2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Explicit Interaction for Fusion-Based Place Recognition

Jingyi Xu[1], Junyi Ma[1], Qi Wu[1], Zijie Zhou[2], Yue Wang[3], Xieyuanli Chen[4], Wenxian Yu[1], and Ling Pei[1*]

***Abstract*— Fusion-based place recognition is an emerging technique jointly utilizing multi-modal perception data, to recognize previously visited places in GPS-denied scenarios for robots and autonomous vehicles. Recent fusion-based place recognition methods combine multi-modal features in implicit manners. While achieving remarkable results, they do not explicitly consider what the individual modality affords in the fusion system. Therefore, the benefit of multi-modal feature fusion may not be fully explored. In this paper, we propose a novel fusion-based network, dubbed EINet, to achieve explicit interaction of the two modalities. EINet uses LiDAR ranges to supervise more robust vision features for long time spans, and simultaneously uses camera RGB data to improve the discrimination of LiDAR point clouds. In addition, we develop a new benchmark for the place recognition task based on the nuScenes dataset. To establish this benchmark for future research with comprehensive comparisons, we introduce both supervised and self-supervised training schemes alongside evaluation protocols. We conduct extensive experiments on the proposed benchmark, and the experimental results show that our EINet exhibits better recognition performance as well as solid generalization ability compared to the state-of-the-art fusion-based place recognition approaches. Our open-source code and benchmark are released at: https://github.com/BIT-XJY/EINet.**

I. Introduction

Place recognition is a trendy technique to provide prior locations for SLAM [1]–[3] and global localization [4]–[6] in autonomous navigation systems, especially in the GPS-denied scenes. Vision-based place recognition (VPR) [7]–[11] has been extensively studied due to the low-cost and simple use of camera sensors. However, camera sensors are vulnerable to environmental factors such as lighting and weather, which can negatively impact descriptor extraction in various environmental changes during place recognition tasks. In contrast, LiDAR-based place recognition (LPR) [12]–[16] overcomes sensitivity to environmental conditions to some degree but lacks the richness of appearance and texture information. As the combination of VPR and LPR, fusion-based place recognition (FPR) [17]–[21] integrates information from multiple modalities, images and point clouds specifically. They preserve the respective strengths of each modality and thus enhance recognition performance. However, the explicit interaction of different modalities remains undiscovered among the existing fusion frameworks that only operate implicit feature fusion, and a new explainable fusion strategy needs to be further explored for FPR tasks.

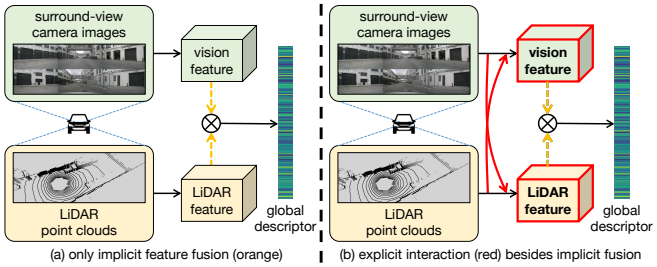

Fig. 1: Different from existing implicit fusion methods, our proposed method leverages the strengths of the LiDAR modality (robustness from range) and camera modality (richness from appearance) using explicit interaction for generating place descriptors.

The two main questions that need to be answered are: *what each modality actually affords positively for possible explicit interaction of place recognition*, and *how to achieve reasonable explicit interaction based on the advantages of the two modalities*. It is clear that these two modalities have distinct affordances: LiDAR provides geometric information that is robust over long periods, while cameras offer appearance and texture information that is more distinctive. The key bottleneck lies in the mutual interaction between LiDAR and camera features, and we advocate that an explicit and explainable fusion can further promote place representation compared to implicit fusion methods such as weights adapting [19] or end-to-end learning [21]. To this end, we utilize LiDAR ranges as sparse depth supervision for camera feature extraction, and meanwhile use camera-image-based appearance rendering for colored point cloud generation. Specifically, a novel explicit interaction network for fusion-based place recognition, dubbed EINet, is proposed to explicitly explore the strengths of both modalities. We integrate two types of interaction, sparse depth supervision and appearance rendering, into a single framework to enhance the descriptive capability of the final global descriptors, as shown in Fig. 1. In addition, we build a new benchmark, namely NUSC-PR, on the nuScenes dataset [22] for general place recognition tasks, to facilitate the advancements in this emerging domain. NUSC-PR takes into account both supervised and self-supervised paradigms, and also provides the standard evaluation protocol to compare baseline methods.

In sum, the main contributions of our work are threefold:

- We propose EINet, a novel fusion-based place recognition network with an explicit and explainable interaction

This work was supported in part by the National Nature Science Foundation of China (NSFC) under Grant No.62273229 and No.61873163 separately.

[1]Jingyi Xu, Junyi Ma, Qi Wu, Wenxian Yu, and Ling Pei are with the Shanghai Jiao Tong University.
[2]Zijie Zhou is with the Beijing Institute of Technology.
[3]Yue Wang is with the Zhejiang University.
[4]Xieyuanli Chen is with the National University of Defense Technology.
*Corresponding author: Ling Pei (ling.pei@sjtu.edu.cn)

mechanism between LiDAR and camera sensors, aiming to jointly leverage the superiority of multi-modal features for generating global descriptors.
- A new benchmark for FPR called NUSC-PR is proposed to provide two standard training schemes alongside evaluation metrics, which can be used as a foundation tool for future place recognition research.
- We conduct comprehensive experiments with detailed analysis to validate that our EINet outperforms the current state-of-the-art methods, including the FPR baselines using implicit feature fusion, and has strong generation ability across different locations.

II. Related Work

The rapid development of place recognition have been comprehensively documented in surveys over the years [23]–[25]. In this section, we review place recognition from the following three aspects: vision-based, LiDAR-based, and fusion-based methods. Compared to traditional place recognition paradigms (mostly depending on hand-crafted features [26]–[29]), deep learning-based approaches show promising ability in both recognition accuracy and running efficiency, which we mainly focus on in the three mentioned aspects.

Vision-based place recognition is commonly considered as an image retrieval problem, wherein the database image most similar to the current query image is retrieved. Some prior works [7], [30], [31] have laid the foundation for VPR. Especially for basic feature aggregation, NetVLAD by Arandjelovic et al. [7] provides a general skill, which is afterward utilized by most descriptor-based methods [8], [13], [32], [33]. For example, MultiRes-NetVLAD [32] aggregates a union of features across multiple resolutions using argumentated NetVLAD representation. Recently, more advanced works have been proposed. R2Former proposed by Zhu et al. [9] uses a novel transformer-based reranking strategy to further improve the recognition accuracy compared to the conventional VPR pipeline. EigenPlaces proposed by Berton et al. [10] extracts descriptors with viewpoint robustness by explicitly clustering the training data. Keetha et al. [8] propose AnyLoc, which utilizes pre-trained foundation models to generate descriptors in structured and unstructured environments. Similarly, DINOv2 SALAD by Izquierdo et al. [11] also integrates the foundation model into the place recognition framework, but replaces the commonly used NetVLAD [7] with a novel feature aggregation approach based on the optimal transport theory.

Compared to the VPR, LiDAR-based place recognition is robust to environmental changes, such as illumination, weather, and seasons. Uy et al. [12] propose PointNetVLAD by combining existing PointNet [34] with NetVLAD to solve the place retrieval problem. LPD-Net by Liu et al. [13] extracts local features, and aggregates them using a graph neural network, to generate global features. OT series including OverlapNet [14], OverlapTransformer [33], SeqOT [35], and CVTNet [15] utilize the transformer [36] to improve recognition performance while maintaining the yaw-angle invariance. BEVPlace by Luo et al. [37] is also designed to generate rotation-invariant descriptors exploiting bird's eye view (BEV), while simultaneously exploring the correlation of global feature distances and geometric distances. Cui et al. [38] build the bag of words based on the devised 3D LiDAR features for place retrieval. Considering texture enhancement in robotics, Xia et al. [39] newly develop a text-to-place retrieval method that first uses the semantic relationship between laser points and texture description for place recognition.

Recently, cross-modal place recognition [40]–[43] has aroused great interest, unifying the cross-modal data into the same modality. While producing remarkable results, directly fusing multi-modal inputs to recognize places rather than modality unifying for query and references still exhibits the most superior performance [21]. One of the early fusion-based approaches, PIC-Net proposed by Lu et al. [17] integrates the global channel attention to directly fuse the features of camera images and LiDAR point clouds. Pan et al. [44] propose CORAL-VLAD, a network fusing the structural features and vision features in the consistent BEV for place recognition. MinkLoc++ by Komorowski et al. [18] introduces a late fusion approach to process each modality separately and fuse them in the final part. AdaFusion by Lai et al. [19] utilizes multi-scale attention to learn the fusion weights between camera and LiDAR modalities in different environments. Compared to AdaFusion, MFF-PR proposed by Liu et al. [20] introduces more comprehensive features including semantic, instance, topological, and image texture features, to further improve the robustness and scene expression abilities for place recognition. More recently, Zhou et al. [21] introduce a novel multi-scale attention-based fusion framework LCPR, to generate discriminative and approximately yaw-rotation invariant global descriptors for place recognition.

To pursue good place recognition performance, our method also adopts the fusion-based stream to generate place descriptors. Different from the above-mentioned FPR methods which advocate implicit integration of multi-modal features, our method instead achieves explicit interaction based on the explainable advantage complementary of multi-modal sensor data.

III. Method

In this section, we first present the overall system of our proposed place recognition approach. Then, we investigate the explicit form of LiDAR and camera interaction in the fusion system. Lastly, we introduce the multiple loss functions used to train EINet.

A. EINet Architecture

The overall architecture of our proposed EINet is illustrated in Fig. 2. EINet uses surround-view camera images and LiDAR point clouds as joint inputs to generate final global descriptors for place retrieval. The camera branch first transforms the multiple input images to pseudo depth maps based on an encoder-decoder structure, under the

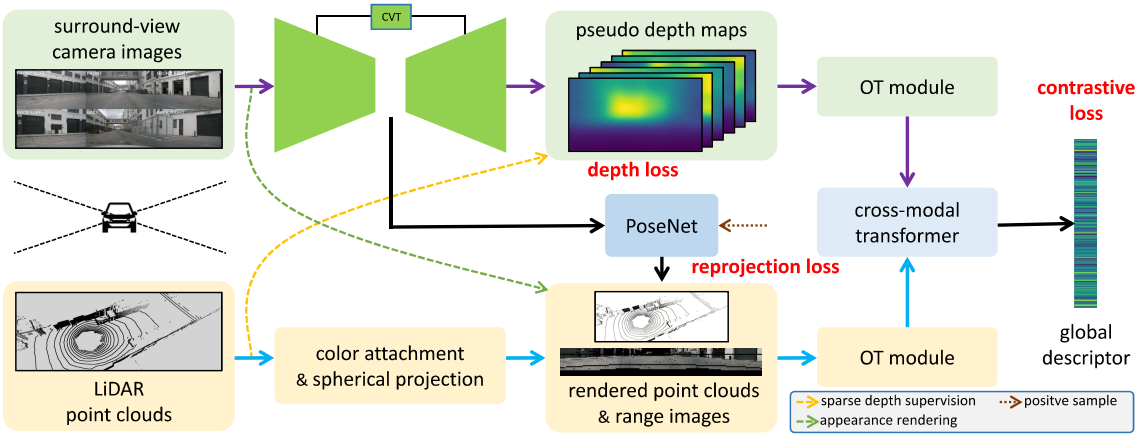

Fig. 2: Pipeline overview of our proposed EINet. The purple and blue arrows represent the camera branches and the LiDAR branches respectively. The camera sensors provide RGB information for rendered point clouds and range images in the LiDAR branch (dotted green arrow), while the LiDAR sensor helps to supervise pseudo depth in the camera branch (dotted orange arrow), achieving explicit interaction in the fusion-based place recognition framework.

sparse depth supervision from LiDAR point clouds. Following SurroundDepth [45], we add cross-view transformer (CVT) blocks between the encoder and decoder, to enhance the correlation between surround-view images. The multiple pseudo depth maps are then transformed into a holistic range image in the LiDAR coordinate system for the following compression and feature fusion. The LiDAR branch first combines the RGB information provided by cameras and spherical projection to produce rendered point clouds as well as range images. The rendered range images from LiDAR and the counterpart transformed from pseudo depth are both compressed by OT modules (OverlapTransformer [33] without Global Descriptor Generator) in the respective branch. Then the two sentence-like features are absorbed by the cross-modal transformer, which contains stacked cross transformer blocks and NetVLAD with MLP [15], to produce the descriptor vector for place recognition. Notably, we further utilize an additional PoseNet to explicitly capture the relative pose between the query and the positive samples. More details can be found in Sec. III-C.

B. Explicit Interaction

Explicit interaction lies in the complementary advantages of the two modalities. LiDAR offers more stable geometric features that can be used to improve the robustness of camera features in long-time gaps. Cameras offer rich appearance and texture information, which can conversely be used to enhance the distinctiveness of point cloud features. Based on the above analysis, we first specify what LiDAR and camera explicitly afford to each other in the cross-modal interaction of our fusion-based place recognition framework.

Sparse depth supervision. We first utilize ranges captured by LiDAR as sparse depth supervision ($\mathcal{I} \oplus \mathcal{P} \rightarrow \mathcal{D}$) for the camera feature extraction. As shown in Fig. 2, LiDAR explicitly helps to guide the image encoder and decoder in the camera branch to output pseudo depth maps $\mathcal{D}$ from the input image $\mathcal{I}$. We achieve this by projecting raw LiDAR point clouds $\mathcal{P}$ to the camera image planes to supervise depth generation with depth loss sparsely (see Sec. III-C). We release the dense constraints proposed by the previous depth estimation works [45], [46] such as photometric reprojection loss, and thus only generate pseudo depth maps instead of common depth maps by strict definition (see Fig. 3). This helps to keep mixed-state features in the camera branch that are geometry-aware as well as appearance-aware to balance robustness and distinctiveness. Notably, the predicted depth maps are scale-aware since we use real distance observed by LiDAR for supervision.

Appearance rendering. Conversely, we propose appearance rendering ($\mathcal{P} \oplus \mathcal{I} \rightarrow \mathcal{P}_{color}$) as what cameras explicitly provide to LiDAR feature generation. Although geometric features directly extracted from point clouds are less affected by environmental disturbance, they are naturally sparse and lack observation distinctiveness. Therefore, we use camera images to attach RGB channels to point clouds $\mathcal{P}$, leading to colored point clouds $\mathcal{P}_{color}$, which are then transformed to rendered range images by spherical projection (see Fig. 3). Color attachment for point clouds can be implemented by the similar operation in sparse depth supervision. We only need to project raw laser points to the camera image planes and sample RGB values respectively. The projection function is the same for different branches, which helps save computing resources in our proposed framework.

The above-mentioned interaction including sparse depth supervision and appearance rendering are the foundation to explicitly complement each other's advantages of the two modalities. We further jointly use them in the unified EINet to achieve explainable fusion-based place recognition. The image features and the point cloud features are tightly coupled by the interaction and then fused by the cross-modal transformer, leading to global descriptors suitable for place retrieval with different time gaps. In the proposed framework, we do not directly fuse vanilla camera and LiDAR features compared to the previous works [19], [21], but let the LiDAR branch assist the camera branch in generating geometry-aware features which are then fused back to the LiDAR features. Similarly, we also let the camera branch assist the LiDAR branch in producing appearance-aware features,

which are then fused with the features in the camera branch. The explicit interaction therefore helps the following transformer to extract correlations and find consistency across modalities. The effectiveness and validity of our proposed explicit interaction are further guaranteed by the setups of training losses, which will be introduced in Sec. III-C.

C. Loss Functions

We use three types of losses to train EINet, harnessing the positive effects of explicit interaction.

Depth loss. For the input point cloud $\mathcal{P}$, a laser point $p_i \in \mathcal{P}$ $(i \in [1, N_{\text{pc}}])$ is first projected to the camera coordinate system by extrinsic parameters, leading to $p_i^{\text{c}} = (x_i^{\text{c}}, y_i^{\text{c}}, z_i^{\text{c}})$ with depth $d_i = z_i^{\text{c}}$. Then p_i^{c} is transformed to the pixel (u_i, v_i) in the corresponding image $\mathcal{I}_j$ $(j \in [1, N_{\text{cam}}])$ by the camera intrinsic parameters. Then we calculate the depth loss using $L1$ distance by

$$\mathcal{L}_{\text{d}} = \sum_{i,j} \mathbb{I}(|d_i - \mathcal{D}_j(u_i, v_i)|), \tag{1}$$

where $\mathcal{D}_j$ is the pseudo depth map of the jth camera. $\mathbb{I}(a) = a$ if $(u_i, v_i) \in \mathcal{I}_j$ and $\mathbb{I}(\cdot) = 0$ otherwise. The depth loss concretely represents how LiDAR ranges function on the camera branch in our proposed EINet.

Contrastive loss. We choose commonly used contrastive loss [13], [15], [33] to supervise global descriptor generation. For each training tuple, we utilize one query descriptor $\mathcal{G}_{\text{q}}$, n_{pos} positive descriptors $\mathcal{G}_{\text{p}}$, and n_{neg} negative descriptors $\mathcal{G}_{\text{n}}$ to calculate lazy triplet loss by

$$\mathcal{L}_{\text{t}} = n_{\text{pos}}(\alpha + \max(\text{dis}(\mathcal{G}_{\text{q}}, \mathcal{G}_{\text{p}}))) - \sum_{n_{\text{neg}}} \text{dis}(\mathcal{G}_{\text{q}}, \mathcal{G}_{\text{n}}), \tag{2}$$

where α is the margin and $\text{dis}(\cdot)$ is squared Euclidean distance to measure similarity of descriptors. Using the triplet loss we can narrow the gap between feature representations for close-place pairs and enlarge the counterparts for distant-place pairs. In Sec. IV, we will introduce different approaches to select positive and negative samples for the query one in our proposed benchmark.

Reprojection loss. We notice that most existing place recognition methods only use the straightforward way, i.e., the abovementioned contrastive loss, to coarsely capture the relationship of similar places. To make the network explicitly aware of the quantized relation between the query and positive samples, we propose an additional reprojection loss based on the devised PoseNet (as shown in Fig. 2). The PoseNet uses the encoded image features of the query sample and a randomly selected positive sample to predict relative poses of the robot T_{e} between the two frames. Then T_{e} is transformed to the relative LiDAR pose T_{L} by extrinsic parameters to compute the reprojection loss by

$$\mathcal{L}_{\text{r}} = |\mathbb{S}(T_{\text{L}}\mathcal{P}_{\text{p}}) - \mathbb{S}(\mathcal{P}_{\text{q}})|, \tag{3}$$

where $\mathbb{S}$ represents spherical projection for point clouds to generate range images. We reduce the difference between the query range image and the positive range image reprojected to the query frame to let the network receive supervision signals related to pose quantification. Note that we substantially use the reprojection loss to optimize the image backbone for point cloud alignment, further enhancing the correlation between the two modalities.

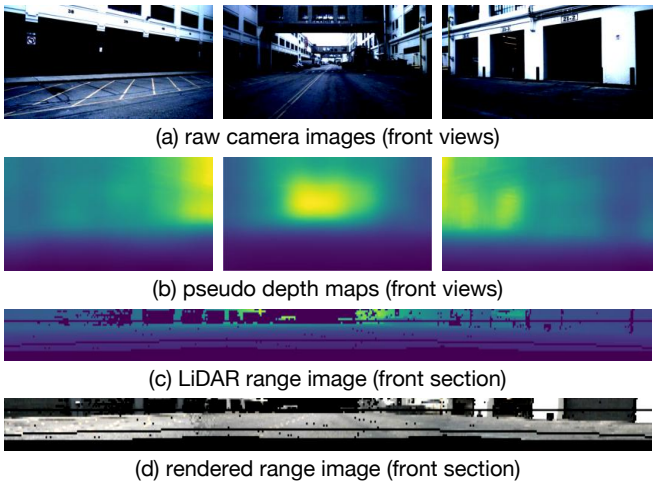

Fig. 3: An example of the raw camera images, pseudo depth maps, raw LiDAR range image, and rendered range image in explicit interaction.

The total loss used to train the EINet is the weighted summation of the above three losses, which is computed by

$$\mathcal{L}_{\text{all}} = \lambda_{\text{d}}\mathcal{L}_{\text{d}} + \lambda_{\text{t}}\mathcal{L}_{\text{t}} + \lambda_{\text{r}}\mathcal{L}_{\text{r}}, \tag{4}$$

where λ_{d}, λ_{t}, and λ_{r} are the weighting parameters.

IV. The NUSC-PR Benchmark

Although place recognition approaches have been widely proposed in recent years, a unified benchmark is lacking to standardize both the supervised and self-supervised learning schemes and evaluation protocols. To this end, we further build a new benchmark, NUSC-PR, based on the publicly available nuScenes dataset [22], with the standardized dataset format and evaluation protocol.

Data organization. We first re-organize the original nuScenes dataset to a new format, to build the training set and the test set. For the test set, the commonly used approach of differentiating samples based on the Euclidean distance between dates [21], [47] is employed to distinguish positive and negative samples. Here we mainly focus on proposing different schemes for building the training sets. Regarding learning-based place recognition methods, the supervised learning scheme is widely exploited based on manually sampling positive and negative places with distance metrics offline [21], [47]. It needs ground-truth positions (always provided by GPS-based positioning systems) to quantify the similarity between the query and reference samples. Our proposed NUSC-PR benchmark continues adopting this supervised learning scheme, for which the data organization is shown in Alg. 1. In addition to the commonly used supervised manner with distance metrics, we also notice that almost all learning-based place recognition models can be trained online using a self-supervised manner, where the positive and negative samples are split by time metrics

Algorithm 1 Data generation pipeline for the **supervised** learning scheme in the NUSC-PR benchmark.

Input: Driving scenes $\{\mathcal{S}\}_{i=0}^{N_s}$, data samples $\{s\}_{j=0}^{N_{\mathcal{S}_i}} \in \mathcal{S}_i$, distance interval δ, date threshold γ, positive distance threshold ρ_{pos}, negative distance threshold ρ_{neg}, the numbers of positive and negative samples n_{pos} and n_{neg} for each query;
Output: The training set $\mathcal{Q}$ and test set $\mathcal{K}$ for supervised learning;

Initialize the database set $\mathcal{A} = \phi$, the training query set $\mathcal{B} = \phi$, the test query set $\mathcal{C} = \phi$.
for each scene $\mathcal{S}_i \in \{\mathcal{S}\}_{i=0}^{N_s}$ **do**
  for each data sample $s_j \in \{s\}_{j=0}^{N_{\mathcal{S}_i}}$ **do**
    $\mu = \min(s_j, \mathcal{A})$;
    if $\mu \geq \delta$ or $\mathcal{A} = \phi$ **then**
      Append s_j to the database set $\mathcal{A}$;
    else if the date of s_j is before γ **then**
      Append s_j to the training query set $\mathcal{B}$;
    else
      Append s_j to the test query set $\mathcal{C}$;
Initialize the training set $\mathcal{Q} = \phi$;
for each train query sample $a_i \in \mathcal{A}$ **do**
  Initialize the positive set $\mathcal{V}_i^{\text{pos}} = \phi$ and the negative set $\mathcal{V}_i^{\text{neg}} = \phi$ for the current query;
  Use K-Nearest Neighbours (KNN) to find the potential positive samples with ρ_{pos}, which are randomly selected to $\mathcal{V}_i^{\text{pos}}$ with the number of n_{pos};
  Find the potential negative samples within the complementary set of KNN results using ρ_{neg}, which are randomly selected to $\mathcal{V}_i^{\text{neg}}$ with the number of n_{neg};
  Append $\{i : [\mathcal{V}_i^{\text{pos}}, \mathcal{V}_i^{\text{neg}}]\}$ to the training set $\mathcal{Q}$;
Initialize the test set $\mathcal{K} = \phi$;
for each test query sample $b_i \in \mathcal{B}$ **do**
  Initialize the ground-truth set $\mathcal{V}_i^{\text{gt}} = \phi$;
  Use KNN to find the ground-truth samples $\mathcal{V}_i^{\text{gt}}$ with ρ_{pos};
  Append $\{i : [\mathcal{V}_i^{\text{gt}}]\}$ to the test set $\mathcal{K}$;
Optional: randomly split validation set out of test set with a preset proportion.
return $\mathcal{Q}$ and $\mathcal{K}$.

Algorithm 2 Data generation pipeline for the **self-supervised** learning scheme in the NUSC-PR benchmark.

Input: Driving scenes $\{\mathcal{S}\}_{i=0}^{N_s}$, data samples $\{s\}_{j=0}^{N_{\mathcal{S}_i}} \in \mathcal{S}_i$, date threshold γ, positive distance threshold ρ_{pos}, negative time threshold σ_{neg}, the numbers of positive and negative samples n_{pos} and n_{neg} for each query;
Output: The training set $\mathcal{Q}$ and test set $\mathcal{K}$ for self-supervised learning;

Initialize the database set $\mathcal{A} = \phi$, the training query set $\mathcal{B} = \phi$, the test query set $\mathcal{C} = \phi$, the old scene set $\mathcal{E}_{\text{old}} = \phi$, the new scene set $\mathcal{E}_{\text{new}} = \phi$.
for each scene $\mathcal{S}_i \in \{\mathcal{S}\}_{i=0}^{N_s}$ **do**
  if the date of $\mathcal{S}_i$ is before γ **then**
    Append $\mathcal{S}_i$ to the old scene set $\mathcal{E}_{\text{old}}$;
  else
    Append $\mathcal{S}_i$ to the new scene set $\mathcal{E}_{\text{new}}$;
Initialize the whole old sample set $\mathcal{V}_*^{\text{old}} = \phi$;
for each scene $\mathcal{S}_i \in \mathcal{E}_{\text{old}}$ **do**
  Initialize the potential negative set $\mathcal{V}_*^{\text{neg}} = \phi$;
  for each data sample $s_j \in \{s\}_{j=0}^{N_{\mathcal{S}_i}}$ **do**
    Append s_j to $\mathcal{V}_*^{\text{old}}$;
    if $j \geq \sigma_{\text{neg}} + n_{\text{pos}} + n_{\text{neg}}$ **then**
      Initialize the positive set $\mathcal{V}_j^{\text{pos}} = \phi$ and the negative set $\mathcal{V}_j^{\text{neg}} = \phi$ for the current query;
      Select $\{s\}_{k=j-n_{\text{pos}}}^{j-1}$ as $\mathcal{V}_j^{\text{pos}}$;
      Randomly select n_{neg} samples in $\mathcal{V}_*^{\text{neg}}$ as $\mathcal{V}_j^{\text{neg}}$;
      Append $s_{j-\sigma_{\text{neg}}}$ to $\mathcal{V}_*^{\text{neg}}$;
      Append $\{j : [\mathcal{V}_j^{\text{pos}}, \mathcal{V}_j^{\text{neg}}]\}$ to the training set $\mathcal{Q}$;
    else
      Append s_j to $\mathcal{V}_*^{\text{neg}}$;
Initialize the test set $\mathcal{K} = \phi$;
for each scene $\mathcal{S}_i \in \mathcal{E}_{\text{new}}$ **do**
  for each data sample $s_j \in \{s\}_{j=0}^{N_{\mathcal{S}_i}}$ **do**
    Initialize the ground-truth set $\mathcal{V}_j^{\text{gt}} = \phi$;
    Use KNN to find the ground-truth samples $\mathcal{V}_j^{\text{gt}}$ with ρ_{pos};
    Append $\{j : [\mathcal{V}_j^{\text{gt}}]\}$ to the test set $\mathcal{K}$;
Optional: randomly split validation set out of test set with a preset proportion.
return $\mathcal{Q}$ and $\mathcal{K}$.

automatically. In this case, the collection time determines whether a sample is positive or negative, and the time period strictly corresponds to the number of samples since the collection frequency is fixed over the dataset. If the time interval between the query sample and the reference counterpart is less than the preset threshold, the reference sample is regarded as positive for its query place. Therefore, the training process can be implemented online even when the autonomous vehicle is driving in a GPS-denied environment since this paradigm does not need high-quality ground-truth positions. However, we still need the ground-truth positions of all the samples to conduct the test set for offline evaluation. The overall data organization for the self-supervised learning scheme is shown in Alg. 2. To summarize, NUSC-PR contains two types of data organization for supervised and self-supervised learning schemes. In Sec. V, we conduct comprehensive experiments based on the datasets built by both Alg. 1 and Alg. 2.

Evaluation metrics. The average recall rate is a widely used metric to represent the proportion of successful retrieval in place recognition [9], [11], [15]. In the NUSC-PR benchmark, we also use the average recall rates (e.g., AR@1, AR@5, AR@10, and AR@20) to evaluate different methods on the test set, which are calculated by

$$\text{AR@}x = \frac{n_{\text{suc}}}{N_{\text{query}}} \times 100\% \tag{5}$$

where N_{query} is the number of test query samples, and n_{suc} is the number of successful recall. Only when there is at least one correct reference sample in the top x retrievals, one successful recall is collected for the numerator of AR@x. We use different test queries and databases to evaluate the supervised learning and self-supervised learning schemes generated by Alg. 1 and Alg. 2, leading to respective evaluation protocols.

V. Experiments

The experiments are conducted to validate the claims that our approach is able to: (i) accurately retrieve previously visited places with large time gaps in driving environments, (ii) exploit explicit interaction of LiDAR and camera modalities to improve the recognition accuracy, (iii) generalize

well into the environments of different locations without fine-tuning, and (iv) be implemented online with efficient inference operation.

A. Experimental Setups

Benchmark details. We first provide the detailed configurations of Alg. 1 and Alg. 2 in our proposed NUSC-PR benchmark. We build datasets for the individual locations of different cities, including Boston-Seaport (BS), Singapore-Onenorth (SON), Singapore-Queenstown (SQ), and Singapore-Hollandvillagee (SHV) respectively. We set the date threshold γ as 105 days following [21], [47], the positive distance threshold ρ_{pos} as 9 m, the numbers of positive and negative samples n_{pos} and n_{neg} as 2 and 4 respectively. We let the distance interval $\delta = 1$ m and the negative distance threshold $\rho_{\text{neg}} = 18$ m in the supervised learning scheme, the negative time threshold $\sigma_{\text{neg}} = 6$ (3 s) in the self-supervised learning scheme. The detailed configurations of the data organization in the NUSC-PR for different locations are also posted in our open-source code.

Baseline setups. We select the existing camera-based methods including NetVLAD [7] and MultiRes-NetVLAD (mrNVLAD) [32], LiDAR-based methods including OverlapTransformer (OT) [33] and CVTNet [15], and fusion-based methods including AdaFusion [19] and LCPR [21] as baselines. We use the default architecture and training parameters of the baseline approaches suggested in their original papers and open sources. We replace the OverlapNetLeg in OT [33] with the point cloud encoder proposed in [15] to adapt 32-beam LiDAR in the nuScenes dataset.

EINet setups. In the camera branch of EINet, we use the resized images with a resolution of 640×352. Following SurroundDepth [45], we set 4 scales in the image encoder and decoder. We use 4 CVT blocks for each scale between the encoder and decoder, instead of 8 blocks to improve inference efficiency. The pseudo depth maps have the same resolution as the input images. The range image from the multiple pseudo depth maps is 352×1056, fed to the following devised OT module. In the LiDAR branch, we project rendered point cloud to the range image with the size of 32×1056, fed to the following OT module with a different structure from the above-mentioned counterpart due to specific input sizes. The cross-model transformer module has 2 stacked cross transformers followed by NetVLAD and one MLP. The global descriptor generated by EINet is a 1×256 vector. The PoseNet for auxiliary training has 3 convolution layers that receive the 6×512×11×20 encoded image features of both a query and a randomly selected positive sample, to produce the 6-DOF relative pose between them. Notably, we discard the PoseNet in the test process since it does not contribute to the final descriptor. The loss weights λ_{d}, λ_{t}, and λ_{r} are set to 0.01, 1.00, and 0.01 respectively. The margin α in $\mathcal{L}_{\text{t}}$ is set to 0.5.

All baselines and our proposed methods are trained with a batch size of 7 on 2 NVIDIA A100 GPUs for 20 epochs. We use the ADAM optimizer to optimize our EINet with an initial learning rate of 1e-5 and weight decay of 0.5 applied every 5 steps. More details about training the baselines, as well as EINet can be found in our open-source repository.

TABLE I: Evaluation of place recognition performance in the supervised learning scheme on the BS split

Methods	Modality[1]	BS split			
		AR@1	AR@5	AR@10	AR@20
NetVLAD [7]	C	73.98	82.33	84.89	85.97
mrNVLAD [32]	C	75.31	84.49	86.40	88.23
OT [33]	L	74.67	84.23	87.12	89.70
CVTNet [15]	L	79.20	87.96	90.35	92.44
AdaFusion [19]	C+L	80.90	87.74	90.08	92.31
LCPR [21]	C+L	85.63	92.21	94.55	95.75
EINet (ours)	C+L	**91.40**	**97.30**	**98.75**	**99.39**

[1] C: Camera-only, L: LiDAR-only, C+L: Camera and LiDAR.

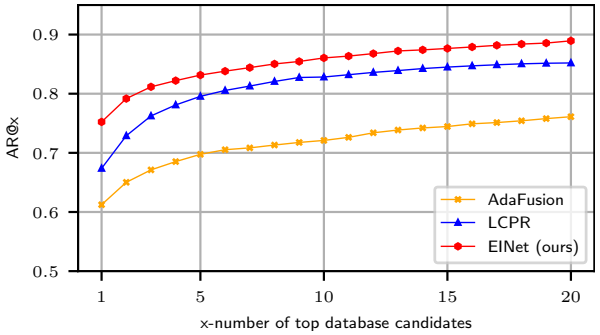

Fig. 4: Evaluation of place recognition performance in the self-supervised learning scheme on the BS split.

B. Assessment on NUSC-PR

The first experiment supports our claim that EINet can accurately retrieve previously visited places with long time spans in large-scale environments. We compare EINet with the baselines in the proposed NUSC-PR benchmark where the largest time gap between test query with the database is over 120 days.

Evaluation on supervised learning scheme. The experimental results are shown in Tab. I, which indicate that our proposed EINet outperforms all the baseline methods using positive and negative samples selected by distance metrics to supervise. Notably, EINet with explicit cross-modal interaction improves the FPR baselines AdaFusion and LCPR using only implicit feature fusion by 10.50% and 5.77% in AR@1. Besides, fusion-based approaches in general have better recognition performance than uni-modal baselines, and LiDAR-based baselines outperform the camera-based counterparts. This demonstrates that fusing multi-model features helps to improve the distinctiveness of global descriptors for place retrieval, and LiDAR can offer stable features with better place description ability than cameras.

Evaluation on self-supervised learning scheme. We compare the fusion-based baselines with EINet in the self-supervised learning scheme of the NUSC-PR. The AR@x is presented in Fig. 4. The recognition performance of all the methods decreases compared with the counterparts posted in Tab. I, because time metrics provide coarser discernment between positive and negative samples than distance metrics. Our EINet still performs better than the baselines over dif-

TABLE II: Ablation study on the explicit interaction

Methods	BS split			
	AR@1	AR@5	AR@10	AR@20
EINet-rL	84.95	93.58	95.92	98.02
EINet-rC	87.29	95.76	97.54	98.83
EINet	**91.40**	**97.30**	**98.75**	**99.39**

ferent recall rates, demonstrating that EINet is more suitable for online self-supervised learning using time metrics.

C. Ablation Study on Explicit Interaction

The experiment supports our claim that our proposed EINet can exploit explicit cross-modal interaction including sparse depth supervision and appearance rendering to enhance recognition performance. In Tab. II, we compare the holistic EINet with the baseline removing LiDAR-based depth supervision (EINet-rL), and the one removing camera-based appearance rendering (EINet-rC). As can be seen, without the feature enhancement from explicit LiDAR ranges, the performance of EINet-rL significantly drops by 6.45% on AR@1 compared to EINet. In contrast, EINet-rC still performs worse than the holistic one but outperforms EINet-rL, which indicates that the interaction of sparse depth supervision yields greater advantages than appearance rendering for fusion-based place recognition. The reason could be geometric information captured by LiDAR is more robust for place retrieval in large-scale environments with long time spans than appearance and texture information from camera images. Note that although we remove any type of interaction in this experiment, we retain the fusion-based framework that helps EINet-rL and EINet-rC achieve better performance than the uni-modal baselines in Tab. I.

D. Study on Generalization Ability

We further implement zero-shot transfer to support the third claim that our EINet has a solid generalization ability across different locations without fine-tuning. We use the supervised scheme in this experiment because the self-supervised scheme can help the place recognition model adaptively fit to new environments through incremental or lifelong learning [48], [49] rather than transfer learning. We train all the fusion-based approaches on the BS split and directly evaluate them on the test sets of other splits from SON, SQ, and SHV locations. The results on AR@1 shown in Tab. III demonstrate that EINet generalizes best into unseen locations even in different countries (US $\rightarrow$ SGP). AdaFusion has the worst generalization ability because the weights learned in one location to balance modalities are hard to ensure rationality regarding other locations. Note that there are counterintuitive results on the unseen SON split where the recognition accuracy is higher than the counterparts in the previously seen BS split, which is because the SON split provides easier query-reference pairs in the test set.

E. Study on Running Efficiency

In this experiment, we provide the inference time of each module in our proposed EINet implemented with Python.

TABLE III: Comparison of the generalization ability of the fusion-based approaches

Methods	Locations			
	BS	SON	SQ	SHV
AdaFusion	80.90	82.24	60.04	62.14
LCPR	85.63	93.76	67.53	69.29
EINet (ours)	**91.40**	**98.29**	**73.84**	**85.25**

TABLE IV: Inference time of each module in our proposed EINet

Module	Image Encoder	Image Decoder	OT-C	OT-L	CMT
Time [ms]	5.14	58.23	2.14	1.39	3.96

The hardware has been explained in Sec. V-A. We calculate the average runtime of the image encoder-decoder, the OT module in the camera branch (OT-C), the OT module in the LiDAR branch (OT-L), the cross-modal transformer (CMT) on the samples of the holistic nuScenes dataset. The results are shown in Tab. IV. As can be seen, the image decoder is the slowest part among the five modules, which costs 58.23 ms due to the utilization of CVT blocks. The summation of the other four modules is only 12.63 ms. Besides, we also calculate the average time (114.70 ms) of the transformation from the pseudo depth maps to the input of the OT module in the camera branch. Note that we do not consider the spherical projection time in the LiDAR branch because some existing LiDAR sensors can directly deliver raw range images [33]. Therefore, the total inference time of EINet is around 186.48 ms (5.36 Hz), which is efficient for online global localization in real-world applications.

VI. Conclusion

In this paper, we proposed explicitly utilizing the explainable interaction of LiDAR and camera modalities for multi-modal fusion-based place recognition. We first proposed the sparse depth supervision and appearance rendering as the specific form of cross-modal interaction and then jointly exploited them in a novel FPR network called EINet. Moreover, we provided a new NUSC-PR benchmark to standardize the training schemes and evaluation protocols in this paper. Extensive experiments conducted on the NUSC-PR demonstrated the effectiveness of the proposed explicit interaction for fusion-based place recognition. The efficient inference process also allows us to deploy EINet in real-world applications such as robots and autonomous driving. Both the implementation of our method and benchmark have been released as open-source to facilitate future research.

References

[1] X. Chen, A. Milioto, E. Palazzolo, P. Giguere, J. Behley, and C. Stachniss, "Suma++: Efficient lidar-based semantic slam," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pp. 4530–4537, 2019.

[2] J. Deng, Q. Wu, X. Chen, S. Xia, Z. Sun, G. Liu, W. Yu, and L. Pei, "Nerf-loam: Neural implicit representation for large-scale incremental lidar odometry and mapping," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8218–8227, 2023.

[3] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Trans. on Robotics (TRO)*, vol. 38, no. 4, pp. 2053–2073, 2022.

[4] F. Cao, H. Wu, and C. Wu, "An end-to-end localizer for long-term topological localization in large-scale changing environments," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 5, pp. 5140–5149, 2022.

[5] X. Chen, I. Vizzo, T. Läbe, J. Behley, and C. Stachniss, "Range image-based lidar localization for autonomous vehicles," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2021.

[6] P. Yin, R. A. Srivatsan, Y. Chen, X. Li, H. Zhang, L. Xu, L. Li, Z. Jia, J. Ji, and Y. He, "Mrs-vpr: a multi-resolution sampling based global visual place recognition method," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2019.

[7] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2016.

[8] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, 2023.

[9] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023.

[10] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2023.

[11] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," *arXiv preprint arXiv:2311.15937*, 2023.

[12] M. A. Uy and G. H. Lee, "Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018.

[13] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu, "Lpd-net: 3d point cloud learning for large-scale place recognition and environment analysis," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2019.

[14] 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 *Proc. of Robotics: Science and Systems*, 2021.

[15] J. Ma, G. Xiong, J. Xu, and X. Chen, "Cvtnet: A cross-view transformer network for lidar-based place recognition in autonomous driving environments," *IEEE Trans. on Industrial Informatics (TII)*, 2023.

[16] D. Kong, X. Li, and Q. Xu, "Sc_lpr: Semantically consistent lidar place recognition based on chained cascade network in long-term dynamic environments," *IEEE Trans. on Image Processing (TIP)*, 2024.

[17] Y. Lu, F. Yang, F. Chen, and D. Xie, "Pic-net: Point cloud and image collaboration network for large-scale place recognition," *arXiv preprint arXiv:2008.00658*, 2020.

[18] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "Minkloc++: lidar and monocular image fusion for place recognition," in *Proc. of the Intl. Conf. on Neural Networks (IJCNN)*, 2021.

[19] 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. 12038–12045, 2022.

[20] W. Liu, J. Fei, and Z. Zhu, "Mff-pr: Point cloud and image multi-modal feature fusion for place recognition," in *In proc. of IEEE Intl. Symp. on Mixed and Augmented Reality (ISMAR)*, 2022.

[21] Z. Zhou, J. Xu, G. Xiong, and J. Ma, "Lcpr: A multi-scale attention-based lidar-camera fusion network for place recognition," *IEEE Robotics and Automation Letters*, 2023.

[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 *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2020.

[23] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. on Robotics (TRO)*, vol. 32, no. 1, pp. 1–19, 2015.

[24] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, and Y. Wang, "A survey on global lidar localization: Challenges, advances and open problems," *arXiv preprint arXiv:2302.07433*, 2023.

[25] X. Hu, Z. Zhou, H. Li, Y. Hu, F. Gu, J. Kersten, H. Fan, and F. Klan, "Location reference recognition from texts: A survey and comparison," *ACM Computing Surveys*, vol. 56, no. 5, pp. 1–37, 2023.

[26] D. Gálvez-López and J. D. Tardos, "Bags of binary words for fast place recognition in image sequences," *IEEE Trans. on Robotics (TRO)*, vol. 28, no. 5, pp. 1188–1197, 2012.

[27] L. He, X. Wang, and H. Zhang, "M2dp: A novel 3d point cloud descriptor and its application in loop closure detection," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2016.

[28] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2018.

[29] K. P. Cop, P. V. Borges, and R. Dubé, "Delight: An efficient descriptor for global localisation using lidar intensities," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2018.

[30] Z. Chen, O. Lam, A. Jacobson, and M. Milford, "Convolutional neural network-based place recognition," *arXiv preprint arXiv:1411.1509*, 2014.

[31] S. Garg, T. Fischer, and M. Milford, "Where is your place, visual place recognition?," *arXiv preprint arXiv:2103.06443*, 2021.

[32] 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.

[33] 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.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2017.

[35] J. Ma, X. Chen, J. Xu, and G. Xiong, "Seqot: A spatial-temporal transformer network for place recognition using sequential lidar data," *IEEE Trans. on Industrial Electronics (TIE)*, 2022.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. of the Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[37] L. Luo, S. Zheng, Y. Li, Y. Fan, B. Yu, S.-Y. Cao, J. Li, and H.-L. Shen, "Bevplace: Learning lidar-based place recognition using bird's eye view images," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision*, 2023.

[38] 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*, vol. 8, no. 5, pp. 2828–2835, 2023.

[39] Y. Xia, L. Shi, Z. Ding, J. F. Henriques, and D. Cremers, "Text2loc: 3d point cloud localization from natural language," *arXiv preprint arXiv:2311.15977*, 2023.

[40] S. Zheng, Y. Li, Z. Yu, B. Yu, S.-Y. Cao, M. Wang, J. Xu, R. Ai, W. Gu, L. Luo, *et al.*, "I2p-rec: Recognizing images on large-scale point cloud maps through bird's eye view projections," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2023.

[41] 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.

[42] H. Yu, W. Zhen, W. Yang, J. Zhang, and S. Scherer, "Monocular camera localization in prior lidar maps with 2d-3d line correspondences," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2020.

[43] P. Yin, L. Xu, J. Zhang, H. Choset, and S. Scherer, "i3dloc: Image-to-range cross-domain localization robust to inconsistent environmental conditions," *arXiv preprint arXiv:2105.12883*, 2021.

[44] Y. Pan, X. Xu, W. Li, Y. Cui, Y. Wang, and R. Xiong, "Coral: Colored structural representation for bi-modal place recognition," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2021.

[45] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Proc. of the Conf. on Robot Learning (CoRL)*, 2023.

[46] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2017.

[47] K. Cai, B. Wang, and C. X. Lu, "Autoplace: Robust place recognition with single-chip automotive radar," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation*, 2022.

[48] J. Cui and X. Chen, "Ccl: Continual contrastive learning for lidar place recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4433–4440, 2023.

[49] J. Knights, P. Moghadam, M. Ramezani, S. Sridharan, and C. Fookes, "Incloud: Incremental learning for point cloud place recognition," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, 2022.