CRPlace: Camera-Radar Fusion with BEV Representation for Place Recognition

Shaowei Fu, Yifan Duan, Yao Li, Chengzhen Meng, Yingjie Wang, Jianmin Ji, Yanyong Zhang*

Abstract— The integration of complementary characteristics from camera and radar data has emerged as an effective approach in 3D object detection. However, such fusion-based methods remain unexplored for place recognition, an equally important task for autonomous systems. Given that place recognition relies on the similarity between a query scene and the corresponding candidate scene, the stationary background of a scene is expected to play a crucial role in the task. As such, current well-designed camera-radar fusion methods for 3D object detection can hardly take effect in place recognition because they mainly focus on dynamic foreground objects. In this paper, a background-attentive camera-radar fusion-based method, named CRPlace, is proposed to generate background-attentive global descriptors from multi-view images and radar point clouds for accurate place recognition. To extract stationary background features effectively, we design an adaptive module that generates the background-attentive mask by utilizing the camera BEV feature and radar dynamic points. With the guidance of a background mask, we devise a bidirectional cross-attention-based spatial fusion strategy to facilitate comprehensive spatial interaction between the background information of the camera BEV feature and the radar BEV feature. As the first camera-radar fusion-based place recognition network, CRPlace has been evaluated thoroughly on the nuScenes dataset. The results show that our algorithm outperforms a variety of baseline methods across a comprehensive set of metrics (recall@1 reaches 91.2%).

I. INTRODUCTION

Place recognition serves as a pivotal function in autonomous driving by addressing the fundamental question of "where am I within a predefined reference map". As an essential component of global localization, it enables the retrieval of a place within a large map database that closely matches the current place [2]. Furthermore, it plays a vital role in Simultaneous Localization and Mapping (SLAM), aiding in the detection of loop closures to rectify drift and tracking errors [3].

Cameras and LiDAR are the most commonly used sensors in place recognition. Visual place recognition [2], [4]–[6] has been widely studied due to the low cost and rich texture information provided by cameras. However, it is susceptible to challenges posed by visual degradation such as night, rain, and direct sunlight [7]. On the other hand, methods based on LiDAR [8]–[12] offer improved robustness against illumination conditions but lack rich texture features. Recent efforts have sought to fuse these two modalities for place recognition [1], [3], [13]–[15], resulting in significant

* The corresponding author.

School of Computer Science and Technology, University of Science and Technology of China, Hefei, 230026, China {fushw, dyf0202, zkdly, czmeng, yingjiewang}@mail.ustc.edu.cn, {jianmin, yanyongz}@ustc.edu.cn.
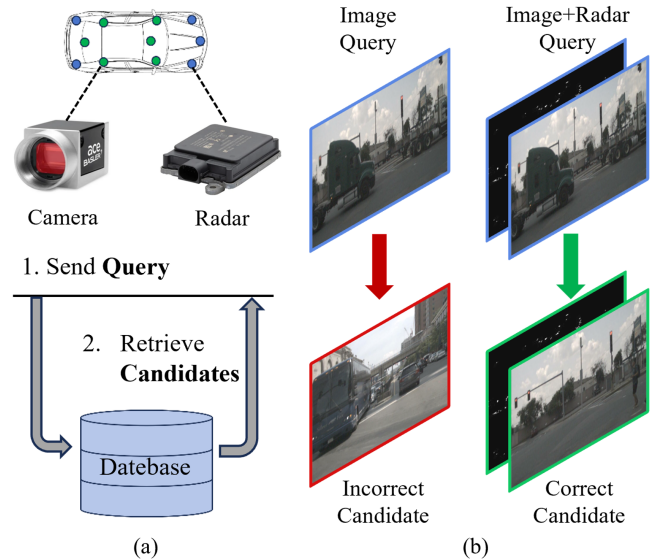
Fig. 1: An illustration of (a) place recognition task with camera and radar fusion and (b) the place recognition results using image-only and fusion-based methods, respectively. Given an image query (marked in blue bounding box) that includes multiple dynamic objects, the image-based place recognition [1] retrieves an incorrect candidate due to the influence of dynamic objects (marked in red bounding box), while our method retrieves the correct candidate successfully (marked in green bounding box) with the image-radar query acquired from the same place.

enhancements in performance and robustness. Nevertheless, these methods still fall short under adverse weather conditions such as rain, snow, and fog [12].

Unlike cameras and LiDAR, millimeter-wave radar (referred to as radar in this paper) remains nearly unaffected by harsh weather conditions and obstacles. It can provide 3D geometry information like LiDAR but is more lightweight and inexpensive. Consequently, radar is becoming an increasingly attractive sensor in autonomous driving. Due to the challenges of sparsity and noisy measurements in radar, it has become a common practice to fuse radar with camera or LiDAR to exploit their complementary characteristics for 3D object detection [16]–[18]. However, these methods mainly focus on pre-defined foreground objects, rendering them unsuitable for place recognition. Specifically, place recognition retrieves candidates by querying the most similar scene in the database (Fig. 1(a)). In this demand, the moving objects from the scene could be misleading instead. We provide an example in Fig. 1(b), the image-based place recognition

method [1] fails to retrieve correct candidates since the disturbance from dynamic objects could not be eliminated. Therefore, *it is necessary to extract global descriptors[1] that focus on stationary background information with the help of the velocity-sensitive radar data.* Although studies have attempted to exploit the unique characteristics of radar data for place recognition [7], [19], [20], how to effectively fuse the complementary characteristics of camera and radar data while focusing on the background remains a challenge.

In this work, we propose CRPlace, a novel place recognition method that fuses the complementary characteristics of multi-view camera images and radar points to generate background-attentive global descriptors. Specifically, CR-Place first engages a Background-Attentive Mask Generation (BAMG) module to adaptively create an attention mask that focuses on the stationary background feature while ignoring the dynamic feature. Guided by the background mask, a Bidirectional cross-attention-based Spatial Fusion (BSF) module is then devised to enable thorough background feature interaction and learn the soft association between camera and radar BEV features. In detail, the Radar-to-Image (R2I) fusion takes each pixel in the camera BEV feature as a query to learn spatial background information from the radar BEV feature, and the Image-to-Radar (I2R) fusion utilizes rich contextual background information from the image feature to enhance the sparse radar feature. Additionally, we establish a baseline for camera-radar fusion-based place recognition by directly combining the feature extraction module of the SOTA fusion-based detection network, i.e., BEVFusion [21], and the global descriptor aggregation module of the SOTA 360-degree visual place recognition network vDiSCO [1]. We refer to this baseline as BEVFusion in the remainder of the paper, which also supports camera-only and radar-only methods.

We evaluate our method on nuScenes dataset [22]. We compare CRPlace with several state-of-the-art camera-based, radar-based, and camera-radar fusion-based methods, all of which do not take into account the influence of dynamic objects. CRPlace outperforms these methods with significant margins (with the relative recall@1 increase of 3.6% to 12.9%). We also validate the robustness of our method in rain conditions, achieving a relative recall@1 increase of 30.1%. In summary, our contributions are:

- We propose a novel and robust background-attentive Camera-Radar fusion-based place recognition method, namely CRPlace, to combine the complementary characteristics of camera and radar in the BEV representation. To the best of our knowledge, this is the first work that effectively fuses multi-view cameras and radars for the task of place recognition.
- We design an adaptive background-attentive mask generation module and a bidirectional cross-attention-based spatial fusion module to learn and interact with stationary background features effectively.
- We conduct extensive experiments on the nuScenes

dataset to validate the merits of our method and show considerably improved performance.

II. RELATED WORK

A. Single-modal Place Recognition

Camera, LiDAR, and radar have all been employed for place recognition.

Camera-based Methods. In camera-based methods, hand-crafted local features [23]–[25] and their vector of locally aggregated descriptors (VLAD) [26] are traditionally used for recognition, but they have been replaced by convolutional neural networks (CNNs) like VGG [27] and AlexNet [28]. NetVLAD [2] is an end-to-end learnable method specifically designed for large-scale place recognition. It first extracts the local feature using VGG/AlexNet, followed by a differentiable VLAD layer used for local feature aggregation, which can be plugged into various learning-based feature extractors and trained through backpropagation. Similarly, Generalized-Mean Pooling (GeM) [4] is an efficient aggregation method, enabling the network to aggregate a compact global descriptor end-to-end.

LiDAR-based Methods. Some LiDAR-based methods project point clouds to 2D structures such as Scan Context [29] and Scan Context++ [30], while MinkLoc3D [12] and Locus [11] directly operate in 3D space by discretizing it into voxel grids. Inspired by NetVLAD, PointNetVLAD [8] combines PointNet [31] and NetVLAD to enable end-to-end training and extraction for the global descriptor from 3D point clouds. OverlapNet [9] and DiSCO [10] try to simultaneously estimate the relative yaw between pairs of scans and their similarity.

Radar-based Methods. In radar-based methods, UnderTheradar [20] uses intermediate features as global descriptors. Kidnappedradar [19] exploits a variant of NetVLAD as the feature extractor to improve rotation invariance. AutoPlace [7] is the first work that uses single-chip automotive radar for place recognition. It first removes the dynamic points from instant Doppler measurement and then extracts spatial-temporal features from radar point clouds with a compact deep neural network. Subsequently, the obtained candidates are re-ranked using Radar Cross Section (RCS) measurement to refine the recognition accuracy.

B. Multi-modal Place Recognition

Many efforts have been made in various works towards LiDAR and camera fusion-based multi-modal place recognition. CORAL [3] first builds the elevation image from 3D points, which is then enhanced with projected RGB image features. In this way, the structural features and visual features are fused in the bird-eye view. MinkLoc++ [14] extracts the global descriptor for LiDAR point cloud and RGD image separately and fuses them in the last channel. Considering the fact that the importance of camera and LiDAR varies as the environment changes, AdaFusion [13] tries to learn the adaptive weights for both image and point cloud features.
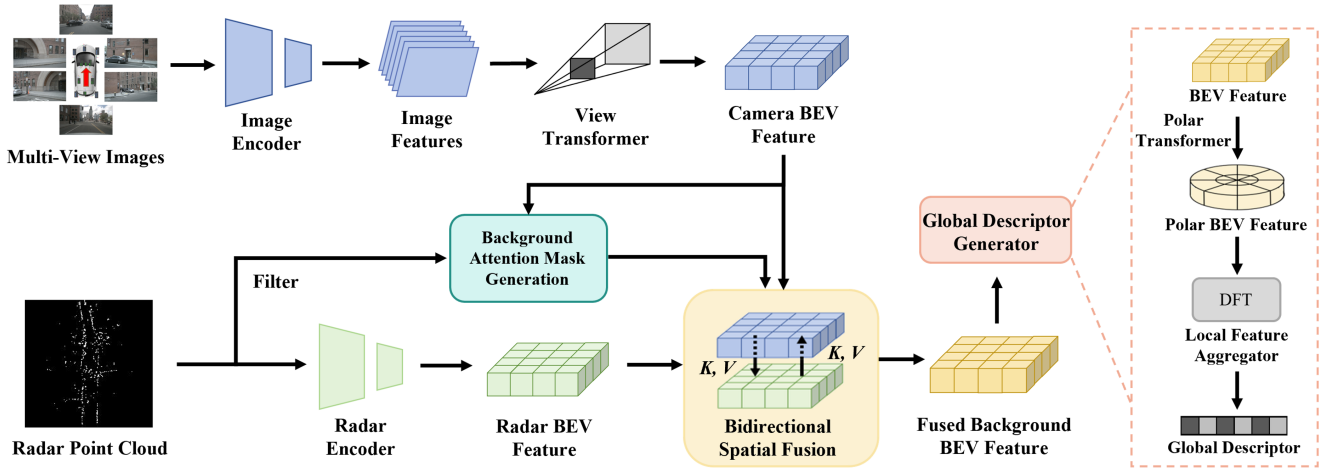
Fig. 2: The network architecture of the proposed CRPlace. Given multi-view images and radar point clouds, two modality-specific streams separately extract features and transform them into the same BEV space at first. Next, the Background-Attentive Mask Generation (BAMG) module uses radar dynamic points and camera BEV features to create a background attention mask adaptively. Then the Bidirectional Spatial Fusion (BSF) module attentively fuses background BEV features from these two modalities. Finally, the Global Descriptor Generator uses the fused BEV features to generate rotation-invariant global descriptors.

However, methods based on camera-radar fusion for place recognition have not yet received any attention.

C. Rotation Invariance

Rotation invariance is crucial in place recognition. Bird's Eye View (BEV) representation has a natural advantage in achieving rotation invariance, which has been widely used in LiDAR-based methods [9], [10], [30]. OverlapNetVLAD [32] is a coarse-to-fine place recognition framework that efficiently uses BEV features to perform loop closure. Previous visual place recognition methods use single-view image to extract the global descriptor, which fail to retrieve the correct candidates when revisiting the same place from different perspectives. Pioneering works LSS [33], BEVFusion [21] and BEVFormer [34] have demonstrated that aggregating features from multi-view images into a unified BEV representation can significantly improve detection and segmentation performance. Inspired by these methods, a recent work, called vDiSCO [1], proposes a method to employ BEV representation for 360-degree visual place recognition, achieving remarkable results. vDiSCO extracts BEV features from multi-view images based on BEVFormer, then combines the polar transformation and the Discrete Fourier transform to aggregate rotation-invariant global descriptors from BEV features. It also supports the vision-LiDAR fusion method. This method has effectively demonstrated the advantages of BEV representation in the task of place recognition.

III. METHOD

In this work, we present CRPlace, a background-attentive camera-radar fusion network for place recognition. As shown in Fig. 2, multi-view images and radar point clouds are separately fed into the camera feature stream and radar feature stream to extract their BEV features. Next, we involve a Background-Attentive Mask Generation (BAMG) module to create a background attention mask adaptively by combining camera BEV features and radar dynamic points. Then a Bidirectional cross-attention-based Spatial Fusion (BSF) module is devised with the guidance of the attention mask to interact with the background features attentively. Subsequently, the rotation-invariant global descriptor is generated according to the method in vDiSCO [1]. Following this way, a place can be represented by a background-attentive global descriptor, denoted as D_p, so that we can generate a global descriptor for each place in a given map to build a database $\{D_i\}$. We also generate the global descriptor for the current query place, say D_q. By comparing the Euclidean distance between the query place descriptor and the place descriptor in the database, the current place can be finally recognized as the one in the map with the minimal difference, denote the i^*th place as:

$$i^* = \arg \min_i \|D_q - D_i\|_2 . \tag{1}$$

Below we will present the details of CRPlace.

A. Modality-Specific BEV Feature Encoding

1) Multi-View Image Feature Encoding: Following BEV-Fusion [21], we use Swin-T [35] as the image backbone to encode the multi-view images into deep features. The Feature Pyramid Network (FPN) [36] is then applied to fuse the multi-scale features. An adaptive average pooling layer is applied to better align these features. To transform these image features from 2D coordinate into 3D ego-car coordinate, we apply the view transformer proposed in [33] to explicitly predict the depth distribution of each pixel. Then, the image-view features can be projected onto the predefined point cloud and a pseudo voxel $V \in R^{X \times Y \times Z \times C}$ can be generated according to camera extrinsic parameters and the predicted image depth. Note that the depth prediction module
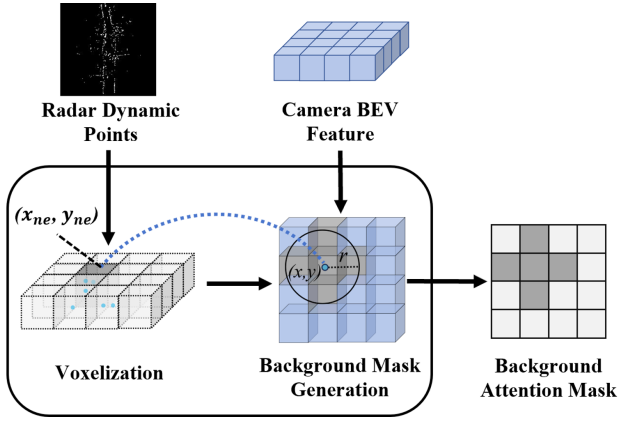
Fig. 3: An illustration of the BAMG module. All dynamic points are selected from radar point clouds and voxelized into a grid. Then the radar voxel grid and camera BEV feature are utilized to generate the background attention mask adaptively according to their positional relationships. (x_{ne}, y_{ne}) denotes the non-empty voxel.

may produce inaccurate depth, leading to the projection of image features to incorrect BEV positions. Therefore, we use ground-truth depth derived from LiDAR point clouds to supervise the depth distribution prediction during training, following BEVDepth [37]. Subsequently, a BEV pooling is applied to reshape the pseudo voxel into a BEV feature map $F_C \in R^{C \times H \times W}$.

2) Radar Feature Encoding: We preprocess the radar point cloud into a feature set containing the 2D coordinate (x, y), radial velocity (v_x, v_y), radar cross-section rcs, dynamic property $dynProp$, cluster validity state $invalid\ state$ and timestamp t. Five frames of radar scans within the same sample are stacked into a full-view radar point cloud in the LiDAR coordinate. To avoid overly sparse inputs, we further accumulate a sequence of radar point clouds into one frame. The dynamic points are identified and removed from the radar point cloud based on radial velocity, as described in [7]. Then, we exploit a PillarFeatureNet [38] to extract radar features which directly converts the radar input to a pseudo image in the BEV space. This encoder considerably alleviates the computation of sparse radar data to traverse the BEV plane in the absence of vertical information. Naturally, the radar BEV features $F_R \in R^{C \times H \times W}$ are generated after a linear transformation, where C, H, and W are equal to F_C.

B. Background-Attentive Mask Generation

Before feature fusion, it is necessary to distinguish which areas in the BEV features represent dynamic objects and which represent stationary backgrounds. In our BAMG module, a background attention mask $M \in \{0, 1\}^{H \times W}$ of the same size as the camera BEV feature is defined to prevent attention for dynamic features, where $H = W = 128$. As shown in Fig. 3, we first retain all dynamic points in the radar point cloud and then perform voxelization to generate a voxel grid $V \in R^{H \times W}$. For each non-empty grid $V(x_{ne}, y_{ne})$, it means that there is a dynamic object near

the position (x_{ne}, y_{ne}). The corresponding position in the camera BEV feature is also considered to have a dynamic object because the voxel grid and the camera BEV feature have been transformed into the same BEV coordinate system. Furthermore, the incorporation of explicit depth supervision during the view transformer process of the camera feature stream further enhances the spatial alignment of the camera BEV feature and radar voxel grid. While we have a general idea of the approximate position of the dynamic object, its spatial extent is not clear. In other words, it is uncertain which pixels belong to the dynamic features. As a result, we set a learnable parameter r to allow the network to adaptively learn the areas occupied by dynamic features. Specifically, for each non-empty voxel grid $V(x_{ne}, y_{ne})$, if the Euclidean distance between the coordinate (x, y) of camera BEV feature pixel $F_C(x, y)$ and (x_{ne}, y_{ne}) is less than the threshold r, it is considered that this feature pixel belongs to the dynamic feature, and the corresponding position of background attention mask $M(x, y)$ will be set to 0. This can be represented using the following formula:

$$M(x,y) = \begin{cases} 0, & \text{if } \|(x,y) - (x_{ne}, y_{ne})\|_2 \leq r \\ 1, & \text{else} \end{cases}, \quad (2)$$

where $x, x_{ne} \in H$, $y, y_{ne} \in W$ and r is initialized to 0.5.

C. Bidirectional Spatial Fusion

While it is possible to identify stationary background features using the generated background mask, simply fusing camera features and radar features through element-wise addition or concatenation would result in spatial misalignment and significant feature wastage due to the sparsity of radar measurements. For this reason, our BSF module aims to fully interact the complementary characteristics and learn the soft-association between camera features and radar features attentively under the guidance of the background attention mask. The BSF module is composed of a stack of 3 identical blocks. As illustrated in Fig. 4, a block consists of four parts: a Self-Attention module, a Radar-to-Image Fusion module, an Image-to-Radar Fusion module, and a convolution-based fusion operation.

1) Radar-to-Image Fusion: Radar-to-Image (R2I) Fusion provides the spatial background information of radar features to image features. The positional embedding operation is applied based on their corresponding BEV spatial coordinates before fusion. To enhance the intrinsic representation capability of camera background features, given a C dimensional camera BEV feature map $F_C \in R^{C \times H \times W}$ as queries Q^I, we first perform the deformable self-attention (DSA) for each query $Q_p^I \in Q^I$ as follow:

$$DSA(Q_p^I) = DefAttn(Q_p^I, p, V^I), \quad (3)$$

where Q_p^I represents the camera BEV query at point $p = (x_p, y_p)$, and $V^I \in Q^I$ is the features with background attention mask $M(x, y) = 1$. Next the radar BEV feature $F_R' \in F_R$ with $M(x, y) = 1$ is used as key and value to
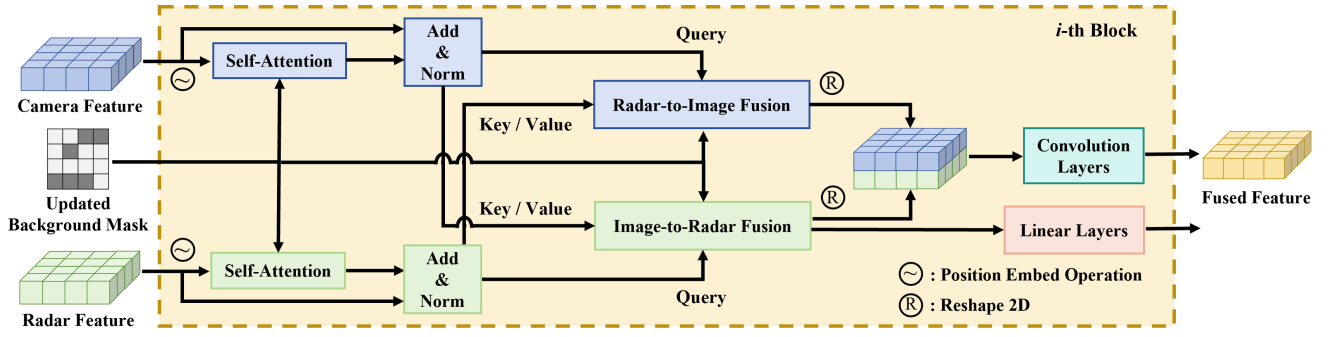
Fig. 4: An illustration of the Bidirectional Spatial Fusion block. Take camera BEV feature, radar BEV feature, and background attention mask as input, a Self-Attention module is first applied to these two features respectively. Then a Radar-to-Image Fusion and an Image-to-Radar Fusion are used for bidirectional spatial interaction. Finally, a convolution-based fusion operation is performed. A linear layer is used to generate the input of radar feature for the next block.

perform the Radar-to-Image deformable cross-attention as follow:

$$R2I(Q_p^I, F_R') = \sum_{V \in F_R'} DefAttn(Q_p^I, p, F_R'). \quad (4)$$

2) Image-to-Radar Fusion: Image-to-Radar (I2R) Fusion module utilizes rich contextual background information from image features to complement sparse radar features. Similarly, a radar BEV feature $F_R \in R^{C \times H \times W}$ is used as query Q_p^R to perform deformable self-attention as follow:

$$DSA(Q_p^R) = DefAttn(Q_p^R, p, V^R), \quad (5)$$

where Q_p^R represents the radar BEV query at point $p = (x_p, y_p)$, and $V^R \in Q^R$ is the features with $M(x, y) = 1$. Then the Image-to-Radar deformable cross-attention is performed as follow:

$$I2R(Q_p^R, F_C') = \sum_{V \in F_C'} DefAttn(Q_p^R, p, F_C'), \quad (6)$$

where $F_C' \in F_C$ is the camera BEV feature with $M(i, j) = 1$. Subsequently, the output of I2R is fed into a linear layer for the next $(i + 1)$-th block.

3) Convolution-based Fusion: Designed for 2D structures, convolution kernels are better at extracting local spatial correlations than 1D attention. Therefore, the outputs of R2I fusion and I2R fusion are transformed to image style again and concatenated along the channel dimension, then sent to the convolution block, expressed as follows:

$$F_{out}^i = H(R2I(Q^I, F_R') \oplus I2R(Q^R, F_C')), \quad (7)$$

where F_{out}^i is the output of the i-th block, and H represents the convolution-based fusion operation. In this way, multiple blocks increase the fitness of F_C and F_R, and the background BEV features can be enhanced gradually.

D. Global Descriptor Generator

Following vDiSCO [1], the fused background BEV features are fed into a Global Descriptor Generator to aggregate rotation-invariant global descriptors. Firstly, the polar transformation is applied to transform BEV features into the polar coordinate system, and then DFT is performed on the polar BEV features to achieve rotation invariance. Specifically, the rotation invariance is realized by the translation invariant property of the magnitude spectrum on polar BEV, where the translation indicates the rotation in the original BEV.

IV. EXPERIMENTS

A. NuScenes Dataset

NuScenes [22] is the first dataset for large-scale environments with multi-modal sensors, including LiDAR, camera, and radar. There are six camera sensors installed in front, front left, front right, back, back left, and back right parts of the vehicle, and five radar sensors installed at the front, left, right, and back, covering a 360° FOV. Following AutoPlace [7], we use the largest split, *Boston* split to train and evaluate our CRPlace. This split is divided into *database set*, *training query set*, *validation query set*, and *test query set* containing 6312, 7075, 924, and 3696 multi-view images and radar point clouds, respectively. See AutoPlace [7] for more details about the dataset.

B. Implementation Details

For multi-view images, we set the image size to 256×704, and the recognition range of the BEV grid is $(-51.2, 51.2)m$ for the X, Y axis, and $(-10, 10)m$ for Z axis. To densify the radar point cloud, we follow typical data pre-processing of [39] to concatenate the nearest six radar point clouds using ground truth ego-motion. The radar voxel size is set to $(0.8, 0.8, 8)m$. In order to convert camera features and radar features into a unified BEV space, we transfer the 2D position and velocity of radar points from the radar coordinates to the LiDAR coordinates.

For the network training, we follow the common practice [1], [7] to adopt metric learning with triplet margin loss. Multi-view images and a corresponding radar point cloud form a mini-batch. Each batch consists of several mini-batches that can be divided into a query, positive and negative samples. Following the scale of AutoPlace [7], we regard places in the database that are within the radius=9 m area to the query as positive samples, while those are outside the

radius=18 m area as negative samples. The loss term is given as:

$$L = \sum_k \max\{\|f(q), f(p)\|_2 - \left\|f(q), f(n^k)\right\|_2 + m, 0\}, \quad (8)$$

where $f(\cdot)$ denotes the network mapping a mini-batch to a feature vector, $\|\cdot\|_2$ means Euclidean distance, q is the query sample, p is the best positive matching sample, n^k is the negative sample, $m = 0.1$ is the predefined margin, and $k = 10$ is the number of negative samples. We use a batch size of 4 and SGD with an initial learning rate of 0.001, momentum of 0.9, and weight decay of 0.001. We decay the learning rate by 0.5 every 5 epochs.

C. Evaluation Metrics

We use *Recall@N* [2], precision-recall curve [40], $max\ F_1$ [19] and average precision (AP) [40] to evaluate the performance of different place recognition methods. *Recall@N* measures the percentage of successfully localized queries using the top N candidates retrieved from the database. Localization is successful if one of the top N retrieved candidates is within d meters of the ground truth. In our experiments, d is set to $9m$.

D. Comparison with State-of-The-Art Methods

We compare our method with both single-modality (camera or radar) and multi-modality (camera-radar) place recognition methods, including:

- Camera-based methods: NetVLAD [2], vDiSCO [1], and BEVFusion [21] (camera-only). The input of NetVLAD is the front-view images, and the others are multi-view images.
- Radar-based methods: AutoPlace [7], and BEVFusion [21] (radar-only).
- Camera-Radar fusion-based method: BEVFusion [21].

We adapt the implementation of the above works to the settings of the nuScenes dataset.

Table I shows the comparison results of place recognition methods on the nuScenes dataset. For visual place recognition, the multi-view methods with BEV representation vDiSCO and BEVFusion outperform the single-view method NetVLAD, achieving 86.0% and 85.5% recall@1, respectively. This is because BEV representation can provide rotation-invariant global descriptors. For radar-based place recognition, they both underperform compared to visual methods because images have stronger representational capabilities. AutoPlace is the SOTA radar-based method as it converts radar point clouds into images. Instead, BEVFusion (R) has inferior performance because it directly extracts features from radar point clouds. Notably, BEVFusion (C+R) outperforms those based on a single modality in all metrics, which indicates the effectiveness of the fusion approach. Our method further improves recall@1, $max\ F_1$, and AP to 91.2%, 0.96, and 0.98, respectively. We can also observe a similar trend from Fig. 5 that camera-radar fusion-based place recognition methods are outperformed by single

TABLE I: Comparison results of place recognition methods on the nuScenes dataset. C denotes camera, and R denotes radar.

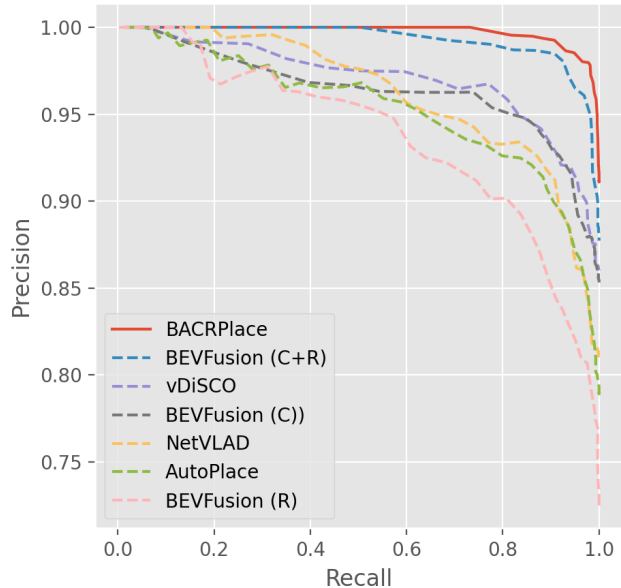| Method	Modality	Recall@1/5/10	$max\ F_1$	AP
NetVLAD [2]	C	80.8/86.2/87.6	0.91	0.95
vDiSCO [1]	C	86.0/88.6/89.2	0.95	0.96
BEVFusion [21]	C	85.5/87.9/88.7	0.94	0.96
AutoPlace [7]	R	77.8/82.3/83.7	0.94	0.97
BEVFusion [21]	R	72.5/79.0/80.8	0.89	0.92
BEVFusion [21]	C+R	88.0/89.9/90.6	0.95	0.96
CRPlace (ours)	C+R	**91.2/92.6/93.3**	**0.96**	**0.98**

Fig. 5: Precision-recall curve of SOTA methods on the nuScenes dataset.

modality-based methods. Still, CRPlace exceeds the others by a significant margin (from 3.6% to 12.9% relative increase of recall@1).

We also provide qualitative analysis in Fig. 6. As we can see, when a query is surrounded by many dynamic objects (first row) and also in the rain conditions (second row), BEVFusion will retrieve a false positive, while CRPlace can still retrieve the correct match.

E. Ablation Study

To understand how each module in CRPlace affects the place recognition performance, we conduct ablation studies by evaluating different groups shown in Table II.

Method (a) is our camera-radar fusion-based baseline BEVFusion, which achieves recall@1 of 88.0%, recall@5 of 89.9%, and recall@10 of 90.6%.

Method (b) extends *(a)* by simply adding the BAMG module. The background mask is directly added to the camera and radar BEV features, and then the original feature fusion module in BEVFusion [21] is performed. However, this method does not yield a significant improvement in

TABLE II: Ablation Study of CRPlace.

Method	BAMG	BSF	Recall@1/5/10	$max\ F_1$	AP
(a)			88.0/89.9/90.6	0.95	0.96
(b)	✓		88.2/89.9/90.5	0.94	0.96
(c)		✓	90.4/91.3/92.2	0.96	0.97
(d)	✓	✓	**91.2/92.6/93.3**	**0.96**	**0.98**

TABLE III: Comparative study of feature aggregation methods on CRPlace.

Aggregation	Recall@1	$max\ F_1$	AP
NetVLAD [2]	87.4	0.95	0.96
GeM [4]	89.7	0.96	0.96
DFT [1]	**91.2**	**0.96**	**0.98**

performance. We believe this is because the original feature fusion module does not effectively leverage the background mask to learn background features.

Method (c) extends *(a)* by replacing the original feature fusion module with our BSF module. Even without explicit guidance from a background mask, this method still improves recall@1 by 2.4%. This shows that our BSF module has powerful feature fusion capabilities.

Method (d) is our CRPlace. By combining the BAMG module and the BSF module, it achieves a gain of 3.2% for recall@1 compared to BEVFusion. This indicates that our BSF module can effectively utilize the background mask to fuse background features attentively.

We also investigate the impact of different feature aggregation methods on CRPlace, including NetVLAD [2], GeM [4], and DFT [1], which are used for generating global descriptors. As shown in Table III, unsurprisingly, DFT outperforms other methods as it generates rotation-invariant global descriptors.

TABLE IV: Comparison results of place recognition methods in rain conditions.

Method	Modality	Recall@1/5/10	$max\ F_1$	AP
NetVLAD [2]	C	65.8/75.4/77.1	0.86	0.93
vDiSCO [1]	C	70.2/76.6/80.4	0.89	0.94
BEVFusion [21]	C	69.9/76.4/80.1	0.88	0.94
AutoPlace [7]	R	75.7/82.0/83.5	0.94	0.96
BEVFusion [21]	R	71.6/77.4/79.7	0.88	0.93
BEVFusion [21]	C+R	83.8/87.3/87.9	0.93	0.96
CRPlace (ours)	C+R	**85.6/88.6/89.1**	**0.95**	**0.96**

F. Comparison in Rain Conditions

To verify the robustness of our method under varying environmental conditions, we filter the rain-affected samples from the *Boston* split in nuScenes dataset as the validation set for additional evaluation. As indicated in Table IV, visual place recognition methods demonstrate considerable performance deterioration in these conditions. For instance, the recall@1 of NetVLAD drops to only 65%, while vDiSCO and BEVFusion (C) show similar diminished effectiveness,
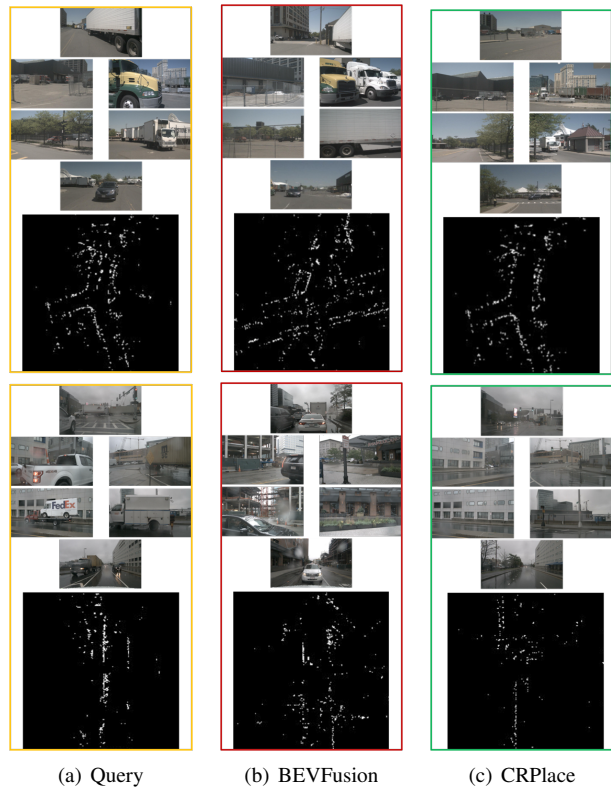

(a) Query   (b) BEVFusion   (c) CRPlace

Fig. 6: Qualitative comparison between BEVFusion and our CRPlace. (a) are queries influenced by (1) dynamic objects (first row) and (2) dynamic objects + rain conditions (second row). (b) is an incorrect retrieval using BEVFusion, and (c) is a correct retrieval using CRPlace.

each achieving around a 70% recall@1 rate. In contrast, radar-based methods and camera-radar fusion-based methods maintain relatively strong performance. Notably, the efficacy of radar-based approaches has now surpassed that of purely visual methods in rainy conditions. Our methodology stands out in this challenging environment, showcasing an improvement in recall@1 ranging between 2.1% to 30.1% compared to other methods. This demonstrates its considerable potential for reliable performance in rain-impacted scenarios.

V. CONCLUSION

In this paper, we introduce CRPlace, a background-attentive bidirectional fusion method that fuses the complementary camera and radar data for improving place recognition. Unlike existing camera-radar fusion schemes that focus on dynamic features in 3D object detection, we leverage the dynamic properties of radar points to adaptively discern which features belong to the stationary background. Subsequently, a bidirectional cross-attention mechanism is employed to interactively fuse background features from both the camera and radar. With our background-attentive bidirectional fusion method, CRPlace outperforms earlier schemes for place recognition on nuScenes dataset.

REFERENCES

[1] X. Xu, Y. Jiao, S. Lu, X. Ding, R. Xiong, and Y. Wang, "Leveraging bev representation for 360-degree visual place recognition," *arXiv preprint arXiv:2305.13814*, 2023.

[2] 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*, 2016, pp. 5297–5307.

[3] Y. Pan, X. Xu, W. Li, Y. Cui, Y. Wang, and R. Xiong, "Coral: Colored structural representation for bi-modal place recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2084–2091.

[4] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[5] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4878–4888.

[6] A. Khaliq, M. Milford, and S. Garg, "Multires-netvlad: Augmenting place recognition training with low-resolution imagery," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3882–3889, 2022.

[7] K. Cai, B. Wang, and C. X. Lu, "Autoplace: Robust place recognition with single-chip automotive radar," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2222–2228.

[8] 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*, 2018, pp. 4470–4479.

[9] X. Chen, T. Läbe, A. Milioto, T. Röhling, O. Vysotska, A. Haag, J. Behley, and C. Stachniss, "Overlapnet: Loop closing for lidar-based slam," *arXiv preprint arXiv:2105.11344*, 2021.

[10] 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.

[11] K. Vidanapathirana, P. Moghadam, B. Harwood, M. Zhao, S. Sridharan, and C. Fookes, "Locus: Lidar-based place recognition using spatiotemporal higher-order pooling," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 5075–5081.

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

[13] H. Lai, P. Yin, and S. Scherer, "Adafusion: Visual-lidar fusion with adaptive weights for place recognition," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 038–12 045, 2022.

[14] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "Minkloc++: lidar and monocular image fusion for place recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–8.

[15] A. J. Lee, S. Song, H. Lim, W. Lee, and H. Myung, "$(lc)^2$: Lidar-camera loop constraints for cross-modal place recognition," *IEEE Robotics and Automation Letters*, 2023.

[16] Y. Wang, J. Deng, Y. Li, J. Hu, C. Liu, Y. Zhang, J. Ji, W. Ouyang, and Y. Zhang, "Bi-lrfusion: Bi-directional lidar-radar fusion for 3d dynamic object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 394–13 403.

[17] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1160–1168.

[18] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, "Crn: Camera radar net for accurate, robust, efficient 3d perception," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 615–17 626.

[19] Ş. Săftescu, M. Gadd, D. De Martini, D. Barnes, and P. Newman, "Kidnapped radar: Topological radar localisation using rotationally-invariant metric learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4358–4364.

[20] D. Barnes and I. Posner, "Under the radar: Learning to predict robust keypoints for odometry estimation and metric localisation in radar," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 9484–9490.

[21] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[23] C. Valgren and A. J. Lilienthal, "Sift, surf and seasons: Long-term outdoor localization using local features," in *3rd European conference on mobile robots, ECMR'07, Freiburg, Germany, September 19-21, 2007*, 2007, pp. 253–258.

[24] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer, 2006, pp. 404–417.

[25] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2564–2571.

[26] R. Arandjelovic and A. Zisserman, "All about vlad," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2013, pp. 1578–1585.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[29] 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)*. IEEE, 2018, pp. 4802–4809.

[30] 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.

[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*, 2017, pp. 652–660.

[32] C. Fu, L. Li, L. Peng, Y. Ma, X. Zhao, and Y. Liu, "Overlapnetvlad: A coarse-to-fine framework for lidar-based place recognition," *arXiv preprint arXiv:2303.06881*, 2023.

[33] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[34] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[36] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[37] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[38] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 697–12 705.

[39] J.-T. Lin, D. Dai, and L. Van Gool, "Depth estimation from monocular images and sparse radar data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 233–10 240.

[40] Y. Hou, H. Zhang, and S. Zhou, "Evaluation of object proposals and convnet features for landmark-based visual place recognition," *Journal of Intelligent & Robotic Systems*, vol. 92, pp. 505–520, 2018.