F. Radenović, G. Tolias, and O. Chum, "CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 3–20.

[44] K. Żywanowski, A. Banaszczyk, and M. R. Nowicki, "Comparison of camera-based and 3D LiDAR-based place recognition across weather conditions," in *Proc. Int. Conf. Control Autom. Robot. Vis.*, 2020, pp. 886–891.

[45] F. Cao, F. Yan, S. Wang, Y. Zhuang, and W. Wang, "Season-invariant and viewpoint-tolerant LiDAR place recognition in GPS-denied environments," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 563–574, Jan. 2021.

[46] F. Cao, H. Wu, and C. Wu, "An end-to-end localizer for long-term topological localization in large-scale changing environments," *IEEE Trans. Ind. Electron.*, vol. 70, no. 5, pp. 5140–5149, May 2023.

[47] K. Burnett et al., "Boreas: A multi-season autonomous driving dataset," *Int. J. Robot. Res.*, vol. 42, no. 1/2, pp. 33–42, 2023.

[48] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[49] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the Oxford robotcar dataset," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 6433–6438.

[50] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 6246–6253.

[51] J. Komorowski, "Improving point cloud based place recognition with ranking-based loss and large batch training," in *Proc. 26th Int. Conf. Pattern Recognit.*, 2022, pp. 3699–3705.

[52] J. Ma, J. Zhang, J. Xu, R. Ai, W. Gu, and X. Chen, "OverlapTransformer: An efficient and yaw-angle-invariant transformer network for LiDAR-based place recognition," *IEEE Robot. Automat. Lett.*, vol. 7, no. 3, pp. 6958–6965, Jul. 2022.

[53] S. H. Cen and P. Newman, "Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 6045–6052.

Weixin Ma (Student Member, IEEE) received the B.S. and M.S. degrees from the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2018 and 2021, respectively. He is currently working toward the Ph.D. degree with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. His current research interests include localization and mapping, robotic perception, and mobile robotics.

Huan Yin (Member, IEEE) received the B.Eng. degree in electronics, and the Ph.D. degree in control engineering from Zhejiang University, Hangzhou, China, in 2016 and 2021, respectively. Dr. Yin was a Visiting Scholar with the University of Technology Sydney, Ultimo, NSW, Australia, and a Research Fellow with the National University of Singapore, Singapore. He is currently a Research Assistant Professor with the Hong Kong University of Science and Technology, Hong Kong. His research interests include robotics, robot perception, SLAM, and autonomous navigation. Dr. Yin was the recipient of two awards named after Zhang Siying and Wu Wenjun. He is an Associate Editor for the International Conference on Robotics and Automation.

Lei Yao (Student Member, IEEE) received the B.S. degree and the M.S. degree from the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2020 and 2023, respectively. He is currently working toward the Ph.D. degree with the Department of Electrical and Eletronic Engineering, The Hong Kong Polytechnic University, Hong Kong. His current research interests includes visual representation learning and reinforcement learning.

Yuxiang Sun (Member, IEEE) received the bachelor's degree from the Hefei University of Technology, Hefei, in 2009, the master's degree from the University of Science and Technology of China, Hefei, China, in 2012, and the Ph.D. degree from The Chinese University of Hong Kong, Shatin, Hong Kong, in 2017. He is currently an Assistant Professor with the Department of Mechanical Engineering, City University of Hong Kong, Hong Kong. His research interests include robotics and AI, autonomous driving, autonomous systems, mobile robots, robotic perception and control, autonomous navigation. He is an Associate Editor for IEEE Transactions on Intelligent Vehicles, IEEE Robotics and Automation Letters, IEEE International Conference on Robotics and Automation, and IEEE/RSJ International Conference on Intelligent Robots and Systems.

Zhongqing Su received the B.Sc. and M.S. degrees from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree from the School of Aerospace, Mechanical and Mechatronic Engineering, The University of Sydney, Sydney, NSW, Australia, in 2004. He is currently a Professor affiliated with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong. He is the author/coauthor of two research monographs, six book chapters, eight edited books/international conference proceedings, and more than 310 archived articles, including 200 articles in top-tier journals. Dr. Su was the recipient of the Structural Health Monitoring–Person of the Year(SHM-POY) in 2012. He is the current Editor in-Chief of *Ultrasonics* journal, and holds the Changjiang Chair Professorship. He is/was an Associate Editor of nine key international journals in his field, including *Journal of Sound and Vibration*, *Structural Health Monitoring: An International Journal, Structural Engineering and Mechanics*, and *Journal of Nondestructive Evaluation* (ASME).