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

2024 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-8457-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICRA57147.2024.10611569

A Coarse-to-Fine Place Recognition Approach using Attention-guided Descriptors and Overlap Estimation

Chencan Fu, Lin Li, Jianbiao Mei, Yukai Ma, Linpeng Peng, Xiangrui Zhao, and Yong Liu*

Abstract— **Place recognition is a challenging but crucial task in robotics. Current description-based methods may be limited by representation capabilities, while pairwise similarity-based methods require exhaustive searches, which is time-consuming. In this paper, we present a novel coarse-to-fine approach to address these problems, which combines BEV (Bird's Eye View) feature extraction, coarse-grained matching and fine-grained verification. In the coarse stage, our approach utilizes an attention-guided network to generate attention-guided descriptors. We then employ a fast affinity-based candidate selection process to identify the Top-*K* most similar candidates. In the fine stage, we estimate pairwise overlap among the narrowed-down place candidates to determine the final match. Experimental results on the KITTI and KITTI-360 datasets demonstrate that our approach outperforms state-of-the-art methods. The code will be released publicly soon.**

I. INTRODUCTION

Place recognition is a crucial task in mobile robots and autonomous driving, as it enables the recognition of previously visited places and provides a basis for loop closure detection. It plays a vital role in Simultaneous Localization and Mapping (SLAM) by identifying loop closures to correct drift and tracking errors. The LiDAR sensor has gained popularity for place recognition due to its robustness to illumination and weather changes and its wide field of view.

Recent LiDAR-based place recognition methods focus on describing the LiDAR point cloud discriminatively using various representation forms, including 3D point clouds [1]–[4], segments [5], [6], range image views [7]–[11], and bird's eye views [12]–[19]. These methods employ hand-crafted or learning-based techniques to generate efficient local or global descriptors. However, compressing the LiDAR data inevitably limits the representation ability of descriptors. Besides, totally different places may have similar descriptions, which can result in failures for loop closure detection. On the other hand, pairwise similarity-based place recognition methods have gained attention [20]–[22]. These methods compare pairs of point clouds using a network to predict their similarity score. Pairwise comparison fully utilizes the information between two point clouds, improving loop closure detection performance. However, pairwise similarity-based methods require exhaustive searches to detect loops, making them time-consuming.

To address these challenges, we propose a novel approach that efficiently combines the advantages of descriptors and pairwise similarity-based methods for coarse-to-fine place recognition. Our approach comprises three parts: BEV (Bird's Eye View) feature extraction, coarse-grained matching, and fine-grained verification. We first voxelize the input point cloud to create a multi-layer BEV representation. Next, we extract BEV features in 2D space and fully utilize them in our coarse-to-fine approach. In the coarse stage, an attention-guided network generates 1D global descriptor vectors from encoded BEV features. We then perform affinity-based candidate selection to obtain the Top-*K* place candidates. Finally, in the fine stage, the candidates are further validated by estimating pairwise overlap using an overlap estimation network. The candidate with the highest overlap score is considered the final match. This combination of efficiency and accuracy highlights the advantages of our proposed method. Our approach effectively reduces the exhaustive search required for pairwise overlap estimation, improving the overall efficiency. Consequently, our method demonstrates superior performance in determining the matching place compared to existing methods. Fig.1 provides an illustration of our proposed approach.

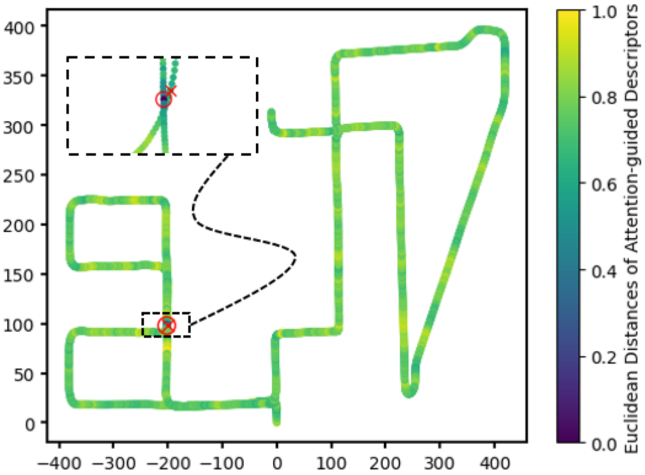


Fig. 1: The visualization of results. The trajectory represents sequence 08 of the KITTI dataset [23], and the color indicates the Euclidean distance of the descriptors to the query. The circle mark represents a query place, and the cross mark represents the matching place found by our approach.

The main contributions of this paper are as follows:

- An attention-guided global descriptor for effective place recognition.
- A coarse-to-fine place recognition approach that effectively uses BEV features to enhance efficiency while maintaining a high level of place recognition ability.
- Comprehensive experiments and detailed ablation studies on the KITTI [23] and KITTI-360 [24] datasets to validate the effectiveness of the proposed approach.

Chencan Fu, Lin Li, Jianbiao Mei, Yukai Ma, Linpeng Peng, Xiangrui Zhao, and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, P. R. China. (*Yong Liu is the corresponding author, email: yongliu@iipc.zju.edu.cn).

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

II. RELATED WORK

Place recognition has been an active area of research for many years, especially in visual place recognition (VPR), where many works have been proposed [25]–[29]. In this paper, we focus on LiDAR-based place recognition (LPR).

Description-based Methods. Description-based methods aim to provide discriminative descriptions of LiDAR point clouds for place recognition. Traditional methods often use hand-crafted features [12]–[15], [30], while more recent approaches leverage the neural networks to achieve powerful feature representation. Some description-based methods take raw point clouds or voxelized and segmented forms as input to the network model [1]–[6]. For example, PointNetVLAD [1] combines PointNet [31] and NetVLAD [32] to generate global descriptors in an end-to-end manner. Locus [6] leverages topological and temporal information and aggregates multi-level features via second-order pooling (O2P). LoGG3D-Net [2] uses O2P followed by Eigen-value Power Normalization to obtain a global descriptor. To improve efficiency, some methods use projections of LiDAR data. M2DP [30], for example, projects a 3D point cloud onto multiple 2D planes to form a descriptor. Range images are also widely used [7]–[11], [20], which are consistent with the characteristics of LiDAR sensors. Thanks to the property of range images, it is easy to obtain orientation invariance theoretically. However, small translations may affect the range image due to perspective transformation. Some methods use bird's eye view forms of data as input [12]–[19], as Cartesian coordinate representation is more consistent with reality. Some methods use a single-layer BEV, such as Scan Context [12], Intensity Scan Context [14]. Xu et al. [16] summarize several BEV representations, including multi-layer occupied BEV, multi-layer density BEV, and single-layer height BEV. The multi-layer BEV can reduce the loss of height information. Furthermore, CVTNet [33] fuses range image views and bird's eye views of LiDAR data and achieves good performance.

Pairwise Similarity-based Methods. Pairwise similarity-based methods focus on comparing two point clouds and scoring their similarity using specialized network structures [5], [20]–[22]. OverlapNet [20], for example, exploits multiple cues, including depth, normal, intensity, and semantic class probability information to predict the overlap score and treats overlap estimation as a regression problem. RINet [21] also takes a similar approach to predict similarity. It combines semantic and geometric features to predict the similarity of descriptor pairs. These methods perform pairwise similarity comparison implicitly. In contrast, our previous work [22] proposed an intuitive approach. It regards overlap prediction as a classification problem of each bin in the bird's eye view and achieves state-of-the-art performance in loop closure detection. However, with the promotion of place recognition ability, the time cost increases significantly to perform exhaustive pairwise overlap estimation.

Cop et al. [34] design an intensity-based local descriptor DELIGHT and propose a two-stage approach to perform place recognition. In the first stage, they search for the N most similar DELIGHT descriptors and subsequently perform geometrical recognition on the identified place candidates. Inspired by their work, we propose a coarse-to-fine approach that combines the advantages of description-based methods and pairwise similarity-based methods. We fully leverage the extracted BEV features to implement the coarse-to-fine process efficiently. In the coarse stage, these features are employed to generate attention-guided descriptors and enable fast affinity-based candidate selection. In the fine stage, we utilize the selected corresponding features to perform pairwise overlap estimation among the narrowed-down place candidates. By doing so, we efficiently detect loop closures and achieve leading performance compared to other state-of-the-art methods.

III. METHODOLOGY

Our approach employs a coarse-to-fine methodology, which consists of three primary modules: BEV feature extraction, coarse-grained matching using attention-guided descriptors, and fine-grained verification through pairwise overlap estimation. Firstly, the encoder extracts BEV features from multi-layer BEVs in 2D space. The features are subsequently used to generate an attention-guided global descriptor. Candidate scans are then rapidly retrieved through affinity-based candidate selection. Finally, we perform pairwise overlap estimation among the narrowed-down place candidates to find the final match. Additional design details, such as the loss function, are also discussed. An overview of our approach is illustrated in Fig. 2.

A. BEV Feature Extraction

The BEV feature extraction process follows our previous work [22]. This process begins by voxelizing the input point cloud to create a voxel grid, where each voxel is assigned a value of 0 or 1 depending on whether any points are present within it. This transformation results in the creation of multi-layer BEVs, denoted as $B \in \mathbb{R}^{H_B \times W_B \times C_B}$. Each layer is regarded as a channel during feature extraction. The feature encoder module is a 2D sparse convolution network with residual blocks. The resulting BEV feature volume is denoted as $f \in H_f \times W_f \times C_f$. f is stored in the database DB_f for later global descriptor generation and pairwise overlap estimation. The BEV features are particularly advantageous for detecting closed loops because they are insensitive to minor translations due to the voxelization operation and exhibit translation invariance due to the convolution operation.

B. Coarse-Grained Matching

Attention-guided Descriptor Generator. The attention-guided descriptor generation network consists of a self-attention module, a NetVLAD layer, and a fully connected layer. Since the BEV feature only contains information in a limited local neighborhood, it results in insufficient learning of patterns from the local feature volume for the global descriptor. Therefore, we utilize a self-attention module [35] to gain contextual information and further contribute to better performance in global descriptors generation.

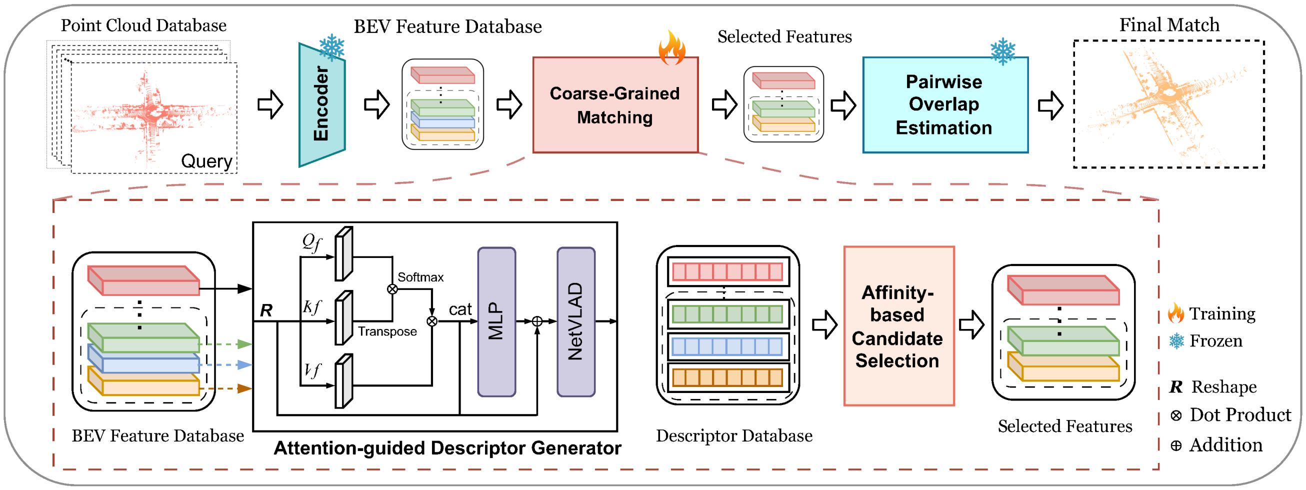


Fig. 2: The pipeline of our approach. The point clouds are first converted to voxel representations and fed into the encoder network to extract BEV features. Then, these features are used to generate global descriptors. In the coarse phase, the Top-K place candidates are selected by affinity-based selection. Then, in the fine phase, the corresponding BEV features are used pairwise to estimate the overlap region between the query scan and the candidates to find the final match.

1) Self-Attention on BEV feature: As the input of the attention module is required to be a sequence, we first reshape the BEV feature f into $f \in \mathbb{R}^{C_f \times L_f}$, where $L_f = H_f \times W_f$. Then f is transformed into three different feature spaces: query feature space $Q_f = W_q f$, key feature space $K_f = W_k f$ and value feature space $V_f = W_v f$, where $W_q, W_k, W_v \in \mathbb{R}^{C_f \times C_f}$ and $Q_f, K_f, V_f \in \mathbb{R}^{C_f \times L_f}$. The generated contextual information $A \in \mathbb{R}^{C_f \times L_f}$ is as follows:

$$A = \text{softmax}(\frac{Q_f K_f^T}{\sqrt{d_k}}) V_f, \tag{1}$$

where $d_k = C_f$. Then the contextual information A and original BEV feature f are used to yield a new feature f':

$$f' = f + \text{MLP}(\text{cat}(f, A)), \tag{2}$$

where MLP$(\cdot)$ denotes a three-layer fully connected network, cat$(\cdot,\cdot)$ means concatenation.

The self-attention module allows each local descriptor in the original feature f to gather contextual information from other local descriptors. By enlarging the receptive field and capturing spatial relations between local descriptors, the attention mechanism enhances the performance of feature extraction in the BEV representation. Furthermore, compared to convolutional networks, self-attention has been shown to improve the detection of reverse loops, as confirmed by our experiments in Section IV-D.

2) Global descriptor generator: Then the feature f' is fed into the NetVLAD layer to generate a global descriptor. NetVLAD is a neural network layer that aggregates local features into a compact and discriminative global descriptor by learning a set of cluster centers and soft assignments. The NetVLAD layer aggregates local descriptors $\mathbf{x}_i$ within the feature f' by summing the residuals between each descriptor and cluster centers, which are weighted by soft-assignment of descriptors to multiple clusters. The output VLAD representation is as follows:

$$V(j,k) = \sum_{i=1}^{N} \overline{a}_k(\mathbf{x}_i)(x_i(j) - c_k(j)), \tag{3}$$

where $\overline{a}_k$ denotes the soft-assignment, $\mathbf{x}_i$ is the i^{th} descriptor in f', and $x_i(j)$ is the j^{th} element of the i^{th} descriptor, c_k is the k^{th} cluster center.

Finally, a fully connected layer is utilized to reduce the dimension of the VLAD representation and computational costs, which follows [1]. Therefore, we obtain the attention-guided global descriptor vector v.

Affinity-based Candidate Selection. Firstly, we build up a descriptor database DB$_v$ from DB$_f$ so that every scan has a discriminative and independent description. Next, we use affinity-based candidate selection to identify loop candidates from the rest in DB$_v$. To assess the affinity between descriptor vectors, we adopt the Euclidean distance, known for its simplicity and efficiency. The affinity is calculated as follows:

$$Af_{PQ} = \left\| v_P - v_Q \right\|_2, \tag{4}$$

where $\left\| \cdot \right\|_2$ denotes the L_2 normalization. Using this simple calculation, we can efficiently select the Top-K candidates with the K closest Euclidean distances, significantly reducing the range for subsequent pairwise overlap estimation.

C. Fine-Grained Verification

In the fine stage, we determine the most similar place among the Top-K candidates by performing pairwise overlap estimation. The overlap estimation module [22] comprises a cross-attention module and a classification head. The cross-attention module effectively enhances feature representation by facilitating information interaction and contextual aggregation. The overlap estimation process is illustrated in Fig.3.

The cross-attention module fuses the relevant feature maps r_P and r_Q from pairwise BEV features f_P and f_Q as follows:

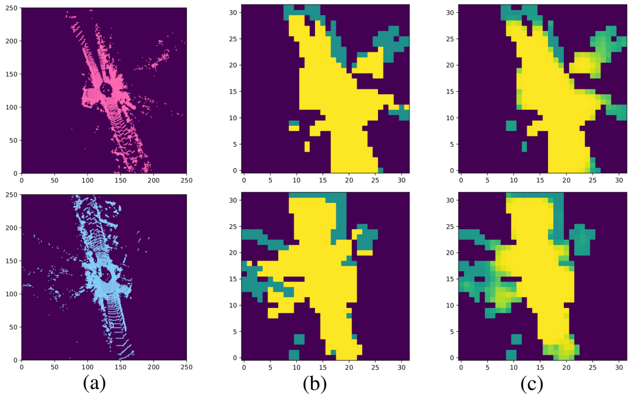


Fig. 3: The figure above depicts the process of overlap estimation, where the top row corresponds to the query scan, and the bottom row corresponds to the candidate. Panel (a) shows the input pairwise point clouds. The yellow region in panel (b) represents the ground truth overlap region and panel (c) shows the predicted overlap region.

$$
\begin{aligned}
r_P &= f_P + \text{MLP}(\text{cat}(f_P, \text{att}(f_P, f_Q, f_Q))) \\
r_Q &= f_Q + \text{MLP}(\text{cat}(f_Q, \text{att}(f_Q, f_P, f_P))),
\end{aligned} \tag{5}
$$

where $\text{MLP}(\cdot)$ denotes a fully connected network, $\text{cat}(\cdot,\cdot)$ means concatenation and $\text{att}(\cdot,\cdot,\cdot)$ is the attention model.

The overlap estimation is regarded as a binary classification problem. The classification head consists of two 3×3 convolutional layers and a sigmoid layer with ReLU as the activation function as follows:

$$
\gamma_t = \text{Sigmoid}(\text{Conv}_{3\times3}(\text{ReLU}(\text{Conv}_{3\times3}(r_t)))), t \in \{P, Q\}. \tag{6}
$$

Thus, we can get the overlap score as follows:

$$
\tau = \frac{1}{2}\left(\frac{\sum \gamma_P}{N_P} + \frac{\sum \gamma_Q}{N_Q}\right), \tag{7}
$$

where N_P and N_Q denote the total number of non-zeros pixels of r_P and r_Q.

By quantifying the overlapping area with the overlap score to evaluate the similarity of pairwise point clouds, we can get the candidate with the highest overlap score as the final match.

D. Loss Function

In our approach, we only train the attention-guided descriptor generation network using lazy triplet loss [1]. The BEV feature extraction network and overlap estimation network use the pre-trained model of our previous work [22]. The lazy triplet loss is defined as follows:

$$
\mathcal{L} = max([m + \delta_{pos_i} - \delta_{neg_j}]_+), \tag{8}
$$

where $[\cdot]_+$ represents the hinge loss, m is a constant margin, and $\delta(\cdot)$ is the descriptor distances between the query and positive/negative samples. We consider point clouds that are less than σ_{pos} meters away as positive samples and those more than σ_{neg} meters away as negative samples.

IV. EXPERIMENTS

To demonstrate the effectiveness and practicality of our proposed approach for LiDAR-based place recognition, we design a series of experiments. We evaluate our method on the widely-used KITTI and KITTI-360 datasets and compare its performance with other state-of-the-art methods. Furthermore, we conduct ablation studies to investigate the impact of the coarse matching stage on the final results. Additionally, runtime experiments are performed to evaluate the computational efficiency of our method.

A. Datasets

KITTI. The KITTI dataset [23] contains point clouds from various urban environments, collected using a Velodyne HDL-64E 3D LiDAR sensor. The dataset includes 22 sequences, but only the first 11 sequences have ground truth poses, so we use the ground truth provided by SemanticKITTI [36]. We train our network on the last 11 sequences and evaluate the performance of loop closing on sequences 00, 02, 05, 06, 07, and 08.

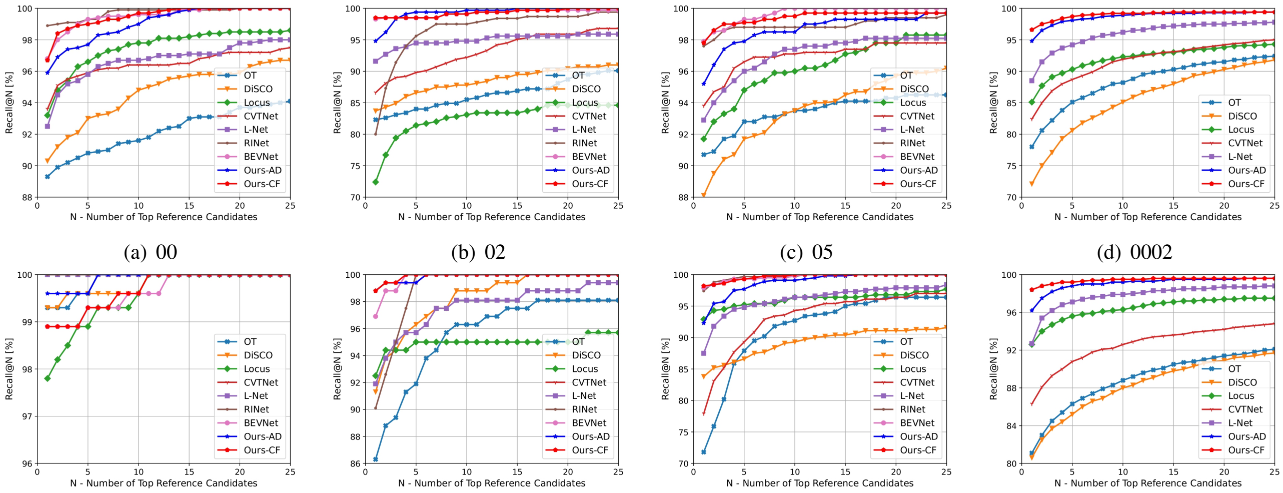
KITTI-360. The KITTI-360 dataset [24] is a suburban driving dataset that consists of 9 sequences, with 6 sequences containing loops. Compared to the KITTI dataset, KITTI-360 contains more scans and more loops and reverse loops. Following [37], we evaluate our approach on sequence 0002 and sequence 0009, which contain the highest number of loop closures. The dataset contains 3D data from a Velodyne HDL-64E and a SICK LMS 200 LiDAR sensor. In our experiments, we use the 3D raw scans captured by the Velodyne scanner.

B. Implementation Details

In the preprocessing stage, we crop each point cloud with a [-50m, 50m] cubic window and [-4m, 3m] in the z axis, then voxelize the input point cloud into a BEV representation with the shape of 256 × 256 × 32. The encoder network subsequently extracts a BEV feature volume with the shape of $32 \times 32 \times 512$ using the SpConv [38] for improved speed. These features are saved in the database DB_f.

For global descriptor generation, we use a NetVLAD layer with an intermediate feature dimension of 512, a maximum number of pooled samples of 1024, and 32 clusters. The final attention-guided global descriptor vector v has a length of 1024. For triplet loss, we set the margin $m = 0.3$, $\sigma_{pos} = 10$, $\sigma_{neg} = 50$, and use 2 positive and 10 negative samples for training. The overlap estimator predicts the pairwise overlap region using a $1 \times 32 \times 32$ tensor.

We only train the attention-guided descriptor generation network while using the pre-trained model from [22] to extract BEV features and perform overlap estimation, which is also trained on the last 11 sequences of the KITTI dataset. For candidate selection in the coarse stage, we use $K = 25$ as the number of candidates. We denote our method with only global descriptor matching as ours-AD and the complete coarse-to-fine method as ours-CF.



(a) 00 (b) 02 (c) 05 (d) 0002

(e) 06 (f) 07 (g) 08 (h) 0009

Fig. 4: Recall@N on KITTI and KITTI-360 datasets.

TABLE I: Recall@1 and Recall@1% on KITTI and KITTI-360 datasets

Methods	KITTI							KITTI-360		
	00	02	05	06	07	08	Mean	0002	0009	Mean
OT [7]	89.3/95.0	82.3/93.9	90.7/94.5	99.3/**100.0**	86.3/96.3	71.8/97.7	86.6/96.2	78.0/97.0	81.1/95.7	79.6/96.4
DiSCO [16]	90.3/97.7	83.7/91.9	88.1/96.2	99.3/**100.0**	91.3/98.8	83.8/93.2	89.4/96.3	72.1/98.1	80.6/96.6	76.4/97.4
Locus [6]	93.2/98.9	72.4/85.2	91.7/98.6	97.8/**100.0**	92.5/95.0	92.9/98.4	90.1/96.0	85.1/97.3	92.6/98.9	88.8/98.1
CVTNet [33]	93.6/99.1	86.6/97.7	93.8/97.8	**100.0/100.0**	**98.8/100.0**	77.9/97.9	91.8/98.8	82.4/97.8	86.3/97.1	84.4/97.4
L-Net [2]	92.5/99.1	91.6/96.5	92.9/98.1	**100.0/100.0**	91.9/98.1	87.5/98.9	92.7/98.4	88.5/99.4	92.7/99.4	90.6/99.4
RINet [21]	**98.9/100.0**	80.0/99.7	97.6/99.6	**100.0/100.0**	90.1/**100.0**	97.4/**100.0**	94.0/99.9	-	-	-
BEVNet [22]	96.8/**100.0**	98.3/**100.0**	**97.9/100.0**	98.9/99.6	96.9/**100.0**	97.9/**100.0**	97.8/99.9	-	-	-
Ours-AD	95.9/**100.0**	94.8/**100.0**	95.2/99.7	99.6/**100.0**	**98.8/100.0**	92.3/**100.0**	96.1/**100.0**	94.8/**99.7**	96.2/**99.8**	95.5/**99.8**
Ours-CF	96.7/**100.0**	**98.5/100.0**	97.8/99.7	98.9/**100.0**	**98.8/100.0**	**98.2/100.0**	**98.1/100.0**	**96.6**/99.4	**98.4**/99.6	**97.5**/99.5

C. Evaluation of Loop Closure Detection

We compare our approach with state-of-the-art methods, including description-based methods: OverlapTransformer [7] (denoted as OT), DiSCO [16], Locus [6], CVTNet [33], LoGG3D-Net [2] (denoted as L-Net), and pairwise similarity-based methods: RINet [21] and BEVNet [22]. For fairness, all these methods trained on other sequences are retrained on sequences 11-21 of the KITTI dataset, except Locus, which does not require training. During the evaluation, we exclude the previous 50 scans of the query scan as they are too close.

We use Recall@1 and Recall@1% as the evaluation metrics. The inference is considered correct if the distance between two scans is less than 10m. As shown in Tab. I and Fig. 4, ours-CF achieves the highest AR@1 on the KITTI dataset. Ours-CF is based on BEVNet, but we improve place recognition ability while significantly reducing the time-consuming pairwise overlap estimation process. In general, pairwise similarity-based methods perform better than description-based methods, where CVTNet achieves the best performance except for ours-AD, as it uses a cross-view transformer network to fuse the features extracted from range image views and bird's eye views images, but the result in sequence 08 shows it does not detect reverse loop closures well, as sequence 08 only has reverse loops. In terms of AR@1%, RINet, BEVNet, Ours-AD, and Ours-CF achieve nearly 100%. The AR@1% of ours-AD indicates that our candidate selection process provides valid and accurate candidates, which is the foundation of our approach.

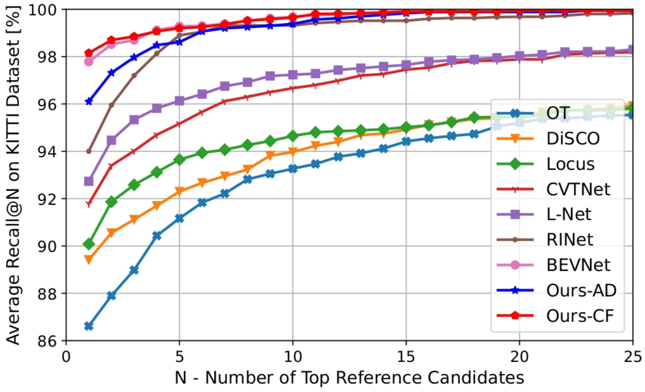


Fig. 5: AR@N of ours-CF on KITTI dataset.

Furthermore, we evaluate the performance of the proposed method using AR@N on the KITTI dataset. As shown in Fig. 5, our method outperforms other methods. The experimental results demonstrate our methods' strong robustness and good generalization ability for place recognition under challenging conditions.

To assess the generalization ability of each method, we conduct direct evaluations on the KITTI-360 dataset without additional training. Since semantic annotations are provided on accumulated point clouds instead of 3D raw scans, we do not evaluate RINet on the KITTI-360 dataset. Furthermore,

TABLE II: Study on candidate selection ability

	00	02	05	06	07	08	Mean
Recall@1	95.9	94.8	95.2	99.6	98.8	92.3	96.1
Recall@10	99.0	99.7	98.5	100.0	100.0	99.1	99.4
Recall@15	99.9	100.0	99.3	100.0	100.0	99.8	99.8
Recall@20	100.0	100.0	99.3	100.0	100.0	100.0	99.9
Recall@25	100.0	100.0	99.7	100.0	100.0	100.0	100.0
Recall@1%	100.0	100.0	99.7	100.0	100.0	100.0	100.0

TABLE III: Study on candidate number of ours-CF

Methods	00	02	05	06	07	08	Mean
Ours-CF-1	95.9	94.8	95.2	99.6	98.8	92.3	96.1
Ours-CF-10	96.5	98.3	96.9	98.9	98.8	97.5	97.8
Ours-CF-15	97.0	98.5	97.1	98.9	98.8	97.5	98.0
Ours-CF-20	96.8	98.5	97.1	98.9	98.8	97.9	98.0
Ours-CF-25	96.7	98.5	97.8	98.9	98.8	98.2	98.1
Ours-CF-1%	96.8	98.3	97.9	98.9	98.8	98.2	98.1

Ours-CF-K means overlap estimation on Top-K candidates proposed by descriptor matching. The result is Recall@1.

we do not evaluate BEVNet as it requires a significant amount of time, as discussed in detail in Sec. IV-E. It can be observed that ours-AD and ours-CF demonstrate consistent recognition ability across different datasets.

D. Ablation Study

In this section, we present our ablation studies on the proposed approach. First, we investigate the impact of the global descriptor matching in the coarse stage. We choose the number of place candidates as $N = 1, 10, 15, 20, 25, 1\%$ to test Recall@N for place recognition using only global descriptor matching, which we refer to as ours-AD. As shown in Tab. II, both AR@25 and AR@1% are nearly 100.0%. This result suggests that 25 candidates are sufficient to achieve outstanding performance.

Next, we further evaluate the effect of candidate number K in the fine stage where the overlap estimation is applied. As shown in Tab. III, Recall@1 increases with the number of candidates, as expected. Ours-CF-25 has a similar performance to ours-CF-1%. Combining these results with those in Tab. II, we choose $K = 25$ as the candidate number for a coarse search. Besides, a fixed candidate number can help stabilize running time.

We also test the impact of the self-attention module, as shown in Tab. IV, in which ours-GD uses two 3×3 convolutional layers and a sigmoid layer with ReLU activation instead of the self-attention module. The results clearly demonstrate that the self-attention module significantly enhances the ability to detect reverse loops.

E. Runtime

In this section, we conduct experiments to evaluate the efficiency of our method. The experiments are performed on a system equipped with an Intel Core i7-7700 CPU and an Nvidia GeForce GTX 1080 Ti GPU, using sequence 00, which contains 4541 scans. First, we evaluate the runtime efficiency of attention-guided descriptor generation and compare it to other methods. The preprocessing steps include range image generation (OverlapTransformer), BEV representation construction (DiSCO and ours-AD), segments extraction (Locus), range and BEV image generation (CVTNet), and voxelization (L-Net). The batch size is set to 1 for all methods. As shown in Tab. V, the global descriptor generation in our approach is quite efficient compared to the other methods, which is the foundation of our proposed coarse-to-fine approach.

TABLE IV: Study on self-attention module (Recall@1)

Methods	00	02	05	06	07	08	Mean
Ours-GD	95.9	91.0	94.3	99.6	97.5	84.7	93.8
Ours-AD	95.9	94.8	95.2	99.6	98.8	92.3	96.1

TABLE V: Runtime of descriptor generation (ms)

Methods	Preprocessing	Description	Total
OT	26.92	2.22	29.14
DiSCO	11.67	3.75	15.42
Locus	1166.49	613.96	1780.45
CVTNet	191.96	13.13	205.09
L-Net	31.64	63.88	95.52
Ours-AD	34.86	43.36	78.22

TABLE VI: Runtime of different modules (ms)

Module	Runtime
Voxelization	34.86
BEV Feature Extraction	41.74
Attention-guided Descriptor Generation	1.62
Affinity-based Candidate Selection	0.04
Pairwise Overlap Estimation	8.78

Tab. VI shows the time consumption of each module in our proposed approach. The candidate selection only takes 0.04ms, while the pairwise overlap estimation takes 8.78ms. Thanks to fast candidate selection, ours-CF performs overlap estimation a limited number of times, whereas BEVNet uses pairwise overlap estimation to search exhaustively. In Sec. IV-C, when evaluating BEVNet on sequence 0002 and sequence 0009, we encounter unacceptable long loop detection times. In sequence 0002, there are 14,122 valid scans and 4,134 loops, which would require 31,408,854 pairwise overlap estimations if exhaustively searching without candidate selection. The time cost of ours-CF is $14122 * (34.86{+}41.74{+}1.62){+}4134{*}(0.04{+}25{*}8.78)ms = 0.56h$. In contrast, the time cost of BEVNet is $14122 * (34.86 + 41.74 + 1.62) + 8.78 * 31408854ms = 76.91h$. It is evident that the time required for BEVNet to detect loop closures is significantly longer than the time for LiDAR data acquisition. In ours-CF, the speed can be further improved by adjusting the value of K according to the actual needs. When $K = 1$, it degenerates into ours-AD, which is still capable of detecting loops effectively. This experiment demonstrates the effectiveness and efficiency of combining descriptor matching and pairwise overlap estimation on shared BEV features.

V. CONCLUSION

In this paper, we propose a novel coarse-to-fine approach for LiDAR-based place recognition. This approach combines global descriptor matching and overlap estimation based on shared BEV features to achieve both speed and accuracy in loop closure detection. While our approach has some limitations, such as the bulky and memory-intensive BEV features, future work will address these challenges and focus on improving efficiency and seeking better results.

References

[1] M. A. Uy and G. H. Lee, “Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4470–4479, 2018.

[2] K. Vidanapathirana, M. Ramezani, P. Moghadam, S. Sridharan, and C. Fookes, “Logg3d-net: Locally guided global descriptor learning for 3d place recognition,” in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2215–2221, 2022.

[3] J. Komorowski, “Improving point cloud based place recognition with ranking-based loss and large batch training,” in *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 3699–3705, 2022.

[4] J. Komorowski, “Minkloc3d: Point cloud based large-scale place recognition,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1790–1799, 2021.

[5] 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.

[6] K. Vidanapathirana, P. Moghadam, B. Harwood, M. Zhao, S. Sridharan, and C. Fookes, “Locus: Lidar-based place recognition using spatiotemporal higher-order pooling,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[7] 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 Robotics and Automation Letters*, vol. 7, no. 3, pp. 6958–6965, 2022.

[8] J. Ma, X. Chen, J. Xu, and G. Xiong, “Seqot: A spatial-temporal transformer network for place recognition using sequential lidar data,” *IEEE Transactions on Industrial Electronics*, pp. 1–10, 2022.

[9] P. Yin, F. Wang, A. Egorov, J. Hou, J. Zhang, and H. Choset, “Seqspherevlad: Sequence matching enhanced orientation-invariant place recognition,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5024–5029, 2020.

[10] S. Zhao, P. Yin, G. Yi, and S. Scherer, “Spherevlad++: Attention-based and signal-enhanced viewpoint invariant descriptor,” *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 256–263, 2023.

[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] 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.

[13] G. Kim, S. Choi, and A. Kim, “Scan context++: Structural place recognition robust to rotation and lateral variations in urban environments,” *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1856–1874, 2021.

[14] H. Wang, C. Wang, and L. Xie, “Intensity scan context: Coding intensity and geometry relations for loop closure detection,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2095–2101, 2020.

[15] 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.

[16] X. Xu, H. Yin, Z. Chen, Y. Li, Y. Wang, and R. Xiong, “Disco: Differentiable scan context with orientation,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2791–2798, 2021.

[17] L. Luo, S.-Y. Cao, B. Han, H.-L. Shen, and J. Li, “Bvmatch: Lidar-based place recognition using bird’s-eye view images,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6076–6083, 2021.

[18] B. Jiang and S. Shen, “Contour context: Abstract structural distribution for 3d lidar loop detection and metric pose estimation,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, (London, United Kingdom), p. 8386–8392, 2023.

[19] L. Lun, Z. Shuhang, L. Yixuan, F. Yongzhi, Y. Beinan, C. Siyuan, and S. Hui-Liang, “BEVPlace: Learning LiDAR-based place recognition using bird’s eye view images,” *arXiv preprint arXiv:2302.14325*, 2023.

[20] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss, “OverlapNet: Loop Closing for LiDAR-based SLAM,” in *Proceedings of Robotics: Science and Systems (RSS)*, 2020.

[21] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, and Y. Liu, “Rinet: Efficient 3d lidar-based place recognition using rotation invariant neural network,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4321–4328, 2022.

[22] L. Li, W. Ding, Y. Wen, Y. Liang, Y. Liu, and G. Wan, “A unified bev model for joint learning of 3d local features and overlap estimation,” *arXiv preprint arXiv:2302.14511*, 2023.

[23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[24] Y. Liao, J. Xie, and A. Geiger, “KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d,” *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[25] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, “Deep learning features at scale for visual place recognition,” in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 3223–3230, IEEE, 2017.

[26] M. Gehrig, E. Stumm, T. Hinzmann, and R. Siegwart, “Visual place recognition with probabilistic voting,” in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3192–3199, IEEE, 2017.

[27] S. Schubert, P. Neubert, and P. Protzel, “Unsupervised learning methods for visual place recognition in discretely and continuously changing environments,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4372–4378, 2020.

[28] L. G. Camara, C. Gäbert, and L. Přeučil, “Highly robust visual place recognition through spatial matching of cnn features,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3748–3755, May 2020.

[29] K. Zhang, Z. Li, and J. Ma, “Appearance-based loop closure detection via bidirectional manifold representation consensus,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6811–6817, IEEE, 2021.

[30] 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, 2016.

[31] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3d classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[32] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, “Netvlad: Cnn architecture for weakly supervised place recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

[33] J. Ma, G. Xiong, J. Xu, and X. Chen, “Cvtnet: A cross-view transformer network for place recognition using lidar data,” *arXiv preprint arXiv:2302.01665*, 2023.

[34] K. P. Cop, P. V. Borges, and R. Dubé, “Delight: An efficient descriptor for global localisation using lidar intensities,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3653–3660, IEEE, 2018.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning*, pp. 7354–7363, PMLR, 2019.

[36] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, “Semantickitti: A dataset for semantic scene understanding of lidar sequences,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

[37] D. Cattaneo, M. Vaghi, and A. Valada, “Lcdnet: Deep loop closure detection and point cloud registration for lidar slam,” *IEEE Transactions on Robotics*, pp. 1–20, 2022.

[38] Y. Yan, Y. Mao, and B. Li, “Second: Sparsely embedded convolutional detection,” *Sensors*, vol. 18, no. 10, p. 3337, 2018.