Spatial-Frequency Domain Aggregation for Visual Place Recognition

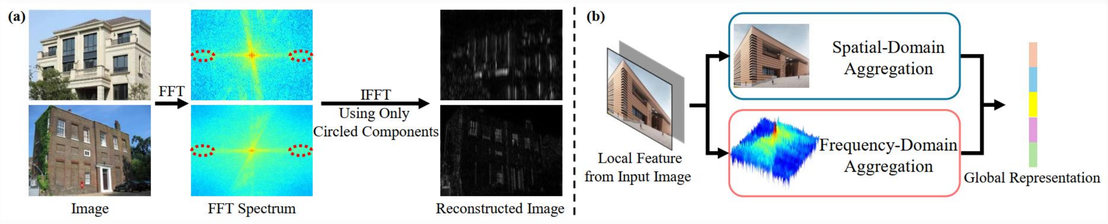
Chaoqun Wang[1], Shaobo Min[2]

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

[2]Data Platform, Tencent, Shenzhen, China

cq23@scnu.edu.cn, mbobo2ml@gmail.com

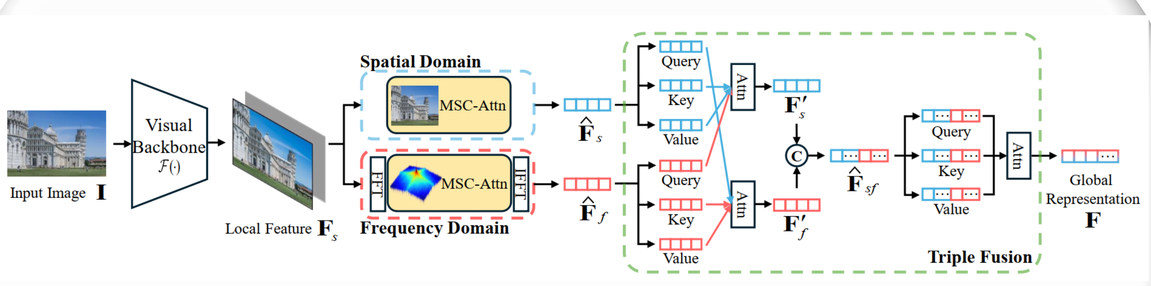
Contributions



(a) Global patterns can be effectively captured in frequency domain. (b) DDA-VPR aggregates features in both domains

- **Dual-Domain Aggregation for Visual Place Recognition** (DDA-VPR): aggregates robust global representations by leveraging complementary cues from both spatial and frequency domains.
- **Triple Fusion Mechanism**: bridges the domain gap between spatial and frequency features to produce highly discriminative representations.
- **Multi-Scale Contextual Attention** (MSC-Attn): explores multiscale information and preserves crucial details during downsampling.

Method



Pipeline of DDA-VPR

- **Triple Fusion:**

integrates different domain features and bridges the inherent domain gap.

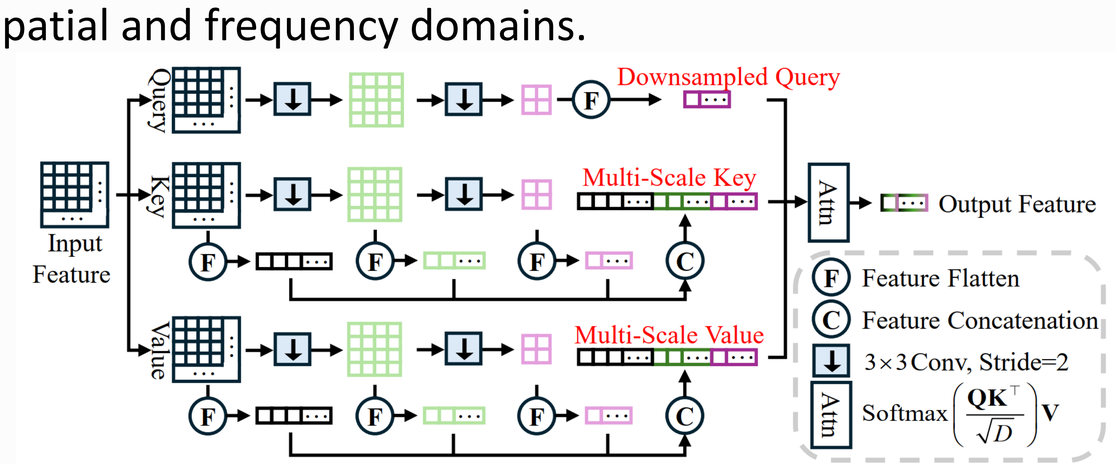
fuses complementary information from spatial and frequency features, ultimately generating robust global representations.

$$\mathbf{F}'_s = \text{Softmax}\left(\frac{q_s(\hat{\mathbf{F}}_f)\cdot k_s(\hat{\mathbf{F}}_s)^\top}{\sqrt{D}}\right) v_s(\hat{\mathbf{F}}_s), \quad \mathbf{F}'_f = \text{Softmax}\left(\frac{q_f(\hat{\mathbf{F}}_s)\cdot k_f(\hat{\mathbf{F}}_f)^\top}{\sqrt{D}}\right) v_f(\hat{\mathbf{F}}_f).$$

$$\mathbf{F} = \text{Softmax}\left(\frac{q_{sf}(\hat{\mathbf{F}}_{sf})k_{sf}(\hat{\mathbf{F}}_{sf})^\top}{\sqrt{D}}\right) v_{sf}(\hat{\mathbf{F}}_{sf}), \quad \text{where } \hat{\mathbf{F}}_{sf} = \text{cat}(\mathbf{F}'_s;\mathbf{F}'_f),$$

- **MSC-Attn:**

integrates multi-scale contextual information and preserves fine-grained details during downsampling in both spatial and frequency domains.



Architecture of MSC-Attn

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

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

$$\hat{\mathbf{K}} = \text{cat}(\mathbf{K};\mathbf{K}^{D_1};\mathbf{K}^{D_2}), \hat{\mathbf{V}} = \text{cat}(\mathbf{V};\mathbf{V}^{D_1},\mathbf{V}^{D_2}).$$

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

Results

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 [1]	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 [7]	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 [2]	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† [32]	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 [3]	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* [3]	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**

Results of VPR on four benchmarks

- **Quantitative Comparisons:**

DDA-VPR consistently outperforms existing methods across datasets and ranks second in R@5 on SPED.

Qualitative comparisons of top-1 retrieval image

- **Qualitative Comparisons:**

DDA-VPR successfully retrieves accurate matches under various challenging conditions.

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**

Effects of each component

- **Effect of the Dual-Domain Pipeline:**

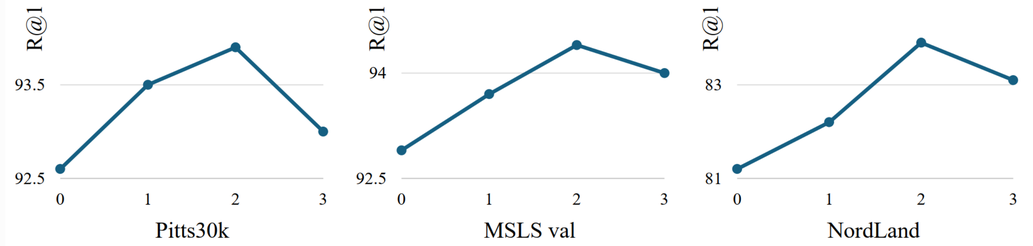
Comparing Models 1, 2, and 3 shows that leveraging information from both spatial and frequency domains brings significant improvements over only using individual domain information.

- **Effect of Multi-Scale Contextual Attention:**

Comparing Models 3 and 4 shows that MSC-Attn ensures the preservation of fine-grained cues essential for distinguishing between visually similar locations during downsampling.

- **Effect of Triple Fusion Mechanism:**

Comparing Model 4 and DDA-VPR shows that triple fusion strategy effectively bridges domain gaps and enhances feature integration.



Effect of different number of downsampling layers

- **Effect of Downsampling Depth in MSC-Attn:**

Multi-scale information enhances feature aggregation. 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.

Limitations

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