Query-Based Adaptive Aggregation for Multi-Dataset Joint Training Toward Universal Visual Place Recognition

Jiuhong Xiao[1,2], Yang Zhou[1], and Giuseppe Loianno[2]

***Abstract*— Visual Place Recognition (VPR)—matching a query image against a geo-referenced image database that serves as a prior map—underpins map-relative localization, SLAM loop closure, and GPS-denied autonomy. Most VPR models are trained on a single dataset and inherit its biases, limiting generalization across the heterogeneous prior maps encountered in real-world deployment. Multi-dataset joint training is a natural fix, but divergences among training datasets can saturate the limited information capacity of feature aggregation layers. We propose Query-based Adaptive Aggregation (QAA), which uses learned queries as reference codebooks to expand aggregation capacity at minimal compute and parameter cost, and Cross-query Similarity (CS), a simple matrix-product paradigm between query-level image features and these codebooks that yields robust descriptors. QAA outperforms state-of-the-art VPR models, balancing generalization across diverse datasets with peak per-dataset performance. Project page: `xjh19971.github.io/QAA`.**

I. Introduction

Visual Place Recognition (VPR) [1], [2], [3] is a fundamental robotic perception task that involves retrieving the top-K most similar images from a database of geo-referenced or pose-annotated images—a *prior map* of the operating environment—given a query image captured onboard. By matching robot observations against this prior map, VPR forms the bridge between a robot and its map-based world model, supporting downstream tasks such as large-scale coarse-to-fine camera localization [4], loop closure in SLAM systems [5], [6], and absolute localization in GPS-denied environments [7], [8]. Learning-based VPR methods [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22] are now the standard recipe for this matching step. However, real-world deployment exposes VPR to highly heterogeneous prior maps—differing in viewpoint, sampling density, geographic coverage, and capture conditions—and to challenges including domain shifts (e.g., day-night variation), changes in viewpoint (e.g., front-view vs. multi-view), occlusion by moving objects, and the absence of prominent landmarks. To address these challenges, researchers have developed several outdoor large-scale VPR datasets [23], [12], [15] to train robust models, each capturing different environmental conditions and scene characteristics. These datasets introduce specific biases in viewpoint, domain, sampling density, and geographic diversity, and a model that overfits to one such bias generalizes poorly when deployed against an unfamiliar prior map.

[1]The authors are with New York University, New York, NY 10012, USA. email: {jx1190, yz5794}@nyu.edu.

[2]The authors are with the University of California Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, CA 94720, USA. email: loiannog@eecs.berkeley.edu.

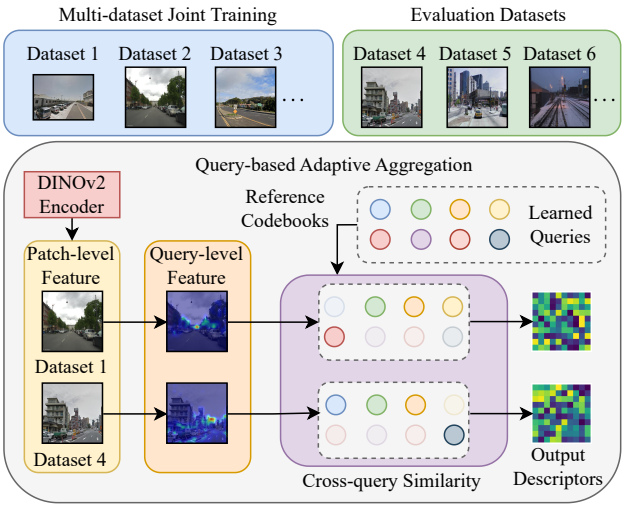

Fig. 1: Query-based Adaptive Aggregation (QAA) for Multi-Dataset Joint Training toward Universal Visual Place Recognition (VPR). QAA calculates cross-query similarity matrices between query-level image features and reference codebooks—constructed by learned queries—for output descriptors, improving the information capacity of aggregation layers and enhancing cross-domain generalization.

Joint training across multiple VPR datasets [13], [24], [25] has emerged as a promising approach to achieving more robust and universally applicable outdoor VPR models capable of operating against the diverse prior maps that arise in real-world robotic autonomy. One pioneering example, SALAD Clique Mining (CM) [13], introduced a clique-based grouping method to cluster densely sampled images from the MSLS dataset [23], facilitating joint training with the sparser GSV-Cities dataset [15]. Our cross-dataset evaluations similarly indicate that models trained exclusively on single datasets become biased toward specific dataset characteristics, resulting in limited generalization capability. Conversely, joint training across multiple diverse datasets consistently yields superior performance in a broad range of benchmarks. Nonetheless, our results also show that the baseline aggregation protocol proposed in [14] occasionally underperforms relative to dataset-specific models, likely due to significant training dataset divergence and limited information capacity in feature aggregation layers.

To address these challenges, we introduce Query-based Adaptive Aggregation (QAA) (Fig. 1). The main contribu-

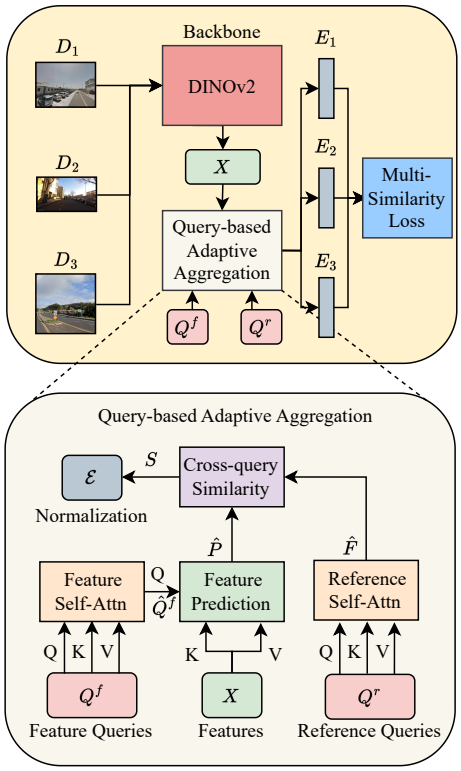

Fig. 2: The training framework (top) and Query-Based Adaptive Aggregation (QAA) architecture (bottom).

tions are:

- We propose **Query-based Adaptive Aggregation (QAA)**, which uses learned queries as an independent reference codebook to scale aggregation capacity at fixed descriptor dimension and minimal overhead.
- We introduce **Cross-query Similarity (CS)**, an aggregation paradigm that constructs similarity matrices between query-level image features and reference codebooks to generate robust descriptors with higher information capacity than score-based aggregation.
- QAA outperforms state-of-the-art VPR methods, delivering balanced generalization across heterogeneous prior maps with peak per-dataset performance.

II. Methodology

A. Framework Overview

We present the training framework (Fig. 2) for multi-dataset training. We denote $\mathcal{D} = \{D_1, D_2, \cdots, D_n\}$ as training datasets, where n is the number of datasets. From each dataset, we sample k places, with each place consisting of m images. The concatenated batch from all datasets is represented as $\mathcal{B} = \{B_1, B_2, \cdots, B_n\}$, where $\mathcal{B}$ contains $n \times k \times m$ images in total as the input image batch.

We use DINOv2-B/14 [26] as the backbone to balance performance and latency. The input has dimensions $N \times C \times H \times W$, where $N = n \times k \times m$ represents the number of images per batch, and C, H, and W are the image channels, height, and width. The output includes a feature map X of size $N \times P \times C_o$, with $P = \frac{H}{14} \times \frac{W}{14}$ and C_o as the output feature map channels. We then employ the Query-based Adaptive Aggregation (QAA) method for feature aggregation. The final outputs $\mathcal{E}$ are descriptors of size $N \times C_d$, where C_d represents the output descriptor dimension. This output can be decomposed as $\mathcal{E} = \{E_1, E_2, \cdots, E_n\}$ for each dataset. The overall framework design allows the flexibility to incorporate additional datasets for joint training.

We train the final blocks of DINOv2 and the QAA module with Multi-Similarity (MS) Loss [27], [15], treating same-place images as positives and other places as negatives.

B. Query-based Adaptive Aggregation (QAA)

QAA introduces two learnable parameters (Fig. 2): reference queries $Q^r \in \mathbb{R}^{N_q \times C_r}$ and feature queries $Q^f \in \mathbb{R}^{N_q \times C_o}$, where N_q is the number of queries and C_r is the channel count of Q^r.

Query-Level Image Features. To generate query-level image features $\hat{P}$, QAA employs two modules: 1) a *Feature Self-Attention (Feature Self-Attn)* mechanism, utilizing a Multi-Head Attention (MHA) module [28], and 2) a *Feature Prediction* module, which integrates an MHA module with a projection layer to reduce the channel dimension to C_f, where C_f represents the number of channels for $\hat{P}$. This process is defined as:

$$\hat{Q}^f = Q^f + \text{Feature-Self-Attn}(Q^f, Q^f, Q^f), \tag{1}$$

$$\hat{P} = \text{Feature-Prediction}(\hat{Q}^f, X, X), \tag{2}$$

where $\hat{P}$ denotes the query-level image features with dimensions $N_q \times C_f$ (omitting the batch dimension) and $\hat{Q}^f$ denotes the self-attention-refined version of Q^f. The patch-level image feature X serves as the key and value in the feature prediction module. The inclusion of the self-attention module, inspired by [18], minimizes the need for substantial modifications to Q^f, thereby enhancing training stability and accelerating convergence. $\hat{Q}^f$ can be cached after training.

Independent Reference Codebook. The reference codebook $\hat{F}$ is derived from the reference queries Q^r using a *Reference Self-Attention (Ref-Self-Attn)* mechanism implemented with an MHA module:

$$\hat{F} = Q^r + \text{Ref-Self-Attn}(Q^r, Q^r, Q^r) \tag{3}$$

where $\hat{F}$ represents the reference codebook with dimensions $N_q \times C_r$. Similar to $\hat{Q}^f$, $\hat{F}$ can be cached after training. By employing these learned queries as the reference codebooks for feature aggregation, QAA enhances the information capacity of the aggregation layers and transforms the predicted features—whose dimensionality scales with the query count—into fixed-dimensional descriptors, ensuring scalability regardless of the input complexity.

Cross-query Similarity. The Cross-query Similarity (CS) matrix S, of dimensions $C_r \times C_f$, quantifies the similarity between image features and the reference codebook. It is computed through matrix multiplication of $\hat{P}$ and $\hat{F}$:

$$S = \hat{F}^\top \hat{P}, \tag{4}$$

where S represents the pairwise similarity matrix between $\hat{F}$ and $\hat{P}$ along the query dimension of N_q, justifying the term

TABLE I: Recall@1 Comparison of State-of-the-Art VPR Methods with Our Results on Multi-view VPR Datasets. The best results are highlighted in **bold** and the rest of the top-3 results are underlined.

	Backbones	C_d	AmsterTime	Eynsham	Pitts250k	Pitts30k	SPED	SF-XL v1	SF-XL v2	Tokyo24/7
NetVLAD [9]	VGG-16	4096	16.3	77.7	85.9	85.0	-	40.0	76.9	69.8
SRFS [10]	VGG-16	4096	29.7	72.3	90.4	89.1	-	50.3	83.8	80.3
Conv-AP [15]	ResNet-50	4096	33.9	87.5	92.4	90.5	80.1	47.5	74.4	76.2
MixVPR [16]	ResNet-50	4096	40.2	89.4	94.2	91.5	85.2	71.1	88.5	85.1
CosPlace [12]	ResNet-50	2048	47.7	90.0	92.3	90.9	75.3	76.4	88.8	87.3
EigenPlace [11]	ResNet-50	2048	48.9	90.7	94.1	92.5	82.4	84.1	90.8	93.0
BoQ [18]	DINOv2-B	12288	63.0	92.2	**96.6**	93.7	**92.5**	91.8	**95.2**	98.1
SALAD CM [14], [13]	DINOv2-B	8448	58.1	91.9	95.2	92.6	89.1	85.6	94.6	96.8
QAA (Ours)	DINOv2-B	8192	**63.7**	**92.9**	**96.6**	**94.4**	91.8	**94.4**	94.6	**98.4**
		4096	61.8	92.9	96.3	93.8	91.1	94.2	94.0	97.8
		2048	61.5	92.7	96.4	94.0	91.1	94.0	94.1	96.5
		1024	59.8	92.5	96.3	93.9	90.8	92.4	94.5	97.1

TABLE II: Recall@1 Comparison of State-of-the-Art VPR Methods with Our Results on Front-view VPR Datasets. The best results are highlighted in **bold** and the rest of the top-3 results are underlined.

	Backbones	C_d	MSLS Val	MSLS Challenge	Nordland*	Nordland**	SVOX Night	SVOX Overcast	SVOX Rain	SVOX Snow	SVOX Sun
NetVLAD [9]	VGG-16	4096	58.9	-	-	13.1	8.0	66.4	51.5	54.4	35.4
SRFS [10]	VGG-16	4096	70.0	-	-	16.0	28.6	81.1	69.7	76.0	54.8
Conv-AP [15]	ResNet-50	4096	83.4	-	38.2	62.9	43.4	91.9	82.8	91.0	80.4
MixVPR [16]	ResNet-50	4096	88.0	64.0	58.4	76.2	64.4	96.2	91.5	96.8	84.8
CosPlace [12]	ResNet-50	2048	87.4	67.5	54.4	71.9	50.7	92.2	87.0	92.0	78.5
EigenPlace [11]	ResNet-50	2048	89.2	67.4	54.2	71.2	58.9	93.1	90.0	93.1	86.4
BoQ [18]	DINOv2-B	12288	93.8	79.0	81.3	90.6	**97.7**	**98.5**	**98.8**	**99.4**	97.5
SALAD CM [14], [13]	DINOv2-B	8448	94.2	82.7	90.7	95.2	95.6	**98.5**	98.4	99.2	98.1
QAA (Ours)	DINOv2-B	8192	97.6	**85.7**	**91.8**	**96.7**	97.2	98.4	98.4	99.1	97.3
		4096	**98.1**	84.8	91.6	96.6	97.2	**98.5**	97.9	99.0	**98.2**
		2048	97.8	84.2	91.4	95.6	96.4	98.4	97.5	98.3	97.1
		1024	97.7	82.1	88.3	92.2	95.0	98.6	97.7	99.0	95.9

Cross-query Similarity. Notably, this mechanism is similar to the similarity computation between keys and queries in the attention mechanism [28]. However, a key distinction is that while attention computes similarity along the channel dimension, QAA performs similarity computation along the query dimension. Equation 4 fundamentally computes the cross-correlation matrix between query-level image features and codebooks, effectively capturing second-order statistics along the query dimension.

The final descriptor $\mathcal{E}$ is obtained by applying intra-L2 normalization on S along the C_r dimension [9], followed by a global L2 normalization. The output descriptor dimension $C_d = C_r \times C_f$. At inference, only Eqs. 2 and 4 and the final normalization are required.

III. Experimental Setup

Datasets. For multi-dataset joint training, we use GSV-Cities [15], MSLS [23], and SF-XL [12]. For evaluation, we additionally include AmsterTime [29], Eynsham [30], Pittsburgh [31], Tokyo24/7 [32], Nordland* [2] and Nordland** [33], SPED [34], and SVOX [35], spanning multi-view, front-view, day-night, seasonal, weather, and historical-grayscale variations.

Evaluation Metrics. We evaluate VPR performance using Recall@1 (R@1), focusing on the proportion of query samples with correct top-1 matches. Following prior works [18], [14], positive matches are defined using a 25 m distance threshold for most datasets, while frame-based thresholds are used for datasets like Nordland [33], [2].

Implementation details. We fine-tune the last two blocks of DINOv2-B/14 [26] alongside the QAA module with the AdamW optimizer [36] and Clique Mining [13]. Full hyperparameters, batching, training schedule, and hardware are reported on the project page.

IV. Results

A. Comparison with Baselines

Tables I and II compare R@1 of QAA against eight one-stage baselines—NetVLAD [9], SRFS [10], Conv-AP [15], MixVPR [16], CosPlace [12], EigenPlace [11], BoQ [18], and SALAD with Clique Mining (CM) [14], [13]—on multi-view and front-view datasets.

BoQ and SALAD CM each carry a training-data bias—BoQ to multi-view (its training distribution) and SALAD CM to front-view—visible in Tables I–II. QAA's joint training and expanded information capacity yield balanced gains on both: it surpasses BoQ on AmsterTime, Eynsham, Pitts30k, SF-XL v1, and Tokyo24/7 at $C_d = 8192$ vs. 12288, and significantly outperforms SALAD CM on MSLS and Nordland while matching it on SVOX.

Performance of Reduced C_d. QAA degrades gracefully as C_d shrinks from 8192 to 1024 (Tables I–II), with only mild drops on AmsterTime, SF-XL v1, MSLS Challenge, Nordland, and SVOX Night/Sun at $C_d = 1024$.

TABLE III: Performance Comparison for Joint Training across Different Datasets with SALAD CM [13] and Our QAA Approaches ($N_q = 256, C_f = 64, C_r = 128, C_d = 8192$). All methods incorporate Clique Mining (CM) [13] and DINOv2-B backbone. The best results are highlighted in **bold**.

Training Datasets	Methods	MSLS Val	Pitts250k Val	SF-XL Val	Nordland*	AmsterTime
GSV-Cities	SALAD CM	**92.7**	94.4	93.8	**70.9**	58.6
	QAA	**92.7**	**95.0**	**95.2**	69.5	**64.6**
MSLS	SALAD CM	97.6	92.2	90.8	88.7	48.5
	QAA	**98.1**	**93.2**	**92.7**	**89.5**	**53.7**
SF-XL	SALAD CM	**94.2**	95.1	94.2	87.4	59.6
	QAA	93.8	**95.5**	**97.0**	**88.6**	**61.7**
GSV-Cities + MSLS + SF-XL	SALAD CM	96.6	95.1	97.0	90.3	60.6
	QAA	**97.6**	**95.4**	**97.8**	**91.8**	**63.7**

TABLE IV: Performance Comparison for Joint Training across Different Settings for Reference Codebooks and Aggregation Methods. The best results are highlighted in **bold**. For inference complexity, GFLOPS is computed for the QAA module using a single 322×322 image.

Methods	GFLOPS	MSLS Val	Pitts250k Val	SF-XL Val	Nordland*	AmsterTime
Softmax	2.29	97.0	94.9	97.7	90.5	61.5
Softmax - Cond	7.43	97.0	94.9	95.7	89.4	58.0
OT	2.29	97.3	95.3	97.5	90.9	62.0
OT - Cond	7.43	97.0	94.9	96.0	88.1	60.0
CS	2.29	**97.6**	**95.4**	**97.8**	**91.8**	**63.7**

TABLE V: Performance Comparison for Joint Training across Different Numbers of Queries N_q with $C_f = 64$ and $C_r = 128$. The best results are highlighted in **bold**. For inference complexity, GFLOPS is computed for the QAA module using a single 322×322 image.

N_q	GFLOPS	MSLS Val	Pitts250k Val	SF-XL Val	Nordland*	AmsterTime
16	1.31	**97.6**	95.1	96.4	88.3	59.2
32	1.38	**97.6**	95.3	97.1	89.6	61.6
64	1.51	**97.6**	95.2	97.7	91.1	60.6
128	1.77	**97.6**	**95.4**	**97.8**	**92.8**	63.5
256	2.29	**97.6**	**95.4**	**97.8**	91.8	**63.7**

TABLE VI: Performance Comparison for Joint Training across Different Numbers of Channels for Learned Queries C_f and C_r with $N_q = 256$.

C_f	C_r	C_d	MSLS Val	Pitts250k Val	SF-XL Val	Nordland*	AmsterTime
64	128	8192	97.6	95.4	97.8	91.8	63.7
64	64	4096	97.7	95.4	97.4	91.0	**63.0**
32	128	4096	**98.1**	**95.5**	**97.7**	**91.6**	61.8
64	32	2048	97.4	**95.2**	**97.2**	90.4	**61.5**
16	128	2048	**97.8**	95.0	97.0	**91.4**	**61.5**
64	16	1024	**97.7**	**95.2**	**96.3**	**89.3**	58.9
8	128	1024	**97.7**	95.1	96.0	88.3	**59.8**

Inference Complexity. At $N_q = 256$, QAA's aggregation module uses 5.1M parameters and 2.29 GFLOPS for a 322×322 image—smaller and faster than attention-based BoQ (8.6M, 8.22 GFLOPS) and competitive with the convolution-based SALAD (1.4M, 0.94 GFLOPS) at higher accuracy.

B. Ablation Study

We ablate QAA on MSLS, Pitts250k, and SF-XL (close to training) and on Nordland* and AmsterTime (out-of-distribution: seasonal and historical-grayscale).

1) Cross-Dataset Evaluation: In Table III, single-dataset models overfit to their own evaluation set; joint training on all three balances performance across in- and out-of-distribution sets. QAA holds MSLS Val under joint training where SALAD CM degrades, and consistently improves on SF-XL Val, Nordland, and AmsterTime.

2) Effectiveness of Independent Reference Codebooks and CS Matrix: Table IV compares the performance of QAA across three aggregation paradigms: (1) Softmax, (2) Optimal Transportation (OT), and (3) Cross-query Similarity (CS). The results yield two main insights: (1) employing an independent reference codebook consistently benefits all paradigms, and (2) CS achieves superior performance, particularly on the MSLS Val, Nordland, and AmsterTime datasets. These findings underscore the advantage of similarity-based aggregation in strengthening VPR representations.

3) Query Scalability: Increasing N_q improves QAA on Pitts250k, SF-XL Val, Nordland, and AmsterTime and saturates near $N_q = 128$–256 (Table V); MSLS Val is stable across N_q due to its fixed front-view pattern.

4) Effect of Channel Numbers: Reducing either C_f or C_r yields a small comparable drop (Table VI); QAA remains robust even at $C_f = 8$, with stability supported by the high-dimensional codebook $\hat{F}$.

V. Conclusions

We introduced Query-based Adaptive Aggregation (QAA) and the Cross-query Similarity (CS) paradigm for multi-dataset VPR training, scaling aggregation capacity at fixed descriptor dimension and minimal compute overhead. QAA delivers balanced generalization across heterogeneous prior maps with peak per-dataset performance. Future work will address performance saturation when N_q is large.

REFERENCES

[1] S. Schubert, P. Neubert, S. Garg, M. Milford, and T. Fischer, "Visual place recognition: A tutorial [tutorial]," *IEEE Robotics & Automation Magazine*, vol. 31, no. 3, pp. 139–153, 2024.

[2] M. Zaffar, S. Garg, M. Milford, J. Kooij, D. Flynn, K. McDonald-Maier, and S. Ehsan, "Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change," *International Journal of Computer Vision*, vol. 129, no. 7, pp. 2136–2174, 2021.

[3] G. Berton, R. Mereu, G. Trivigno, C. Masone, G. Csurka, T. Sattler, and B. Caputo, "Deep visual geo-localization benchmark," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5396–5407.

[4] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 12 716–12 725.

[5] C. Masone and B. Caputo, "A survey on deep visual place recognition," *IEEE Access*, vol. 9, pp. 19 516–19 547, 2021.

[6] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[7] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for uav," *Robotics and Autonomous Systems*, vol. 135, p. 103666, 2021.

[8] J. Xiao, D. Tortei, E. Roura, and G. Loianno, "Long-range uav thermal geo-localization with satellite imagery," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 5820–5827.

[9] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5297–5307.

[10] Y. Ge, H. Wang, F. Zhu, R. Zhao, and H. Li, "Self-supervising fine-grained region similarities for large-scale image localization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer, 2020, pp. 369–386.

[11] G. Berton, G. Trivigno, B. Caputo, and C. Masone, "Eigenplaces: Training viewpoint robust models for visual place recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 080–11 090.

[12] G. Berton, C. Masone, and B. Caputo, "Rethinking visual geo-localization for large-scale applications," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4878–4888.

[13] S. Izquierdo and J. Civera, "Close, but not there: Boosting geographic distance sensitivity in visual place recognition," *arXiv preprint arXiv:2407.02422*, 2024.

[14] ——, "Optimal transport aggregation for visual place recognition," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 17 658–17 668.

[15] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, pp. 194–203, 2022.

[16] A. Ali-Bey, B. Chaib-Draa, and P. Giguere, "Mixvpr: Feature mixing for visual place recognition," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 2998–3007.

[17] F. Lu, L. Zhang, X. Lan, S. Dong, Y. Wang, and C. Yuan, "Towards seamless adaptation of pre-trained models for visual place recognition," in *The Twelfth International Conference on Learning Representations*, 2024.

[18] A. Ali-bey, B. Chaib-draa, and P. Giguère, "BoQ: A place is worth a bag of learnable queries," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 17 794–17 803.

[19] I. Tzachor, B. Lerner, M. Levy, M. Green, T. B. Shalev, G. Habib, D. Samuel, N. K. Zailer, O. Shimshi, N. Darshan, and R. Ben-Ari, "EffoVPR: Effective foundation model utilization for visual place recognition," in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] J. Xiao, G. Zhu, and G. Loianno, "Vg-ssl: Benchmarking self-supervised representation learning approaches for visual geo-localization," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, February 2025, pp. 6667–6677.

[21] M. Leyva-Vallina, N. Strisciuglio, and N. Petkov, "Data-efficient large scale place recognition with graded similarity supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 487–23 496.

[22] G. Trivigno, G. Berton, J. Aragon, B. Caputo, and C. Masone, "Divide&classify: Fine-grained classification for city-wide visual geo-localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 11 142–11 152.

[23] F. Warburg, S. Hauberg, M. López-Antequera, P. Gargallo, Y. Kuang, and J. Civera, "Mapillary street-level sequences: A dataset for lifelong place recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2623–2632.

[24] G. Berton and C. Masone, "Megaloc: One retrieval to place them all," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, June 2025, pp. 2861–2867.

[25] B. Liu, P. Zhang, L. He, H. Chen, S. Guo, Y. Wu, J. Cui, and H. Zhang, "Superplace: The renaissance of classical feature aggregation for visual place recognition in the era of foundation models," *arXiv preprint arXiv:2506.13073*, 2025.

[26] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[27] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5022–5030.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] B. Yildiz, S. Khademi, R. M. Siebes, and J. Van Gemert, "Amstertime: A visual place recognition benchmark dataset for severe domain shift," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2749–2755.

[30] M. J. Cummins and P. Newman, "Highly scalable appearance-only slam - fab-map 2.0," in *Robotics: Science and Systems*, 2009.

[31] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 883–890.

[32] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1808–1817.

[33] N. Sünderhauf, P. Neubert, and P. Protzel, "Are we there yet? challenging seqslam on a 3000 km journey across all four seasons," in *Proc. of workshop on long-term autonomy, IEEE international conference on robotics and automation (ICRA)*, 2013, p. 2013.

[34] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep learning features at scale for visual place recognition," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3223–3230.

[35] G. M. Berton, V. Paolicelli, C. Masone, and B. Caputo, "Adaptive-attentive geolocalization from few queries: A hybrid approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2021, pp. 2918–2927.

[36] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.