2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

iBoW3D: Place Recognition Based on Incremental and General Bag of Words in 3D Scans

Yuxiaotong Lin, Jiming Chen, and Liang Li

***Abstract*— Existing methods for place recognition in 3D point clouds either ignore partial structure information by converting 3D scans to 2D images or construct constrained bag-of-words (BoW) representations reliant on specific feature extraction algorithms. In this paper, we propose a novel method based on incremental and general bag of words. Incorporating an adaptable keypoint and 3D local feature extraction method, we employ an incremental BoW model that is updated regularly. This enables a coarse-to-fine candidate selection from the database. And a revisit can be identified following geometric verification. In addition, we propose a new supplementary metric that addresses the leaving-out issue of the conventional metric, enhancing the identification of true loops. Employing a state-of-the-art (SOTA) keypoint and feature extraction algorithm, we evaluate our method as well as SOTA place recognition methods using diverse datasets with varying qualities. Experimental results demonstrate that our method outperforms the baselines across all three datasets, showcasing robust performance and notable generalization capabilities.**

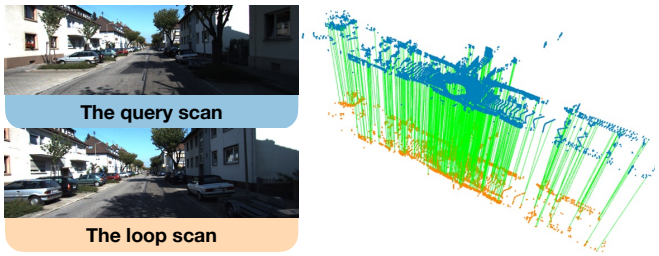
I. Introduction

Place recognition, *i.e.*, the ability to detect a revisited place, is an important issue in many robotic tasks. It plays a key role in robot relocalization and loop detection of Simultaneous Localization and Mapping (SLAM). As odometry which only estimates relative pose transformation has accumulated drifts inevitably, loop detection and back-end optimization are indispensable in modern SLAM systems.

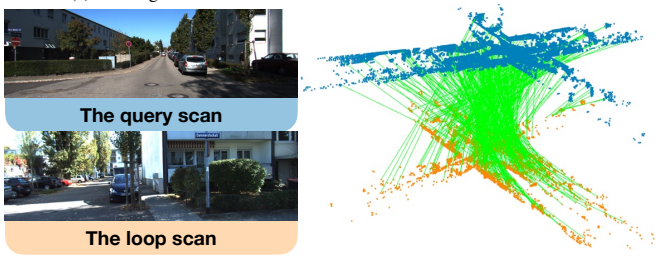
Bag of words (BoW) is commonly used for loop closure detection in visual SLAM as it can retrieve images efficiently [1]–[5]. However, visual recognition based on images is vulnerable to illumination, short-term (*e.g.*, moving objects), and long-term (*e.g.*, seasons) changes. To solve these problems, 3D scans captured by LiDARs have drawn attention because of their robustness to the above changes. Some works focus on methods of extracting global point cloud features to accomplish place recognition in 3D scans [6]–[8]. Recently, BoW has also utilized in some works in 3D scans [9], [10] for its high retrieval efficiency. However, the existing methods either transform 3D scans to 2D images and then apply offline BoW on images [9], or rely on a specific feature extraction method to construct a BoW dictionary [10]. The former does not totally exploit the information of 3D structure while the latter, even though based on a BoW framework, is constrained by a designated feature extraction method, thus lacks the generalization.

In this paper, we propose a novel place recognition method which employs an **i**ncremental and general **bag-of-words** in **3D** scans, called **iBoW3D**. Two examples of recognition are shown in Fig. 1. This framework (shown in Fig. 2) is constructed using local features extracted from 3D scans and is not limited by specific feature extraction algorithms. For each input scan, keypoint and feature extraction are applied first. Then a coarse-to-fine candidate selection is employed by leveraging BoW. The fittest candidate that passes the geometric verification would be identified as a loop. As the number of processed scans increases, both the feature database and the BoW dictionary undergo continuous updates. As a result, our BoW framework works in an online and incremental manner, avoiding the pre-prepared work before recognition of conventional offline and pre-trained approaches. Additionally, we find that the conventional metric based on a fixed threshold is so strict that some loops would be left out when identifying true loops. To address this problem, we propose a new metric in addition to the conventional one. To give a comprehensive evaluation, we conduct extensive experiments to compare our method with other SOTA methods on three different datasets. The results show that our method can achieve better performance and performs well across datasets with various data qualities, which verifies the generalization ability of our proposed method. In summary, our main contributions are:

- We propose a novel place recognition method based on an incremental bag of words and local point cloud



(a) A recognized result of KITTI 00 which is in the same direction



(b) A recognized result of KITTI 08 which is in a reverse direction

Fig. 1. Two place recognition results in the same direction from KITTI 00 and in a reverse direction from KITTI 08 using our method.

This work is supported by the National Natural Science Foundation of China (62088101/62203383). The authors are with the College of Control Science and Engineering, Zhejiang University, Hangzhou, 310027, China.

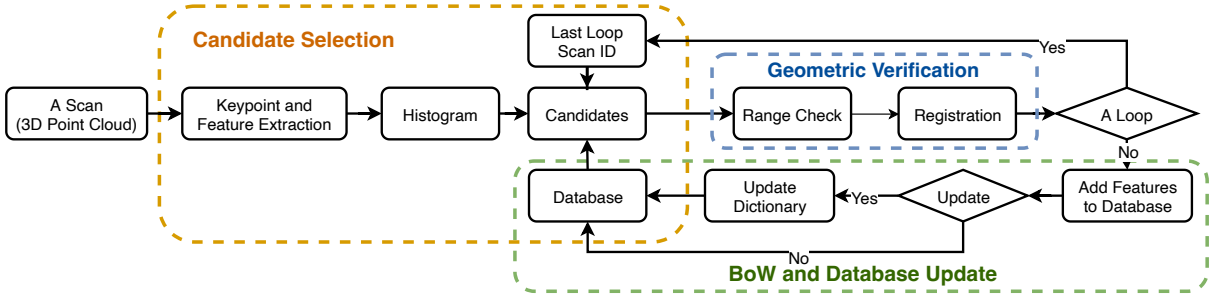


Fig. 2. The method overview. Our method mainly consists of three parts: (a) Candidate Selection, (b) Geometric Verification, and (c) BoW and Database Update

features. The method achieves better performance on datasets with different data qualities. The code is available at https://github.com/NeSC-IV/iBoW3D.

- Our method is not limited by specific methods in terms of keypoint and local feature extraction, showing the generalization of our proposed method. As a result, a better extraction algorithm can replace the current one without affecting other parts of our method.
- In addition to the conventional metric, we propose a new metric to identify true loops. This new metric can avoid leaving out true loops and aligns with the purpose of detecting revisits for better back-end optimization.

II. Related Work

Visual place recognition has been extensively studied in recent decades, with the BoW approach being the most commonly employed method for this purpose. Different feature extraction methods have been utilized within this framework, such as the SIFT-based method [1], the SURF-based FAB-MAP [2], and the BRIEF-based algorithm [3]. Nonetheless, the vulnerability to illumination, short-term, and long-term changes in images remains a challenge for visual place recognition.

To overcome these condition changes, the utilization of 3D LiDAR in the field of SLAM has drawn much attention recently. Within this context, a number of researchers have concentrated their efforts on crafting comprehensive and distinguishing global features tailored for effective place recognition. On the other hand, given the paper's focus on place recognition in 3D point clouds, we are limiting our review to this particular aspect. Muhammad et al. [6] propose a global feature based on histograms of local surface normals. Key histograms are selected and matched with histograms of the input scan to detect loops. M2DP [7] projects the 3D point cloud to various 2D planes and derives a signature for every plane. The derived signatures are then integrated into a signature matrix, which is used to match point clouds. In contrast to the aforementioned methods, Scan Context [8] directly encodes the structure of a 3D scan into a matrix as its global feature. And more location and structure information are preserved in this way. LiDAR Iris [11] converts 3D scans into LiDAR-Iris images, preserving the internal structure of the point clouds. Focusing on structure information as well, STD [12] extracts keypoints and encodes them into triangular descriptors as the global features. Scan Context, LiDAR Iris, and STD all rely on global feature matching to accomplish place recognition. Furthermore, certain studies incorporate semantic information during the generation of global features, *e.g.*, SGPR [13] and SSC [14].

Despite the design of search algorithms, detecting loops through the matching of global features remains inefficient. Thus, some researchers turn to the utilization of bag of words in 3D scans. Steder et al. [9] propose a method based on an offline bag of words in 3D scans. However, their method converts 3D scans to 2D range images but does not directly exploit 3D information. BoW3D [10], which is similar to our method, directly extracts local features from 3D scans and constructs bag of words on the features. Nevertheless, its bag-of-words construction relies on a specific feature extraction method, which is not general.

In contrast to other approaches, our bag-of-words framework stands out for its regular updates, enabling it both online and incremental. Notably, our method has a higher degree of generality, remaining effective by the replacement of the feature extraction algorithm. As a result, any method capable of extracting more discriminative keypoints and local features can be integrated into our method easily.

III. Methodology

In this section, we present our approach for place recognition by the integration of BoW and locally extracted features from point clouds. As illustrated in Fig. 2, our method consists of three key components: candidate selection, geometric verification, and database update.

A. System Overview

Given a scan database and a new input scan, our method aims to identify if there is a loop and where the loop occurs. It is assumed that there are no loops for the first N_d scans, in order to use them to establish the initial database and BoW dictionary. After receiving a new input scan (3D point cloud), we first extract the keypoints and local features that represent the scan's peculiarity. In order to efficiently select candidates where a loop is likely to happen, we apply BoW and utilize the information of the last occurred loop. Then a geometric verification is applied for all candidates and the fittest one that passes the verification is identified as

the loop scan. The last occurred loop information will also be updated. If no candidates pass the verification, no loop occurs and the current local features will be added to the database. If the number of newly added scan features reaches N_u, the BoW dictionary (a set of words) will be updated to keep discrimination, achieving an online and incremental dictionary. The above process would be applied to every new input scan. In the following section, we will elaborate on the method in detail.

B. Candidate Selection

One of the primary objectives of this paper is to formulate a versatile place recognition algorithm for 3D point clouds, accommodating any keypoint detection and descriptor construction methods. In this paper, we leverage D3Feat [15], a SOTA deep learning-based neural network, for both keypoint detection and description. This serves as the cornerstone of our feature extraction process. Furthermore, we conduct assessments using alternative non-learning-based methods, and an in-depth comparison and analysis can be found in Section IV-B. When a new scan comes (referred to as the query scan), after voxel downsampling, the keypoints and features are then extracted through the D3Feat framework to characterize the query scan. Subsequently, a coarse-to-fine candidate selection procedure based on the characterization is employed to identify scans that potentially correspond to loops. Given N_k local features corresponding to the keypoints (referred to as key local features), the word distribution of the query scan can be obtained by classifying each feature into a word of the N_w-word BoW dictionary based on the Euclidean distance:

$$\mathbf{d}(f_m, w_i) = \|(f_m - w_i)\|_2 \tag{1}$$

where f_m represents the m-th key local features and w_i represents the i-th word. Then the word closest to the feature f_m can be identified as the classification result:

$$\mathbf{i}_\mathbf{m}^* = \underset{i \in \{1,2,\cdots,N_w\}}{\arg\min} \mathbf{d}(f_m, w_i) \tag{2}$$

Hence, the query scan can be described by a N_w-bin histogram, in which the number of each bin represents the number of features classified as the word.

Since not all words have the same discrimination, we employ the Term Frequency-Inverse Document Frequency (TF-IDF) on the histogram. TF-IDF is defined as:

$$\eta_i^j = TF_i^j \times IDF_i = \frac{n_i^j}{n^j} \times \log \frac{n}{n_i} \tag{3}$$

where n^j and n represent the total number of features of the j-th scan and the dictionary, respectively, while n_i^j and n_i represent the appearance times of word w_i in the j-th scan and the dictionary, respectively. A higher TF indicates a word appears more frequently in a scan, whereas a higher IDF means a word appears less frequently in the dictionary. Both of these show the discrimination of the word. As a result, a new N_w-bin histogram with a TF-IDF value η_i^q in the i-th bin can be obtained as the description of the query scan. For better comparison, we normalize the N_w-bin histogram, denoted as H_q.

After preprocessing the query scan, a coarse-to-fine selection is adopted for quick retrieval. However, we exclude the most recent N_g scans from consideration during this process, as they are closely temporally aligned and are therefore less likely to form a meaningful loop. Given the histogram H_q and the histogram database D_h, the coarse selection is implemented based on the fraction of word co-occurrence. The scans with fractions that surpass a given threshold λ_w are retained while others are discarded. All scans passed the coarse selection constitute the selected scan set $\mathbf{S_{qc}}$.

Subsequent to the coarse selection, a fine selection is further implemented to identify candidates by computing the histogram similarity through the Euclidean distance. Inspired by [3], in light of temporal coherence, scans that are aligned temporally closely are clustered together. Prior to clustering, the scan set $\mathbf{S_{qc}}$ is reorganized into a new set $\mathbf{S_{qn}}$ in descending order of similarity, facilitating subsequent processing. Among all the clusters, clusters with the highest number of scans and the largest average similarity are selected. Additionally, if loop scans have been identified, the cluster containing the most recent loop scan is also chosen, which can help recognize new loops as a revisit often lasts for a time, manifesting as sequential loops. Within each selected cluster, the N_s scans having the highest similarity are finally chosen as the outcomes of the fine selection process, thereby forming the candidate scan set. The fine selection algorithm is summarized in Algorithm 1.

Algorithm 1: The Fine Selection

Input: The selected scan set $\mathbf{S_{qc}}$, the query scan s_q, the threshold for cluster N_{near}, and the number of selected scan in each cluster N_s.
Output: The candidate scans $\mathbf{S_{cand}}$.

```
for s_i in S_qc do
    d_i = Distance(s_q, s_i);
    Add d_i into DisSet;
S_qn = Rearrange(S_qc, DisSet);
ClusterSet = Cluster(S_qn, N_near);
if Last loop exists then
    CL_Most, CL_Largest, CL_Loop =
      ClusterChoose(ClusterSet);
    S_Most, S_Largest, S_Loop = ScanChoose(CL_Most,
      CL_Largest, CL_Loop, N_s);
    S_cand = Combine(S_Most, S_Largest, S_Loop);
else
    CL_Most, CL_Largest =
      ClusterChoose(ClusterSet);
    S_Most, S_Largest = ScanChoose(CL_Most,
      CL_Largest, N_s);
    S_cand = Combine(S_Most, S_Largest);
```

Result: $\mathbf{S_{cand}}$

C. Geometric Verification

After the candidate selection, a geometric verification is conducted between each candidate scan and the query scan to ensure geometric consistency. In some cases, scans captured in narrow scenarios may confound the registration (shown in Fig. 3), even when a substantial difference exists between the source and target scans. To mitigate this, a range check is employed prior to registration. Assuming that scans captured in a specific location have similar point distributions, we compare the range difference between the two scans after removing the outliers to mitigate the impact of noise. Let x^c_{max}, x^c_{min}, y^c_{max}, y^c_{min}, x^q_{max}, x^q_{min}, y^q_{max}, and y^q_{min} denote the maximum and minimum values in x and y directions for both the candidate and the query scan in their respective coordinate systems. The range check passes if:

$$\lambda_r^1 < \frac{x^c_{max}}{x^q_{max}}, \frac{x^c_{min}}{x^q_{min}}, \frac{y^c_{max}}{y^q_{max}}, \frac{y^c_{min}}{y^q_{min}} < \lambda_r^2 \tag{4}$$

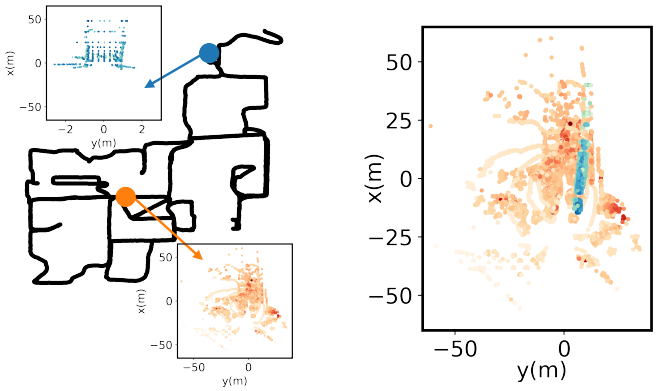
where λ_r^1 and λ_r^2 are two thresholds to identify the degree of closeness.

For candidates that pass the range check, a registration procedure is applied. Given a correspondence set of points identified based on the local features, we utilize RANSAC [16] for point cloud registration. Two metrics, *e.g.*, the fitness and the RMSE (Root Mean Squared Error) of all inlier correspondences, can be also obtained. Representing the overlapping area, a higher fitness value indicates better alignment. In contrast, a lower RMSE value represents more accurate registration. Therefore, a candidate scan successfully passes geometric verification if its fitness surpasses the threshold λ_f and its RMSE is below the threshold λ_r. From the pool of scans that have passed verification, the one with the highest fitness or the lowest RMSE is selected as the loop scan. Additionally, the ID of the most recent loop scan is updated.

D. BoW and Database Update

The first N_d scans are utilized to initiate the databases and the BoW dictionary. For each scan, local features and keypoints are extracted, which help to construct two distinct databases. The first one is the all-feature database D_f formed by aggregating all local features extracted from the initial N_d scans, which is used for registration. The other one, referred to as the key feature database D_{kf}, specifically includes key local features for candidate selection. A k-medians clustering with the k-means++ seeding [17] is then applied to D_{kf}, yielding N_w clusters. Each cluster corresponds to a word. Thus, the BoW dictionary is obtained and the IDF can also be calculated.

If there is no loop for the query scan, its features are incorporated into the two databases. However, with the increase of databases, the discrimination of the BoW dictionary decreases. Therefore, the dictionary needs to be updated periodically, and the number of words should also be augmented to fit the increase of the databases. To accomplish the update, a repetition of the k-medians clustering process and the calculation of IDF is necessary. This highlights that our BoW approach is not only online and incremental, but it also avoids the need for pre-training. The update process is shown in Algorithm 2.



(a) scans and locations (b) registration result

Fig. 3. An example of narrow condition from NCLT 20120526. (a) shows the source (blue) and target (orange) with their locations. The source is scanned in a narrow scenario. (b) shows the result after transforming the source by the registration result. Although no loop exists, confounded by the narrow scan, the fitness and RMSE results after registration are 0.999 and 1.297, which can pass the verification.

Algorithm 2: BoW and Database Update

Input: All feature database D_f, key feature database D_{kf}, and feature addition time counter cnt.
Output: The cluster result $Dict$ and the list of IDF values $IDFList$

```
if no loop exists then
    Add features to D_f and D_kf;
    cnt++;
    if cnt == N_u then
        cnt = 0;
        N_w += Δ_w;
        Dict = K-medians(D_kf, N_w);
        IDFList = IDFCalculate(Dict);
```

Result: $Dict$, $IDFList$

IV. Experiments

In this section, we first introduce a supplementary metric of identifying true loops, and then we provide a comprehensive result and analysis of experiments conducted on three diverse datasets, each presenting varying levels of detection challenges. We also conduct a comparative analysis against SOTA methods and present a thorough performance evaluation.

A. Metrics and Experiment Settings

1) Metrics: For existing methods, when the ground truth distance $\mathbf{d}_i^{gt}$ between the query scan and the i-th candidate scan falls below a predetermined threshold $\mathbf{d}_0$, the association is classified as a true loop. If a candidate is identified successfully by the place recognition algorithm, it is regarded as a true positive detection result. Nevertheless, relying solely

on a fixed threshold can potentially omit certain valid loops. This happens due to the fact that if the distance between two scans slightly exceeds the threshold, they are not recognized as a true loop, even though they are in close proximity to each other. Therefore, in addition to the aforementioned conventional metric, we propose a supplementary metric for true loop identification. Based on the distance $\mathbf{d}_i^{reg}$, the second metric is defined as Eq. 6:

$$\mathbf{T}_i^{reg} = \begin{bmatrix} \mathbf{R}_i^{reg} & \mathbf{t}_i^{reg} \\ \mathbf{0} & \mathbf{1} \end{bmatrix}, \quad \mathbf{t}_i^{reg} = \left(\mathbf{t}_{i_x}^{reg}, \mathbf{t}_{i_y}^{reg}, \mathbf{t}_{i_z}^{reg}\right)^\top \tag{5}$$

$$\mathbf{LC}_t = \{\mathbf{LC}_i | \mathbf{d}_i^{reg} = \|(\mathbf{t}_{i_x}^{reg}, \mathbf{t}_{i_y}^{reg})\|_2 < \mathbf{d}_1\} \tag{6}$$

where $\mathbf{T}_i^{reg} \in SE(3)$ is the transformation matrix from registration by aligning the query scan to the i-th candidate scan, and $\mathbf{d}_1$ is a predetermined threshold.

Given that the primary objective of place recognition is revisitation detection independently of odometry, this secondary metric holds practical significance for optimizing back-end processes. By determining the relative pose solely through their features, this metric contributes to effective back-end optimization.

2) Datasets and Experiment Settings: To encompass a variety of 3D LiDAR sensor types, loop occurrence directions, and environmental complexities, we employ the KITTI dataset [18], the NCLT dataset [19], and the Complex Urban (CU) dataset [20] for the comprehensive evaluation of our method. In our experiments, we set $N_d = 400$, $N_u = 200$, and $N_g = 250$. As for parameters related to BoW, we set $N_k = 20$, the initial $N_w = 50$, and $\Delta_w = 10$. These values are not set too large because we find that using more local features and a dictionary with more words may decrease the discrimination, which is also stated in [9]. Additionally, we set $N_{near} = 8$, $N_s = 5$, and $\lambda_w = 0.2$ based on experimental results. λ_r^1 and λ_r^2 are set as different values for different datasets. For KITTI, $\lambda_r^1 = 0$, $\lambda_r^2 = \infty$. For NCLT, $\lambda_r^1 = \frac{1}{3}$, $\lambda_r^2 = 3$. For Complex Urban, $\lambda_r^1 = \frac{1}{4}$, $\lambda_r^2 = 4$. Two thresholds for identifying true loop $\mathbf{d}_0$ and $\mathbf{d}_1$ are both set as $4m$.

B. Results and Performance Evaluation

For keypoint and feature extraction, we conduct experiments on KITTI 00 to compare various methods, including FPFH [21], FCGF [22], and D3Feat [15]. D3Feat can extract keypoints and local features simultaneously, while the other methods can only extract local features. Therefore, when employing FPFH and FCGF, we first downsample the scan using Normal Space Sampling (NSS) [23] and then extract keypoints via Intrinsic Shape Signatures (ISS) [24]. We adopt the pre-trained models of KITTI for both FCGF and D3Feat, with a voxel downsampling of $30cm$. The results in Fig. 4 show that D3Feat performs better. Moreover, D3Feat can extract keypoints and give the corresponding local features, which is convenient in our scenario. As a result, D3Feat is employed for the extraction of local features and keypoints in subsequent experiments across all three datasets, utilizing the same pre-trained model with a $30cm$ voxel downsampling, consistent across the paper.

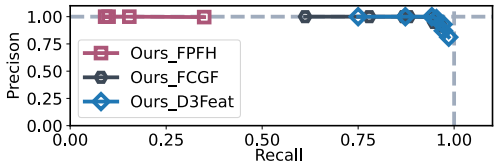
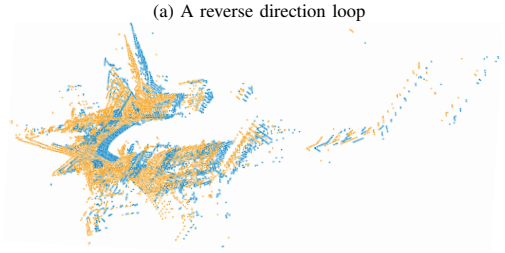


Fig. 4. Precision-Recall curves on KITTI 00 using different keypoint and local feature extraction methods.

(a) A reverse direction loop



(b) A same direction loop

Fig. 5. Two examples of different direction revisits. Each example shows two scans of one place.

To analyze the performance of our method, we conduct a series of experiments, plot the Precision-Recall (PR) curves for each sequence, and give the max F1-score results. The F1-score is defined as:

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

where P and R are precision and recall respectively. As a harmonic average, the F1-score shows a comprehensive consideration of precision and recall.

Our method is implemented in two distinct modes: one prioritizes the candidate scan with the highest fitness as the loop scan, while the other favors the selection of the candidate scan with the lowest RMSE. Furthermore, we conduct a comparative analysis of our method against two baseline methods, *i.e.*, Scan Context (SC) [8] and BoW3D [10]. The results are shown in Fig. 6 and Table I.

As can be seen from Fig. 6 and Table I, our method can achieve better performance across all three datasets varying in quality. As the supplementary metric is applied on all three methods, our method contributes to the improvement. Our approach achieves higher recall values without significantly compromising precision. Notably, the recall even attains a perfect score of 1.000 on KITTI 06, 07, and 08. Comparing the two loop-selecting approaches within our method, the PR curves do not differ a lot, with the exception of KITTI 02. The result of KITTI 02 can be attributed to the fact that, within certain clusters of candidate scans, fitness values tend to be close, thereby reducing the discriminative power of fitness as a selection criterion. Consequently, selecting loops based on fitness in this scenario leads to an increase in

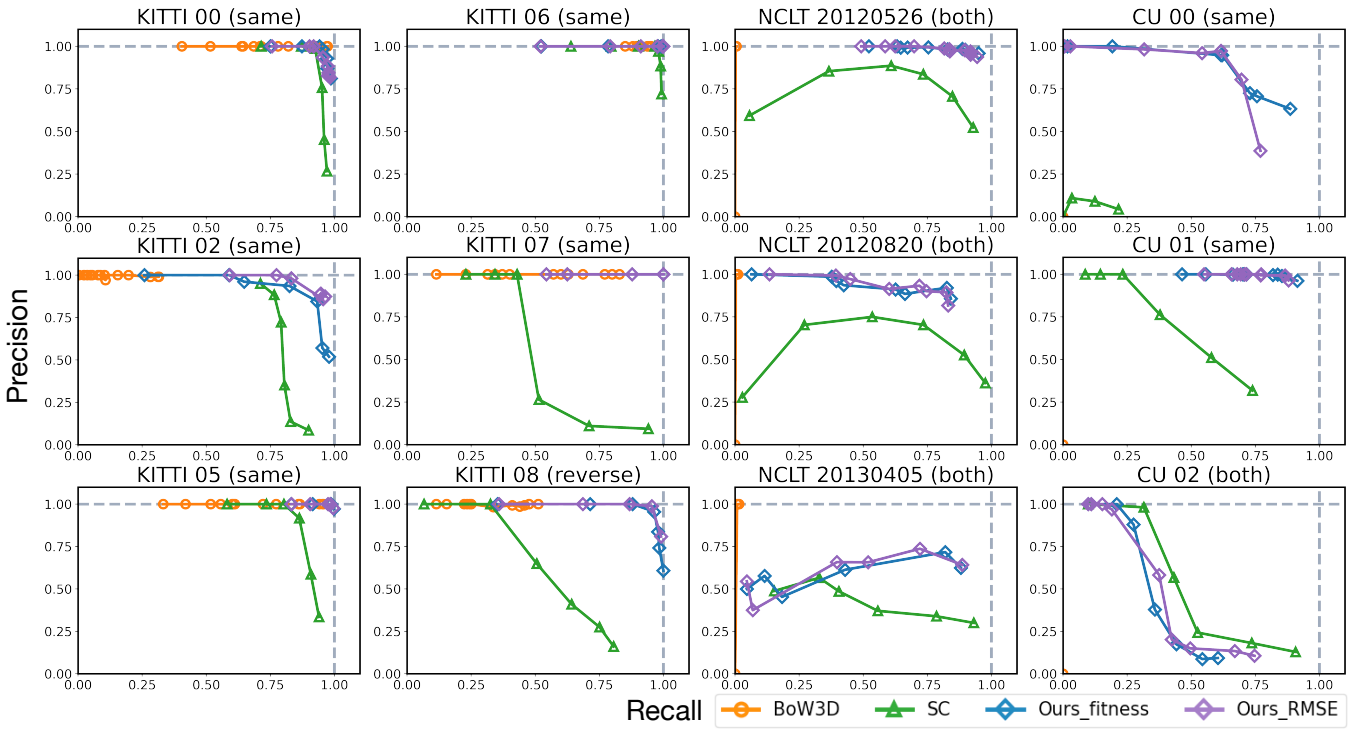


Fig. 6. Precision-Recall curves on three datasets using different methods.

TABLE I Max F1-score results on three datasets. The best results are marked in bold and the second best ones are underlined.

Method	KITTI						NCLT			Complex Urban		
	00	02	05	06	07	08	20120526	20120820	20130405	00	01	02
SC [8]	0.956	0.820	0.891	0.983	0.600	0.569	0.782	0.719	0.474	0.104	0.543	**0.491**
BoW3D [10]	**0.987**	0.478	0.978	0.975	0.906	0.683	0.010	0.022	0.030	0.000	0.000	0.000
Ours_fitness	0.970	0.887	**0.993**	**1.000**	**1.000**	0.960	**0.954**	**0.871**	**0.765**	**0.777**	**0.931**	0.419
Ours_RMSE	0.958	**0.918**	0.992	**1.000**	**1.000**	**0.971**	0.943	0.857	0.745	0.754	0.925	0.456

false positives. Conversely, the RMSE values exhibit greater variation, enabling superior performance when employed for loop selection due to their higher discriminative capacity.

Compared with BoW3D [10], our method demonstrates superior generalization capabilities. While BoW3D performs well on the KITTI dataset, it fails to perform effectively on the NCLT and Complex Urban datasets. Furthermore, our method consistently outperforms Scan Context in terms of precision and recall, except for the case of Complex Urban 02 where our method's performance is constrained by the nature of the scans themselves. The LiDARs utilized in the Complex Urban dataset are mounted obliquely, capturing only partial scenes. In Complex Urban 02, certain locations are revisited in reverse directions. Consequently, for these reverse revisits, the overlap between new and old scans can be minimal, posing a challenge for correspondence-based registration. Fig. 5 illustrates two examples: one showcasing a reverse-direction revisit and the other featuring a revisit in the same direction. Notably, in the case of the same-direction revisit, the scans exhibit substantial overlap, facilitating registration. However, in scenarios involving reverse-direction revisits, scan overlap is minimal, even when considering the same plane as indicated by the red circle in Fig. 5 (a), thereby resulting in intricate registration challenges.

V. Conclusions

In this paper, we present a novel approach to place recognition in 3D scans, utilizing an incremental and general BoW framework. Notably, our BoW model remains unpre-trained and is adaptable to diverse keypoint and local feature extraction algorithms. Our method is structured into three core components: candidate selection, geometric verification, and BoW and database update. The first component employs a coarse-to-fine selection strategy to enhance efficiency. Furthermore, we propose a novel supplementary metric alongside the conventional one to identify true loops. This strategic addition prevents the oversight of true loops, which may occur due to the limitations of traditional fixed thresholds. Experimental results compared with state-of-the-art place recognition methods highlight the superior performance and heightened generalization capability of our approach. In future work, we plan to focus on refining accuracy and computational efficiency and integrate our method into a 3D LiDAR SLAM system.

REFERENCES

[1] A. Angeli, D. Filliat, S. Doncieux, and J.-A. Meyer, "Fast and incremental method for loop-closure detection using bags of visual words," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1027–1037, 2008.

[2] M. Cummins and P. Newman, "Appearance-only slam at large scale with fab-map 2.0," *The International Journal of Robotics Research*, vol. 30, no. 9, pp. 1100–1123, 2011.

[3] D. Gálvez-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.

[4] R. Mur-Artal and J. D. Tardós, "Fast relocalisation and loop closing in keyframe-based slam," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 846–853, IEEE, 2014.

[5] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[6] N. Muhammad and S. Lacroix, "Loop closure detection using small-sized signatures from 3d lidar data," in *2011 IEEE International Symposium on Safety, Security, and Rescue Robotics*, pp. 333–338, IEEE, 2011.

[7] L. He, X. Wang, and H. Zhang, "M2dp: A novel 3d point cloud descriptor and its application in loop closure detection," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 231–237, IEEE, 2016.

[8] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4802–4809, IEEE, 2018.

[9] B. Steder, M. Ruhnke, S. Grzonka, and W. Burgard, "Place recognition in 3d scans using a combination of bag of words and point feature based relative pose estimation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1249–1255, IEEE, 2011.

[10] Y. Cui, X. Chen, Y. Zhang, J. Dong, Q. Wu, and F. Zhu, "Bow3d: Bag of words for real-time loop closing in 3d lidar slam," *IEEE Robotics and Automation Letters*, 2022.

[11] Y. Wang, Z. Sun, C.-Z. Xu, S. E. Sarma, J. Yang, and H. Kong, "Lidar iris for loop-closure detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5769–5775, IEEE, 2020.

[12] C. Yuan, J. Lin, Z. Zou, X. Hong, and F. Zhang, "Std: Stable triangle descriptor for 3d place recognition," *arXiv preprint arXiv:2209.12435*, 2022.

[13] X. Kong, X. Yang, G. Zhai, X. Zhao, X. Zeng, M. Wang, Y. Liu, W. Li, and F. Wen, "Semantic graph based place recognition for 3d point clouds," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8216–8223, IEEE, 2020.

[14] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, "Ssc: Semantic scan context for large-scale place recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2092–2099, IEEE, 2021.

[15] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, "D3feat: Joint learning of dense detection and description of 3d local features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6359–6367, 2020.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[17] D. Arthur and S. Vassilvitskii, "K-means++ the advantages of careful seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[18] 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*, pp. 3354–3361, IEEE, 2012.

[19] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of michigan north campus long-term vision and lidar dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[20] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *International Journal of Robotics Research*, vol. 38, no. 6, pp. 642–657, 2019.

[21] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE International Conference on Robotics and Automation*, pp. 3212–3217, IEEE, 2009.

[22] C. Choy, J. Park, and V. Koltun, "Fully convolutional geometric features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8958–8966, 2019.

[23] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pp. 145–152, IEEE, 2001.

[24] Y. Zhong, "Intrinsic shape signatures: A shape descriptor for 3d object recognition," in *IEEE International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 689–696, 2009.