Neurocomputing 668 (2026) 132404

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

APR-BiCA: LiDAR-based absolute pose regression with bidirectional cross attention and gating unit

Jianlong Dai, Hui Wang*, Yuqian Zhao, Zhihua Liu

School of Automation, Central South University, Changsha 410083, Hunan, China

HIGHLIGHTS

- A LiDAR-based absolute pose regression network is proposed, which accurately and efficiently regress 6-DoF poses.
- A dual-branch architecture is designed to extract complementary information from raw point clouds and range images.
- A Feature Fusion Module is adopted to capture correlations between two types of data and achieve efficient feature fusion.
- Extensive experiments on the Oxford RobotCar dataset and the NCLT dataset, show that our network achieves good performance.

ARTICLE INFO

Communicated by J. Yu

Keywords:
LiDAR-based localization
Absolute pose regression
Attention
Gating unit
Range image

ABSTRACT

LiDAR localization is a critical component in fields like intelligent robotics and autonomous driving. Absolute pose regression (APR) techniques directly infer global poses from input point clouds through end-to-end regression, achieving superior computational efficiency. However, APR struggles with dynamic objects and environmental noise in large-scale scenarios. To address this, we propose an APR network called APR-BiCA to fuse complementary information from raw point clouds and range images, which aims to improve the localization accuracy of robots in large-scale autonomous driving scenarios. The APR-BiCA incorporates two distinct branches: one extracts features from the raw point cloud to capture key features and build global point relationships, while the other processes the range image derived from the point cloud to extracts robust structural features. Additionally, a bidirectional cross attention mechanism combined with a gating unit-based fusion module is designed to facilitate effective inter-modal feature interaction, thereby enhancing the feature representational capability to support efficient handling of large-scale environments. Experimental results on the Oxford RobotCar and NCLT datasets demonstrate the superior performance of APR-BiCA, while maintaining exceptional efficiency. This well-balanced combination of accuracy and efficiency underscores its potential to advance LiDAR-based localization technology and drive its practical application in real-world autonomous driving systems.

1. Introduction

In recent years, LiDAR has emerged as a crucial sensor for environmental perception, owing to its robustness against lighting conditions, weather, and other environmental factors, as well as its capability to provide precise depth information. LiDAR-based localization, which involves estimating the position and orientation from LiDAR point clouds, plays a pivotal role in a variety of applications, including autonomous driving [1], place recognition [2,3], and robot navigation [4].

A typical approach for LiDAR-based localization involves map-based retrieval methods that estimate a 6-DoF pose by matching features between a query point cloud and a 3D database or pre-built map, offering accurate global pose estimation [1,5–9]. These methods usually require high costs for map storage and communication, and are more appropriate for offline situations than for real-time mobile applications. Different from that, absolute pose regression(APR) [10–13] directly estimates the 6-DoF pose of the 3D query point cloud using deep regression networks, enabling LiDAR-based localization with a single forward pass. This approach helps alleviate the computational and storage burdens associated with map-based retrieval methods, enabling real-time pose regression to meet the localization needs of robotics, mobile apps, and autonomous vehicles.

* Corresponding author.
Email address: 224601027@csu.edu.cn (H. Wang).

https://doi.org/10.1016/j.neucom.2025.132404
Received 16 June 2025; Received in revised form 27 October 2025; Accepted 10 December 2025
Available online 15 December 2025

Although absolute pose regression methods have achieved initial success, they are susceptible to interference from dynamic objects and noise in autonomous driving scenarios. To solve this problem, the robustness of the model to disturbances needs to be further improved, especially in outdoor environments. The point cloud representation plays a key role. Various representation formats of LiDAR data have been utilized as the inputs for localization, including voxels [7], normal distributions transform (NDT) cells [14], bird's eye views [15,16], and range images [5,12,16,17]. Each representation has its unique advantages in capturing different characteristics of point cloud data. Therefore, how to choose the appropriate representation is crucial to improve the localization accuracy and robustness in large-scale autonomous driving scenarios. In this study, we try to explore the complementarity of multi-modal data and integrate their respective strengths. By fusing complementary modalities, we aim to enhance the algorithm's robustness against dynamic objects, sensor noise, and environmental variations in large-scale scenarios, addressing a critical limitation of single-modal APR methods.

Moreover, the efficiency–accuracy trade-off is a barrier to real-world applicability in large-scale scenarios. High-resolution LiDAR processing retains details but causes prohibitive computational overhead, while lightweight representations reduce latency but lose critical features, leading to localization drift. Inspired by prior work [47], we further design a lightweight multi-modal fusion framework to balance the expressive power of multi-modal data with optimized feature extraction, ensuring the algorithm meets on-board computing constraints while maintaining high precision.

Against this backdrop, the selection of appropriate LiDAR representations becomes a pivotal link in addressing both motivations. It serves as the foundation for multi-modal feature fusion and as a lever for controlling computational overhead.

In this paper, we propose APR-BiCA, an absolute pose regression network that leverages both raw point cloud and range image as complementary inputs to achieve accurate and efficient localization. The raw point cloud preserves fine-grained 3D geometric relationships, and range images capture intuitive 2D structural and depth cues. Therefore, we design a dual branch structure to process these two kinds of data. The point cloud branch uses a point-based network and self-attention to capture key features and global point relationships, while the range image branch employs a residual network to extract robust structural features. Eventually, a fusion module with bidirectional cross-attention and a gating unit enables effective inter-modal interaction, presenting fused features for pose regression. By fusing the strengths of these two inputs, APR-BiCA effectively mitigates the shortcomings of single-modal representation, while avoiding the efficiency–accuracy tradeoff of conventional multi-modal methods. This design enables the network to achieve robust, high-precision localization specifically tailored to the complexity of large-scale autonomous driving scenarios.

Our key contributions are summarized as follows:

- A lightweight, efficient multi-modal pose regression framework named APR-BiCA is proposed for LiDAR localization, addressing the tradeoff between real-time performance and robustness in large-scale scenarios.
- APR-BiCA adopts a dual-branch architecture: one extracts point cloud features via an attention-based module to dynamically weight point-wise features and model global point relationships, while the other extracts range image features to capture structural and depth cues. A bidirectional cross-attention mechanism combined with a gating unit-based fusion module fuses these heterogeneous features, enabling effective inter-modal interaction and reducing multi-source redundancy.
- Extensive experiments on two representative autonomous driving datasets, Oxford RobotCar and NCLT, demonstrate not only APR-BiCA's competitive localization performance but also its notable efficiency.

2. Related work

2.1. Map-based LiDAR localization

Map-based LiDAR localization methods strive to recover the global pose of a query point cloud against a pre-built map by establishing a correspondence matching, which can be divided into local feature-based [18–20] global feature-based [15,16], and semantic-based ones [21–24]. The local feature-based methods extract and match features between the query point cloud and the map, and these methods are primarily suitable for scenarios where the scale of the query point cloud and the map are relatively consistent, such as indoor environments. The global feature-based methods such as BEVplace [15] and BEVPlace++ [16], first use place recognition to identify the most suitable keyframe or submap for the query, and then calculate the pose transformation matrix. The semantic-based methods first conduct object detection or semantic segmentation, and then match the semantic features to achieve localization. For example, in [23], long-term static objects in the scene, such as poles and trunks, are extracted and registered into a semantic cluster map to enhance localization accuracy. However, the considerable expenses for map storage and communication present a barrier to the large-scale deployment of map-based methods.

2.2. Attention for pose regression

The attention mechanism achieves great success in the field of NLP [25,26] and shows great potential in the field of computer vision [27–30,48] due to its strong capability for aggregating features. PointLoc [11] is a pioneering work in the field of LiDAR-based absolute pose regression. This work adopts PointNet++ to learn global features from query point clouds and leverages a self-attention mechanism to further refine these features. TransAPR [31] designs spatial and temporal fusion modules based on Transformer architecture to enable comprehensive feature interaction among the neighboring images in the input sequence, achieving more robust localization under challenging environments. Similarly, STCLoc [34] employs attention-based feature aggregation to enhance features by integrating adjacent frames. The diverse viewpoints from different frames provide complementary contextual cues, thereby enhancing discriminative feature learning and reducing scene ambiguities. The paper [32] employs two separate Transformers to focus on position and orientation features derived from a convolutional backbone. AtLoc [10] proves that the attention can be used to force the network to focus on objects and features that are more geometrically robust. DiffLoc [50] uses attention to construct the basic estimation module, thereby achieving position regression based on conditional generation of poses at the cost of extremely time-consuming training. However, compared to the aforementioned visual-based absolute pose regression, there are still limited studies on LiDAR-based absolute pose regression with attention.

2.3. Absolute pose regression (APR)

Absolute pose regression uses deep learning networks to directly regress poses end-to-end without a pre-built map. Early APR methods primarily concentrated on the visual domain, leveraging convolutional neural networks to extract features from RGB images and directly regress camera poses [33]. However, with the widespread adoption of LiDAR technology and the advancements in point cloud data processing, LiDAR-based APR methods are increasingly emerging as a prominent research focus. PointLoc [11] and the method in [13] are the earliest LiDAR-based global pose regression frameworks. PointLoc's contribution also lies in collecting and creating a new indoor LiDAR-visual sensor dataset, named vReLoc. The paper [13] introduces a universal encoding mechanism to eliminate the need for redundant retraining of the entire network from scratch. STCLoc [34] uses attention-based feature aggregation to capture the correlation in LiDAR sequences and uses a classification task to regularize regression in the spatial dimension. SGLoc [35] separates the pose estimation process into two distinct tasks: regressing

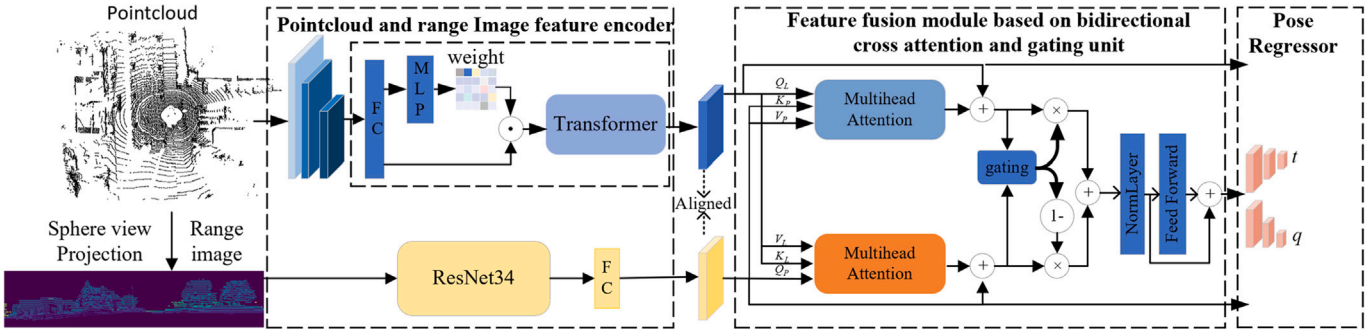

Fig. 1. Overview of the proposed APR-BiCA, consisting of a point cloud and range image feature encoder to extract and align features, a feature fusion module based on bidirectional cross attention and gating unit to extract correlations among the features of two different inputs, and a pose regressor to return 6-DoF poses.

point cloud correspondences and estimating the pose based on these correspondences. NIDALoc [36] incorporates neurobiologically inspired mechanisms into deep networks to enhance the regularization of APR. These studies have demonstrated remarkable achievements, however, compared to map-based localization methods, they still have significant room for improvement in accuracy.

3. Methodology

3.1. Overall framework

In this section, we provide a detailed description of our framework. An overview of the APR-BiCA architecture is illustrated in Fig. 1. It consists of three modules: a point cloud and range image feature encoder (PRFE), a feature fusion module based on bidirectional cross attention and gating unit (FFM), and a pose regressor. PRFE has two branches that extract features from the point cloud and range image, respectively. FFM extracts correlations among the features of two different inputs to achieve effective feature fusion. The pose regressor generates 6-DoF poses.

3.2. Point cloud and range image feature encoder

Point cloud encoding module: We use PointNet++ [34] to encode the point cloud into high-level features. PointNet++ resolves point cloud permutation invariance through symmetric function operators. Its hierarchical architecture contains three components: (1) Sampling layer, using iterative farthest point sampling to choose key points; (2) Grouping layer, forming clusters of points based on a radius around the sampled points; (3) PointNet [35] layer, extracting local features from each cluster. All extracted local features are then aggregated into a global feature F. To further refine the point cloud features extracted by PointNet++, we propose an attention feature learning module that enhances point cloud global feature representations. This module operates in two stages. A multilayer perceptron (MLP) with shared weights first computes importance scores via Sigmoid activation, which are then multiplied element-wise with the original point features. Then we use a transformer [25] to learn the inter-feature dependencies captured by the self-attention mechanism. Attention feature learning module can be expressed as follows:

$$F_L = \mathrm{Transformer}(\sigma(\mathrm{W}_2\mathrm{ReLU}(\mathrm{W}_1\mathrm{F}))\mathrm{F}) \tag{1}$$

where W_1 and W_2 are weight matrices of the MLP, σ denotes the sigmoid function.

Spherical view projecting and encoding module: Range images, which are dense but lightweight, can effectively preserve the distance information of the scene and perform well in many computer vision tasks [5,17,37,38]. A range image is generated from the point cloud through spherical projection [17], which converts the Cartesian coordinates of the point cloud into spherical coordinates, and map points to the pixel positions of the range image. A point $p = (x, y, z) \in R^3$ can be projected to a range image $I \in R^2$ by Eq. (2), and its corresponding pixel value can be computed by Eq. (3).

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan 2(y, x)/\pi\right] w \\ \left[1 - \left(\arcsin(z/r) + f_{up}\right)/f\right] h \end{pmatrix} \tag{2}$$

$$I(u, v) = \sqrt{x^2 + y^2 + z^2} \tag{3}$$

where w, h, and r are the width, height, and depth of the range image, respectively. The vertical field-of-view of the LiDAR sensor is defined as $f = f_{up} + f_{down}$, where f_{up} and f_{down} are the upward and downward field-of-view angles, respectively. Mature CNN architectures can be applied to quickly extract features from range images. Therefore, we use a pre-trained ResNet34 [39] as the backbone to extract features.

It is worth noting that the features generated by the point cloud encoding module and ResNet34 have different sizes. This dimensional mismatch not only hinders the establishment of direct feature correspondence but also compromises the representational consistency required for robust modality interaction. To guarantee the effective subsequent cross-modal interaction and feature fusion, we introduce a linear transformation layer to map the range image features to the size same as the point cloud features. The alignment process preserves the original feature distribution while enabling subsequent cross-attention mechanisms to operate in a unified feature space.

3.3. Feature fusion module based on bidirectional cross attention and gating unit

Direct concatenation of point cloud and range image features may lead to feature redundancy or conflicts, especially when there is a significant difference in the distribution of the two features. An efficient fusion strategy is needed to simultaneously integrate point cloud and range image features. Inspired by the application of cross attention in multimodal fusion [42–44,49], we propose a feature fusion module based on bidirectional cross attention and a gating unit (FFM). The FFM works collaboratively through bidirectional cross attention and an adaptive gating mechanism, using cross attention to capture the interdependence between point cloud features and range image features, and dynamically weighting the features through a gating unit to obtain more robust feature representations, avoiding redundancy caused by direct concatenation or other simple methods.

Given the LiDAR point cloud features $F_L \in R^{N\times D}$ and the range image features $F_P \in R^{N\times D}$ extracted by the PRFE, the bidirectional cross attention computation for the point cloud features is formalized as:

$$h_i^{L\to P} = \mathrm{softmax}\left(\frac{(F_L W_i^Q)(F_P W_i^K)^T}{\sqrt{d/h}}\right)(F_P W_i^V) \tag{4}$$

$$F'_L = F_L + \text{Concat}\left(h_1^{(L\to P)}, \ldots, h_h^{(L\to P)}\right) W^o \tag{5}$$

and that for range image features is calculated as:

$$h_i^{P\to L} = \text{softmax}\left(\frac{(F_P W_i^Q)(F_L W_i^K)^\top}{\sqrt{d/h}}\right)(F_L W_i^V) \tag{6}$$

$$F'_P = F_P + \text{Concat}\left(h_1^{(P\to L)}, \ldots, h_h^{(P\to L)}\right) W^o \tag{7}$$

where W_i^Q, W_i^K, W_i^V are the weight matrices of the i-th attention head, W_i^O is the output weight matrix, h is the number of attention heads, and d denotes the feature dimension.

To dynamically balance contributions from both modalities, a learnable gating unit is introduced. It concatenates the bidirectional cross attention outputs and generates adaptive weights by a sigmoid-activated linear layer to balance the contributions from both modalities:

$$\rho = \sigma\left(w_g(F'_L \odot F'_P) + b_g\right) \tag{8}$$

where w_g is the gating weight matrix, σ denotes the sigmoid function, and $\odot$ represents feature concatenation.

We use gating weights to combine the enhanced features F'_L and F'_P, enabling the network to automatically emphasize reliable features and suppress conflicting signals:

$$F_f = \rho F'_L + (1-\rho)F'_P \tag{9}$$

The fused features F_f are further processed by Layer Normalization (LN) and Feed-Forward Network (FFN):

$$F_F = LN\left(F_f\right) + FFN\left(LN\left(F_f\right)\right) \tag{10}$$

The FFM module identifies the connections between the point cloud features and the image features by exploiting their spatial relationships, which significantly boosts the performance of our method.

3.4. Pose regressor and loss function

The pose regressor contains three regression heads for point cloud features F_L, range image features F_P, and fused features F_F, respectively. Each regression head contains a position regression branch and an orientation regression branch. Each branch consists of multiple fully connected layers, activation function layers and dropout layers. The parameters within the network are optimized using L1 loss based on the following loss function.

$$l = \left\|\mathrm{t} - \hat{\mathrm{t}}\right\| \mathrm{e}^{-\alpha} + \alpha + \left\|\log q - \log\hat{q}\right\| \mathrm{e}^{-\beta} + \beta \tag{11}$$

For the regression results obtained from each of the three regression heads in the pose regressor, the loss is calculated separately to obtain the total loss of the absolute pose regression network, which is calculated as follows

$$l_{total} = l_L + l_P + l_F \tag{12}$$

where (t, q) represents the ground truth pose and $(\hat{t}, \hat{q})$ the predicted one; α and β are learnable factors to jointly learn translation and rotation.

4. Experiments

4.1. Benchmark datasets

The Oxford RobotCar dataset [40] is an outdoor autonomous driving dataset collected by the University of Oxford. It was collected in January 2019 over 32 times on a route in Oxford city center for a total of 280 km in various weather, traffic and lighting conditions. It includes image data, LiDAR data, GPS and INS data, and processed odometry data. The 3D LiDAR point cloud is captured by a Velodyne HDL-32 E, and ground truth pose is obtained through interpolation from the INS.

Table 1
Dataset details on the Oxford dataset and the NCLT dataset.

Scene	Date	Sky	Use as	Scene	Date	Sky	Use as
Full-1	2019–01–11–14-02–26	sun	Train	Seq-1	2012–01-22	cloud	Train
Full-2	2019–01-14–12-05–52	overcast		Seq-2	2012–02-02	sunny	
Full-3	2019–01-14–14-48–55	overcast		Seq-3	2012–02-18	sunny	
Full-4	2019–01-18–15-20–12	overcast		Seq-4	2012–05-11	sunny	
Full-6	2019–01-10–11-46–21	rain	Test	Seq-6	2012–02-12	sunny	Test
Full-7	2019–01-15–13-06–37	overcast		Seq-7	2012–02-19	cloudy	
Full-8	2019–01-17–14-03–00	sun		Seq-8	2012–03-31	cloudy	
Full-9	2019–01-18–14-14–42	overcast		Seq-9	2012–05-26	sunny	

Table 2
Detailed parameters of Pointnet++ encoder.

Layer	Point num	Radius	Sample	MLP
SA1	384	0.2	24	[64,64,128]
SA2	64	0.4	48	[128,128,256]
SA3	64	–	–	[256,512,1024]

The NCLT dataset [41] is a large-scale, long-term dataset collected by a Segway mobile robot at the University of Michigan North Campus, which contains one year's worth of data, covering 27 different trajectories in a fixed route with varying seasons, traffic conditions, lighting, vegetation, and weather conditions, both in indoor and outdoor environments. It consists of data from the 3D LiDAR, planar LiDAR, omnidirectional camera, and GPS. The 3D LiDAR point cloud is collected by a Velodyne HDL-32 E, with ground truth pose obtained through the interpolation of SLAM.

We use the same training and testing sequences as the previous work on the two datasets[31][33][44], as shown in Table 1.

4.2. Implementation details

The network is implemented with Pytorch [45] using an Adam [46] optimizer, an initial learning rate of 1×10^{-4}, and a weight decay of 5×10^{-4}. All of our experiments are conducted on NVIDIA RTX 4090 GPU. We reduce the point cloud to 4096 points using farthest point sampling, and use all points within 80 metres to generate a range image. For the transformer module, we set the embedding dimension to 512 and the number of heads to 4. For the bidirectional cross attention, the embedding dimension is set to 512 and the number of heads to 8. The detailed parameters of the PointNet++ encoder for APR-BiCA are shown in Table 2. We train the network for 100 epochs. It takes about 9 minutes to train one epoch on the Oxford dataset and about 7 minutes on the NCLT dataset.

4.3. Baselines

We compare APR-BiCA with several recent absolute pose regression methods to demonstrate its validity and performance, including PointLoc [11], PosePN [13], PosePN++ [13], PoseSOE [13], PoseMinkLoc [13], STCLoc [34], NIDALoc [36], and Diffloc [50]. All results are from the original paper or reported in the same environment using the source code.

4.4. Main results

Same as the previous APR methods, we report the mean absolute position error (m) and mean absolute orientation error (°) on the Oxford dataset and NCLT dataset. To enable a more equitable comparison of performance across different models, we have additionally reported each algorithm's parameter count and training duration.

Results on the Oxford Dataset: The comparison of the proposed APR-BiCA with other methods on the Oxford dataset test set is presented in Table 3 (where the best results are highlighted in bold and the second best are underlined). On this dataset, the DiffLoc achieves

Table 3
Mean absolute position error (m) and orientation error (°) on the Oxford dataset. The best results are highlighted in bold and the second best are highlighted with an underline. APR-BiCA achieves the second-best performance in most metrics.

Methods	Training Time	Params	Full-6	Full-7	Full-8	Full-9	Average	Ranks
PointLoc	125 h	13 M	13.81 m,1.53°	9.81 m,1.27°	11.51 m,1.34°	9.51 m,1.07°	11.16 m,1.30°	6/4
PosePN	5 h	4 M	16.32 m,2.43°	14.32 m,3.06°	13.48 m,2.60°	9.14 m,1.78°	13.32 m,2.47°	8/7
PosePN + +	11 h	5 M	10.64 m,1.78°	9.59 m,1.92°	9.01 m,1.51°	8.44 m,1.71°	9.42 m,1.73°	5/5
PoseSOE	15 h	5 M	8.81 m,2.04°	7.59 m,1.94°	12.35 m,2.46°	7.27 m,1.87°	8.22 m,1.99°	4/6
PoseMinkLoc	7 h	13 M	11.20 m,2.62°	14.69 m,2.90°	12.35 m,2.46°	10.06 m,2.15°	12.08 m,2.53°	7/8
STCLoc	20 h	9 M	7.68 m,1.30°	6.93 m,1.48°	7.44 m,1.24°	6.13 m,1.15°	7.05 m,1.29°	3/3
DiffLoc	145 h	40 M	**3.65 m,0.68°**	**3.57 m,0.88°**	**4.03 m,0.70°**	**2.86 m,0.60°**	**3.53 m,0.72°**	1/1
APR-BiCA	15 h	33 M	8.48 m,1.21°	7.05 m,1.08°	6.38 m,0.99°	3.92 m,0.69°	6.46 m,0.99°	2/2

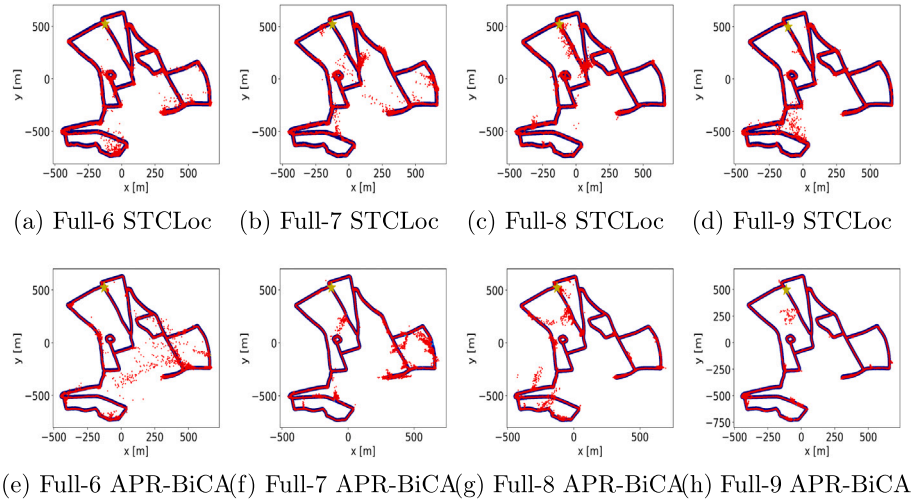

Fig. 2. Trajectories of APR-BiCA and STCLoc on the Oxford dataset. The ground truth and the predicted results are shown in blue and red, respectively, with the star indicating the starting position.

the top performance across all metrics, with an average positional error of 3.53 m and an average angular error of 0.72°. Meanwhile, APR-BiCA achieves second-place rankings in all average metrics, with an average positional error of 6.68 m and an average angular error of 1.04°. This performance demonstrates strong competitiveness compared to the third-ranked STCLoc, which reports an average positional error of 7.05 m and an average angular error of 1.29°. It is worth noting that the STCLoc outperforms our method in two positional errors. We attribute this to its spatio-temporal attention framework which we believe can eliminate environmental noise by comparing inter-frame consistency, supplement sparse texture information through historical frame features, and therefore directly reduce position prediction errors. Notably, APR-BiCA maintains this competitive accuracy while balancing model efficiency. It has 33 M parameters and requires only 15 hours of training time, which is far less than DiffLoc (40 M parameters, 145 hours of training) and slightly more efficient than STCLoc (9 M parameters, 20 hours of training).

The estimated trajectories of APR-BiCA and STCLoc on the Oxford dataset are shown in Fig. 2. The ground truth and predictions are shown in blue and red, respectively, with a star indicating the starting position. The APR-BiCA predicted trajectories overlap closely with the ground truth, demonstrating the accuracy of the localization.

Results on the NCLT Dataset: We also evaluated the APR-BiCA on the NCLT dataset, with results reported in Table 4 (following the same bold/underline annotation to indicate the top and second-best results). Similar to the Oxford dataset, DiffLoc demonstrates the best performance, with an average positional error of 1.06 m and an average angular error of 2.32°. The APR-BiCA again ranks second in all metrics, retaining strong accuracy with an average positional error of 2.56 m and an average angular error of 3.15°. Compared to key competitor NIDALoc, the third-best method, with an average positional error of 4.47 m and an average angular error of 3.53°), the APR-BiCA improves position estimation accuracy by 42.7 % and orientation estimation accuracy by 10.8 %. Importantly, the APR-BiCA maintains this high localization accuracy while excelling in efficiency. The APR-BiCA requires only 12 hours of training time and has 33 M parameters, far more efficient than DiffLoc (100 hours of training, 40 M parameters).

Fig. 3 shows the estimated trajectories of APR-BiCA and NIDALoc on the NCLT dataset. It can be seen that our predictions are closer to the true values than the second best performing method, NIDALoc.

As the point cloud can provide 3D spatial information while its corresponding range image can provide depth information, we use both of them as inputs, enabling more precise capture of object geometry. The feature fusion module based on bidirectional cross attention and gating unit enhances the accuracy and robustness of absolute pose regression, achieving or even exceeding the performance of multi-frame sequential techniques. The above results demonstrate that APR-BiCA is effective for localization in large-scale outdoor environments and mixed indoor and outdoor scenes.

Overall, these results confirm that APR-BiCA achieves reliable localization across both the Oxford dataset (outdoor urban scenes) and the NCLT dataset (mixed indoor-outdoor scenes), and its balanced accuracy and efficiency meet the practical deployment needs of real-world autonomous driving systems.

4.5. Ablation study

We validate the effectiveness of each module through ablation experiments. The results are shown in Table 5. Each module contributes to the improvement of estimation accuracy, especially the bidirectional

Table 4
Mean absolute position error (m) and orientation error (°) on the NCLT dataset. The best results are highlighted in bold and the second best are highlighted with an underline. APR-BiCA achieves the second-best performance in all metrics.

Methods	Training Time	Params	Seq-6	Seq-7	Seq-8	Seq-9	Average	Ranks
PointLoc	125 h	13 M	7.23 m,4.88°	6.13 m,3.89°	6.71 m,4.32°	10.02 m,5.32°	7.57 m,4.60°	7/6 W
PosePN	5 h	4 M	9.45 m,7.47°	6.15 m,5.05°	5.79 m,5.28°	13.47 m,7.77°	8.72 m,6.39°	8/9
PosePN + +	11 h	5 M	4.97 m,3.75°	3.68 m,2.65°	4.35 m,3.38°	9.59 m,4.49°	5.65 m,3.57°	5/4
PoseSOE	15 h	5 M	13.09 m,8.05°	6.16 m,4.51°	5.24 m,4.56°	12.60 m,7.67°	9.27 m,6.20°	8/8
PoseMinkLoc	7 h	13 M	6.24 m,5.03°	4.87 m,3.94°	4.23 m,4.03°	10.32 m,6.52°	6.42 m,4.88°	6/7
STCLoc	20 h	9 M	4.91 m,4.34°	3.25 m,3.10°	3.75 m,4.04°	8.67 m,5.23°	5.15 m,4.18°	4/5
NIDALoc	14 h	8 M	4.48 m,3.59°	3.14 m,2.52°	3.67 m,3.46°	6.60 m,4.56°	4.47 m,3.53°	3/3
DiffLoc	100 h	40 M	**0.99 m,2.40°**	**0.92 m,2.14°**	**0.98 m,2.27°**	**1.36 m,2.48°**	**1.06 m,2.32°**	1/1
APR-BiCA	12 h	33 M	2.19 m,3.52°	2.02 m,2.50°	1.90 m,2.92°	4.16 m,3.27°	2.57 m,3.05°	2/2

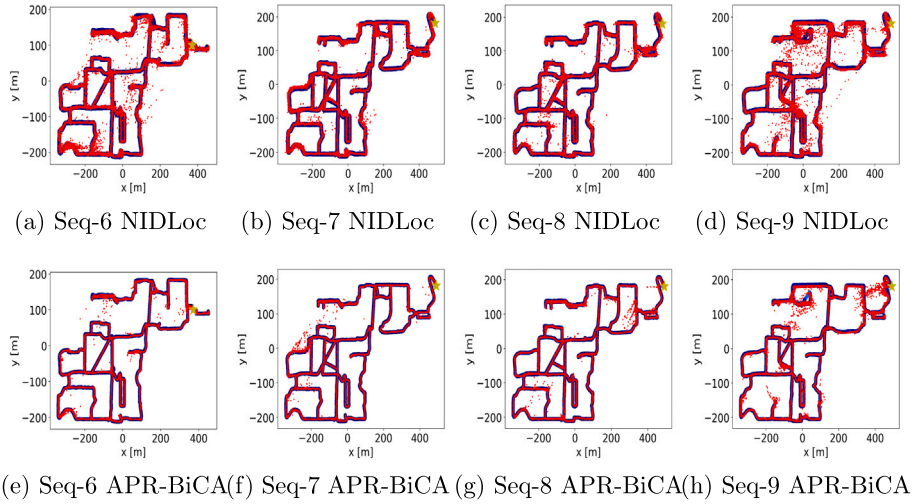

Fig. 3. Trajectories of our network on the NCLT dataset. The ground truth and the predicted results are represented in blue and red, respectively, with the star indicating the starting position.

Table 5
Ablation study of different proposed modules. The average errors of all trajectories on the Oxford datasets are reported.

No.	Attention Feature Learning Module	Bidirectional Cross Attention	Gating Unit	Mean Error (m, °)
1		✓	✓	6.89 m, 1.09°
2	✓		✓	6.93 m, 1.38°
3	✓	✓		6.60 m, 1.18°
4	✓			7.80 m, 1.39°
5		✓		7.54 m, 1.24°
6			✓	7.10 m, 1.39°
7	✓	✓	✓	**6.46 m, 0.99°**

cross attention module which brings significant improvement in both position and orientation accuracies.

We also conduct ablation experiments on different branches. The results are shown in Table 6. The cumulative distributions of position and orientation errors on Full-8 trajectory are shown in Fig. 4. The cumulative distributions of position and orientation errors on Seq-8 trajectory are shown in Fig. 5. We can observe that the localization performance achieved with fused features outperforms that achieved with only point cloud or range image features. This demonstrates the effectiveness of the feature fusion module and highlights the fact that the point cloud and range image focus on different aspects of information representation. The combination of both enhances the richness of feature representation.

To verify the necessity of our multi task loss strategy, we conducted ablation experiments and compared different loss configurations, as shown in Table 7. The results of Experiments No. 1, 2, and 3 collectively clarify the critical role of range image branch loss supervision. Experiments No. 1 and No. 2 both failed to converge effectively due to the lack of range image branch supervision, while Experiment No. 3 achieved stable convergence. This contrast confirms that range image supervision is indispensable for extracting high-quality 2D structural cues. Additionally, Experiment No. 4 which incorporates all three losses attained the optimal performance, with an average position/orientation error of 6.46 m/0.99°. This highlights that multitasking strategies ensure hierarchical optimization. A single loss refines modality specific features, while the fusion loss guides the FFM to integrate these high-quality

Table 6
Ablation study of different proposed branches.

Different branches	Oxford	NCLT
point cloud feature	12.01 m,1.67°	8.85 m,6.56°
range image feature	8.57 m,1.99°	2.57 m,3.76°
fused feature	**6.38 m,0.99°**	**1.90 m,2.92°**

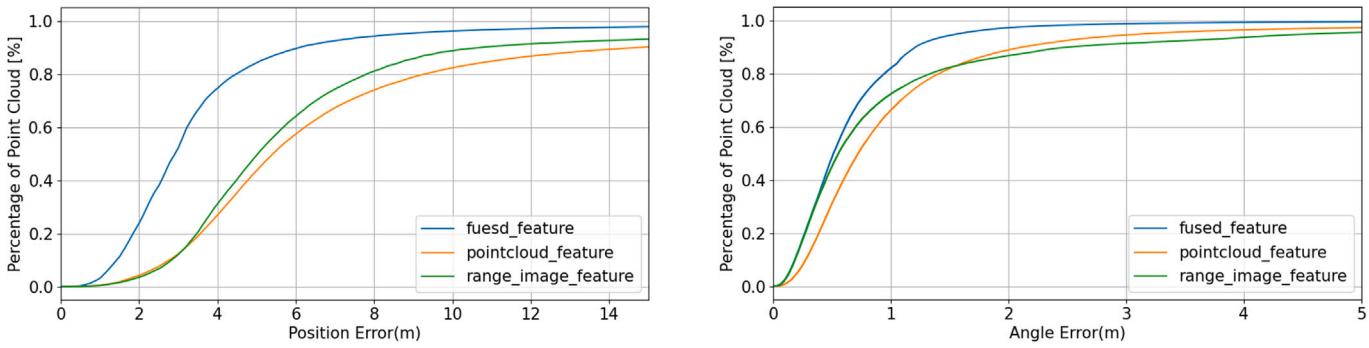

Fig. 4. Cumulative distributions of position errors (m) and angle errors (°) on Oxford dataset. The x-axis represents the error, and the y-axis represents the percentage of point clouds with errors less than the value.

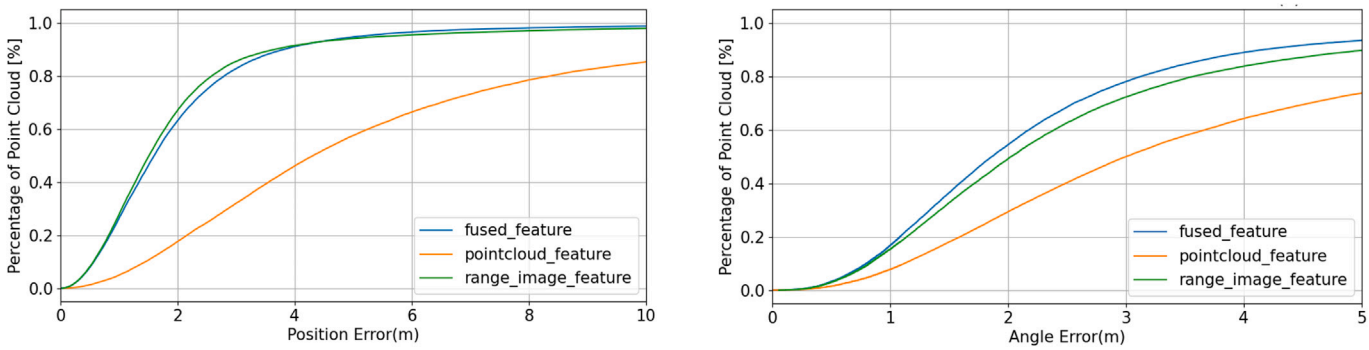

Fig. 5. Cumulative distributions of position errors (m) and angle errors (°) on NCLT dataset. The x-axis represents the error, and the y-axis represents the percentage of point clouds with errors less than the value.

Table 7
Ablation study of different loss combinations. The average errors of all trajectories on the Oxford datasets are reported.

No.	Fussion loss	Lidar loss	Range image loss	Average	Training Convergence
1	✓			434.55 m,90.25°	Failed
2	✓	✓		434.48 m,89.48°	Failed
3	✓		✓	7.60 m,1.84°	Converged
4	✓	✓	✓	6.46 m,0.99°	Converged

Table 8
Ablation study of different hyperparameters. The average errors of all trajectories on the Oxford datasets are reported.

No.	Number head	Number sampled	Average
1	16	4096	7.20 m,1.10°
2	4	4096	7.65 m,1.08°
3	8	4096	6.46 m,0.99°
4	8	2048	8.03 m,1.82°

single-modal features via bidirectional cross-attention and gating unit. The results clearly confirm that multi task supervision is indispensable for stable convergence and accurate attitude regression.

To enhance the APR-BiCA's robustness and interpretability, we conducted ablation experiments on the Oxford RobotCar dataset, focusing on two core hyperparameters: the FFM's attention head count for inter-modal fusion and the PRFE downsampled LiDAR point count for geometric detail retention. The results are shown in Table 8.

With the number of PRFE downsampled points fixed at 4096, the configuration with 16 attention heads in FFM results in redundant inter-modal correlation calculations, while the 4-head configuration fails to capture adequate inter-modal correlations. In contrast, the 8-head configuration balances correlation granularity and computational efficiency, achieving the lowest average error (6.46 m/0.99°) in this experimental group.

With the number of FFM attention heads fixed at 8, the configuration with 2048 PRFE downsampled points suffered from insufficient geometric details, resulting in an average error of 8.03 m/1.82°. The 8192-point configuration introduced redundant noise, leading to an average error of 7.89 m/1.74°. Different from that, the 4096-point configuration balanced geometric detail retention and noise suppression.

4.6. Computational time

The runtime is crucial for the localization framework and it is recommended that the processing time per frame be shorter than the acquisition interval of the sensor. The LiDAR scanning frequencies for the Oxford and NCLT datasets are 20 Hz [40] and 10 Hz [41], respectively. Our network takes only 1.3 ms per frame to run from input to output on the Oxford dataset, and 1.5 ms on the NCLT dataset (note that this is for frames sampled to 4096 points). Thus, APR-BiCA can achieve real-time localization with good performance.

5. Conclusion

This study presents a LiDAR-based absolute pose regression (APR) framework APR-BiCA for real-time and robust autonomous driving localization, mainly consisting of a dual-branch Point Cloud and Range Image Feature Encoder (PRFE), a Feature Fusion Module (FFM), and a

multi-head pose regressor. The PRFE is designed to address the incompleteness of single-modal feature representation through a dual-branch structure. The FFM captures correlations between two types of data and balances the contribution of each modality to avoid conflicting signals. The multi-head pose regressor paired with L1 loss further optimizes the accuracy of 6-DoF pose estimation by jointly learning translation and rotation. Ablation experiments on the Oxford dataset fully validate each core module's effectiveness. Extensive experiments on two benchmark datasets confirm its effectiveness. On the Oxford dataset, it achieves an average position error of 6.68 m and an orientation error of 1.04°. On the NCLT dataset, these errors drop to 2.56 m and 3.15°, respectively. Notably, the APR-BiCA maintains high accuracy with low computational costs as its training time is one ninth that of DiffLoc. This balance of performance and efficiency validates its practical value for real-world autonomous driving systems, while also laying a foundation for future research on cross-scene APR generalization. It is worth noting that our algorithm still has significant room for improvement in terms of error performance, which will be the focus of our further research in the future.

CRediT authorship contribution statement

Jianlong Dai: Writing – review & editing, Writing – original draft, Methodology. **Hui Wang:** Writing – review & editing, Validation. **Yuqian Zhao:** Writing – review & editing. **Zhihua Liu:** Writing – original draft, Methodology.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (Grant nos. U23B2063 and 62076256), the Key R&D Plan Program of Hunan Province (Grant No. 2023GK2021 and 2024JK2028), the Scientific Research Project of Hunan Provincial Education Department (23A0011), and 111 Project (Grant No. B18059). We are grateful for the resources from the High Performance Computing Center of Central South University.

Data availability

Data will be made available on request.

References

[1] H. Yin, X. Xu, S. Lu, X. Chen, R. Xiong, S. Shen, C. Stachniss, Y. Wang, A survey on global lidar localization: challenges, advances and open problems, Int. J. Comput. Vis. 132 (8) (2024) 3139–3171.
[2] J. Zhang, Y. Zhang, L. Rong, R. Tian, S. Wang, MVSE-net: a multi-view deep network with semantic embedding for lidar place recognition, IEEE Trans. Intell. Transp. Syst. 25 (2024) 17174–17186.
[3] Z. Hou, Y. Yan, C. Xu, H. Kong, Hitpr: hierarchical transformer for place recognition in point cloud, in: Proc. IEEE Int. Conf. Robot. Autom, *ICRA)*, 2022, pp. 1–2.
[4] Q. Zou, Q. Sun, L. Chen, B. Nie, Q. Li, A comparative analysis of Lidar SLAM-based indoor navigation for autonomous vehicles, IEEE Trans. Intell. Transp. Syst. 23 (7) (Jul. 2022) 6907–6921.
[5] J. Ma, et al., Cvtnet: a cross-view transformer network for lidar-based place recognition in autonomous driving environments, IEEE Trans. Ind. Informat. 20 (3) (2023) 4039–4048.
[6] L. Hui, H. Yang, M. Cheng, J. Xie, J. Yang, Pyramid Point cloud transformer for large-scale place recognition, in: Proc. IEEE Int. Conf. Comput. Vis, *ICCV)*, 2021, pp. 6098–6107.
[7] J. Komorowski, Minkloc3D: point cloud based large-scale place recognition, in: Proc. IEEE Winter Conf. Appl. Comput. Vis, *(WACV)*, 2021, pp. 1790–1799.
[8] D. Cattaneo, M. Vaghi, A. Valada, Lcdnet: deep loop closure detection and point cloud registration for Lidar SLAM, IEEE Trans. Robot. 38 (4) (Aug. 2022) 2074–2093.
[9] J. Zhu, K. Yang, Y. Zhang, Y. Peng, Y. Peng, APFN: adaptive perspective-based fusion network for 3-d place recognition, IEEE Trans. Instrum. Meas. 73 (2024) 1–10.
[10] B. Wang, C. Chen, C. Lu, P. Zhao, N. Trigoni, A. Markham, Atloc: attention guided camera localization, in: Proc. AAAI Conf. Artif. Intell (AAAI), 2020, pp. 10393–10401.
[11] W. Wang, B. Wang, P. Zhao, C. Chen, R. Clark, B. Yang, A. Markham, N. Trigoni, Pointloc: deep pose regressor for lidar point cloud localization, IEEE Sens. J. 22 (1) (2021) 959–968.
[12] S. Wang, Q. Kang, R. She, W. Wang, K. Zhao, Y. Song, W.P. Tay, Hypliloc: towards effective lidar pose regression with hyperbolic fusion, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, *CVPR)*, 2023, pp. 5176–5185.
[13] S. Yu, C. Wang, C. Wen, M. Cheng, M. Liu, Z. Zhang, X. Li, Lidar-based localization using universal encoding and memory-aware regression, Pattern Recognit. 128 (2022) 108915.
[14] Z. Zhou, C. Zhao, D. Adolfsson, S. Su, Y. Gao, T. Duckett, L. Sun, NDT-transformer: large-scale 3d point cloud localization using the normal distribution transform representation, in: Proc. IEEE Int. Conf. Robot. Autom, *ICRA)*, 2021.
[15] L. Luo, et al., Bevplace: learning lidar-based place recognition using bird's eye view images, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, *ICCV)*, 2023.
[16] L. Luo, et al., BEVPlace++: Fast, Robust, and Lightweight LiDAR Global Localization for Unmanned Ground Vehicles, arXiv preprint arXiv:2408.01841, 2024).
[17] P. Yin, L. Xu, J. Zhang, H. Choset, Fusionvlad: a multi-view deep fusion networks for viewpoint-free 3d place recognition, IEEE Robot. Autom. Lett. 6 (2) (2021) 2304–2310.
[18] Y. Wang, J.M. Solomon, Deep closest point: learning representations for point cloud registration, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, *ICCV)*, 2019, pp. 3523–3532.
[19] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, K. Schindler, PREDATOR: registration of 3d point clouds with low overlap, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, *CVPR)*, Jun. 2021, pp. 4267–4276.
[20] L. Li, W. Tang, T. Xue, et al., "iterative btreenet: unsupervised learning for large and dense 3d point cloud registration, Neurocomputing 506 (2022) 336–354.
[21] L. Li, X. Kong, X. Zhao, T. Huang, W. Li, F. Wen, H. Zhang, Y. Liu, Rinet: efficient 3d lidar-based place recognition using rotation invariant neural network, in: IEEE Robot. Autom. Lett, 2022, pp. 1.
[22] G. Pramatarov, D. Martini, M. Gadd, P. Newman, Boxgraph: semantic place recognition and pose estimation from 3d lidar, in: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst, *IROS)*, 2022, pp. 7004–7011.
[23] K. Wei, P. Ni, X. Li, Y. Hu, W. Hu, Lidar-based semantic place recognition in dynamic urban environments, IEEE Sens. J. 24 (2024) 28397–28408.
[24] Z. Wang, et al., Pole-like objects mapping and long-term robot localization in dynamic urban scenarios, in: Proc. IEEE Int. Conf. Robot, *Biomimetics (ROBIO)*, 2021.
[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. Conf. North Amer. Chapter Assoc. Comput, *Hum. Lang. Technol*, *Linguistics*, 2019, pp. 4171–4186.
[26] A. Vaswani, N.M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. Adv. Neural Inf. Process. Syst, 2017.
[27] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proc. Eur. Conf. Comput. Vis, 2020, pp. 213–229.
[28] A. Dosovitskiy, et al., An image is worth 16x16 words: transformers for image recognition at scale, in: Proc. Int. Conf. Learn. Representations, 2021.
[29] Y. Hu, B. Li, C. Xu, S. Saydam, W. Zhang, "featsync: 3d point cloud multiview registration with attention feature-based refinement, Neurocomputing 600 (2024) 128088.
[30] L. Guo, Z. Chen, S. Cheng, et al., "learning compact and overlap-biased interactions for point cloud registration, Neurocomputing 598 (2024) 127949.
[31] C. Qiao, Z. Xiang, Y. Fan, T. Bai, X. Zhao, J. Fu, Transapr: absolute camera pose regression with spatial and temporal attention, IEEE Robot. Autom. Lett. 8 (8) (Aug. 2023) 4633–4640.
[32] Y. Shavit, R. Ferens, Y. Keller, Coarse-to-fine multi-scene pose regression with transformers, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 14222–14233.
[33] A. Kendall, M. Grimes, R. Cipolla, Posenet: a convolutional network for real-time 6-dof camera relocalization, in: Proc. IEEE Int. Conf. Comput. Vis, 2015, pp. 2938–2946.
[34] S. Yu, C. Wang, Y. Lin, C. Wen, M. Cheng, G. Hu, Stcloc: deep lidar localization with spatio-temporal constraints, IEEE Trans. Intell. Transp. Syst. 24 (2023) 489–500.
[35] W. Li, S. Yu, C. Wang, G. Hu, S. Shen, C. Wen, Sgloc: scene geometry encoding for outdoor lidar localization, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, *CVPR)*, 2023, pp. 9286–9295.
[36] S. Yu, X. Sun, W. Li, C. Wen, Y. Yang, B. Si, G. Hu, C. Wang, Nidaloc: neurobiologically inspired deep lidar localization, IEEE Trans. Intell. Transp. Syst. 25 (2024) 4278–4289.
[37] L. Fan, X. Xiong, F. Wang, N.-L. Wang, Z. Zhang, Rangedet: in defense of range view for lidar-based 3d object detection, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, *ICCV)*, 2021, pp. 2898–2907.
[38] Z. Liang, Z. Zhang, M. Zhang, X. Zhao, S. Pu, Rangeioudet: range image based real-time 3d object detector optimized by intersection over union, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, *CVPR)*, 2021, pp. 7136–7145.
[39] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, 2016, pp. 770–778.
[40] D. Barnes, M. Gadd, P. Murcutt, P. Newman, I. Posner, The Oxford Radar Robotcar dataset: a radar extension to the Oxford Robotcar dataset, in: Proc. IEEE Int. Conf. Robot. Autom, *ICRA)*, 2020, pp. 6433–6438.

[41] N. Carlevaris-Bianco, A.K. Ushani, R.M. Eustice, University of Michigan North campus long-term vision and lidar dataset, Int. J. Robot. Res. 35 (9) (2016) 1023–1035.
[42] J. Shen, Y. Chen, Y. Liu, et al., Icafusion: iterative cross-attention guided feature fusion for multispectral object detection, Pattern Recognit. 145 (2024) 109913.
[43] C.K. Yang, M.H. Chen, Y.Y. Chuang, et al., 2d-3d interlaced transformer for point cloud segmentation with scene-level supervision, in: Proc. IEEE/CVF Int. Conf. Comput. Vis, *ICCV)*, 2023.
[44] X. Wei, T. Zhang, Y. Li, Y. Zhang, F. Wu, Multi-modality cross attention network for image and sentence matching, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit, *CVPR)*, 2020, pp. 10938–10947.
[45] A. Paszke, et al., Pytorch: an imperative style, high-performance deep learning library, in: Proc. Adv. Neural Inf. Process. Syst, 2019, pp. 8024–8035.
[46] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proc. Int. Conf. Learn, *Representations (ICLR)*, 2015, pp. 1–15.
[47] Y. Zhang, B. Liu, J. Bao, Q. Huang, M. Zhang, J. Yu, Learnability matters: active learning for video captioning, Adv. Neural Inf. Process. Syst. 37 (2024) 37928–37954.
[48] F. Cui, Y. Zhang, X. Wang, X. Tian and J. Yu, "Enhancing Target-unspecific Tasks through a Features Matrix", arXiv preprint, arxiv:2505.03414, 2025.
[49] J. Hu, Y. Bai, Z. Li, Q. Jin, C. Peng, Multi-source attention autoencoder network for hyperspectral unmixing with lidar data, Neurocomputing 623 (2025) 129380.
[50] W. Li, Y. Yang, S. Yu, G. Hu, C. Wen, M. Cheng, C. Wang, "diffloc: diffusion model for outdoor lidar localization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15045–15054.2024.

Author biography

Jianlong Dai received the master's degree from the School of Mechanical Engineering, Central South University, Changsha, China, in 2013. He is currently pursuing a Ph.D. degree in the School of Automation. His research interests include sensing and environmental modeling, navigation and positioning, and robot control.

Hui Wang received B. S. degree from School of Automation, Central South University, China, in 2021. Now he is currently pursuing the Ph. D degree. His research interests are salient object detection, object detection, computer vision and deep learning.

Yuqian Zhao received the M.Eng. degree in Automatic Control Engineering and the Ph.D. degree in Computer Science and Technology from Central South University, Changsha, China, in 2002 and 2006, respectively. He engaged in postdoctoral research at New Jersey Institute of Technology, Newark, New Jersey, USA, from 2009 to 2010. He is currently a Full Professor in Central South University. His research interests include image processing, pattern recognition, computer vision, video analysis, and intelligent sensing.

Zhihua Liu is currently pursuing the master's degree in the School of Automation, Central South University, China. Her research interests include LiDAR-based place recognition, global localization, and deep learning.