2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

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*— Deep learning methods for Visual Place Recognition (VPR) have advanced significantly, largely driven by large-scale datasets. However, most existing approaches are trained on a single dataset, which can introduce dataset-specific inductive biases and limit model generalization. While multi-dataset joint training offers a promising solution for developing universal VPR models, divergences among training datasets can saturate the limited information capacity in feature aggregation layers, leading to suboptimal performance. To address these challenges, we propose Query-based Adaptive Aggregation (QAA), a novel feature aggregation technique that leverages learned queries as reference codebooks to effectively enhance information capacity without significant computational or parameter complexity. We show that computing the Cross-query Similarity (CS) between query-level image features and reference codebooks provides a simple yet effective way to generate robust descriptors. Our results demonstrate that QAA outperforms state-of-the-art models, achieving balanced generalization across diverse datasets while maintaining peak performance comparable to dataset-specific models. Ablation studies further explore QAA's mechanisms and scalability. Visualizations reveal that the learned queries exhibit diverse attention patterns across datasets. 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 given a query image. Learning-based VPR methods [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17] are widely applied in robotics and computer vision, particularly in tasks such as large-scale coarse-to-fine camera localization [18], loop closure in SLAM systems [19], [20], and absolute localization in GPS-denied environments [21], [22]. However, outdoor VPR still faces challenges, including significant variations due to 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], [7], [10] to train robust models, each capturing different environmental conditions and scene characteristics. These datasets introduce specific biases influenced by factors such as camera viewpoint, domain, sampling density, and geographic diversity (Table I).

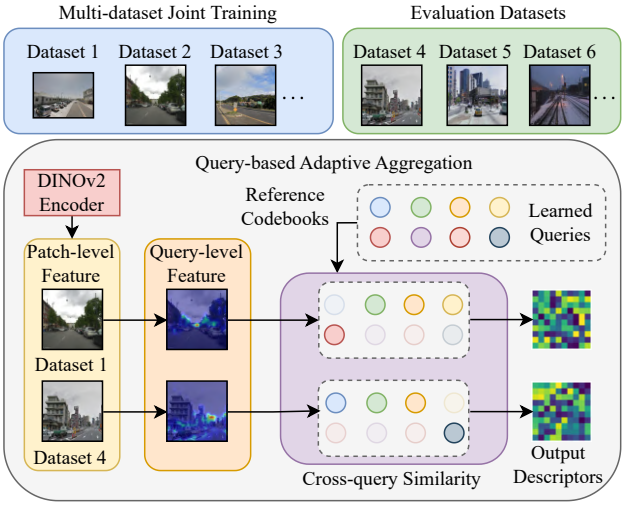

Fig. 1: Query-based Adaptive Aggregation (QAA) for Multi-Dataset Joint Training toward Universal Visual Place Recognition (VPR). QAA calculates cross-query similarity matrices between query-level image features and reference codebooks—constructed by learned queries—for output descriptors, improving the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

Joint training across multiple VPR datasets [8], [24], [25] has emerged as a promising approach to achieving more robust and universally applicable outdoor VPR models. One pioneering example, SALAD Clique Mining (CM) [8], 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 [10]. 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 [9] 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), a novel feature aggregation method aimed at enhancing multi-dataset joint training performance. QAA employs learned queries as

[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.
This work was supported by the NSF CPS Grant CNS-2603416, the NSF CAREER Award 2546659, the DARPA YFA Grant D22AP00156-00, the DEVCOM ARL Grant SARA W911NF-24-2-0057, and NYU IT High Performance Computing resources, services, and staff expertise.

reference codebooks to efficiently expand memory within aggregation layers, thereby increasing information capacity and maintaining strong performance even with low-dimensional descriptors. This approach strengthens cross-domain generalization while preserving peak performance comparable to dataset-specific models. By utilizing the cross-query similarity matrix between query-level image features and the independent reference codebook, QAA effectively models robust geographic descriptors that capture the relative spatial relationships between images. The adaptive learned queries enable the generation of diverse feature representations across datasets while introducing minimal computational and parameter overhead. The main contributions are:

- We propose the **Query-based Adaptive Aggregation (QAA)** approach that utilizes learned queries as the independent reference codebook for aggregation. QAA demonstrates its strengths in capturing the global context for query-level image features and reference codebooks, handling scalable queries without increasing output descriptor dimensions, and maintaining minimal computational and parameter overhead.
- We introduce **Cross-query Similarity (CS)**, a simple yet effective aggregation paradigm that constructs similarity matrices between image features and reference codebooks to generate robust geographic descriptors. We analyze its information capacity through coding rate [26], [27], providing insights into its performance.
- Extensive evaluations demonstrate that QAA outperforms state-of-the-art VPR methods, achieving balanced generalization across diverse datasets and peak performance compared with models trained on specific datasets. Ablation studies and visualizations provide insight into the mechanism and scalability of QAA, demonstrating the enhanced information capacity of aggregation layers for better performance and diverse attention patterns in various datasets.

II. Related Works

Learning-based VPR methods can be broadly categorized into one-stage [4], [5], [6], [7], [8], [9], [10], [11], [16] and two-stage approaches [28], [29], [12], [14], [17]. One-stage methods typically employ a CNN [30] or ViT [31] model as the feature extractor, generating local feature maps, followed by a feature aggregation module to convert the 2D feature map into a 1D global descriptor. The query image descriptor is then compared against a database to identify the top-K similar images and retrieve their geo-referenced information. Two-stage methods, in contrast, re-rank these top-k candidates by further utilizing the 2D feature maps, achieving enhanced localization performances with higher computational costs. In VPR, the output dimensionality of descriptors plays a crucial role, with a linear impact on both memory usage and matching cost.

Score-based feature aggregation methods [4], [9], also known as cluster-based aggregation, compute global descriptors by weighting and summing patch-level image features based on predicted scores. For instance, NetVLAD [4] uses softmax scores to weight the distances between image features and cluster centroids. Similarly, SALAD [9] weights patch features by scores normalized with the Sinkhorn optimal transport algorithm [32], [33]. In contrast, our QAA method eliminates explicit score prediction. Instead, it generates robust descriptors by computing a cross-query similarity matrix between learned query-level image features and an independent reference codebook, achieving comparable or reduced dimensionality.

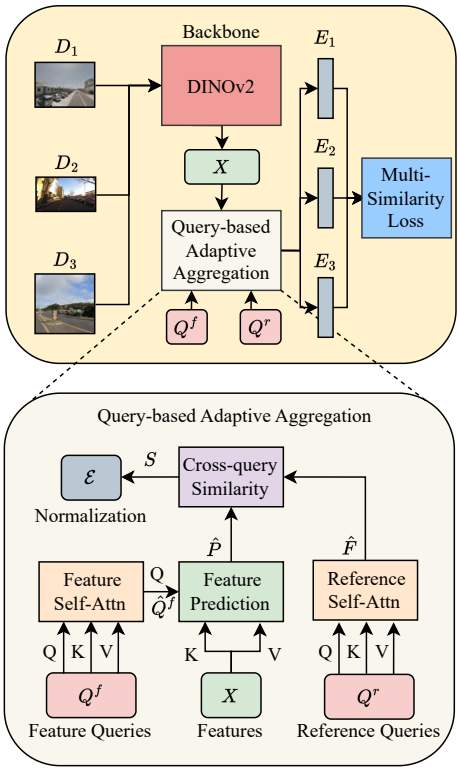

Fig. 2: The training framework (top) and Query-Based Adaptive Aggregation (QAA) architecture (bottom).

Recent VPR methods leverage foundational models like DINOv2 [34] for better generalization. For example, AnyLoc [35] leverages general-purpose representations from DINOv2, SelaVPR adds lightweight adapters to DINOv2, and EffoVPR uses DINOv2's self-attention features for reranking. The most relevant work, BoQ [13], utilizes learned queries to project image features from a DINOv2 backbone. Unlike BoQ's concatenation approach, which combines all learned query outputs and reduces dimensionality with linear layers, our QAA method computes cross-query similarity with independent reference codebooks. This approach prevents the output dimension from increasing with number of queries, enabling scalable query use with a fixed output size.

III. 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 [34] 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 employ Multi-Similarity (MS) Loss [36], [10] to train the final blocks of the DINOv2 model along with the QAA module. Images from the same place are grouped and treated as positive examples, while images from different places (groups) serve as negative samples.

B. Query-based Adaptive Aggregation (QAA)

Figure 2 illustrates the Query-based Adaptive Aggregation (QAA) approach. We define the learnable parameters—reference queries Q^r and feature queries Q^f—with dimensions $N_q \times C_r$ and $N_q \times C_o$, respectively, where N_q represents the number of queries, C_r represents the number of channels for Q^r. These queries are learned via backpropagation during training, without test-time adaptation.

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 [37], 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 [13], 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 *Cross-query Similarity*. Notably, this mechanism is similar to the similarity computation between keys and queries in the attention mechanism [37]. 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 [4], followed by a global L2 normalization of the entire vector. The output descriptor dimension C_d is given by $C_r \times C_f$. During inference, the aggregation process involves only the computations from Equations 2 and 4, along with the final normalization. This ensures that QAA maintains minimal computational and parameter complexity.

Our intuition behind the CS paradigm is to preserve the information capacity of $\hat{P}$. Unlike score-based paradigms that compress the output space into the range $[0, 1]$, the CS paradigm avoids such projection, retaining more information for interaction with the reference codebook. To quantify the information retained in $\hat{P}$, we leverage the coding rate concept from information theory [26], [27]. The coding rate is computed as:

$$R(\hat{P}^\top, \epsilon) = \frac{1}{2} \log\det\left(I + \frac{C_f}{N_q \epsilon^2} \hat{P}^\top \hat{P}\right) \tag{5}$$

where $R(\hat{P}^\top, \epsilon)$ represents the coding rate of the random variable $\hat{p}$, with finite samples given by $\hat{P}^\top = \left[\hat{p}_1, ..., \hat{p}_{N_q}\right]$, having dimensions $C_f \times N_q$ and a prescribed precision ϵ, which is set to 0.001. This equation utilizes the covariance matrix $\hat{P}^\top \hat{P}$ to quantify the information capacity of $\hat{P}$. Our coding rate analysis demonstrates that the coding rate of $\hat{P}$ in the CS paradigm exceeds that of OT and Softmax, despite sharing the same dimensionality. This result validates the superior information capacity of CS. This enhanced capacity enables CS to generate more informative descriptors, ultimately leading to improved performance.

CS aggregation paradigm represents a significant advancement in retrieval-based VPR, as it demonstrates—for the first time—that informative geographic descriptors can be formulated directly from the similarity matrix between query-

TABLE I: Overview of large-scale VPR datasets for joint training. *Sampling density refers to the spacing between sampled locations, with each location potentially containing multiple images.

Datasets	GSV-Cities [10]	MSLS [23]	SF-XL [7]
Viewpoint	multi-view	front-view	multi-view
Domain	urban	mostly urban	mostly urban
Sampling Density*	sparse	dense	dense
Number of Cities	multiple	multiple	single
Season Variation	✓	✓	✓
Day-night Change		✓	
Weather Conditions	✓	✓	✓

level image features and an independent reference codebook. This eliminates the need for explicit score prediction techniques [4], [9] or implicit linear projection. Therefore, CS aggregation enhances the interpretability of descriptor generation in retrieval-based VPR.

IV. Experimental Setup

Datasets. For multi-dataset joint training, we utilize GSV-Cities [10] with sparse place sampling, MSLS [23] incorporating place cluster information, and SF-XL [7], which is grouped by location and orientation. These large-scale datasets encompass a wide range of variations essential for outdoor VPR, as outlined in Table I.

For evaluation, we incorporate a wide range of VPR datasets. The AmsterTime dataset [38] matches historical grayscale images with current RGB images, testing robustness to long-term temporal changes. The Eynsham dataset [39] focuses on grayscale images for VPR. Pittsburgh [40], Tokyo24/7 [41], and SF-XL [7] datasets utilize Google Street View as the source of database images and primarily assess viewpoint variations. MSLS [23] collects images from a forward-facing car-mounted camera. Nordland**[42] examines seasonal changes between summer and winter conditions, using a 10-frame threshold for positive samples, while its variant, Nordland*[2], employs a 1-frame threshold, demanding higher accuracy. The SPED dataset [43] captures seasonal and day-night variations using surveillance footage. SVOX [44] poses varying illumination and weather conditions.

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 [13], [9], positive matches are defined using a 25 m distance threshold for most datasets, while frame-based thresholds are used for datasets like Nordland [42], [2].

Implementation details. For multi-dataset joint training, we set the number of places per batch to $k = 30$ and the number of images per place to $m = 4$ for each dataset, resulting in a batch size of 90 places and 360 images. Training images are resized to 224×224 and evaluation images to 322×322. We fine-tune the last two blocks of the DINOv2 model [34] alongside the QAA module, selecting models based on MSLS val performance. We optimize models using the AdamW optimizer [45] with a learning rate of 4×10^{-5} for aggregation layers. Training spans at most 80 epochs, with approximately 2000 iterations per epoch and a 4000-iteration linear learning rate warmup. A weight decay of 1×10^{-3} is applied. Clique Mining (CM) [8] is used to group place data, with cliques being recomputed during training. The implementation is based on PyTorch Lightning, and the training is conducted on a single NVIDIA A100 GPU, taking about 35 hours.

V. Results

A. Comparison with Baselines

In this section, we evaluate the R@1 performance of our proposed QAA method in comparison with state-of-the-art VPR methods. Results for multi-view datasets are presented in Table II, while those for front-view datasets are shown in Table III. Our comparison includes leading baselines such as NetVLAD [4], SRFS [5], Conv-AP [10], MixVPR [11], CosPlace [7], EigenPlace [6], BoQ [13], and SALAD with Clique Mining (CM) [9], [8]. These methods represent some of the most effective and widely recognized one-stage approaches in the field. Notably, recent baselines like BoQ and SALAD CM have demonstrated superior accuracy and efficiency than two-stage methods.

Tables II and III reveal distinct strengths for BoQ and SALAD CM on multi-view and front-view datasets, respectively, due to their specialized training. BoQ's training on a multi-view dataset creates a bias for such data, while SALAD CM's training on both, combined with a limited model capacity, leads to overfitting on front-view characteristics. Our QAA method addresses these limitations by efficiently enhancing information capacity and extensive joint training. This strategy enables QAA to learn more generalized representations that are effective across both multi-view and front-view datasets, demonstrating superior generalization capabilities compared to BoQ and SALAD CM.

Specifically, as presented in Table II, the QAA method surpasses BoQ on the AmsterTime, Eynsham, Pitts30k, SF-XL v1, and Tokyo 24/7 datasets, which are multi-view in nature. On the Pitts250k, SPED, and SF-XL v2 datasets, QAA achieves performance comparable to BoQ despite utilizing a significantly smaller output dimension (8192 vs. 12288). The reduced output dimension not only indicates computational efficiency but also suggests that QAA captures essential features without the need for a larger output descriptor dimension. These results highlight the effectiveness of our method on multi-view datasets and underscore its competitive advantage over state-of-the-art baselines.

Similarly, as shown in Table III, QAA significantly outperforms both SALAD CM and BoQ on the MSLS and Nordland evaluation sets. It also achieves performance on par with SALAD CM and BoQ on the SVOX evaluation sets. Notably, despite SALAD CM's training bias toward front-view datasets, our method delivers superior overall performance on front-view datasets. This demonstrates that QAA yields balanced and optimal performance to different dataset types, validating its robustness and adaptability.

TABLE II: 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 [4]	VGG-16	4096	16.3	77.7	85.9	85.0	-	40.0	76.9	69.8
SRFS [5]	VGG-16	4096	29.7	72.3	90.4	89.1	-	50.3	83.8	80.3
Conv-AP [10]	ResNet-50	4096	33.9	87.5	92.4	90.5	80.1	47.5	74.4	76.2
MixVPR [11]	ResNet-50	4096	40.2	89.4	94.2	91.5	85.2	71.1	88.5	85.1
CosPlace [7]	ResNet-50	2048	47.7	90.0	92.3	90.9	75.3	76.4	88.8	87.3
EigenPlace [6]	ResNet-50	2048	48.9	90.7	94.1	92.5	82.4	84.1	90.8	93.0
BoQ [13]	DINOv2-B	12288	63.0	92.2	**96.6**	93.7	**92.5**	91.8	**95.2**	98.1
SALAD CM [9], [8]	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 III: 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 [4]	VGG-16	4096	58.9	-	-	13.1	8.0	66.4	51.5	54.4	35.4
SRFS [5]	VGG-16	4096	70.0	-	-	16.0	28.6	81.1	69.7	76.0	54.8
Conv-AP [10]	ResNet-50	4096	83.4	-	38.2	62.9	43.4	91.9	82.8	91.0	80.4
MixVPR [11]	ResNet-50	4096	88.0	64.0	58.4	76.2	64.4	96.2	91.5	96.8	84.8
CosPlace [7]	ResNet-50	2048	87.4	67.5	54.4	71.9	50.7	92.2	87.0	92.0	78.5
EigenPlace [6]	ResNet-50	2048	89.2	67.4	54.2	71.2	58.9	93.1	90.0	93.1	86.4
BoQ [13]	DINOv2-B	12288	93.8	79.0	81.3	90.6	**97.7**	**98.5**	**98.8**	**99.4**	97.5
SALAD CM [9], [8]	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

TABLE IV: Performance Comparison for Joint Training across Different Datasets with SALAD CM [8] and Our QAA Approaches ($N_q = 256, C_f = 64, C_r = 128, C_d = 8192$). All methods incorporate Clique Mining (CM) [8] 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**

Performance of Reduced C_d. Tables II and III present the performance of reduced C_d. For $C_d = 4096$ and 2048, QAA overall maintains strong performance, particularly on the Eynsham, Pitts250k, Pitts30k, SPED, SF-XL, Tokyo24/7, MSLS Val and Challenge, Nordland, and SVOX evaluation sets. A mild performance degradation is observed for $C_d = 1024$ on AmsterTime, SF-XL v1, MSLS Challenge, Nordland, SVOX Night, and SVOX Sun, while performance remains competitive on other datasets. These results highlight the robustness of the resource-efficient QAA variant.

Inference Complexity. We compare the computational and parameter complexity of DINOv2 SALAD, BoQ, and QAA. SALAD employs 1.4M parameters for its aggregation module and, for a 322×322 image, requires 0.94 GFLOPS with convolution-based aggregation. For attention-based aggregation, BoQ, with 64 queries, utilizes 8.6M parameters and 8.22 GFLOPS. Compared with BoQ, QAA, despite using 256 queries, requires only 5.1M parameters and 2.29 GFLOPS, demonstrating its superior efficiency while maintaining a reduced output size.

B. Ablation Study

We evaluate the effectiveness and examine the underlying mechanism of our QAA approach, aiming to achieve balanced and optimal performance across a wide range of datasets. We assess validation R@1 performance on MSLS, Pitts250, and SF-XL, as their performance is closely tied to the training datasets. To evaluate generalization performance

TABLE V: 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 VI: 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**

across markedly different data attributes, we include the Nordland* and AmsterTime datasets in our study. These datasets introduce significant domain variations, such as seasonal transitions and historical grayscale images.

1) Cross-Dataset Evaluation: Table IV presents the cross-dataset evaluation results using the DINOv2 SALAD CM approach [8] and our QAA approaches. The findings reveal that models trained on individual datasets achieve the highest accuracy on their respective evaluation sets but struggle on others, indicating an inductive bias toward the specific characteristics of the training data. In contrast, joint training on GSV-Cities, MSLS, and SF-XL leads to more balanced performance across all three validation sets, as well as on significantly different datasets such as Nordland and AmsterTime. While joint training degrades MSLS Val performance for both models, SALAD CM exhibits a sharper decline, whereas QAA maintains performance comparable to MSLS-only training. Moreover, our QAA approaches consistently outperform the SALAD CM method across various training configurations, with particularly strong improvements on MSLS Val, SF-XL Val, Nordland, and AmsterTime, underscoring QAA's superior accuracy.

2) Effectiveness of Independent Reference Codebooks and CS Matrix: Table V 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.

We further investigate the use of a conditional reference codebook (Cond) in the Softmax and OT paradigms, aiming to capture inter-image relationships within the codebook. It is worth noting that CS does not support this extension, as incorporating a conditional codebook would reduce it to dual mappings of the same features. Our analysis highlights two key findings: (1) the conditional codebook incurs higher computational cost, since it requires image features as input, and (2) in query-driven mechanisms, it not only fails to enhance performance but instead leads to degradation compared to independent reference codebooks.

3) Coding Rate Analysis: We analyze the information capacity of $\hat{P}$ by examining the histogram of coding rates per query using the MSLS Val in Fig. 4. The results indicate that $\hat{P}$ generated by CS exhibits a $\sim 2\times$ coding rate with reduced variance compared to Softmax and OT, both of which compress the output space and restrict information content. In contrast, CS preserves more information in query-level image features, enabling richer interactions with the reference codebook. This enhanced information retention facilitates the generation of highly informative descriptors, ultimately leading to superior performance.

4) Query Scalability: Table VI explores the relationship between the number of queries, N_q, and the performance of QAA. The results indicate that increasing N_q enhances QAA's performance on Pitts250k, SF-XL Val, Nordland, and AmsterTime, with performance gains saturating at $N_q = 128$ and $N_q = 256$. The MSLS Val performance remains stable across different N_q, indicating that this evaluation set follows a consistent pattern from the fixed front-viewpoint. Therefore, fewer queries are required to capture the pattern effectively. This highlights the importance of a larger N_q in optimizing QAA's effectiveness.

5) Effect of Channel Numbers: Table VII analyzes performance variations across different values of C_f and C_r, which influence the final output dimension C_d. We observe that reducing either C_f or C_r results in a slight but comparable performance degradation. Interestingly, the model remains robust **even when C_f is extremely reduced to** 8, despite the information bottleneck in the query-level image feature $\hat{P}$, whose dimension shrinks to 256×8. This resilience is attributed to the support of the high-dimensional codebook $\hat{F}$, which helps maintain performance stability.

C. Qualitative Results

Figure 3 presents attention maps corresponding to different vectors in Q^f for the feature prediction model, evaluated across the MSLS Val, Pitts250k, and Tokyo24/7 datasets:

- **Global Context Capture**: The multi-attention module

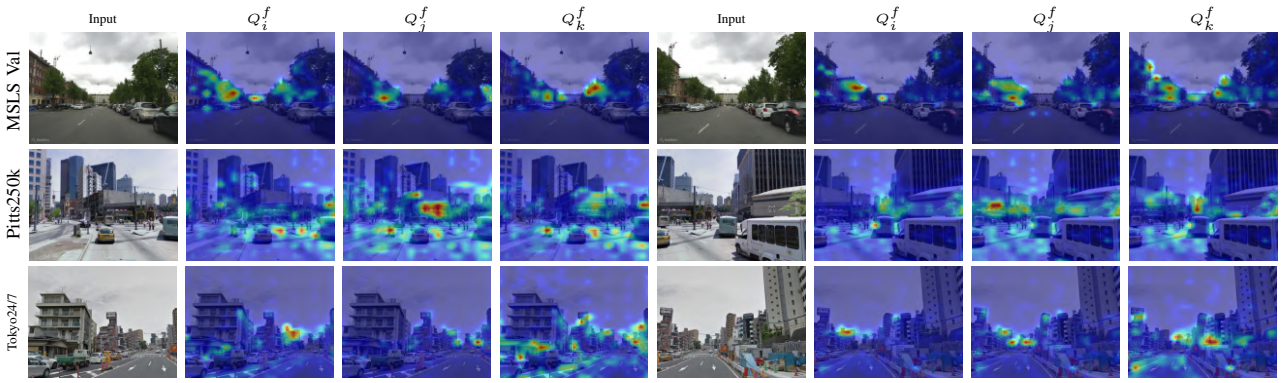

Fig. 3: Attention Maps Corresponding to Different Query Vectors (Q_i^f, Q_j^f, and Q_k^f) in Q^f from The Feature Prediction Model for A Front-view Dataset (MSLS Val) and Multi-view Datasets (Pitts250k and Tokyo24/7). Each pair of images represents the same location within the same dataset but from different viewpoints.

TABLE VII: 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**

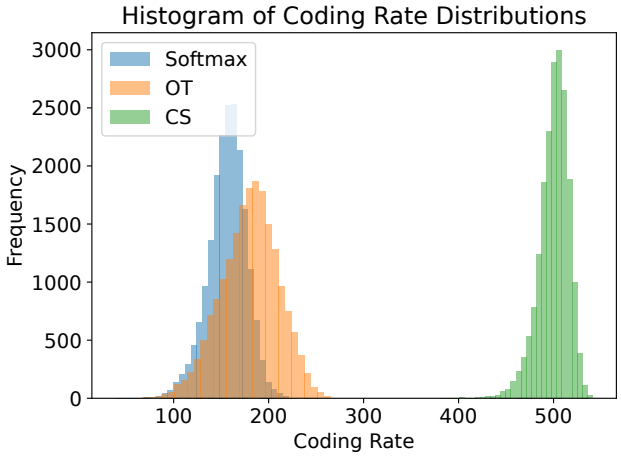

Fig. 4: Histogram of the coding rate distributions for different aggregation paradigms.

allows attention maps to span the entire image, capturing global context rather than relying solely on patch-level feature combinations. This broader perspective likely contributes to performance gains in experiments.

- **Diverse Attention Patterns**: Different query vectors exhibit unique attention patterns. For instance, some queries focus on distant objects in the foreground, while others emphasize nearby roads or structures. This diversity may further enhance model performance.
- **Consistency Across Viewpoints**: Each pair of attention maps for the same location, taken from different viewpoints, highlights similar landmarks, while the overall patterns adjust based on the changes in viewpoint.

VI. Conclusions

This work introduces the Query-based Adaptive Aggregation (QAA) method with the Cross-query Similarity (CS) paradigm for enhancing VPR multi-dataset training performance. Extensive evaluations demonstrate that QAA consistently achieves high performance across diverse evaluation datasets, excelling in: (1) capturing global context for query-level image features and independent reference codebooks, (2) handling scalable queries without increasing output dimensionality, and (3) enabling efficient query training with minimal computational and parameter overhead. Moreover, we demonstrate for the first time that the CS matrix, with better informational capacity than score-based aggregation, can generate robust geographical descriptors for VPR. These findings further highlight the broad potential of QAA and the CS matrix for tasks requiring enhanced information capacity or robust feature representations. Future work will address the challenge of 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] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] ——, "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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[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] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 9, pp. 1546–1562, 2007.

[27] Y. Yu, K. H. R. Chan, C. You, C. Song, and Y. Ma, "Learning diverse and discriminative representations via the principle of maximal coding rate reduction," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 9422–9434.

[28] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R2former: Unified retrieval and reranking transformer for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 19 370–19 380.

[29] R. Wang, Y. Shen, W. Zuo, S. Zhou, and N. Zheng, "Transvpr: Transformer-based place recognition with multi-level attention aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 13 648–13 657.

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[32] R. Sinkhorn and P. Knopp, "Concerning nonnegative matrices and doubly stochastic matrices," *Pacific Journal of Mathematics*, vol. 21, no. 2, pp. 343–348, 1967.

[33] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," *Advances in neural information processing systems*, vol. 26, 2013.

[34] 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.

[35] N. Keetha, A. Mishra, J. Karhade, K. M. Jatavallabhula, S. Scherer, M. Krishna, and S. Garg, "Anyloc: Towards universal visual place recognition," *IEEE Robotics and Automation Letters*, 2023.

[36] 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.

[37] 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.

[38] 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.

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

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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