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

RaLF: Flow-based Global and Metric Radar Localization in LiDAR Maps

Abhijeet Nayak*, Daniele Cattaneo*, and Abhinav Valada

***Abstract*— Localization is paramount for autonomous robots. While camera and LiDAR-based approaches have been extensively investigated, they are affected by adverse illumination and weather conditions. Therefore, radar sensors have recently gained attention due to their intrinsic robustness to such conditions. In this paper, we propose RaLF, a novel deep neural network-based approach for localizing radar scans in a LiDAR map of the environment, by jointly learning to address both place recognition and metric localization. RaLF is composed of radar and LiDAR feature encoders, a place recognition head that generates global descriptors, and a metric localization head that predicts the 3-DoF transformation between the radar scan and the map. We tackle the place recognition task by learning a shared embedding space between the two modalities via cross-modal metric learning. Additionally, we perform metric localization by predicting pixel-level flow vectors that align the query radar scan with the LiDAR map. We extensively evaluate our approach on multiple real-world driving datasets and show that RaLF achieves state-of-the-art performance for both place recognition and metric localization. Moreover, we demonstrate that our approach can effectively generalize to different cities and sensor setups than the ones used during training. We make the code and trained models publicly available at `http://ralf.cs.uni-freiburg.de`.**

I. Introduction

Localization is pivotal for any autonomous robot, whether it operates in controlled environments such as factory floors or human-centric environments such as pedestrian zones and sidewalks. It is particularly important for the latter case, where the safety of other road users is of utmost importance. While Global Navigation Satellite Systems (GNSSs) are widely used for outdoor localization, their accuracy and reliability strongly deteriorate in urban canyons. Therefore, localization systems that rely on alternative modalities are essential for autonomous robots operating in such environments.

While vision-based localization [1]–[4] has been extensively studied, their performance suffers in adverse conditions such as night and rain. To overcome these limitations, methods based on LiDARs have been proposed [5], [6] due to their high precision and robustness to illumination changes. However, LiDARs are also affected by extreme weather conditions such as fog and snow, which can drastically reduce their range and accuracy. Moreover, their high cost and large size make them unsuitable for large-scale deployment. Due to these factors, recent work investigates exploiting radar scans for localization [7], [8]. Radars are intrinsically robust to both weather and illumination conditions, making them a promising alternative to cameras and LiDARs. Most of existing radar-based localization methods [7], [8] compare the onboard radar measurement with a pre-built radar map of the environment. As radar maps are not readily available as of today, these methods require a first mapping phase of each environment in which the robot will be deployed. LiDAR maps, on the other hand, are becoming increasingly available, thanks to the growing demand for high-definition maps.

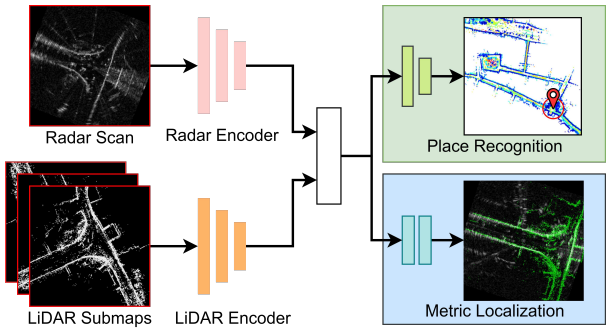


Fig. 1: Our proposed RaLF localizes a radar scan within a LiDAR map both at a global (place recognition) and metric scale.

A small number of works have been proposed to address the cross-modal task of localizing a radar scan within a LiDAR map, however, they only focus on a part of the localization problem, either place recognition or metric localization. On the one hand, place recognition methods [7] provide global localization, but with inaccurate precision. On the other hand, metric localization methods [9], [10] provide accurate localization but require an initial coarse position as input. Therefore, for a complete localization system, a place recognition approach has to be combined with a metric localization method, introducing inefficiencies, as sensor data has to be processed by two separate approaches.

In this paper, we propose RaLF, a novel Deep Neural Network (DNN)-based method for radar localization in prior LiDAR maps. Differently from existing radar-LiDAR localization methods, RaLF is, to the best of our knowledge, the first method to jointly address both place recognition and metric localization. We reformulate the metric localization task as a flow estimation problem, where we aim at predicting pixel-level correspondences between the radar and LiDAR samples, which are subsequently used to estimate a 3-DoF transformation. For place recognition, we leverage a combination of same-modal and cross-modal metric learning to learn a shared embedding space where features from both modalities can be compared against each other. We evaluate place recognition and metric localization performance of our approach on three real-world driving datasets, namely,

*These authors contributed equally to this work.
Department of Computer Science, University of Freiburg, Germany.
This work was funded by the German Research Foundation (DFG) Emmy Noether Program grant number 468878300.

Oxford Radar Robotcar [11], MulRan [12], and Boreas [13]. We compare our method against methods, and show that RaLF achieves state-of-the-art performance.

The main contributions of this work are as follows:

1) We propose the novel RaLF for radar localization in prior LiDAR maps, addressing both place recognition and metric localization tasks.
2) We propose to solve the metric localization task by predicting pixel-level matches in the form of a flow field between the radar and the LiDAR Bird's-Eye-View (BEV) images.
3) We extensively evaluate RaLF against state-of-the-art place recognition and metric localization methods on three real-world datasets.
4) We investigate the generalization ability of our method by evaluating it in a different city and using a different sensor setup than the ones used for training.
5) We release the code and trained models at `http://ralf.cs.uni-freiburg.de`.

II. Related Work

In this section, we discuss related work on LiDAR/radar place recognition and metric localization, for both the same-modality and cross-modality settings.

Place Recognition: Place recognition has been extensively explored over recent decades. Scan Context [14] and M2DP [15] investigate place recognition with LiDAR data using hand-crafted descriptors. With the advent of Convolutional Neural Networks (CNNs), DiSCO [16] employs a learning-based approach for LiDAR-based place recognition using Scan Context as input. On the other hand, Gadd *et al.* [17] tackle radar-based place recognition through unsupervised contrastive loss learning. They additionally highlight the importance of carefully selecting positive and negative samples to facilitate training the network with a contrastive loss function. Additionally, Cait *et al.* [18] leverage data from a single-chip automotive radar (point cloud) for place recognition. While these works primarily focus on single sensor modalities, there have been notable efforts involving cross-modal data.

In the context of cross-modal place recognition, addressing the disparities between diverse sensor modalities requires the establishment of a cohesive embedding space. Various strategies have emerged to achieve this goal. Recent research efforts employ shared networks to create joint embeddings, unifying data from different modalities [19], [20]. Conversely, Yin *et al.* [10] suggest leveraging Generative Adversarial Networks (GANs) to transform data from one modality into a newly generated sample that resembles the other modality. Furthermore, Radar-to-lidar [7] employs shared network processing of radar and LiDAR Scan Contexts to generate feature embeddings, subsequently employing KD-Trees for clustering. In this application, a shared encoder is employed under the assumption that Scan Contexts for both radar and LiDAR modalities exist within the same embedding space.

Our approach distinguishes itself by employing BEV images of radar and LiDAR modalities for place recognition. Our solution hinges on a novel CNN-based method that crafts joint-space embeddings between the two modalities. Notably, the use of BEV images facilitates metric localization as well, a feat unattainable through alternative data representations such as Scan Context, which only provides an estimation of the angle between two samples.

Metric Localization pertains to precisely estimating the position of a robot within a map of the environment. Historically, metric localization was addressed with classical robotics techniques involving probabilistic updates to adjust the likelihood of a robot's location on a map [21]–[23]. Recent progress has enabled learning-based techniques to achieve accurate metric localization. LCDNet [5] detects loop closures in LiDAR point clouds and estimates relative scan-to-map poses. On the other hand, OverlapNet [24] gauges the overlap between LiDAR scans and uses ICP-based techniques for relative pose estimation.

Past endeavors have seen the application of cross-modal strategies in metric localization. CMRNet [25], [26] employs CNN-based techniques to localize a camera image onto a pre-existing LiDAR map. In the method proposed by Tang *et al.* [27], LiDAR scans are localized against aerial satellite images. To manage the distinct modalities, an encoder-decoder network creates an occupancy map from the overhead images. These maps are subsequently transformed into point clouds and registered against LiDAR scans using point-based methods. Tang *et al.* [28] propose to estimate relative poses in a self-supervised manner by identifying the augmented sample with optimal rotation and translation noise that aligns most closely with the current overhead image. RaLL [9] proposes the use of a differentiable measurement model to localize radar samples on a pre-existing LiDAR map. This measurement model is then applied to a Kalman filter, thus making the whole learning system differentiable. In contrast to previous approaches, we tackle the metric localization challenge as a flow estimation task. We compute flow vectors between radar and LiDAR BEV images to establish initial pixel correspondences. Subsequently, we employ Random Sample Consensus (RANSAC) [29] to estimate the accurate relative transformations between the two inputs.

III. Technical Approach

In this section, we describe RaLF, our proposed approach for place recognition and metric radar localization in LiDAR maps. An overview of RaLF is illustrated in Fig. 2. The architecture of our approach is built upon RAFT [30], a state-of-the-art network for optical flow estimation. RaLF comprises three main components: feature extraction, place recognition head, and metric localization head. In the rest of this section, we detail each of the components and the respective loss functions, followed by a description of the inference procedure.

A. Feature Extraction

The architecture of the two encoders, namely the radar encoder and LiDAR encoder, is based on the feature encoder of RAFT [30], which consists of a convolutional layer with stride equal to two, followed by six residual layers with

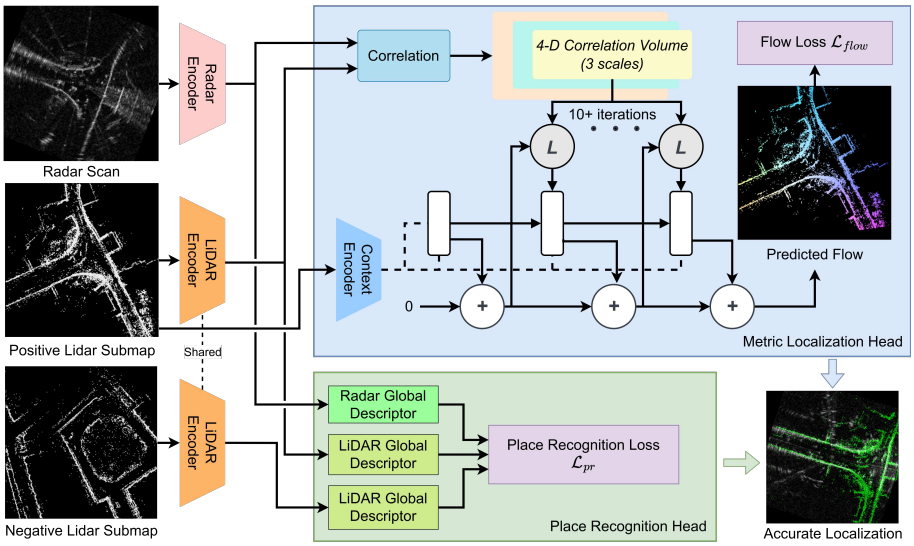


Fig. 2: Overview of RaLF for joint place recognition and metric localization of radar scans in a LiDAR map. It consists of feature encoders, a place recognition head to extract global descriptors, and a metric localization head to estimate the 3-DoF pose of the query radar scan within the LiDAR map.

downsampling after the second and fourth layer. Differently from the original feature encoder of RAFT which shares weights between the two input images, RaLF employs separate feature extractors for each modality due to the distinct nature of radar and LiDAR data. Formally, given a radar BEV image $\mathbf{R} \in \mathbb{R}^{H \times W \times 1}$ and a LiDAR BEV image $\mathbf{L} \in \mathbb{R}^{H \times W \times 1}$, the two encoders g_r and g_l extract features at one-eight of the original resolution $g_r, g_l : \mathbb{R}^{H \times W \times 1} \rightarrow \mathbb{R}^{H/8 \times W/8 \times D}$. The features extracted by the two encoders are shared between the place recognition head and the metric localization head.

B. Place Recognition Head

The place recognition head has a twofold purpose: firstly, it aggregates the feature maps from the feature extractor into a global descriptor. Secondly, it maps features from radar and LiDAR data, which naturally lie in different embedding spaces, into a shared embedding space, where global descriptors of radar scans and LiDAR submaps can be compared against each other. The architecture of the place recognition head is a shallow CNN composed of four convolutional layers with feature sizes (256, 128, 128, 128), respectively. Each convolutional layer is followed by batch normalization and ReLU activation. Differently from the feature encoders, the place recognition head is shared between the radar and LiDAR modalities.

To train the place recognition head, we use the well-known triplet technique [31], where triplets composed of (anchor, positive, negative) samples are selected to compute the triplet loss. The positive sample is a BEV image depicting the same place as the anchor sample, while the negative sample is a BEV image of a different place. While typically this technique is employed to compare triplets of samples of the same modality, in our case the samples can be generated from different modalities. For instance, given an anchor radar scan $\mathbf{R}^a$, a positive LiDAR submap $\mathbf{L}^p$, and a negative LiDAR submap $\mathbf{L}^n$, we define the triplet loss $\mathcal{L}_{tr}^{RLL}$ as

$$\mathcal{L}_{tr}^{RLL} = \max\left\{d(\mathbf{F}_R^{\mathrm{a}}, \mathbf{F}_L^{\mathrm{p}}) - d(\mathbf{F}_R^{\mathrm{a}}, \mathbf{F}_L^{\mathrm{n}}) + m, 0\right\}, \quad (1)$$

where $\mathbf{F}_R^a$, $\mathbf{F}_L^p$, and $\mathbf{F}_L^n$, are the global descriptors of $\mathbf{R}^a$, $\mathbf{L}^p$, $\mathbf{L}^n$, respectively; m is the triplet margin, and $d(\cdot)$ is a given distance function. The superscript RLL of $\mathcal{L}_{tr}^{RLL}$ represent the modalities of the (anchor, positive, negative) samples, in this case (radar, LiDAR, LiDAR). We apply the same loss to all the eight possible combination of modalities, leading to the final place recognition loss:

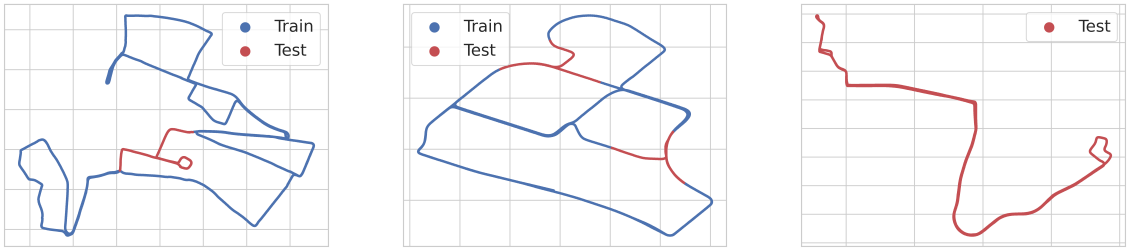
$$\begin{aligned}\mathcal{L}_{pr} = &\mathcal{L}_{tr}^{RRR} + \mathcal{L}_{tr}^{RLL} + \mathcal{L}_{tr}^{RLR} + \mathcal{L}_{tr}^{RRL} + \\ &\mathcal{L}_{tr}^{LLL} + \mathcal{L}_{tr}^{LRR} + \mathcal{L}_{tr}^{LRL} + \mathcal{L}_{tr}^{LLR}.\end{aligned} \quad (2)$$

To select the triplets that compose a batch, we first randomly sample a positive sample for each anchor sample. We define a sample to be positive of an anchor sample if the distance between their position is smaller than a positive threshold τ_p. Furthermore, we select the hardest negative sample from the batch of samples currently being processed by the network, making sure that its position is farther away from the anchor than a negative threshold τ_n. This technique is known as online hardest negative mining.

C. Metric Localization Head

For metric localization of radar scans against a LiDAR map $\mathcal{M}$, we propose to learn pixel-wise matches in the form of flow vectors. The intuition behind this decision is that a radar BEV image and a LiDAR BEV image taken at the same position should be well aligned, as depicted in the bottom right part of Fig. 2. Therefore, for every pixel in the LiDAR BEV image, our metric localization head predicts the corresponding pixel in the radar BEV image.

More formally, given a radar BEV image $\mathbf{R}$, and the initial coarse pose T_{init} predicted by the place recognition head, we generate a LiDAR BEV image $\mathbf{L}$ centered around T_{init}. The metric localization head takes the two BEV images $\mathbf{R}$ and $\mathbf{L}$ as input, and predicts a dense flow map $\mathbf{f}$ that aligns the two images. Each pixel (u, v) in $\mathbf{f}$ contains the flow vector $(\Delta u, \Delta v)$ that maps the pixel $\mathbf{L}_{(u,v)}$ to the pixel $\mathbf{R}_{(u+\Delta u, v+\Delta v)}$. The architecture of our metric localization



(a) Oxford Radar Robotcar (b) MulRan-Kaist (c) Boreas

Fig. 3: Train-test split of the three datasets used in our experiments. The blue and red trajectories represent the train and test splits, respectively.

head is based on RAFT [30], which first computes a 4-D correlation volume between the features extracted by the two encoders as described in Sec. III-A. The correlation volume is then fed into a Gated Recurrent Unit (GRU) that iteratively refines the estimated flow map. Each iteration i of the GRU update outputs a flow update $\Delta\mathbf{f}_i$, which is added to the previous flow estimate $\mathbf{f}_{i-1}$ to obtain the updated flow map $\mathbf{f}_i$. Following [30], we employ an additional context encoder that extracts features only from the LiDAR BEV image, which is additionally fed to the GRU. To generate the ground truth flow map $\mathbf{f}_{GT}$, we first transform the LiDAR map points in the initial pose T_{init}, and compute their pixel position in the relative BEV image $\mathbf{L}$ as follows:

$$[u_{init}, v_{init}] = f_{bev}(T_{init} \cdot \mathcal{M}), \tag{3}$$

where f_{bev} is the function that projects a 3D point cloud into the BEV image. Similarly, we compute the pixel position of the projection when transforming the map using the ground truth pose T_{GT}:

$$[u_{GT}, v_{GT}] = f_{bev}(T_{GT} \cdot \mathcal{M}). \tag{4}$$

Finally, we compute the ground truth flow map $\mathbf{f}_{GT}$ by comparing the projected points using the initial pose and the ground truth pose as

$$\mathbf{f}_{GT} = [u_{GT} - u_{init}, v_{GT} - v_{init}]. \tag{5}$$

To train the metric localization head, we use the loss function originally proposed in RAFT [30], which supervises the predicted flow maps $\mathbf{f}_i$ at each iteration of the GRU as

$$\mathcal{L}_{flow} = \sum_{i=1}^{N} \gamma^{N-i} ||\mathbf{f}_{GT} - \mathbf{f}_i||_1, \tag{6}$$

where $\gamma = 0.8$ gives exponentially increasing weights to later iterations. Due to the sparse nature of the BEV images, we only compute the loss on the non-zero pixels in $\mathbf{L}$.

The final loss function of RaLF is the sum of the individual loss functions defined in Eq. (2) and Eq. (6):

$$\mathcal{L}_{total} = \mathcal{L}_{pr} + \mathcal{L}_{flow}. \tag{7}$$

D. Inference

Before deployment, we split the LiDAR map $\mathcal{M}$ into multiple overlapping submaps $\mathcal{M}_i$, with the relative poses T_i, and we generate a LiDAR BEV image $\mathbf{L}_i$ for each submap. We then generate the global descriptors $\mathbf{F}_L^i$ for each submap $\mathbf{L}_i$ using the LiDAR encoder and the place recognition head.

During inference, given a query radar scan $\mathbf{R}$, we first compute its global descriptor $\mathbf{F}_R$, and we compare it against all the submap descriptors $\mathbf{L}_i$. We then select the submap $\mathbf{L}_k$ with the highest similarity to $\mathbf{R}$:

$$k = \text{argmin}_i \|\mathbf{F}_R - \mathbf{F}_L^i\|_2 \ . \tag{8}$$

We then feed $\mathbf{F}_R$ and $\mathbf{L}_k$ to the metric localization head to predict the flow map $\mathbf{f} = [\mathbf{f}^u, \mathbf{f}^v]$, which we use to generate a warped LiDAR BEV image $\mathbf{L}_k^{warp}$ as

$$\mathbf{L}_k^{warp}(u + \mathbf{f}^u_{(u,v)}, v + \mathbf{f}^v_{(u,v)}) = \mathbf{L}_k(u, v). \tag{9}$$

Subsequently, we convert the two images $\mathbf{L}_k$ and $\mathbf{L}_k^{warp}$ into point clouds $\mathbf{P}_k$ and $\mathbf{P}_k^{warp}$, by multiplying each pixel location by the pixel resolution of the BEV image, and setting the height to zero. Finally, we use RANSAC to estimate the query radar pose T_{pred} that minimizes the distance between the two point clouds:

$$T_{pred} = \text{argmin}_T \|\mathbf{P}_k^{warp} - T \cdot \mathbf{P}_k\|_2 \ . \tag{10}$$

IV. Experimental Evaluation

In this section, we first describe the datasets that we use for training and evaluation, followed by details on our training protocol. We then present results from evaluating RaLF against the state-of-the-art in both place recognition and metric localization. Finally, we perform multiple ablation studies to validate the design choices of our method.

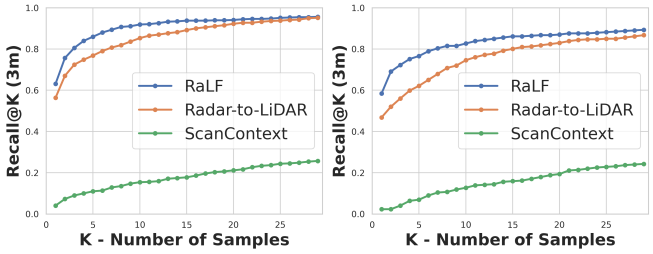
A. Datasets

We evaluate our proposed approach on three real-world driving datasets, namely Oxford Radar Robotcar [11], MulRan [12] and Boreas [13]. We use the Oxford and the MulRan datasets for training and evaluation, while we use the Boreas dataset for testing the generalization ability of RaLF in a different city and a different sensor setup than those used for training. For the Robotcar dataset, we use the train-test split proposed in [10], while for MulRan, we use the KAIST sequences and selected two random geographical areas as the test set. Fig. 3 shows the train-test split of the three datasets*.

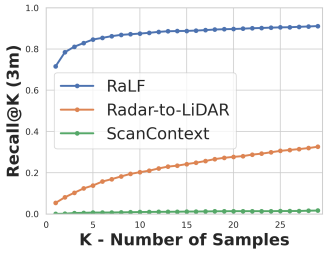
*On Oxford, we used the sequences 2019-01-18-12-42-34, 2019-01-18-14-14-42, 2019-01-18-14-46-59, and 2019-01-18-15-20-12 for training. For testing, we used the sequence 2019-01-18-15-20-12 as a map and the sequence 2019-01-18-14-46-59 as a query. On KAIST, we used sequences 02 and 03 for training, and for testing, sequence 02 was used as a query against sequence 03. On Boreas, the sequence 2021-05-06-13-19 serves as a query while the sequence 2021-05-13-16-11 as the map.

TABLE I: Comparison with the state of the art in terms of recall@1 (3 m).

Approach	Oxford [11]			MulRan-Kaist [12]			Boreas [13]		
	R2R	L2L	R2L	R2R	L2L	R2L	R2R	L2L	R2L
M2DP [15]	-	0.20	-	-	0.34	-	-	0.74	-
Scan Context [14]	0.87	0.97	0.016	**0.97**	0.97	0.02	0.96	0.89	0.002
RaPlace [8]	0.78	-	-	0.83	-	-	0.87	-	-
DiSCO [16]	0.90	0.96	0.013	**0.97**	**0.98**	0.05	0.96	0.93	0.001
Radar-to-LiDAR [7]	0.85	0.93	0.56	0.90	0.89	0.46	0.96	0.92	0.05
RaLF (Ours)	**0.97**	**0.98**	**0.63**	0.88	0.89	**0.58**	**0.99**	**0.99**	**0.71**



(a) Oxford Radar Robotcar (b) MulRan-Kaist



(c) Boreas

Fig. 4: Recall@k (3 m) at different values of k.

It is important to note that we do not train a separate model for each dataset, instead, we train a single model on the combined training split of the Oxford and MulRan datasets and evaluate it on the test splits of all three datasets.

B. Training Details

We use the PyTorch deep learning library for model training and evaluation on a machine equipped with an Intel i5-6500@3.2GHz processor and two NVIDIA TITAN-X GPUs. We use images of size 256 $\times$ 256 at a resolution of 0.5 m/pixel for the BEV projection. For data augmentation, we apply random rotations and translations in the range $\pm 30^\circ$ and ± 5 m, respectively, for both radar and LiDAR samples. We use the AdamW optimizer and the OneCycle learning rate scheduler with a learning rate of $5 \cdot 10^{-4}$. The duration of the learning rate increase is 10% of the whole training time, which is about $2 \cdot 10^5$ iterations. We train the network with a batch size of 15 (anchor, positive) pairs per iteration. During training, we set $m = 0.5$, $\tau_p = 2$ m and $\tau_n = 80$ m, and the distance function $d(\cdot)$ as the L2 distance.

C. Place Recognition

We compare the place recognition performance of RaLF against handcrafted methods M2DP [15], Scan Context [14], and RaPlace [8] as well as learning-based methods DiSCO [16] and Radar-to-lidar [7]. For all the above methods, we use the implementation provided by the respective authors. To provide a fair comparison, we retrained all learning-based methods on the same training set used to train RaLF. Radar-to-lidar is the only existing method that specifically tackles the cross-modal radar-LiDAR place recognition task. RaPlace, on the other hand, focuses on radar-radar place recognition. All the other methods were originally proposed for LiDAR place recognition, although some of them can easily be adapted to radar data. We use the recall@1 (3m) metric, which considers a query as correctly localized if the pose of the most similar database sample is within three meters from the pose of the query.

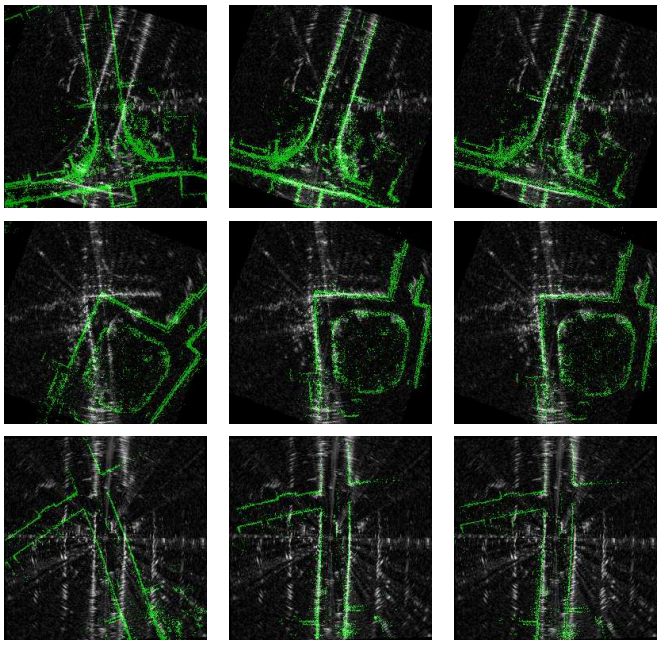
Tab. I summarizes the results from this experiment. We observe that our method outperforms all the other methods in the radar-LiDAR place recognition task. In particular, we achieve a recall@1 of 0.63, 0.58, and 0.71 on the Oxford, MulRan, and Boreas datasets, respectively. This is a significant improvement over the state-of-the-art method Radar-to-lidar [7], which achieves a recall@1 of 0.56, 0.46, and 0.05 on the same datasets. We also observe that RaLF achieves state-of-the-art performance in the radar-radar and LiDAR-LiDAR place recognition tasks on both the Oxford and Boreas datasets. Fig. 4 shows the recall@k (3 m) at different values of k for all the datasets. We observe that our method outperforms the state-of-the-art at all values of k, especially at smaller values of k.

TABLE II: Comparison of relative pose errors (rotation and translation) between positive pairs on the Oxford Radar Robotcar dataset.

Approach	δx (m) $\downarrow$	δy (m) $\downarrow$	$\delta\theta$ (Deg) $\downarrow$
RaLL [9]	1.04	**0.69**	1.26
RaLF (Ours), (RaLL params)	**0.97**	0.96	**1.15**
RaLF (Ours), ($\pm 30^\circ, \pm 5$ m)	1.07	1.03	1.26

D. Metric Localization

In Tab. II, we report the rotation and translation errors between the estimated transformation and the ground truth transformation. The results show that our method outperforms RaLL [9] while using the same data augmentation scheme ($\pm 6^\circ, \pm 6$ m) during evaluation. We achieve a mean rotation error of 1.26° and translation errors of 1.07 m and 1.03 m along X and Y-directions respectively. In Fig. 5, we report the qualitative results of metric localization, where the query radar scans are overlaid with LiDAR submaps using the initial coarse pose T_{init}, the pose estimated by RaLF T_{pred}, and the ground truth pose T_{GT}.



Initial Pose RaLF alignment GT alignment

Fig. 5: Qualitative results of radar scans (grayscale) aligned with the LiDAR submaps (green) using our proposed method.

TABLE III: Ablation study on the encoder architecture.

Encoder	# Params (M)	Recall@1 (3 m) ↑	δx (m) ↓	δy (m) ↓	$\delta\theta$ (Deg) ↓
VGG-19	21	0.54	1.10	1.11	1.49
ResNet-18	6.4	0.52	1.28	1.25	1.60
ResNet-34	7.8	0.52	1.17	1.03	1.54
ResNet-50	8.9	0.60	1.26	1.04	1.42
RAFT-C	6.4	0.04	5.71	3.28	14.5
RAFT-S	2.0	0.66	1.32	1.30	1.51
RAFT	7.4	**0.67**	**1.07**	**1.03**	**1.26**

E. Ablation Studies

We perform multiple ablation studies to validate the design choices that we made while developing RaLF. All the experiments reported in this section are trained and tested on the Oxford Radar RobotCar dataset.

Encoder Choice: In the feature extraction module, we employ different encoder architectures, including ResNet [32], VGGNet [33], and RAFT encoders [30], to process BEV images and generate feature embeddings. Tab. III summarizes the results of this study which shows that the RAFT encoder performs best for our approach. The smaller RAFT-S version also achieves similar recall@1 with just 2 Million parameters. Furthermore, when we used a common encoder for both the radar and LiDAR images in RAFT-C (Common), the network was not able to learn a meaningful representation for the two heads, emphasizing the need to process different sensor modalities with separate encoders.

Choice of Place Recognition Head: We performed experiments with various CNN architectures, each with a different number of convolutional layers, for the place recognition head. Additionally, we evaluated the NetVLAD layer, which has shown success in Visual Place Recognition (VPR) tasks.

TABLE IV: Ablation study on the place recognition architecture.

PR Head	Recall@1 (3 m) ↑
NetVLAD	0.12
CNN-2	0.57
CNN-3	0.62
CNN-4	**0.67**
CNN-5	0.64

TABLE V: Ablation study on the place recognition loss function.

$\mathcal{L}_{pr}$	Recall@1 (3 m) ↑
Contrastive Loss	0.19
NPairs Loss	0.26
Quadruplet Loss	0.43
Triplet Loss - Cosine	0.58
Triplet Loss - L1	0.51
Triplet Loss - L2	**0.67**

The results of this experiment are presented in Tab. IV. Interestingly, we observed that NetVLAD underperforms in comparison to CNN-based heads in our setting. We hypothesize that this can be attributed to the sparsity of features extracted from the BEV images and the cross-modal nature of feature comparison. Furthermore, the CNN architectures show increasingly better performance while increasing the number of layers up to four.

Choice of Metric Loss Function: Finally, we experiment with different loss functions for the place recognition head. We evaluate the contrastive loss, N-pairs loss, triplet margin loss, and quadruplet loss. For the triplet loss, we additionally evaluate different distance functions $d(\cdot)$ in Eq. (1): Cosine Similarity, L1 and L2. As this change only affects the place recognition head, we exclusively report the recall@1 scores for this experiment. Tab. V shows the results from this experiment. We observe that the triplet margin loss outperforms all the other losses, while the quadruplet loss is the next best loss function for our use case. This is probably due to the hard-negative mining strategy that we use while mining for triplets. We also observe that the L2 distance function achieves the best performance.

V. Conclusion

In this paper, we proposed RaLF, a novel approach to simultaneously address place recognition and metric localization of radar scans in LiDAR maps. To the best of our knowledge, RaLF is the first approach that combines both these tasks in an end-to-end manner. Our method is composed of feature encoders, a place recognition head, and a metric localization head based on flow vectors. We presented extensive experiments on the Oxford Radar Robotcar and MulRan datasets, and demonstrated that it achieves state-of-the-art performance for both place recognition and metric localization. Furthermore, we demonstrated that our method can effectively generalize to different cities and sensor setups than the ones used during training by evaluating it on the Boreas dataset. To foster future research, we made the code for RaLF and the trained models publicly available.

REFERENCES

[1] N. Vödisch, D. Cattaneo, W. Burgard, and A. Valada, "Covio: Online continual learning for visual-inertial odometry," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 2463–2472.

[2] A. L. Ballardini, D. Cattaneo, and D. G. Sorrenti, "Visual localization at intersections with digital maps," in *Int. Conf. on Robotics and Automation*, 2019, pp. 6651–6657.

[3] N. Vödisch, D. Cattaneo, W. Burgard, and A. Valada, "Continual slam: Beyond lifelong simultaneous localization and mapping through continual learning," in *Int. Symposium of Robotics Research*, 2022, pp. 19–35.

[4] A. Valada, N. Radwan, and W. Burgard, "Deep auxiliary learning for visual localization and odometry," in *Int. Conf. on Robotics and Automation*, 2018, pp. 6939–6946.

[5] D. Cattaneo, M. Vaghi, and A. Valada, "Lcdnet: Deep loop closure detection and point cloud registration for lidar slam," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2074–2093, 2022.

[6] J. Arce, N. Vödisch, D. Cattaneo, W. Burgard, and A. Valada, "Padloc: Lidar-based deep loop closure detection and registration using panoptic attention," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1319–1326, 2023.

[7] H. Yin, X. Xu, Y. Wang, and R. Xiong, "Radar-to-lidar: Heterogeneous place recognition via joint learning," *Frontiers in Robotics and AI*, vol. 8, p. 661199, 2021.

[8] H. Jang, M. Jung, and A. Kim, "Raplace: Place recognition for imaging radar using radon transform and mutable threshold," *arXiv preprint arXiv:2307.04321*, 2023.

[9] H. Yin, R. Chen, Y. Wang, and R. Xiong, "Rall: end-to-end radar localization on lidar map using differentiable measurement model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6737–6750, 2021.

[10] H. Yin, Y. Wang, J. Wu, and R. Xiong, "Radar style transfer for metric robot localisation on lidar maps," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 1, pp. 139–148, 2023.

[11] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset," in *Int. Conf. on Robotics and Automation*, 2020, pp. 6433–6438.

[12] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "Mulran: Multimodal range dataset for urban place recognition," in *Int. Conf. on Robotics and Automation*, 2020, pp. 6246–6253.

[13] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y. Leung, *et al.*, "Boreas: A multi-season autonomous driving dataset," *Int. Journal of Robotics Research*, vol. 42, no. 1-2, pp. 33–42, 2023.

[14] G. Kim and A. Kim, "Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map," in *Int. Conf. on Intelligent Robots and Systems*, 2018, pp. 4802–4809.

[15] L. He, X. Wang, and H. Zhang, "M2dp: A novel 3d point cloud descriptor and its application in loop closure detection," in *Int. Conf. on Intelligent Robots and Systems*, 2016, pp. 231–237.

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

[17] M. Gadd, D. De Martini, and P. Newman, "Contrastive learning for unsupervised radar place recognition," in *Int. Conf. on Advanced Robotics*, 2021, pp. 344–349.

[18] K. Cait, B. Wang, and C. X. Lu, "Autoplace: Robust place recognition with single-chip automotive radar," in *Int. Conf. on Robotics and Automation*, 2022, pp. 2222–2228.

[19] I. A. Barsan, S. Wang, A. Pokrovsky, and R. Urtasun, "Learning to localize using a LiDAR intensity map," in *Conf. on Robot Learning*, vol. 87, 2018, pp. 605–616.

[20] D. Cattaneo, M. Vaghi, S. Fontana, A. L. Ballardini, and D. G. Sorrenti, "Global visual localization in lidar-maps through shared 2d-3d embedding space," in *Int. Conf. on Robotics and Automation*, 2020, pp. 4365–4371.

[21] V. Fox, J. Hightower, L. Liao, D. Schulz, and G. Borriello, "Bayesian filtering for location estimation," *IEEE pervasive computing*, vol. 2, no. 3, pp. 24–33, 2003.

[22] D. Fox, S. Thrun, W. Burgard, and F. Dellaert, "Particle filters for mobile robot localization," in *Sequential Monte Carlo methods in practice*, 2001, pp. 401–428.

[23] S. Chen, "Kalman filter for robot vision: a survey," *IEEE Transactions on industrial electronics*, vol. 59, no. 11, pp. 4409–4420, 2011.

[24] 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 *Robotics: Science and Systems*, 2020.

[25] D. Cattaneo, M. Vaghi, A. L. Ballardini, S. Fontana, D. G. Sorrenti, and W. Burgard, "Cmrnet: Camera to lidar-map registration," in *2019 IEEE intelligent transportation systems conference (ITSC)*. IEEE, 2019, pp. 1283–1289.

[26] D. Cattaneo, D. G. Sorrenti, and A. Valada, "Cmrnet++: Map and camera agnostic monocular visual localization in lidar maps," *Int. Conf. on Robotics & Automation Workshop on Emerging Learning and Algorithmic Methods for Data Association in Robotics*, 2020.

[27] T. Y. Tang, D. De Martini, and P. Newman, "Get to the point: Learning lidar place recognition and metric localisation using overhead imagery," *Robotics: Science and Systems*, 2021.

[28] T. Y. Tang, D. De Martini, S. Wu, and P. Newman, "Self-supervised learning for using overhead imagery as maps in outdoor range sensor localization," *Int. Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1488–1509, 2021.

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

[30] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Europ. Conf. on Computer Vision*, 2020, pp. 402–419.

[31] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.