FGF-Net: Frequency-Guided Fusion for Hybrid Voxel–Point LiDAR-Based Place Recognition

Lineng Chen[1,2,3], Haiying Xia[1,2], Shuxiang Song[1,2,*]
[1]*Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, School of Computer Science and Engineering, Guangxi Normal University, Guilin 541004, China*
[2]*Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, School of Electronic and Information Engineering, Guangxi Normal University, Guilin 541004, China*
[3]*Guangxi Key Laboratory of Automatic Detecting Technology and Instruments, Guilin University of Electronic Technology, Guilin 541004, China*

***Abstract*—LiDAR-based place recognition is fundamental to large-scale outdoor simultaneous localization and mapping (SLAM) and long-term autonomous navigation. Voxel-based networks are efficient and robust but suffer from information loss induced from quantization, whereas point-based networks preserve geometric details but are sensitive to sensor noise. Recent hybrid voxel–point approaches typically fuse these two representations in the spatial domain, overlooking their complementary properties in the frequency domain. In this paper, we propose FGF-Net, a lightweight frequency-guided fusion network for LiDAR-based place recognition with hybrid voxel-point representations. Specifically, we propose a Frequency-Guided Fusion Module that applies local discrete cosine transform to decompose point-wise features into low- and high-frequency bands and performs an energy-based dual-stream gating mechanism. In this way, voxel-wise features provide semantic priors that enhance reliable high-frequency details in the point branch, while low-frequency cues from the point branch refine voxel-wise representations for stable structural encoding. Extensive experiments on multiple large-scale benchmarks show that FGF-Net achieves competitive performance with only 0.22M parameters and exhibits strong generalization across diverse environments and adverse weather conditions, validating the effectiveness of frequency-guided hybrid fusion for LiDAR-based place recognition.**

***Index Terms*—Place recognition, LiDAR, frequency-guided, global descriptor**

I. Introduction

Place recognition is usually treated as an instance retrieval problem that aims to identify whether the current observation corresponds to a previously visited location. It is a fundamental component of long-term simultaneous localization and mapping (SLAM) [1], loop closure detection [2], and large-scale autonomous navigation [3]. Robust place recognition enables a system to maintain global consistency, reduce accumulated drift, and operate reliably over extended periods and environments. As a result, it has become a core problem in robotics and autonomous driving, particularly in large-scale outdoor scenarios [4]. Compared with cameras, 3D LiDAR sensors provide direct geometric information of the environment and are largely robust to illumination changes, weather variations, and appearance changes [5]. These properties make LiDAR-based place recognition (LPR) especially suitable for large-scale outdoor environments with significant lighting variation and challenging weather. However, the inherent irregularity and sparsity of 3D point clouds make it non-trivial to extract features that are simultaneously robust and discriminative. Existing 3D LPR methods can be roughly divided into voxel-based, point-based, and more recently hybrid voxel-point approaches. Voxel-based methods [6], [7] first discretize raw point clouds into regular voxels and then apply 3D sparse convolutions (SPConvs) [8] to extract features. The voxel structure facilitates efficient large-scale processing and naturally aggregates information within the voxels, making voxel-wise features stable and robust to noise and small perturbations. However, this discretization process inevitably induces quantization errors and blurs fine-grained geometry details [9]. As a result, voxel-wise features tend to capture global consistency and low-frequency structures of the scene, but may miss high-frequency local cues that are essential to discriminate geometrically similar places.

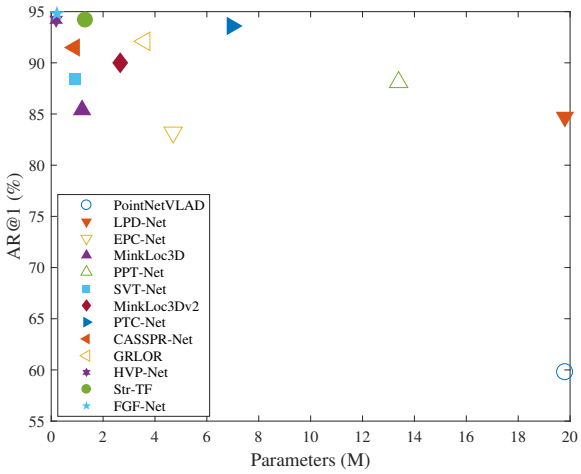

Fig. 1. Comparison of parameters and AR@1 results of different methods.

In contrast, point-based methods [10], [11], operate directly

This work was supported in part by the Guangxi Natural Science Foundation for Youths (2025GXNSFBA069391), in part by Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips (BCIC-24-Z3), in part by 2025 National Natural Science Foundation of China Joint Cultivation Project of GXNU (2025PY047), and in part by Guangxi Key Laboratory of Automatic Detecting Technology and Instruments (YQ26201)

*Corresponding author: Shuxiang Song (songshuxiang@gxnu.edu.cn)

on raw points [12] or local neighborhoods [13]. By avoiding voxelization, they preserve fine-grained geometric details and capture sharp structures such as edges and corners, leading to informative, high-frequency-rich features that are crucial to distinguish places differing only in subtle local patterns. However, point-based pipelines are more sensitive to sensor noise, density variation, and may also amplify undesirable high-frequency noise together with informative high-frequency structures.

To leverage the complementary strengths of voxel- and point-based methods, recent works have explored hybrid voxel–point architectures [14], [15] that combine the stability and global context of voxel-wise features with the fine-grained, discriminative nature of point-wise features. However, there are few methods that investigate the fusion strategy between voxel and point representations for LPR task. Common strategies simply add the two aligned features [15], or directly apply cross-attention mechanisms [14]. These approaches neglect the fact that these features exhibit inherently different frequency characteristics that the voxel features predominantly encode low-frequency, smooth components due to spatial aggregation and down-sampling, whereas point-wise features provide high-frequency local variations that are both fine-grained and noise-prone. Ignoring this spectral difference may lead to either over-smoothing of discriminative high-frequency details or uncontrolled amplification of noisy high-frequency responses during fusion.

To address these limitations, we propose FGF-Net, a lightweight frequency-guided fusion network for hybrid voxel–point LPR. As shown in Fig. 1, our method achieves a more favorable trade-off between recognition accuracy and computational efficiency. Following HVP-Net [15], FGF-Net adopts a dual-branch architecture to extract both voxel- and point-wise. In the voxel branch, the input point cloud is first quantized into a sparse voxel representation and processed by a stack of sparse convolutions (SPConvs) to sequentially produce multi-scale voxel-wise features. These voxel features are then transformed into point representations by a unidirectional feature transformation unit (FTU) [9] and fed into the point branch. Meanwhile, in the point branch, we employ farthest point sampling (FPS) and construct k-nearest neighbor (k-NN) graphs over the sampled points, on which graph convolutions followed by 1D convolutions are applied to learn multi-scale point-wise features.

To effectively fuse voxel- and point-wise features, we propose a Frequency-Guided Fusion Module (FGFM), which exploits frequency-domain characteristics to guide cross-representation feature fusion. Specifically, FGFM applies a 1D discrete cosine transform (DCT) to point-wise features to decompose them into high- and low- frequency bands, and performs fusion with a dual-stream energy-aware gating mechanism. On one hand, the Semantic-Aware High-Frequency Enhancement (SA-HFE) module uses voxel-wise features as semantic priors to predict gates that modulate the high-frequency coefficients of the point-wise features, enabling FGF-Net to selectively enhance reliable local details while suppressing unstable high-frequency noise. On the other hand, the Consistency-Aware Low-Frequency Enhancement (CA-LFE) module exploits the low-frequency consistency between voxel- and point-wise features to refine voxel-wise features in a structure-aware manner, avoiding excessive smoothing in regions with strong local variations. After obtaining the frequency-guided fused features, an efficient Transformer block [15] is employed to further capture global contextual dependencies. Multi-scale features in the point branch are then aggregated via up-sampling and multi-layer perceptrons (MLPs), and a lightweight generalized mean (GeM) pooling layer [16] is used to produce the final global descriptor of the point cloud for retrieval.

From a frequency-domain perspective, FGFM acts as a learnable cross-modal equalizer that balances the low-frequency stability of the voxel representations and the high-frequency expressiveness of the point representations, leading to more robust and discriminative LiDAR place descriptors. The main contributions of our work are summarized as follows:

- We propose FGF-Net, a lightweight hybird voxel-point LiDAR place recognition network that explicitly exploits the complementary frequency characteristics of voxel- and point-wise features.
- We propose the Frequency-Guided Fusion Module, which performs local DCT-based spectral decomposition and dual-stream energy-guided gating. Within FGFM, the Semantic-Aware High-Frequency Enhancement and Consistency-Aware Low-Frequency Enhancement modules selectively enhance reliable high-frequency point details and low-frequency voxel context, respectively.
- We conduct extensive experiments on large-scale LiDAR place recognition benchmarks, showing that FGF-Net achieves competitive performance compared with state-of-the-art methods in a highly parameter-efficient manner, and exhibits strong generalization to diverse environments and adverse weather conditions.

II. Related Works

A. Point-Based Place Recognition

The key challenge of point-based methods lies in extracting discriminative features from the inherent irregularity and sparsity of raw 3D point clouds. PointNetVLAD [10], the first end-to-end LPR network, combines PointNet [12] for local feature extraction with NetVLAD [17] for global descriptor aggregation. PCAN [18] enhances PointNetVLAD with a contextual attention mechanism, but PointNet-based architectures still largely ignore the local spatial distribution of points, limiting their performance in large-scale outdoor scenes. To obtain more geometric structure, LPD-Net [19] enhance local features with handcrafted geometric descriptors, while EPC-Net [20] uses proxy point convolutions to aggregate multi-scale local geometry.

Recently, transformer-based architectures have been explored to capture long-range contextual dependencies. SOE-Net [21], PPT-Net [11], HiTPR [22], HOTFormerLoc [23], and

RIA-Net [24] employ various self-attention and hierarchical schemes to model global context over local features. These methods are effective at preserving fine-grained geometric details and, from a frequency-domain perspective, retain rich high-frequency components associated with sharp structures and subtle local variations. However, their direct reliance on raw points makes them sensitive to sensor noise and density variation, which may amplify undesirable high-frequency noise and limit robustness in complex outdoor environments.

B. Voxel-Based Place Recognition

To handle the irregular format of point clouds, voxel-based methods discretize the 3D space into regular voxela, enabling efficient sparse 3D convolutions and scalable processing. Built upon MinkowskiNet [8], MinkLoc3D [25] quantizes point clouds into sparse voxels, uses SPConvs to extract local voxel-wise features, and aggregates them with generalized mean pooling [16] into a global descriptor. MinkLoc3Dv2 [6] further introduces channel attention and a truncated Smooth-AP loss [26] with multistage backpropagation [27] for improved training efficiency and accuracy. SVT-Net [7] combines SPConvs with atom- and cluster-based sparse voxel transformers to jointly encode local and long-range semantic relationships.

Voxel representations naturally aggregate information in local neighborhoods and, in the frequency domain, behave like a low-pass filter, that they suppress high-frequency noise and emphasize smooth, low-frequency structures, leading to robust and computationally efficient descriptors. However, sparse voxelization inevitably introduces quantization errors and blurs fine-grained details, degrading high-frequency geometric cues that are crucial for discriminating geometrically similar places. To alleviate this, LCD-Net [28] introduces voxel-to-keypoint encoding inspired by PV-RCNN [29], and PTC-Net [9] projects voxel-wise features back to points to partially restore spatial details. Nonetheless, since point features are derived from voxel representations, these methods remain fundamentally voxel-centric and still suffer from information loss induced by voxelization.

C. Hybrid Voxel-Point-Based Place Recognition

Hybrid methods aim to jointly exploit voxel-wise features, which provide robust semantic context efficiently, and point-wise features, which preserve discriminative fine-grained geometry, via dual-branch architectures. However, research on hybrid designs for LPR is still limited. CASSPR [14] employs a sparse voxel branch for low-resolution feature aggregation and a point-wise branch for local detail encoding, and fuses them through a series of cross-attention modules without explicit feature transformation. Although effective, the heavy cross-attention incurs considerable computational overhead. More recently, HVP-Net [15] combines SPConvs for robust voxel-wise features with lightweight grouped efficient attention for global point-wise features, and fuses them in an interactive manner. DAH-Net [30] further incorporates density-driven adaptive modules on the point branch to mitigate the effects of sparsity and noise variations in large-scale LiDAR point clouds.

Overall, existing hybrid methods mainly perform fusion in the spatial domain, typically by attention or simple additive operations, without explicitly accounting for the distinct frequency characteristics of voxel- and point-wise features. As a result, they do not fully exploit the low-frequency stability of voxel-wise features and the high-frequency richness of point-wise features, motivating our frequency-guided hybrid fusion strategy.

III. Method

In this section, we introduce the proposed FGF-Net for large-scale LiDAR-based place recognition. We first give an overview of the proposed architecture, followed by a detailed explanation of the frequency-guided fusion module, which is the core component of our method.

A. Overview

As illustrated in Fig. 2, FGF-Net adopts a dual-branch architecture to extract both voxel- and point-wise features. For the voxel branch, the input point cloud is first quantized into sparse voxel representations, followed by multiple sparse convolution (SPConv) layers to sequentially extract multi-scale local voxel-wise features. For the point branch, the point cloud is down-sampled using the farthest point sampling algorithm. Graph convolutions [13] are then applied on the k-nearest neighbor graph of the down-sampled points. Once the voxel- and point-wise features are aligned by the feature transformation unit, we propose the frequency-guided fusion module (FGFM) to combine these features by considering the frequency domain information of the features. Next, we feed the fused representations into an efficient Transformer [15], which models long-range dependencies and strengthens global contextual perception. To further enhance feature discriminability, the multi-scale outputs from the Transformer are merged via up-sampling operations followed by multi-layer perceptrons (MLPs). Finally, a lightweight generalized mean (GeM) pooling layer aggregates the enhance features into a single global descriptor for the input point cloud.

B. Frequency-Guided Fusion Module

To effectively fuse cross-representations between voxel- and point-wise features, we propose a Frequency-Guided Fusion Module (FGFM), as shown in Fig. 3. Point-wise features encode fine-grained local geometry, which mainly corresponds to high-frequency signals, whereas voxel-wise features provide smoother and more stable semantic context due to the voxelization process, corresponding to low-frequency components. Conventional fusion strategies, such as element-wise summation in HVP-Net [15] or cross-attention in CASSPR [14], operate purely in the spatial domain and may fail to exploit this spectral complementarity. In contrast, FGFM explicitly splits frequency bands via a discrete cosine transform (DCT) and performs feature fusion with a dual-stream energy-guided efficient gating mechanism.

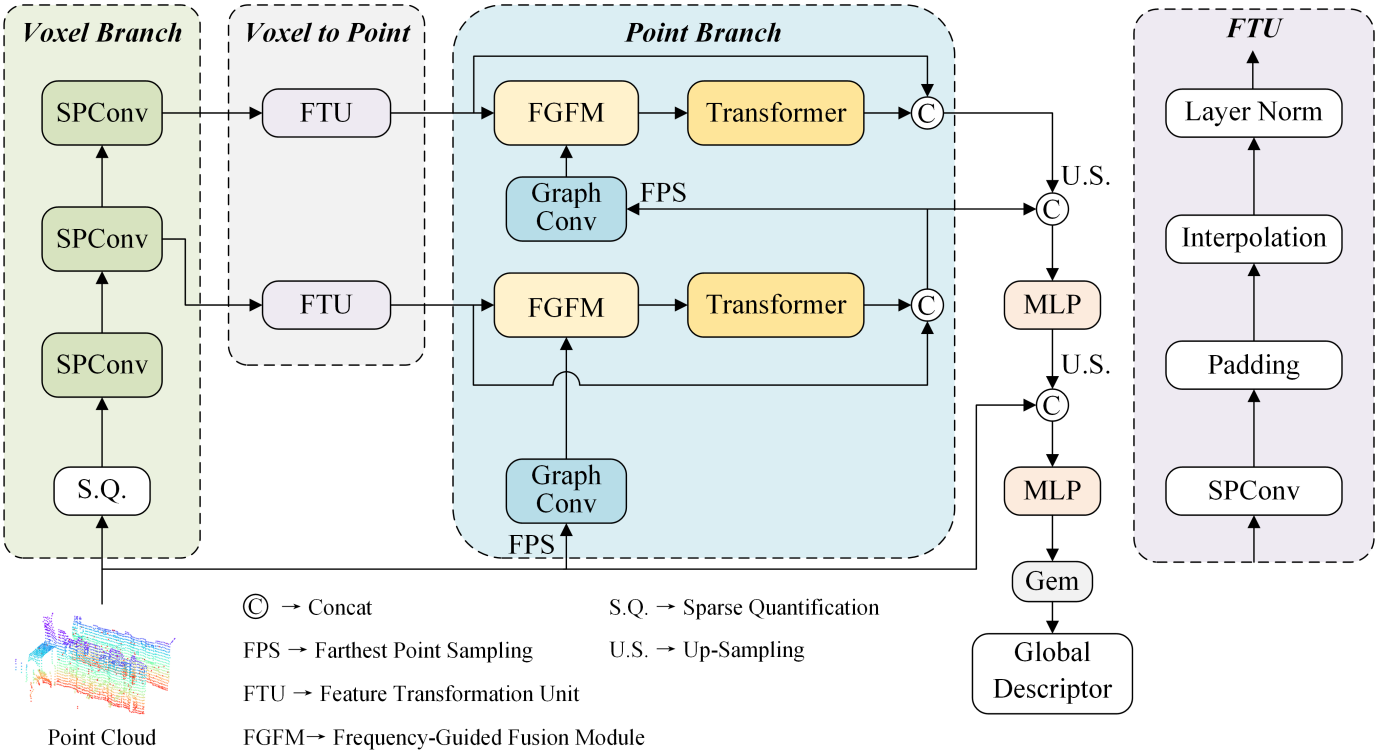

Fig. 2. **Overview of the proposed FGF-Net**. The proposed model integrates voxel- and point-wise features through a dual-branch architecture. In the voxel branch, multi-scale voxel-wise features are first extracted using sparse convolutions. These features are then transformed into point representations by a feature transformation unit to ensure spatial alignment between the voxel- and point-wise features. In the point branch, the point cloud is initially down-sampled using farthest point sampling. Then, graph convolutions are applied on the k-nearest neighbor graph of the down-sampled points. Afterward, the voxel- and point-wise features are fused by a frequency-guided fusion module (FGFM). The fused features are then passed through an efficient transformer to capture long-range contextual dependencies. Multi-scale features from transformers are further fused by up-sampling operations and multi-layer perceptrons. Finally, a generalized mean pooling layer aggregates the enhanced point-wise features into a final global descriptor for place recognition.

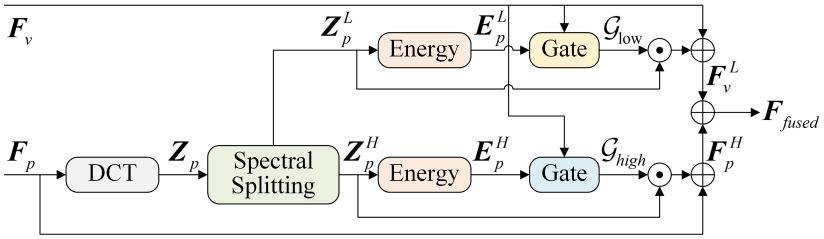

Fig. 3. The overview of frequency-guided fusion module.

Spectral Splitting. Let $\mathbf{F}_p \in \mathbb{R}^{B\times C\times N}$ and $\mathbf{F}_v \in \mathbb{R}^{B\times C\times N}$ denote the aligned point-wise and voxel-wise features, respectively, where B is the batch size, C is the number of channels, and N is the number of points. For each point, we construct a local neighborhood $\mathcal{N}_p \in \mathbb{R}^{B\times C\times N\times K}$ that contains K neighbor points for the point-wise features. To analyze local variations in the frequency domain, we apply a 1D DCT along the neighborhood dimension K and obtain the spectral representation $\mathbf{Z}_p \in \mathbb{R}^{B\times C\times N\times K}$:

$$\mathbf{Z}_p = \mathcal{DCT}(\mathcal{N}_p). \tag{1}$$

The DCT decouples slowly varying (low-frequency) components from rapidly changing (high-frequency) structures around each point. We further split $\mathbf{Z}_p$ into low- and high-frequency bands:

$$\mathbf{Z}_p^L = \mathbf{Z}_p[:,:,:,\mathcal{L}], \tag{2}$$
$$\mathbf{Z}_p^H = \mathbf{Z}_p[:,:,:,\mathcal{H}], \tag{3}$$

where $\mathcal{L}$ and $\mathcal{H}$ denote the index sets of low- and high-frequency coefficients, respectively. This separation allows FGFM to process low-frequency structural information and high-frequency local details in a decoupled manner.

Semantic-Aware High-Frequency Enhancement. Considering that high-frequency bands in point clouds contain both sharp geometric details and sensor noise, indiscriminate enhancement is therefore suboptimal. To selectively emphasize informative high-frequency components, we first estimate a per-point high-frequency energy using the ℓ_1 norm:

$$\mathbf{E}_p^H(n) = \frac{1}{C|\mathcal{H}|}\sum_{c=1}^{C}\sum_{k\in\mathcal{H}}\left|\mathbf{Z}_p^H(c,n,k)\right|. \tag{4}$$

One of the core innovations of FGFM is using the semantically stable voxel-wise features to validate the high-frequency details of the point-wise features. We hypothesize that high-frequency details should only be emphasized when they are consistent with the semantic context. To this end, we introduce a semantic-aware high-frequency enhancement (SA-HFE), where voxel-wise features serve as context priors to decide where and how much high-frequency detail from the point branch should be injected. The gating weights $\mathcal{G}_{\text{high}}$ are computed as

$$\mathcal{G}_{\text{high}} = \sigma\left(\phi_h\left([\mathbf{F}_v; \mathbf{E}_p^H]\right)\right), \tag{5}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation, ϕ_h means a MLP layer, and σ is the sigmoid function. The enhanced point-wise features are then obtained as

$$\mathbf{F}_p^H = \mathbf{F}_p + \text{Agg}\left(\mathcal{G}_{\text{high}} \odot \mathbf{Z}_p^H\right), \tag{6}$$

where $\odot$ denotes element-wise multiplication and $\text{Agg}(\cdot)$ aggregates along the neighborhood dimension. In this way, the voxel-wise features $\mathbf{F}_v$ act as a semantic anchor, suppressing high-frequency noise where semantic confidence is low while preserving important and genuine geometric details for the LPR task.

Consistency-Aware Low-Frequency Enhancement. Conversely, to enhance the voxel-wise features, we propose a consistency-aware low-frequency enhancement (CA-LFE) that exploits the low-frequency consistency of the voxel- and point-wise features. We analogously estimate the low-frequency energy $\mathbf{E}_p^L(n)$ from $\mathbf{Z}_p^L$ and obtain its normalized form $\tilde{\mathbf{E}}_p^L(n)$ using the same normalization as above. The voxel-wise features are then modulated by a gate $\mathcal{G}_{\text{low}}$ derived from the voxel-wise features and the low-frequency energy of the point-wise features:

$$\mathcal{G}_{\text{low}} = \sigma\left(\phi_l\left([\mathbf{F}_v; \mathbf{E}_p^L]\right)\right), \tag{7}$$

$$\mathbf{F}_v^L = \mathbf{F}_v + \text{Agg}\left(\mathcal{G}_{\text{low}} \odot \mathbf{Z}_p^L\right), \tag{8}$$

where ϕ_l is another MLP layer. In this way, low-frequency energy highlights regions with stable structural support, guiding the voxel-wise features to focus on consistent geometric patterns. Finally, we obtain the fused features by summing the two enhanced features of different representations:

$$\mathbf{F}_{\text{fused}} = \mathbf{F}_p^H + \mathbf{F}_v^L. \tag{9}$$

This frequency-guided dual-stream fusion enables $\mathbf{F}_{\text{fused}}$ to simultaneously inherit the robust semantic context of voxels and the fine-grained geometric details captured by points.

IV. Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of our proposed FGF-Net method. We compare our method with state-of-the-art place recognition methods, including PointNetVLAD [10], PCAN [18], EPC-Net [20], LPD-Net [19], MinkLoc3D [25], SOE-Net [21], PPT-Net [11], SVT-Net [7], MinkLoc3Dv2 [6], CASSPR [14], GRLoR [31], PTC-Net [9], Str-TF [32], and HVP-Net [15]. The Recall@N metric is adopted to evaluate the proposed method. Specifically, we report Average Recall at the top 1% (AR@1%) and Average Recall at the top 1 (AR@1) metrics.

A. Datasets Overview

To comprehensively evaluate the performance of our method in realistic scenarios, we conduct experiments on the following LPR datasets: benchmark datasets (Oxford RobotCar [33] and three In-house datasets [10]), KITTI [34], NCLT [35], and an in-house campus dataset, NJUST [15]. These datasets span a wide range of conditions, including different LiDAR sensors, different point densities, long-term temporal variations, and diverse weather conditions such as sunny and snowy scenes. This combination forms a comprehensive benchmark for evaluating place recognition in large-scale outdoor environments subject to complex and realistic real-world changes.

B. Implementation details

In all experiments, the quantization step of point coordinates in the voxel branch is set to 0.01. For the point branch, the number of sampled points for FPS is fixed to 1024, and the number of neighbors K is set to 32 in both stages. For spectral splitting, the frequency bands is divided evenly, so that the sizes of the low- and high-frequency bands satisfy $|\mathcal{L}| = |\mathcal{H}| = 16$. The dimension of the final global descriptor for each input point cloud is 256. Following HVP-Net [15], we employ the same data augmentation strategies, including random jittering, random translation, random point removal, and random erasing, to mitigate overfitting. We adopt the truncated Smooth-AP loss and the multistage backpropagation strategy from MinkLoc3Dv2 [6] to enable training with a large batch size of 512. All experiments are conducted on a server equipped with an NVIDIA RTX 4090 GPU (24 GB memory), and the model is implemented using PyTorch 2.4.1.

C. Recognition Performance Analysis

In this part, we conduct experiments to comprehensively evaluate the proposed method. Following the compared methods, we train the model only on a subset of the Oxford dataset and evaluate them on the remaining Oxford sequences as well as on the three In-house datasets. For all compared methods, the reported results are directly adopted from the original papers to ensure a fair comparison.

As shown in Table I, FGF-Net achieves the best performance on most test datasets, which demonstrates the effectiveness of our hybrid architecture with frequency-aware fusion strategy. On the Oxford dataset, where the testing environment is consistent with the training data, FGF-Net further improves upon the second best method, increasing AR@1 by 0.6%, which demonstrates the effectiveness of our method in modeling the underlying data distribution. For the three unseen In-house datasets (U.S., R.A., and B.D.), which are collected with a different LiDAR sensor, FGF-Net exhibits strong generalization to unseen environments with different point cloud distributions, especially in terms of AR@1 where it surpasses other methods by a large margin. Compared with

TABLE I

AR@1% AND AR@1 RESULTS OF COMPARED METHODS TRAINED ON THE **OXFORD** DATASET. THE **BOLD** AND UNDERLINED NUMBERS INDICATE THE BEST AND SECOND BEST RESULTS,RESPECTIVELY.

Methods	Parameters (M)	Oxford		U.S.		R.A.		B.D.		Mean	
		AR@1%	AR@1	AR@1%	AR@1	AR@1%	AR@1	AR@1%	AR@1	AR@1%	AR@1
PointNetVLAD [10]	19.8	80.3	62.8	72.6	63.2	60.3	56.1	65.3	57.2	69.6	59.8
PCAN [18]	20.40	83.8	69.1	79.1	62.4	71.2	56.9	66.8	58.1	75.2	61.6
EPC-Net [20]	4.70	94.7	86.2	96.5	88.2	88.6	80.2	84.9	78.1	91.2	83.2
LPD-Net [19]	19.8	94.9	86.3	96	87	90.5	83.1	89.1	82.5	92.6	84.7
MinkLoc3D [25]	1.19	97.9	93.0	95	86.7	91.2	80.4	88.5	81.5	93.2	85.4
SOE-Net [21]	19.39	96.4	-	93.2	-	91.5	-	88.5	-	92.4	-
PPT-Net [11]	13.39	98.1	93.5	97.5	90.1	93.3	84.1	90	84.6	94.7	88.1
SVT-Net [7]	0.90	97.8	93.7	96.5	90.1	92.7	84.3	90.7	85.5	94.4	88.4
MinkLoc3Dv2 [6]	2.66	98.9	96.3	96.7	90.9	93.8	86.5	91.2	86.3	95.1	90.0
CASSPR [14]	0.90	98.5	95.6	97.9	92.9	94.8	89.5	92.1	87.9	95.8	91.5
GRLoR [31]	3.6	99.1	96.9	98.7	93.1	95.7	89.2	94.3	89.9	97.0	92.1
PTC-Net [9]	6.97	98.8	96.4	98.5	95.0	97.0	92.2	94.5	90.7	97.2	93.6
HVP-Net [15]	**0.19**	99.0	96.7	99.4	**96.6**	97.1	92.7	94.2	90.9	97.5	94.2
Str-TF [32]	1.31	98.7	95.6	**99.6**	**96.6**	97.2	92.9	**95.0**	91.7	97.6	94.2
FGF-Net (ours)	0.22	**99.2**	**97.2**	99.1	96.0	**97.6**	**93.6**	94.8	**92.3**	**97.7**	**94.8**

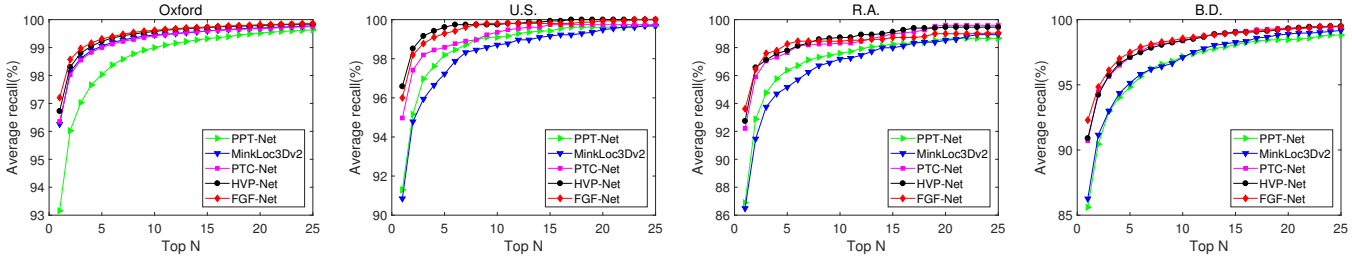

Fig. 4. Mean AR@N of different place recognition methods trained on the Oxford dataset.

TABLE II

COMPARISON OF NETWORK PERFORMANCE TRAINED ON THE OXFORD DATESET.

Methods	Parameters (M)	Inference Time (ms)	Mean AR@1
MinkLoc3Dv2 [6]	2.66	17.7	90.0
PTC-Net [9]	6.97	27.4	93.6
Str-TF [32]	1.31	$>$ 27.4	94.2
HVP-Net [15]	**0.19**	35.2	94.2
FGF-Net (ours)	0.22	**11.4**	**94.8**

the point-based method Str-TF, FGF-Net achieves competitive AR@1% results while reducing more than 1 million parameters, indicating the efficient of our frequency-guided fusion strategy. Compared to the hybrid method HVP-Net, FGF-Net improves AR@1 by 0.6% with only 0.03 million additional parameters. This suggests that the high-frequency enhanced features produced by our fusion module are particularly effective in alleviating perceptual aliasing in geometrically similar regions.

In Fig. 4, we show the top-N recall curves of the compared methods on the test datasets. It can be found that our FGF-Net can retrieve more correct point clouds in most cases when returning the same number of candidates, highlighting its superior retrieval capability.

D. Efficiency analysis

To evaluate the efficiency of our model, we compare FGF-Net with representative methods in terms of both the number of parameters and inference time, as shown in Table II. For Str-TF, we report the relative inference times provided in the original manuscript, since PTC-Net has been shown to have superior runtime performance. For the voxel-based MinkLoc3Dv2, although it contains 2.66M parameters, it still achieves relatively low inference time, benefiting from the high computational efficiency of SPConvs. For the voxel-based PTC-Net, the use of four PTC layers for multi-scale feature extraction leads to the largest parameter count among all compared methods. For the point-based Str-TF, although the parameter count is reduced compared with PTC-Net, its multi-stage cross-attention mechanism is computationally demanding and results in higher inference time. For the hybrid HVP-Net, despite having the fewest parameters (0.19M), its interactive feature fusion strategy incurs the longest inference time. By contrast, our FGF-Net slightly increases the parameter count to 0.23M, but improves the mean AR@1 by 0.7% while reducing the inference time down to 11.4 ms. These results indicate that the proposed frequency-guided fusion strategy is an effective and parameter-efficient way to exploit cross-representation complementarity.

E. Generalization Performance Analysis

To further evaluate the generalization ability of our method, we conduct experiments on the In-house, KITTI, and NJUST datasets using models trained on the Oxford dataset. The training sets and in-house datasets consist of submaps built from multiple frames, while KITTI and NJUST datasets use single-frame point clouds. As shown in Table III, FGF-Net outperforms all other methods on most test datasets, demonstrating its superior generalization ability to unseen environments and different point cloud densities.

We further evaluate the generalization of our method under snow weather conditions. As shown in Table IV, experiments

TABLE III
AR@1 RESULTS OF COMPARED METHODS TRAINED ON THE **OXFORD** DATASET AND TEST ON THE **IN-HOUSE**, **KITTI** AND **NJUST** DATASETS.

Methods	In-house				KITTI					NJUST			
	U.S.	R.A.	B.D.	Mean	00	02	05	06	Mean	2m	4m	8m	Mean
MinkLoc3Dv2 [6]	90.9	86.5	86.3	87.9	93.5	84.5	93.5	97.0	92.1	96.9	89.1	78.6	88.2
PTC-Net [9]	95.0	92.2	90.7	92.6	94.3	81.9	**95.1**	98.1	92.3	97.9	91.0	77.3	88.7
HVP-Net [15]	**96.6**	92.7	90.9	93.4	94.6	84.1	92.8	97.8	92.3	97.6	90.4	78.3	88.8
FGF-Net (ours)	96.0	**93.6**	**92.3**	**94.0**	**94.9**	**86.1**	92.4	**99.6**	**93.3**	**99.0**	**95.9**	**87.8**	**94.2**

TABLE IV
AR@1 RESULTS OF METHODS TRAINED ON THE **OXFORD** DATASET AND TEST ON THE **NCLT** DATASET, USING SEQUENCE 2012-02-05 AS THE DATABASE SET. SEQUENCES MARKED WITH * **INDICATE SNOW WEATHER CONDITION** ON THE CORRESPONDING DAYS. MEAN DEGRADATION REFERS TO THE PERFORMANCE DROP FROM SUNNY SEQUENCES TO SNOWY ONES.

Methods	2012-01-08	2012-02-02	2012-11-17	Mean	2012-01-15*	2012-01-22*	2013-02-23*	Mean	Mean degradation
MinkLoc3Dv2 [6]	79.9	78.2	68.8	75.6	80.0	77.2	67.1	74.8	-0.8
PTC-Net [9]	73.7	67.3	56.4	65.8	70.1	66.1	53.2	64.1	-1.7
HVP-Net [15]	80.9	**78.4**	70.7	76.7	80.3	79.3	65.3	75.0	-1.7
FGF-Net (ours)	**83.7**	77.8	**72.5**	**78.0**	**81.3**	**82.8**	**69.7**	**77.9**	**-0.1**

TABLE V
RESULTS OF FGF-NET UNDER DIFFERENT SETTINGS OF FREQUENCY-GUIDED FUSION MODULE.

Methods	SA-HFE		CA-LFE		Mean	
	$\mathbf{F}_v$	$\mathbf{E}_p^H$	$\mathbf{F}_v$	$\mathbf{E}_p^L$	AR1%	AR@1
A					97.4	94.2
B	✓				97.5	94.4
C	✓	✓			97.5	94.7
D			✓		97.6	94.4
E			✓	✓	97.4	94.5
FGF-Net	✓	✓	✓	✓	**97.7**	**94.8**

on the NCLT dataset, which includes sequences collected under different weather conditions, show that FGF-Net performs best under sunny conditions. However, in snowy weather, all methods experience some performance degradation due to snow-induced noise in LiDAR data. Notably, MinkLoc3Dv2, which relies on voxel-wise features, shows less degradation because voxel quantization provides robustness to noise. In contrast, methods like PTC-Net and HVP-Net, which fuse voxel- and point-wise features through simple element-wise summation, suffer more due to ineffective cross-representation modeling. As for FGF-Net, with its frequency-guided fusion strategy, it exhibits the least performance drop in snowy conditions, as the strategy enables more effective fusion between voxel- and point-wise features by leveraging their complementary frequency-domain characteristics.

F. Ablation Studies

In this section, we conduct ablation studies to evaluate the effectiveness of the core components in FGF-Net, including the proposed FGFM and its two submodules, SA-HFE and CA-LFE. All variants are trained on the Oxford dataset, and we report the mean AR over the benchmark test sets.

As shown in Table V, method A denotes the baseline FGF-Net without FGFM, where voxel- and point-wise features are directly fused by element-wise summation. Compared with state-of-the-art methods in Table I, method A achieves a similar level of AR@1, which demonstrates the effectiveness of our hybrid voxel-point architecture even without the frequency-guided fusion strategy. Method B introduces a gating mechanism but does not explicitly model high-frequency information of point-wise features. In this case, the enhanced features lack sufficient high-frequency geometric details, leading to a drop in AR@1 compared with the frequency-aware variants, indicating that simple spatial-domain gating is not enough. Method C activates the SA-HFE module, i.e., voxel-wise features are used as semantic priors to modulate the high-frequency components of point-wise features.

Method E employs only CA-HFE in FGFM and still outperforms Method A, confirming the importance of low-frequency–aware enhancement. Compared with Method D, which only considers the low-frequency information of voxel-wise features in the spatial domain, Method E achieves comparable results, suggesting that the low-frequency information captured by these two representations is largely redundant.

When both SA-HFE and CA-LFE are applied, i.e., the full FGFM with both high- and low-frequency awareness is used, the performance gains become the most significant, which demonstrates the effectiveness of our proposed frequency-guided feature fusion strategy that jointly exploits high-frequency point details and low-frequency voxel context.

V. CONCLUSION

In this paper, we proposed FGF-Net, a lightweight frequency-guided fusion network for LiDAR-based place recognition with hybrid voxel-point representations. Motivated by the complementary strengths and weaknesses of voxel- and point-wise features, we explicitly exploited their frequency-domain characteristics instead of fusing them only in the spatial domain. To this end, we proposed the FGFM, which applies a local DCT to decompose point-wise features into low- and high-frequency bands and then performs an energy-aware dual-stream gating mechanism. Through this design, voxel-wise features serve as semantic priors that selectively enhance reliable high-frequency details in the point branch, while low-frequency cues from the point branch refine voxel-wise features into more stable structural representations. This frequency-guided hybrid fusion enables FGF-Net to effectively

balance robustness to noise, preservation of geometric details, and computational efficiency.

Extensive experiments on the benchmark, KITTI, NCLT, and NJUST datasets demonstrate that FGF-Net achieves competitive place recognition performance with only 0.22M parameters. Furthermore, it shows strong generalization across diverse environments and challenging weather conditions. These results validate the effectiveness of our FGF-Net for practical LPR.

REFERENCES

[1] W. Deng, Y. Tie, and L. Qi, "Eca-slam: A visual slam utilizing depth features and attention mechanism," in *Proceedings of the International Joint Conference on Neural Networks*, 2024, pp. 1–8.

[2] M. Usuelli, M. Frosi, P. Cudrano, S. Mentasti, and M. Matteucci, "Radarlcd: Learnable radar-based loop closure detection pipeline," in *Proceedings of the International Joint Conference on Neural Networks*, 2024, pp. 1–7.

[3] 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," *International Journal of Computer Vision*, vol. 132, no. 8, pp. 3139–3171, 2024.

[4] P. Yin, J. Jiao, S. Zhao, L. Xu, G. Huang, H. Choset, S. Scherer, and J. Han, "General place recognition survey: Towards the real-world autonomy age," *IEEE Transactions on Robotics*, vol. 41, pp. 3019–3038, 2025.

[5] K. Luo, H. Yu, X. Chen, Z. Yang, J. Wang, P. Cheng, and A. Mian, "3d point cloud-based place recognition: A survey," *Artificial Intelligence Review*, vol. 57, no. 4, p. 83, 2024.

[6] J. Komorowski, "Improving point cloud based place recognition with ranking-based loss and large batch training," in *Proceedings of the International Conference on Pattern Recognition*, 2022, pp. 3699–3705.

[7] Z. Fan, Z. Song, H. Liu, Z. Lu, J. He, and X. Du, "Svt-net: Super light-weight sparse voxel transformer for large scale place recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 551–560.

[8] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3070–3079.

[9] L. Chen, H. Wang, H. Kong, W. Yang, and M. Ren, "PTC-Net: Point-Wise Transformer With Sparse Convolution Network for Place Recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3414–3421, 2023.

[10] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4470–4479.

[11] L. Hui, H. Yang, M. Cheng, J. Xie, and J. Yang, "Pyramid Point Cloud Transformer for Large-Scale Place Recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 6078–6087.

[12] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 77–85.

[13] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph Cnn for learning on point clouds," *ACM Transactions on Graphics*, vol. 38, no. 5, 2019.

[14] Y. Xia, M. Gladkova, R. Wang, Q. Li, U. Stilla, J. F. Henriques, and D. Cremers, "Casspr: Cross attention single scan place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8427–8438.

[15] L. Chen, H. Kong, H. Wang, W. Yang, J. Lou, F. Xu, and M. Ren, "Hvp-net: A hybrid voxel- and point-wise network for place recognition," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 395–406, 2024.

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 936–944.

[17] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[18] W. Zhang and C. Xiao, "PCAN: 3D attention map learning using contextual information for point cloud based retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 428–12 437.

[19] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y. Liu, "LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2831–2840.

[20] L. Hui, M. Cheng, J. Xie, J. Yang, and M. M. Cheng, "Efficient 3D Point Cloud Feature Learning for Large-Scale Place Recognition," *IEEE Transactions on Image Processing*, vol. 31, pp. 1258–1270, 2022.

[21] Y. Xia, Y. Xu, S. Li, R. Wang, J. Du, D. Cremers, and U. Stilla, "SOE-Net: A Self-Attention and Orientation Encoding Network for Point Cloud based Place Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 343–11 352.

[22] Z. Hou, Y. Yan, C. Xu, and H. Kong, "HiTPR: Hierarchical Transformer for Place Recognition in Point Cloud," in *Proceedings of the International Conference on Robotics and Automation*, 2022, pp. 2612–2618.

[23] E. Griffiths, M. Haghighat, S. Denman, C. Fookes, and M. Ramezani, "Hotformerloc: Hierarchical octree transformer for versatile lidar place recognition across ground and aerial views," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025, pp. 6648–6658.

[24] W. Hao, W. Zhang, and H. Su, "Ria-net: Rotation invariant aware 3d point cloud for large-scale place recognition," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5014–5021, 2024.

[25] J. Komorowski, "MinkLoc3D: Point cloud based large-scale place recognition," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 1789–1798.

[26] A. Brown, W. Xie, V. Kalogeiton, and A. Zisserman, "Smooth-AP: Smoothing the Path Towards Large-Scale Image Retrieval," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 677–694.

[27] J. Revaud, J. Almazan, R. Rezende, and C. D. Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5106–5115.

[28] 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.

[29] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 526–10 535.

[30] S. Wang, Y. Zhang, J. Zhang, Y. Shen, D. Shan, and Z. Zhang, "Density-driven adaptive hybrid network for large-scale place recognition," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–15, 2025.

[31] W. Liu, S. Ao, Y. Zhang, H. Wang, and Y. Guo, "Grlor: A unified global retrieval and local reranking framework for 3-d place recognition," *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, 2024.

[32] S. So, S. Kim, S. Lee, I. Yun, and S. Lee, "Query-centric structural transformer for cross-source robust place recognition," *IEEE Access*, vol. 13, pp. 184 531–184 542, 2025.

[33] W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000 km: The Oxford RobotCar dataset," *International Journal of Robotics Research*, vol. 36, no. 1, pp. 3–15, 2017.

[34] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[35] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and lidar dataset," *International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.