Spatial-Frequency Domain Aggregation for Visual Place Recognition

Chaoqun Wang[1]
cq23@scnu.edu.cn
Shaobo Min[2]
mbobo2ml@gmail.com

[1] School of Artificial Intelligence
South China Normal University
Guangzhou, China

[2] Data Platform
Tencent,
Shenzhen, China

Abstract

Visual place recognition (VPR) aims to identify an image by matching its global representation with those of reference images. Existing methods primarily aggregate features in the spatial domain, usually ignoring critical global or structural information embedded in the frequency domain. To address this limitation, we propose Dual-Domain Aggregation for Visual Place Recognition (DDA-VPR), which jointly integrates spatial and frequency domain features to enhance global representation aggregation. The core motivation is that the frequency domain inherently captures global and structural patterns, offering complementary and discriminative cues that are difficult to capture in the spatial domain. To bridge the domain gap between spatial and frequency features, a triple fusion strategy is introduced to facilitate cross-domain interaction and fuse spatial and frequency domain features into a unified and robust global representation. Additionally, to refine feature abstraction in both domains, a multi-scale contextual attention module is designed to leverage multi-scale information and preserve critical details during downsampling. Consequently, DDA-VPR generates more discriminative representations than methods that aggregate features solely in the spatial domain. Extensive experiments on challenging benchmarks demonstrate the superior performance of DDA-VPR, validating the effectiveness of dual-domain aggregation in VPR.

1 Introduction

Visual place recognition (VPR) aims to recognize a query image from a database with geotagged images [17, 20, 52], supporting applications such as autonomous vehicles, mobile robotics, and augmented reality [15, 16, 31]. Despite its significance, VPR remains challenging due to serious appearance variations caused by changes in illumination, occlusion, seasons, and viewpoints.

Typically formulated as an image retrieval task, VPR relies on aggregating local features into global representations for similarity matching. The aggregation strategy significantly impacts representation quality and retrieval accuracy, and numerous studies have sought to enhance aggregation strategies. For instance, VLAD [22] aggregates features by computing their differences with clustered centroids. Generalized-mean (GeM) pooling [39] achieves

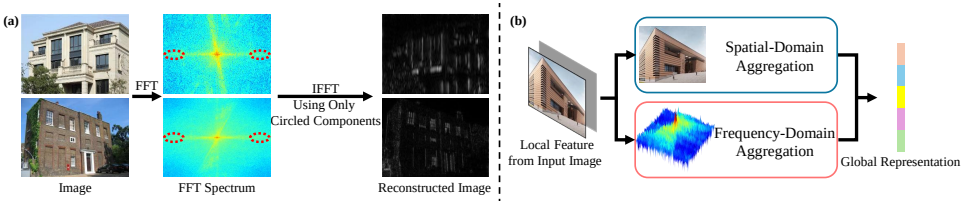

Figure 1: (a) Reconstructing image only using components circled in FFT Spectrum, which demonstrates that global patterns in scene images, *i.e.,* building windows, can be effectively captured in frequency domain. (b) Instead of focusing only on spatial features, DDA-VPR aggregates features in both spatial and frequency domains, yielding more discriminative global representations.

a balance between computational efficiency and feature quality, as demonstrated in recent works [8, 26, 32]. SALAD [20] employs optimal transport for feature aggregation using the Sinkhorn algorithm. Despite these advancements, existing aggregation methods are predominantly confined to spatial-domain processing, ignoring the rich information present in the frequency domain. With global receptive field, the frequency domain inherently encodes an image's global characteristics [25], enabling the effective capture of overall patterns and structural information. In VPR, scene images usually contain repetitive textures and shapes, such as the regular arrangement of building windows, as shown in Fig. 1 (a). These patterns result in a concentrated distribution in the frequency domain, *e.g.,* building windows in both images correspond to the same regions highlighted by red dashed lines in FFT Spectrum in Fig. 1 (a). This facilitates more efficient processing compared to the spatial domain.

In this paper, we propose Dual-Domain Aggregation for Visual Place Recognition (DDA-VPR), a novel method designed to enhance global representations by jointly aggregating critical information from both spatial and frequency domains. As shown in Fig. 1 (b), DDA-VPR adopts a dual-branch architecture, with each branch processing features in a specific domain. The frequency-domain features contain global and structural information that is complementary to the spatial domain. In each branch, we introduce a Multi-Scale Contextual Attention (MSC-Attn) module to preserve critical local details when downsampling features for information extraction. By exploiting multi-scale contexts, MSC-Attn ensures comprehensive feature aggregation. Finally, a triple fusion mechanism is designed to integrate different domain features and bridge the inherent domain gap. This mechanism fuses complementary information from spatial and frequency features, ultimately generating robust global representations. Compared to spatial-only aggregation, DDA-VPR produces more informative and discriminative global representations for VPR.

We evaluate DDA-VPR on several challenging public datasets, demonstrating its superior performance in VPR. The main contributions of this work are summarized as follows:

1. We introduce Dual-Domain Aggregation for Visual Place Recognition (DDA-VPR), a novel method that aggregates robust global representations by leveraging complementary cues from both spatial and frequency domains.

2. A novel triple fusion mechanism is designed to bridge the domain gap between spatial and frequency features to produce highly discriminative representations. Additionally, a Multi-Scale Contextual Attention (MSC-Attn) module is designed to explore multi-scale information and preserve crucial details during downsampling.

3. Extensive experiments on multiple benchmarks demonstrate the effectiveness and generalization of DDA-VPR, achieving state-of-the-art performance under challenging conditions.

2 Related Work

2.1 Visual Place Recognition

Visual place recognition (VPR) involves extracting features from query and reference images, and aggregating these features into global representations. For local feature extraction, Early VPR methods [17, 42] rely on handcrafted features [4, 6, 30].With the advent of deep learning, CNN [8, 24] and Transformer [20, 46] have been widely adopted to extract representative local features. Recent visual foundation models, such as DINOv2 [37], have demonstrated strong generalization capabilities in VPR [20, 26]. To enhance performance, SALAD [20] fine-tunes the final layers of DINOv2.

Aggregating these local features into compact global representations is a critical step, as it directly impacts the retrieval accuracy. VLAD-based methods [5, 22, 33], such as VLAD [22] and NetVLAD [5], cluster local features and compute differences between these features and cluster centroids to generate global descriptors. However, these methods often generate high-dimensional global descriptors, increasing computational cost. Generalized-mean (GeM) pooling [39] has been widely adopted [8, 29] as a compact alternative. SALAD [20] utilizes optimal transport-based aggregation via the Sinkhorn Algorithm. Conv-AP [1] introduces a 1×1 convolution followed by average pooling, and MixVPR [2] employs a fully MLP-based architecture for feature aggregation.

2.2 Image Representations in Frequency Domain

Various vision tasks [13, 14, 19, 21, 38, 45, 50] utilize frequency-domain operations to enhance feature extraction. Some methods [11, 18, 19, 28, 40] replace conventional spatial operations with frequency-based alternatives. For example, FFC [11] introduces Fourier Units to replace convolution operation to extend receptive fields. Other works [35, 44] incorporate explicit frequency-domain operations alongside spatial-domain networks. For example, Vien *et al.* [44] design an explicit frequency-domain module to fuse multi-scale spatial features.

To effectively combine spatial and frequency features, simple fusion strategies such as addition [35, 36] or concatenation [23] have been explored. However, due to the inherent domain gap, advanced fusion mechanisms [27, 34, 51] have been developed to better integrate complementary information. Foe instance, SFIIN [51] applies spatial and channel attention for effective fusion. SFCFusion [9] employs convolutional sub-networks to achieve deeper integration.

Despite its utilization in other vision tasks, frequency-domain processing remains underexplored in VPR. Zhang *et al.* [49] enhance mid-frequency components for data augmentation but do not incorporate frequency information into feature extraction or aggregation. Existing VPR methods operate exclusively in the spatial domain, limiting their ability to leverage global and structural characteristics inherent in the frequency domain.

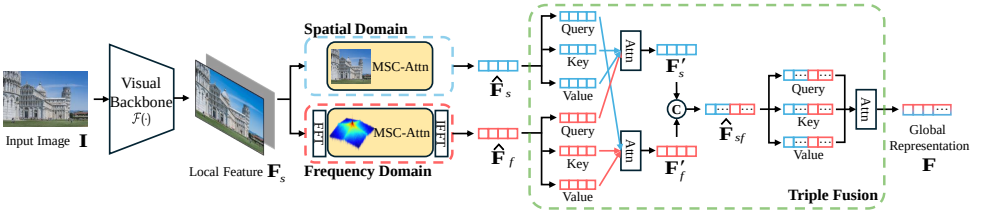

Figure 2: Pipeline of DDA-VPR, which processes local features jointly in spatial and frequency domains for robust feature aggregation.

3 Method

Visual place recognition (VPR) aims to identify the location of an input image $\mathbf{I} \in \mathbb{R}^{H^I \times W^I \times 3}$ by retrieving the most similar reference image $\hat{\mathbf{I}}_r$ from a database $\mathcal{B}$. The retrieval process can be formalized as:

$$\hat{\mathbf{I}}_r = \arg\max_{\mathbf{I}_r \in \mathcal{B}} s(f_{agg}(\mathcal{F}(\mathbf{I})), f_{agg}(\mathcal{F}(\mathbf{I}_r))), \tag{1}$$

where $s(\cdot,\cdot)$ denotes the similarity metric, such as cosine similarity. $\mathcal{F}(\cdot)$ represents the visual backbone for extracting local features from the input image. $f_{agg}(\cdot)$ is the function used to aggregate local features into a compact global representation.

As indicated in Eq. (1), the performance of VPR is strongly influenced by the quality of feature aggregation. To this end, we introduce a novel feature aggregation method, termed Dual-Domain Aggregation for Visual Place Recognition (DDA-VPR), aimed at enhancing the robustness and accuracy of global representations. As shown in Fig. 2, DDA-VPR consists of two parallel branches focusing on spatial and frequency domains, respectively. A triple fusion module is designed to integrate the two domain features. This enables the effective capture of critical and complementary information from local features. During the aggregation process in both domains, a Multi-Scale Contextual Attention (MSC-Attn) module is introduced to refine both local and global information simultaneously.

3.1 Dual-Domain Aggregation Framework

DDA-VPR leverages spatial and frequency domain information to jointly extract critical and complementary information from local features. The spatial domain captures intricate spatial details, while the frequency domain reveals global patterns and structural characteristics that are complementary and challenging to obtain in the spatial domain.

Given an input image $\mathbf{I}$, a visual backbone $\mathcal{F}(\cdot)$, such as DINOv2 [57], is first used to extract the initial local feature map by:

$$\mathbf{F}_s = \mathcal{F}(\mathbf{I}), \tag{2}$$

where $\mathbf{F}_s \in \mathbb{R}^{H \times W \times D}$ represents the local feature map. H and W denote the height and width, respectively. D is the feature dimension. Intuitively, $\mathbf{F}_s$ contains the spatial signals to describe the visual content of $\mathbf{I}$. To further emphasize latent structural and global characteristics in $\mathbf{F}_s$, DDA-VPR transforms $\mathbf{F}_s$ into the frequency domain via 2D Fast Fourier

Transform (FFT) by:

$$\mathbf{F}_f = \mathrm{FFT}(\mathbf{F}_s), \tag{3}$$

where $\mathbf{F}_f \in \mathbb{R}^{H\times W\times D}$ is the frequency-domain feature map, which highlights important frequency-specific information. With $\mathbf{F}_s$ and $\mathbf{F}_f$, DDA-VPR then refines them via the proposed MSC-Attn module in parallel:

$$\hat{\mathbf{F}}_s = \text{MSC-Attn}(\mathbf{F}_s), \quad \hat{\mathbf{F}}_f = \mathrm{IFFT}(\text{MSC-Attn}(\mathbf{F}_f)), \tag{4}$$

where $\mathrm{IFFT}(\cdot)$ transforms the frequency-domain feature back to spatial domain. The detailed architecture of MSC-Attn$(\cdot)$ will be illustrated in Sec. 3.2, which is a downsampling and refinement layer. The downsampled features $\hat{\mathbf{F}}_s$ and $\hat{\mathbf{F}}_f$ are more discriminative and compact than $\mathbf{F}_s$ and $\mathbf{F}_f$. Notably, $\hat{\mathbf{F}}_s, \hat{\mathbf{F}}_f \in \mathbb{R}^{S\times D}$, where $S = (H/4)\times(W/4)$.

Next, DDA-VPR aggregates $\hat{\mathbf{F}}_s$ and $\hat{\mathbf{F}}_f$ into a unified global representation in a triple-fusion manner, effectively addressing the domain gap issue between the spatial and frequency features. As shown in Fig. 2, DDA-VPR reorganizes the spatial feature $\hat{\mathbf{F}}_s$ according to the frequency feature $\hat{\mathbf{F}}_f$ via a cross-attention operation:

$$\mathbf{F}'_s = \mathrm{Softmax}\left(\frac{q_s(\hat{\mathbf{F}}_f)\cdot k_s(\hat{\mathbf{F}}_s)^\top}{\sqrt{D}}\right) v_s(\hat{\mathbf{F}}_s), \tag{5}$$

where $q_s(\cdot)$, $k_s(\cdot)$, and $v_s(\cdot)$ are linear projections for query, key, and value, respectively. $\mathbf{F}'_s$ contains reorganized spatial cues that have similar signal structure with $\hat{\mathbf{F}}_f$. Similarly, $\mathbf{F}'_f$ is obtained by:

$$\mathbf{F}'_f = \mathrm{Softmax}\left(\frac{q_f(\hat{\mathbf{F}}_s)\cdot k_f(\hat{\mathbf{F}}_f)^\top}{\sqrt{D}}\right) v_f(\hat{\mathbf{F}}_f). \tag{6}$$

Both $\mathbf{F}'_s$ and $\mathbf{F}'_f$ are reorganized to be compatible with each other, reducing the domain gap between $\hat{\mathbf{F}}_s$ and $\hat{\mathbf{F}}_f$. Building on the implicitly aligned $\mathbf{F}'_s$ and $\mathbf{F}'_f$, DDA-VPR further employs a self-attention mechanism to fuse them into a unified representation:

$$\mathbf{F} = \mathrm{Softmax}\left(\frac{q_{sf}(\hat{\mathbf{F}}_{sf})k_{sf}(\hat{\mathbf{F}}_{sf})^\top}{\sqrt{D}}\right) v_{sf}(\hat{\mathbf{F}}_{sf}), \tag{7}$$

$$\text{where } \hat{\mathbf{F}}_{sf} = \mathrm{cat}(\mathbf{F}'_s; \mathbf{F}'_f), \tag{8}$$

where $\mathrm{cat}(\cdot;\cdot)$ denotes the concatenation operation for two features.

Consequently, the aggregated representation $\mathbf{F}$ contains important information from both spatial and frequency domains, providing a more discriminative and compact description of the image compared to the original local feature $\mathbf{F}_s$. This dual-domain aggregation pipeline ensures that the final representation effectively captures distinguishing cues for visual place recognition.

3.2 Multi-Scale Contextual Attention

Feature aggregation involves downsampling operations to refine and compact features. However, conventional downsampling operations, such as average and GeM pooling, directly aggregate neighboring features, discarding critical details and limiting adaptability to varying scales. To this end, we design a novel Multi-Scale Contextual Attention (MSC-Attn) module,

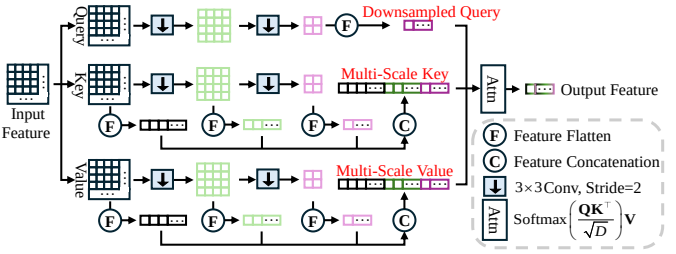

Figure 3: Implementations of Multi-Scale Contextual Attention (MSC-Attn) module.

which can effectively integrate multi-scale contextual information and preserve fine-grained details during downsampling in both spatial and frequency domains. The implementation of MSC-Attn is illustrated in Fig. 3.

Taking the spatial domain as an example, MSC-Attn begins by projecting the local feature map $\mathbf{F}_s$ into query, key, and value matrices, denoted as $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ with the same resolution $H \times W$. To capture multi-scale information, these matrices are downsampled hierarchically:

$$\mathbf{Q}^{D_1} = \mathcal{D}(\mathbf{Q}), \quad \mathbf{K}^{D_1} = \mathcal{D}(\mathbf{K}), \quad \mathbf{V}^{D_1} = \mathcal{D}(\mathbf{V}), \tag{9}$$

$$\mathbf{Q}^{D_2} = \mathcal{D}(\mathbf{Q}^{D_1}), \quad \mathbf{K}^{D_2} = \mathcal{D}(\mathbf{K}^{D_1}), \quad \mathbf{V}^{D_2} = \mathcal{D}(\mathbf{V}^{D_1}), \tag{10}$$

where $\mathcal{D}(\cdot)$ represents a 3×3 convolution with a stride of 2, achieving a downsampling ratio of 2×2. Compared to conventional pooling, convolution-based downsampling adaptively selects critical neighboring information for enhanced features.

To incorporate multi-scale information, key and value tensors of different scales are flattened and concatenated to form multi-scale key and value features $\hat{\mathbf{K}} = \text{cat}(\mathbf{K}; \mathbf{K}^{D_1}; \mathbf{K}^{D_2})$, $\hat{\mathbf{V}} = \text{cat}(\mathbf{V}; \mathbf{V}^{D_1}, \mathbf{V}^{D_2})$. The downsampled query $\mathbf{Q}^{D_2}$ interacts with the multi-scale keys and values, to compute the output refined feature:

$$\hat{\mathbf{F}}_s = \text{Softmax}\left(\frac{\mathbf{Q}^{D_2} \cdot \hat{\mathbf{K}}^\top}{\sqrt{D}}\right)\hat{\mathbf{V}}. \tag{11}$$

Eq. (11) exploits multi-scale complementary information to enhance the accuracy of feature extraction.

Compared to general downsampling operations, *e.g.,* average pooling or GeM pooling in VPR, MSC-Attn integrates contextual information at multiple scales, ensuring the preservation of discriminative details across resolutions in both spatial and frequency domains.

3.3 Overall Training

We adopt the multi-similarity loss [47] to train DDA-VPR:

$$\begin{aligned}\mathcal{L}_{ms} = \frac{1}{N}\sum_{i=1}^{N}\{\frac{1}{\alpha}\log[1+\sum_{j\in\mathcal{P}_i} e^{-\alpha(s(\mathbf{F}^i,\mathbf{F}^j)-\lambda)}] \\ +\frac{1}{\beta}\log[1+\sum_{k\in\mathcal{N}_i} e^{-\beta(s(\mathbf{F}^i,\mathbf{F}^k)-\lambda)}]\},\end{aligned} \tag{12}$$

Table 1: Comparison with state-of-the-art visual place recognition methods on public datasets. "Dim" denotes the global representation dimension. The best result is bolded, and the second best is underlined.

Method	Backbone	Dim	Pitts30k			MSLS val			SPED			Tokyo24/7		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Conv-AP [8]	ResNet50	8,192	90.4	95.2	96.4	83.4	90.5	92.3	80.1	90.3	93.6	73.1	84.8	87.6
CosPlace [8]	ResNet50	512	88.4	94.5	95.7	82.8	89.7	92.0	75.3	85.9	88.6	81.9	90.2	92.7
GCL [29]	ResNet50	2,048	80.7	91.5	93.9	79.5	88.1	90.1	49.4	65.1	71.0	69.5	81.0	85.1
MixVPR [8]	ResNet50	4,096	91.5	95.5	96.3	88.0	92.7	94.6	85.2	92.1	94.6	86.7	92.1	94.0
EigenPlaces [8]	ResNet50	2,048	92.5	96.8	97.6	89.1	93.8	95.0	69.9	82.9	87.6	93.0	96.2	97.5
CricaVPR† [53]	DINOv2	4,096	91.8	95.9	96.9	88.9	95.0	95.7	87.5	92.9	94.9	90.5	95.2	96.8
BoQ [8]	ResNet50	16,384	92.4	96.5	97.4	91.4	94.5	96.1	86.2	94.4	96.1	90.8	95.6	96.5
BoQ* [8]	DINOv2	12,288	93.7	97.1	97.9	93.8	96.8	97.0	92.5	95.9	96.7	96.5	98.1	98.7
SALAD [20]	DINOv2	8,448	92.4	96.3	97.4	92.2	96.4	97.0	92.1	**96.2**	96.5	94.6	97.5	97.8
DDA-VPR (Ours)	DINOv2	4,096	**93.9**	**97.7**	**98.8**	**94.4**	**97.4**	**97.8**	**92.9**	95.9	**97.2**	**98.7**	**98.7**	**99.1**

Table 2: Results of visual place recognition on more challenging datasets.

Method	MSLS challenge			NordLand		
	R@1	R@5	R@10	R@1	R@5	R@10
Conv-AP [8]	54.2	66.6	71.5	38.2	54.8	61.2
CosPlace [8]	61.4	72.0	76.6	54.4	69.8	75.9
GCL [29]	57.9	70.7	75.7	8.8	15.5	18.7
MixVPR [8]	64.0	75.9	80.6	58.4	74.6	80.0
EigenPlaces [8]	67.4	77.1	81.7	54.2	68.0	73.9
CricaVPR† [53]	67.5	79.6	83.0	59.7	77.2	82.2
BoQ [8]	73.0	82.9	86.0	74.4	86.1	89.8
BoQ* [8]	79.0	90.3	92.0	81.3	92.5	94.8
SALAD [20]	75.0	88.8	91.3	76.0	89.2	92.0
DDA-VPR (Ours)	**79.8**	**90.7**	**92.6**	**83.9**	**94.8**	**96.9**

where $\mathbf{F}^i$ denotes the global representation of the i-th training image I_i. $\mathcal{P}_i$ and $\mathcal{N}_i$ respectively represent positive and negative sample sets for I_i. The terms $s(\mathbf{F}^i, \mathbf{F}^j)$ and $s(\mathbf{F}^i, \mathbf{F}^k)$ represent similarities between global representations of image pairs, and α, β, and λ are hyperparameters.

4 Experimental Results

In this section, we first describe the experimental settings. We then compare our DDA-VPR with state-of-the-art visual place recognition methods on several challenging public datasets. Finally, we analyze the contribution of each component on the representative Pitts30k and NordLand datasets.

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on several public challenging VPR datasets, including Pitts30k [42], MSLS [48], SPED [10], Tokyo24/7 [43], and NordLand [41]. Pitts30k dataset consists over 6,000 query images and 10,000 database images with significant viewpoint variations. MSLS dataset comprises over 1.6 million images from 30 major cities across six continents. SPED dataset comprises 607 query images and 607 reference images, with seasonal and day-to-night variations. Tokyo24/7 dataset includes more than 75,000 database images and 315 query images, presenting day-to-night variations. NordLand dataset contains over 2,700 query images and more than 27,000 database images, captured from a train driver's perspective over four seasons.

Figure 4: Qualitative comparisons of top-1 retrieval image between our DDA-VPR and state-of-the-art methods. The correct retrieval is framed in green, while the incorrect is in red.

4.1.2 Evaluation Metrics

We adopt the widely-used metrics Recall@N (R@N, N=1, 5, 10) in our experiments. R@N measures the proportion of query images for which at least one accurate matching image is identified within the top N retrieved images from the database, with a predefined distance threshold of 25 meters. For the NordLand dataset, this threshold is set to 1 frame.

4.1.3 Implementation Details

We use DINOv2 [37] based on ViT-B/14 [12] as visual backbone $\mathcal{F}(\cdot)$. Inspired by existing methods [20], we fine-tune the last three layers of the backbone while freezing the remaining parameters. The feature dimension D is 384. The final representation is projected to a dimension of $4,096$ using a FC layer. Following existing methods [3, 20, 32], the model is trained on GSV-Cities [1] dataset and evaluated on different benchmarks. Input images are resized to 280×280. We adopt the multi-similarity loss [47] and AdamW optimizer to train the model, with an initial learning rate of $2e-4$. Following [32], hyper-parameters α, β, λ in Eq. (12) are set to 1.0, 50.0, and 0, respectively. We train the model for 40 epochs with a batch size of 60 locations, each represented by four images. All experiments are performed on NVIDIA Tesla V100 GPU.

4.2 Comparison with State-of-the-Art Methods

To evaluate the performance of DDA-VPR, we compare it against state-of-the-art visual place recognition methods across several challenging public datasets. Results are provided in Tables 1 and 2. Notably, CricaVPR [32] leverages information from other query images within the inference batch, causing performance to vary depending on the inference batch size. For a fair comparison, the inference batch size of CricaVPR is set to 1, which is labeled as "CricaVPR†" in the tables. Additionally, BoQ [3] provides results using the DINOv2 backbone on their GitHub repository, which is labeled as "BoQ*" in the tables.

As shown, our DDA-VPR consistently outperforms existing methods across six datasets and ranks second in R@5 on SPED. The superior results are attributed to DDA-VPR's ability to effectively integrate complementary spatial and frequency domain information. Compared to BoQ [3], the existing best-performing method, DDA-VPR demonstrates better results due to the incorporation of frequency-domain aggregation, which captures critical cues ignored by BoQ. Moreover, the proposed triple fusion and MSC-Attn enhances feature aggregation, contributing to the robustness and discrimination of representations.

Further, we provide visualizations of top-1 retrieval results to qualitatively evaluate the performance of DDA-VPR in Fig. 4. As shown, our DDA-VPR successfully retrieves accu-

Table 3: Effect of components in DDA-VPR.

Method	Spatial	Frequency	MSC Attn	Triple Fusion	Pitts30k			NordLand		
					R@1	R@5	R@10	R@1	R@5	R@10
Model 1	✓				91.4	95.8	95.5	79.7	91.0	92.9
Model 2		✓			90.9	95.4	94.8	78.6	90.4	92.5
Model 3	✓	✓			92.7	96.6	97.3	81.3	92.2	93.6
Model 4	✓	✓	✓		93.4	97.1	97.9	82.1	93.2	94.7
DDA-VPR	✓	✓	✓	✓	**93.9**	**97.7**	**98.8**	**83.9**	**94.8**	**96.9**

rate matches under various challenging conditions. For instance, in the first row, the query and corresponding reference images represent the same location but exhibit significant variations in season and illumination. While existing methods fail to produce correct matches under such circumstances, DDA-VPR generates robust representations via the dual-domain aggregation strategy, accurately identifying the location.

Both qualitative and quantitative comparisons with SOTA methods across several challenging benchmarks demonstrate the effectiveness and generalization of our DDA-VPR.

4.3 Ablation Study

4.3.1 Effect of the Dual-Domain Pipeline

DDA-VPR introduces a dual-domain framework that processes local features in both spatial and frequency domains, enabling the extraction of complementary information. We conduct experiments to evaluate its effectiveness, with results presented in Table 3. "Model 1" is the baseline with only spatial features and uses average pooling followed by self-attention to downsample local features. "Model 2" uses only frequency features, while other operations are the same with "Model 1". "Model 3" aggregates features from both spatial and frequency domains, and fuses them via the simple concatenation and FC operations. It can be seen that leveraging information from both spatial and frequency domains brings significant improvements over only using individual domain information. This underscores the importance of incorporating the frequency domain in VPR, as it provides complementary cues to the spatial domain, significantly enhancing the final performance.

4.3.2 Effect of Multi-Scale Contextual Attention

We propose a novel Multi-Scale Contextual Attention (MSC-Attn) module to improve feature aggregation in both spatial and frequency domains. As shown in Table 3, "Model 4" uses MSC-Attn to capture multi-scale information during downsampling, which obtains better performance than "Model 3". MSC-Attn ensures the preservation of fine-grained cues essential for distinguishing between visually similar locations during downsampling. This highlights its crucial role in improving feature quality for VPR.

4.3.3 Effect of Triple Fusion Mechanism

To facilitate the integration of features from spatial and frequency domains, DDA-VPR designs a triple fusion mechanism. This mechanism uses cross-attention between domain-specific features, followed by self-attention, to produce a robust and unified representation that emphasizes critical information from both domains. The effect of this mechanism is

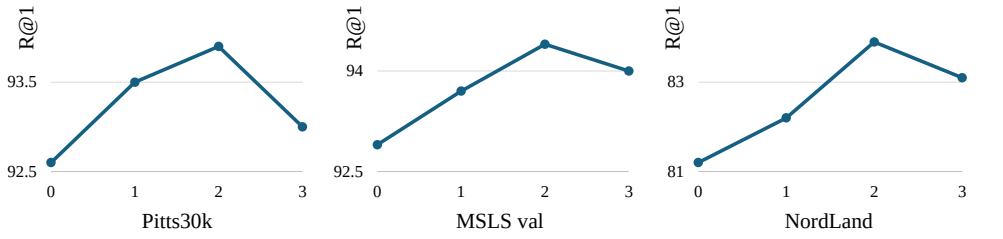

Figure 5: Effect of different number of downsampling layers in MSC-Attn.

evaluated by incorporating it into "Model 4", resulting in the fully-developed model DDA-VPR in Table 3. The results indicate further performance improvements, demonstrating that the triple fusion strategy effectively bridges domain gaps and enhances feature integration. By emphasizing complementary information, DDA-VPR produces a more discriminative and robust final representation, ultimately contributing to state-of-the-art VPR performance.

4.3.4 Effect of Downsampling Depth in MSC-Attn

In MSC-Attn, query, key, and value matrices are downsampled, enabling multi-scale interaction. We evaluate the impact of different number of downsampling layers, with results shown in Fig. 5. Performance improves as the number of downsampling layers increases from 0 to 2, reaching its peak at 2 layers. This indicates that multi-scale information enhances feature aggregation despite the reduced resolution. However, further increasing the number of layers beyond 2 leads to performance degradation, likely due to the loss of critical information at small feature resolutions.

5 Conclusion

In this paper, we propose Dual-Domain Aggregation for Visual Place Recognition (DDA-VPR), a novel method that enhances feature aggregation by jointly leveraging spatial and frequency domain information. Specifically, DDA-VPR designs a Multi-Scale Contextual Attention (MSC-Attn) module for feature abstraction in both domains, which effectively incorporates multi-scale information and preserves critical details during downsampling. More importantly, a triple fusion mechanism is introduced to mitigate domain gaps and produce a robust global representation. Extensive experiments on several challenging benchmarks demonstrate the state-of-the-art performance and effectiveness of DDA-VPR.

However, the effectiveness of frequency-domain features may degrade with poor image quality (*e.g.*, heavy compression), which can distort spectral characteristics and reduce robustness.

Acknowledgements

This work was supported by Guangdong Basic and Applied Basic Research Foundation under grants 2023A1515110695 and 2024A1515140109.

References

[1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguere. GSV-Cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022.

[2] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. MixVPR: Feature mixing for visual place recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 2998–3007, 2023.

[3] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. BoQ: A place is worth a bag of learnable queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17794–17803, 2024.

[4] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918, 2012.

[5] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *European Conference on Computer Vision*, pages 404–417, 2006.

[7] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022.

[8] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. Eigenplaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 11080–11090, 2023.

[9] Hanrui Chen, Lei Deng, Zhixiang Chen, Chenhua Liu, Lianqing Zhu, Mingli Dong, Xitian Lu, and Chentong Guo. SFCFusion: Spatial-frequency collaborative infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[10] Zetao Chen, Adam Jacobson, Niko Sünderhauf, Ben Upcroft, Lingqiao Liu, Chunhua Shen, Ian Reid, and Michael Milford. Deep learning features at scale for visual place recognition. In *IEEE International Conference on Robotics and Automation*, pages 3223–3230, 2017.

[11] Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. *Annual Conference on Neural Information Processing Systems*, 33:4479–4488, 2020.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

[13] Yongzhi Fan, Xin Du, Lun Luo, and Jizhong Shen. Fresco: Frequency-domain scan context for lidar-based place recognition with translation and rotation invariance. In *International Conference on Control, Automation, Robotics and Vision*, pages 576–583, 2022.

[14] Chaowei Fang, Dingwen Zhang, Liang Wang, Yulun Zhang, Lechao Cheng, and Junwei Han. Cross-modality high-frequency transformer for mr image super-resolution. In *Proceedings of ACM International Conference on Multimedia*, pages 1584–1592, 2022.

[15] Quan Fang, Jitao Sang, and Changsheng Xu. Giant: Geo-informative attributes for location recognition and exploration. In *Proceedings of ACM International Conference on Multimedia*, pages 13–22, 2013.

[16] Sourav Garg, Tobias Fischer, and Michael Milford. Where is your place, visual place recognition? In *International Joint Conferences on Artificial Intelligence*, volume 8, pages 4416–4425, 2021.

[17] Petr Gronat, Guillaume Obozinski, Josef Sivic, and Tomas Pajdla. Learning and calibrating per-location classifiers for visual place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 907–914, 2013.

[18] Yan He, Bing Tu, Bo Liu, Jun Li, and Antonio Plaza. 3DSS-Mamba: 3D-spectral-spatial mamba for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Zheng-Jun Zha, Yan Lu, and Baining Guo. Adaptive frequency filters as efficient global token mixers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6049–6059, 2023.

[20] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17658–17668, 2024.

[21] Hyesu Jang, Minwoo Jung, and Ayoung Kim. Raplace: Place recognition for imaging radar using radon transform and mutable threshold. In *International Conference on Intelligent Robots and Systems*, pages 11194–11201, 2023.

[22] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010.

[23] Bo Jiang, Jinxing Li, Huafeng Li, Ruxian Li, David Zhang, and Guangming Lu. Enhanced frequency fusion network with dynamic hash attention for image denoising. *Information Fusion*, 92:420–434, 2023.

[24] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2136–2145, 2017.

[25] Yitzhak Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

[26] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. AnyLoc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023.

[27] Manish Kumar, Suman Kumar Maji, and Anirban Saha. DRSFANet: Dual-path CNN with residual and frequency attention for image denoising. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[28] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. FNet: Mixing tokens with fourier Transforms. In *Proceedings of Conference of the Association for Computational Linguistics*, 2021.

[29] María Leyva-Vallina, Nicola Strisciuglio, and Nicolai Petkov. Data-efficient large scale place recognition with graded similarity supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 23487–23496, 2023.

[30] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[31] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2015.

[32] Feng Lu, Xiangyuan Lan, Lijun Zhang, Dongmei Jiang, Yaowei Wang, and Chun Yuan. CricaVPR: Cross-image correlation-aware representation learning for visual place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16772–16782, 2024.

[33] Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. SuperVLAD: Compact and robust image descriptors for visual place recognition. In *Advances in Neural Information Processing Systems*, pages 5789–5816, 2024.

[34] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16317–16326, 2021.

[35] Xintian Mao, Yiming Liu, Fengze Liu, Qingli Li, Wei Shen, and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1905–1913, 2023.

[36] Ju-Hyeon Nam, Nur Suriza Syazwany, Su Jung Kim, and Sang-Chul Lee. Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11480–11491, 2024.

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[38] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. FcaNet: Frequency channel attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 783–792, 2021.

[39] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning CNN image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018.

[40] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. In *Annual Conference on Neural Information Processing Systems*, pages 980–993, 2021.

[41] Niko Sünderhauf, Peer Neubert, and Peter Protzel. Are we there yet? challenging SeqSLAM on a 3000 km journey across all four seasons. In *IEEE International Conference on Robotics and Automation Workshop*, page 2013, 2013.

[42] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 883–890, 2013.

[43] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817, 2015.

[44] An Gia Vien, Hyunkook Park, and Chul Lee. Dual-domain deep convolutional neural networks for image demoireing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 470–471, 2020.

[45] Chenxi Wang, Hongjun Wu, and Zhi Jin. Fourllie: Boosting low-light image enhancement by fourier frequency information. In *Proceedings of ACM International Conference on Multimedia*, pages 7459–7469, 2023.

[46] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022.

[47] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019.

[48] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2626–2635, 2020.

[49] Qieshi Zhang, Zhenyu Xu, Zhiyong Yang, Ziliang Ren, Shuai Yuan, and Jun Cheng. Enhancing visual place recognition using discrete cosine transform and difference-based descriptors. *IEEE Transactions on Circuits and Systems–II*, 2024.

[50] Man Zhou, Jie Huang, Chongyi Li, Hu Yu, Keyu Yan, Naishan Zheng, and Feng Zhao. Adaptively learning low-high frequency information integration for pan-sharpening. In *Proceedings of ACM International Conference on Multimedia*, pages 3375–3384, 2022.

[51] Man Zhou, Jie Huang, Keyu Yan, Hu Yu, Xueyang Fu, Aiping Liu, Xian Wei, and Feng Zhao. Spatial-frequency domain information integration for pan-sharpening. In *European Conference on Computer Vision*, pages 274–291, 2022.

[52] Yingying Zhu, Jiong Wang, Lingxi Xie, and Liang Zheng. Attention-based pyramid aggregation network for visual place recognition. In *Proceedings of ACM International Conference on Multimedia*, pages 99–107, 2018.