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

TinyVPR: Distilling Correct and Confusing Knowledge for Lightweight Visual Place Recognition

Zhuochen Yang[1,2], Runheng Zuo[3], Xu Yang[2], Runjiang Dou[2], Zhe Wang[2], Liyuan Liu[1,2], Shuangming Yu[2*]

***Abstract*— Visual Place Recognition (VPR) is a key technology in autonomous driving, robotics, and augmented reality, requiring efficient and robust localization in large-scale environments. However, most existing methods rely on heavy deep models that are computationally expensive and difficult to deploy on edge devices, limiting their practical use. While model compression techniques such as compact model fine-tuning and traditional knowledge distillation have shown some promise, they often fall short in visual retrieval tasks. Inspired by the teaching principle that emphasizes both reinforcing correct knowledge and correcting errors, we propose an online positive-negative sample contrastive distillation framework. This approach enables the student model to learn more discriminative features by simultaneously distilling the relationships among positive and negative samples. We also design a cross-attention based feature alignment operator to better align intermediate feature representations between teacher and student models after feature extraction, improving feature consistency and distillation efficiency. Experimental results demonstrate that our method achieves a favorable trade-off between accuracy and efficiency on multiple visual localization benchmarks, significantly outperforming existing lightweight approaches in several scenarios. These advantages make it well-suited for deployment on resource-constrained edge devices.**

I. INTRODUCTION

Visual Place Recognition (VPR) is essential for autonomous robot localization, augmented reality (AR), and autonomous driving [1]–[4]. Most existing methods [5]–[7] adopt a two-stage paradigm: extracting visual features and aggregating them into global descriptors, treating VPR as an image retrieval task. Recent approaches commonly employ Vision Transformer (ViT) [8] and its variants [9] as backbones, coupled with structured aggregation operators [3, 10]–[12] to enhance recognition accuracy.

However, these one or two stage methods [13]–[19] typically involve large models and high computation costs, limiting deployment on platforms with constrained resources (e.g., mobile devices, autonomous vehicles, AR headsets). To tackle this, lightweight architectures like Tiny-ViT [20] have emerged. While effective on classification tasks such as ImageNet, these models often suffer considerable recall performance drops when applied to VPR, even with fine-tuning.

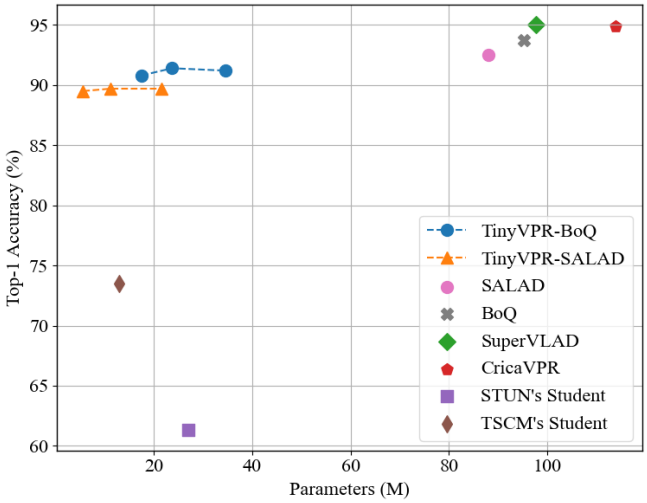

Fig. 1. Accuracy and model size comparisons between our method and other compact visual place recognition models, evaluated on the Pitts30k-test dataset using Recall@1.

Current distillation frameworks [21, 22] for VPR mainly rely on triplet loss and are trained on datasets Pittsburgh30k [23], which poorly reflect the complexity of modern datasets such as GSV [24]. Moreover, they are not aligned with contrastive learning practices and perform inadequately in challenging urban scenes with frequent distractors.

To explore a more scalable solution, we applied basic distillation strategies (e.g., MSE, KL divergence) in a mini-batch contrastive learning setting on the GSV-Cities dataset. However, these methods yielded limited improvement, likely because they fail to transfer the teacher's ability to distinguish between positive and negative samples—an essential factor in retrieval-based VPR. As a result, the student model lacked robustness, especially when faced with hard negative distractors.

To address this limitation, we propose a knowledge distillation framework that retains the overall architecture of the teacher model while replacing its high-capacity ViT feature extractor with a lightweight Tiny-ViT. Drawing inspiration from real-world teaching, where effective instruction involves not only providing correct answers but also addressing common misconceptions, we introduce a contrastive distillation strategy based on online mining of positive and negative samples. Built upon the Multi-Similarity (MS) loss [25], our

*Corresponding Author: Shuangming Yu. (yushuangming@semi.ac.cn)
[1]Zhuochen Yang and Liyuan Liu are with the School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing 101408, China.
[2]Zhuochen Yang, Xu Yang, Runjiang Dou, Zhe Wang, Liyuan Liu and Shuangming Yu are with the State Key Laboratory of Semiconductor Physics and Chip Technologies, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China.
[3]Runheng Zuo is with the School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China.
This work was supported by National Key Research and Development Program of China (2024YFE0201500) and National Natural Science Foundation of China under Grant 62274154.

method simultaneously distills "correct knowledge" (alignment with positives) and "confusing knowledge" (separation from hard negatives) into the student, thereby significantly enhancing its robustness in complex visual scenarios.

To further alleviate the representation gap between teacher and student, we design a Cross-Feature Attention alignment module that employs spatial-channel attention to align intermediate features, enhancing semantic and structural consistency without introducing significant parameters.

We apply our framework to BoQ [16] and SALAD [15], which utilize distinct feature aggregation strategies, and adapt our student model accordingly. Experimental results verify the effectiveness of our approach and indicate its potential generalizability to various VPR architectures.

Extensive experiments on standard benchmarks indicate that our method achieves a well-balanced trade-off between model efficiency and recognition accuracy. As illustrated in Figure 1, the distilled Tiny-ViT retains competitive R@1 performance on the Pitts30k dataset, while reducing parameter count by more than 5× compared to previous student models such as STUN and TSCM. These results suggest that our approach holds practical potential for deployment in resource-constrained VPR applications.

In summary, this work addresses two key challenges in VPR distillation—limited robustness to hard negatives and insufficient teacher-student feature alignment by introducing the following contributions:

1) A contrastive distillation strategy that leverages online mining of positives and hard negatives, integrating Multi-Similarity loss to transfer both "correct" and "confusing" knowledge from the teacher model, thereby enhancing the student's discriminative capacity;

2) A cross-attention feature alignment module that explicitly bridges the representational gap between teacher and student via spatial-channel mappings;

3) A unified and adaptable distillation framework validated on two representative teacher architectures, BoQ and SALAD, enabling effective compression to lightweight Tiny-ViT students without significant loss of retrieval performance.

II. Related Works

Visual Place Recognition: VPR has evolved from methods relying on global descriptors [26], which struggle with appearance variations, to deep learning-based approaches including CNNs [11], MLPs [5], and Vision Transformers (ViTs) [27], offering better performance through learned feature aggregation and global context modeling. Recent VPR approaches primarily focus on compressing high-dimensional features into compact descriptors while preserving essential geometric information. SALAD [15] reformulates NetVLAD [11]'s soft assignment as an optimal transport problem for aggregating DINOv2 [28] features, while SuperVlad [17] simplifies VLAD by removing cluster centers and using fewer clusters, improving domain generalizability and enabling highly compact, high-performing descriptors. CricaVPR [14] encodes cross-image relationships through a multi-scale pyramid correlation structure. BoQ [16] introduces learnable global queries that leverage cross-attention to aggregate input features, yielding consistent and interpretable representations for place recognition.

Despite the performance gains of existing models, their large parameter sizes typically limit deployment on storage-constrained edge devices. To further optimize the efficiency-accuracy trade-off, existing works have explored lightweight feature aggregation techniques such as pruning and quantization [5, 29, 30]. Compared to these coarse-grained compression methods at the structural or numerical level, our proposed method addresses this challenge by leveraging knowledge distillation to transfer knowledge explicitly and achieve competitive accuracy with significantly reduced model size, making it particularly suitable for real-world applications in resource-limited environments.

Knowledge Distillation: KD [31], originally introduced for image classification, reduces computational cost while preserving performance through a teacher-student learning framework and has been extensively studied in CNNs [32] and ViTs [20, 33, 34]. In the specific domain of Visual Place Recognition, StructVPR [35] enhances the robustness of RGB-based representations by incorporating scene geometry features, while LSD-Net [36] uses dual-pathway designs to facilitate effective transfer of teacher feature patterns. STUN [21] equips student models with both location prediction and uncertainty estimation capabilities. In terms of computational efficiency, TSCM [22] introduces a cross-metric distillation mechanism to reduce the teacher-student gap, and TeTRA-VPR [37] applies ternary quantization with attention modules to compress Transformer embeddings. ASHT-KD [38] further explores a multi-teacher collaborative scheme, dynamically integrating geometry and texture knowledge to improve generalization.

However, these methods are often constrained in two key aspects: they are typically not implemented within modern mini-batch contrastive training frameworks, and they rarely focus on explicitly enhancing the student model's ability to distinguish between positive and negative samples. In particular, the discriminative knowledge embedded in hard negatives, a crucial component for retrieval-based VPR, is largely underutilized.

To the best of our knowledge, this work is the first to propose a contrastive distillation framework for VPR that explicitly incorporates hard negative mining in a mini-batch training setting. By transferring both the teacher's knowledge of correct matches and commonly confused distractors, our method significantly enhances the student's discriminative capacity and robustness in complex environments.

III. Methodology

A. Overview

Visual Place Recognition (VPR) typically follows a "feature extraction – feature aggregation – similarity computation" pipeline. Backbones such as DINOv2 (ViT-based) or ResNet-50 extract semantically rich features, which are aggregated into compact global descriptors using modules

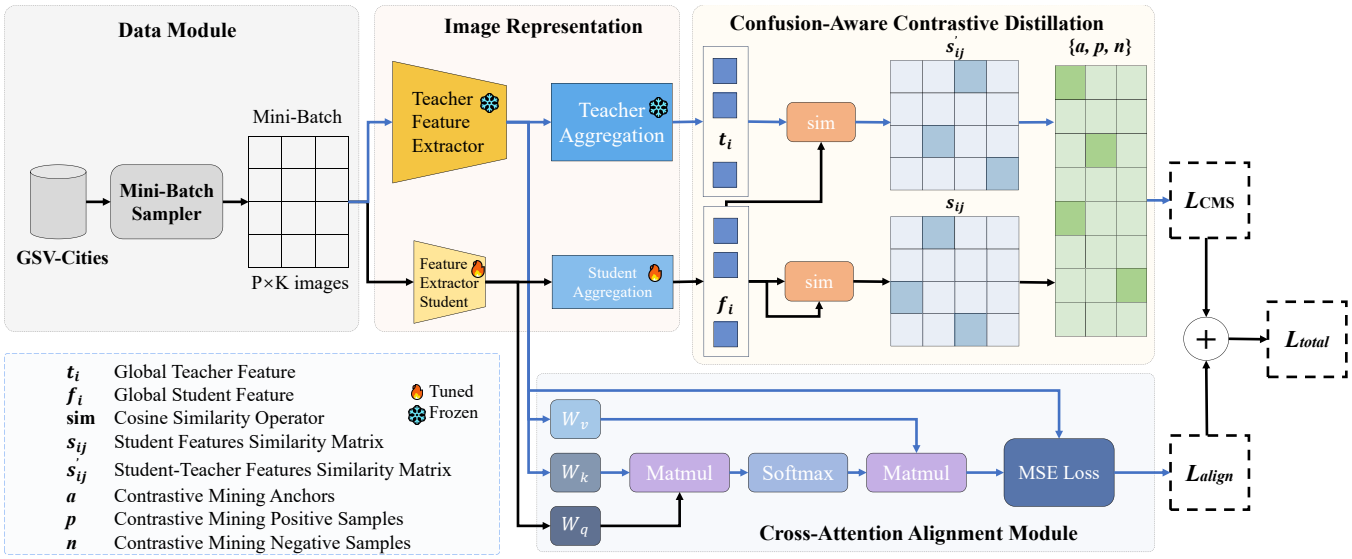

Fig. 2. **Overview of our proposed distillation framework.** A batch of images is sampled from the GSV-Cities dataset and passed through both teacher and student encoders. We apply a cross-attention feature alignment module to align intermediate features, optimized with MSE loss. Global descriptors are then aggregated, and two similarity matrices are constructed: one within the student (S_{ij}), and one between teacher and student (S'_{ij}). Multi-Similarity loss is applied over mined hard triplets to transfer both correct and confusing knowledge. The final loss jointly supervises structural and semantic alignment, enhancing the student model's robustness and discriminative ability under resource constraints.

like NetVLAD or transformer-based aggregators. Models are trained on datasets like Google Street View (GSV) with contrastive losses (e.g., Multi-Similarity Loss), aiming to increase similarity among positive pairs while reducing similarity among negatives. Online hard negative mining is further used to enhance discrimination by selecting feature-space-close negatives.

To achieve lightweight yet discriminative VPR, we propose a distillation-based Tiny-ViT framework, as illustrated in Figure 2. A Cross-Attention Alignment Module aligns intermediate features of the student and teacher, guiding the student to inherit the teacher's spatial-semantic structure. We also adapt feature aggregation (e.g., SALAD, BoQ) to Tiny-ViT's reduced channel depth and patch resolution for better structural compatibility. Finally, a multi-path contrastive distillation strategy is introduced:

1) Aligning student-student and student-teacher similarities on positive pairs.

2) Minimizing student-teacher similarities on negatives.

3) Integrating online positive-negative mining to capture confusing negatives identified by the teacher, enabling the student to learn both correct associations and subtle misclassification boundaries.

B. Cross-Attention Alignment Module

To effectively transfer representational knowledge from the teacher's feature extractor, we introduce a *Cross-branch Feature Alignment* module (CrossAlignAggregator) that bridges the heterogeneous feature spaces between teacher and student models. This module reconstructs the teacher's intermediate features via an attention mechanism, enforcing structural consistency in the student's feature space and enhancing the student's representational capacity.

Specifically, let the intermediate features extracted by the teacher model be denoted as $T \in \mathbb{R}^{B\times N_1\times C_1}$ and those from the student model as $S \in \mathbb{R}^{B\times N_2\times C_2}$, where B is the batch size, N is the number of tokens, and C is the channel dimension. The module first projects the teacher features into keys (K) and values (V) through linear transformations, while the student features serve as queries (Q). Cross-attention scores are then computed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the scaling factor. The output $\hat{T}$ represents the aligned teacher features, which are then compared with the student features S using an L_2 normalized mean squared error:

$$\mathcal{L}_{\text{align}} = \left\|\text{normalize}(\hat{T}) - \text{normalize}(S)\right\|_2^2 \tag{2}$$

This alignment strategy effectively mitigates the mismatch in token and channel dimensions between teacher and student, and enables intermediate-level distillation to improve the generalization of the student network.

C. Feature Aggregation Adaptation

We adopt two strong VPR models, SALAD and BoQ as teacher models due to their proven effectiveness in large-scale place recognition. However, due to architectural differences, direct application of their aggregation modules to the Tiny-ViT student is infeasible. We thus design specific adaptations:

a) BoQ Integration.: To accommodate the Tiny-ViT backbone within the BoQ aggregation framework, we adjust the input interface of BoQ to align with the output feature dimensions of Tiny-ViT. BoQ expects inputs of size

16×16, while Tiny-ViT outputs feature maps of 7×7 with channel dimensions varying across model sizes. We use bilinear interpolation to upscale the spatial dimensions of Tiny-ViT outputs to 16×16, ensuring compatibility without modifying channel depth. We also modify BoQ's input and projection channels to match the output channel size of Tiny-ViT, ensuring seamless compatibility without altering its aggregation structure. This adaptation enables the BoQ module to effectively process high-level features extracted by the lightweight backbone while maintaining its query-based aggregation mechanism.

b) SALAD Integration.: To adapt the SALAD aggregation module to our lightweight Tiny-ViT architecture, we make structural adjustments to both the token representations and the aggregation input format. Specifically, patch tokens from Tiny-ViT are first passed through a linear projection layer to increase their channel dimension, followed by LayerNorm normalization. These tokens are then reshaped into 2D spatial feature maps to match the expected input format of the convolutional aggregation in SALAD. Meanwhile, the class token is separately extracted and used as a global semantic reference during aggregation.

These adaptations ensure that aggregation modules remain structurally compatible and semantically meaningful within the student framework, laying the foundation for effective feature alignment and knowledge transfer.

D. Confusion-aware Contrastive Distillation

1) Multi-Similarity Loss: Given a batch of normalized feature vectors $\{f_i\}_{i=1}^N$ with corresponding label set $\{y_i\}_{i=1}^N$, we define the positive and negative sets for each anchor feature f_i as follows:

$$\mathcal{P}_i = \{j \mid y_j = y_i,\ j \neq i\}, \quad \mathcal{N}_i = \{k \mid y_k \neq y_i\} \tag{3}$$

Using cosine similarity or dot product as the similarity metric, we denote:

$$s_{ij} = \mathrm{sim}(f_i, f_j) \tag{4}$$

The Multi-Similarity (MS) loss for anchor f_i is computed as:

$$\mathcal{L}_i = \frac{1}{\alpha} \log \left(1 + \sum_{j \in \mathcal{P}_i} e^{-\alpha(s_{ij} - \lambda)} \right) + \frac{1}{\beta} \log \left(1 + \sum_{k \in \mathcal{N}_i} e^{\beta(s_{ik} - \lambda)} \right) \tag{5}$$

The final loss is the average over all samples:

$$\mathcal{L}_{\mathrm{MS}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_i \tag{6}$$

2) Confusion-aware Distillation Loss: While MSLoss effectively pulls positive samples closer and pushes negatives apart, it may be insufficient for lightweight models to capture fine-grained semantics. We propose a **Confusion-aware Distillation** strategy that transfers the teacher's knowledge about "correct" and "confusable" cases to enhance student discrimination.

a) Motivation.: Inspired by real-world pedagogy where instructors not only provide correct answers but also highlight confusing wrong choices, we encourage the student model to distinguish not only the matching target but also hard negatives. This is achieved by leveraging both student-teacher similarity and **online hard sample mining** guided by the teacher.

Let $\{f_i\}_{i=1}^N$ be the student feature vectors, $\{t_i\}_{i=1}^N$ the teacher outputs, and $\{y_i\}_{i=1}^N$ their labels. We define similarities:

$$s_{ij} = \mathrm{sim}(f_i, f_j), \quad s'_{ij} = \mathrm{sim}(f_i, t_j) \tag{7}$$

b) Online Positive-Negative Mining.: We define four sets for each anchor f_i:

$$\mathcal{P}_i = \{j \mid y_j = y_i,\ s_{ij} < \max(s_{ik} \mid y_k \neq y_i) + \epsilon\}, \tag{8}$$

$$\mathcal{N}_i = \{j \mid y_j \neq y_i,\ s_{ij} > \min(s_{ik} \mid y_k = y_i) - \epsilon\}, \tag{9}$$

$$\mathcal{P}\infty_i = \left\{j \mid y_j = y_i,\ s'_{ij} < \max\left(s'_{ik} \mid y_k \neq y_i\right) + \epsilon\right\}, \tag{10}$$

$$\mathcal{N}\infty_i = \left\{j \mid y_j \neq y_i,\ s'_{ij} > \min\left(s'_{ik} \mid y_k = y_i\right) - \epsilon\right\}. \tag{11}$$

where hard pairs are selected *relative to the hardest positive/negative in the batch* with margin $\epsilon = 0.1$.

c) Final Loss.: The combined loss for f_i is:

$$\mathcal{L}_{i'} = \frac{1}{\alpha} \log \left(1 + \sum_{j \in \mathcal{P}_i} e^{-\alpha(s_{ij} - \lambda)} + \sum_{j \in \mathcal{P}\infty_i} e^{-\alpha(s'_{ij} - \lambda)} \right) + \frac{1}{\beta} \log \left(1 + \sum_{k \in \mathcal{N}_i} e^{\beta(s_{ik} - \lambda)} + \sum_{k \in \mathcal{N}\infty_i} e^{\beta(s'_{ik} - \lambda)} \right) \tag{12}$$

The overall Confusion-aware MS loss (CMS loss) is:

$$\mathcal{L}_{\mathrm{CMS}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{i'} \tag{13}$$

The final training objective combines the CMS loss and the feature alignment loss:

$$\mathcal{L}_{\mathrm{total}} = \eta \cdot \mathcal{L}_{\mathrm{CMS}} + (1 - \eta) \cdot \mathcal{L}_{\mathrm{align}} \tag{14}$$

where $\eta \in [0, 1]$ is a weighting factor that balances structural alignment and similarity supervision.

This formulation combines student internal contrast and teacher-guided supervision, improving the student's ability to distinguish true positives from confusing negatives. Empirically, it enhances robustness in large-scale and ambiguous urban scenes.

IV. Experiments

A. Implementation Details

a) Training Setup.: We conduct all training and distillation experiments on 3 NVIDIA RTX 4090 GPUs with a total batch size of 60. The student model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-4}. We adopt the Multi-Similarity Loss (MS Loss) as the primary supervision objective, with hyperparameters set to $\alpha = 1.0$, $\beta = 50$, and $\lambda = 0.0$. Our proposed contrastive distillation loss (CMS Loss) follows a similar formulation and shares the same hyperparameter configuration.

To mine informative pairs within each mini-batch, we employ a Multi-Similarity Miner for online hard sample minin. During training, we linearly combine the CMS Loss and the feature alignment loss from the CrossAlign module with a weight ratio of $0.9 + 0.1$, emphasizing contrastive supervision while preserving structural consistency between teacher and student representations.

B. Benchmarks and Performance Evaluation

We evaluate our models on several standard VPR benchmark datasets, primarily Tokyo24/7, Pitts250k, and MSLS, with additional testing on SVOX. Tokyo24/7 [39] contains around 76,000 database images and 315 queries captured in urban settings, with each scene covering three viewpoints and three time-of-day variations, resulting in substantial illumination changes. Pitts250k [23] is derived from Google Street View panoramas and features notable viewpoint shifts, moderate appearance changes, and limited dynamic content. MSLS (Mapillary Street-Level Sequences) [40] is a large-scale dataset comprising over 1.6 million images from diverse environments—including urban, suburban, and natural scenes—designed to assess model robustness under varying lighting, weather, and dynamic object conditions. SVOX [41] focuses on cross-weather and cross-illumination challenges, with queries sourced from the Oxford RobotCar dataset. Following established practice [16], we use Recall@N as the primary metric, applying a 25-meter threshold for positive matches on Tokyo24/7, Pitts30k, and MSLS to evaluate retrieval performance across diverse scenarios.

C. Comparisons with state-of-the-art Methods

We comprehensively evaluate the proposed TinyVPR models against state-of-the-art (SOTA) visual place recognition methods across multiple benchmark datasets. As shown in Table I, despite reducing parameter counts by a factor of 4× to 20× compared to full-sized teacher models, TinyVPR variants still maintain competitive retrieval performance.

To better assess the trade-off between performance and model size, we introduce the **Avg/M** metric, which normalizes the average retrieval accuracy (R@1 and R@5) by the model's parameter count (in millions). As shown in Table I, our TinyVPR models exhibit superior efficiency compared to all baselines. Notably, TinyVPR-5M-SALAD achieves the highest R@1/M and R@5/M values of **14.571** and **16.154**, respectively, significantly outperforming the second-best MixVPR (7.295 and 7.778). This highlights the outstanding retrieval accuracy per parameter that TinyVPR provides.

In summary, benefiting from a lightweight architecture and effective knowledge distillation, TinyVPR offers highly competitive performance with drastically fewer parameters. This makes it an ideal solution for real-world deployment in resource-constrained environments.

D. Ablation Study on Core Components

E. Ablation Study on Core Components

This ablation study, along with the subsequent module replacement experiments, uses BoQ as the teacher model, and adopts the original feature aggregation module from BoQ for compatibility. Results based on SALAD are provided in the supplementary material, which show similar overall trends to those of BoQ, indicating that the conclusions are generally consistent across architectures.

To comprehensively evaluate the independent effectiveness of each component in our proposed distillation framework, we conduct ablation experiments under three different student backbone capacities (22M / 11M / 5M). We isolate the effects of the feature alignment module (CrossAlign) and the contrastive distillation loss (CMS Loss).

Table II summarizes the ablation results under three student model sizes (22M, 11M, and 5M) across four benchmark datasets. We observe the following key findings:

1) **Baseline shows stable but limited generalization.** Without distillation, the MS Loss-based Baseline performs well on standard datasets like Pitts30k, but drops notably on harder ones such as Tokyo24/7 and SVOX-Night, suggesting insufficient robustness to challenging scenarios.
2) **CrossAlign alone may hinder performance.** Using CrossAlign without contrastive guidance results in inconsistent or degraded performance, especially on SVOX-Night, indicating that alignment without supervision may interfere with feature aggregation.
3) **CMS Loss significantly improves robustness.** The proposed CMS Loss consistently boosts accuracy across all student sizes, especially on open-world datasets like Tokyo24/7 and MSLS, by transferring fine-grained relational knowledge from the teacher.
4) **Combining CrossAlign and CMS yields the best results.** The full configuration outperforms all others, demonstrating that structural alignment and contrastive distillation are complementary and mutually reinforcing.
5) **Smaller models benefit more from distillation.** In some cases, the 11M model surpasses the 22M variant, likely due to better generalization from structural regularization under distillation.

In summary, CrossAlign enables structural knowledge transfer, but its full benefit emerges when coupled with CMS Loss. The synergy between alignment and contrastive supervision is key to the success of our framework.

F. Module Comparison and Substitution

To further validate the effectiveness of our proposed feature alignment strategy and distillation loss, we conduct two sets of replacement experiments as detailed below.

(1) Feature Alignment Replacement. We replace only the CrossAlign module while keeping the CMS Loss unchanged, comparing it against a conventional multi-layer perceptron (MLP) projection head. As shown in Table III, our CrossAlign consistently outperforms the MLP counterpart across all model sizes and datasets. Notably, the performance gains are more significant on challenging datasets such as SVOX-Night and MSLS, where the models must handle nighttime conditions and large-scale urban variation. This

TABLE I

COMPARISON TO STATE-OF-THE-ART METHODS ON FOUR BENCHMARKS (R@1 AND R@5 REPORTED). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**. **AVG/M** DENOTES THE AVERAGE RETRIEVAL ACCURACY (R@1 OR R@5) DIVIDED BY THE NUMBER OF MODEL PARAMETERS (IN MILLIONS, M), REFLECTING THE ACCURACY EFFICIENCY PER UNIT OF MODEL CAPACITY. NOTE THAT *Rain*, *Snow*, *Sun*, *Night*, AND *Overcast* ARE ALL SUBSETS FROM THE SVOX DATASET, REPRESENTING DIVERSE WEATHER AND ILLUMINATION CONDITIONS.

Method	Pitts30k		Tokyo24/7		Rain		Snow		Sun		Night		Overcast		MSLS		Avg/M	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1/M	R@5/M
SuperVLAD NeurIPS' 2025	**95.0**	**97.4**	95.2	97.8	96.1	98.6	99.2	99.7	96.7	**99.5**	94.9	98.3	98.2	99.2	92.2	96.6	0.983	1.008
SALAD CVPR' 2024	92.5	96.4	94.6	97.5	98.5	99.7	98.9	99.7	97.2	99.4	95.4	99.3	98.3	99.3	92.2	96.4	1.090	1.119
BoQ CVPR' 2024	93.7	97.1	**98.1**	**98.1**	**98.8**	**99.7**	**99.4**	**99.7**	**97.5**	99.4	**97.7**	**99.5**	**98.5**	**99.3**	**93.8**	**96.8**	1.021	1.037
CricaVPR CVPR' 2024	94.9	97.3	93.0	97.5	94.8	98.5	96.0	99.2	93.8	98.1	86.3	95.3	96.7	99.0	90.0	95.4	0.818	0.856
MixVPR WACV' 2023	91.5	95.5	86.7	92.1	91.5	97.2	97.0	98.3	84.8	93.2	64.4	79.2	96.2	98.2	88.2	93.0	**7.295**	**7.778**
Tinyvpr22M_BoQ	91.2	**95.8**	**89.8**	**94.0**	92.8	98.3	94.8	98.6	91.6	97.2	**78.4**	**89.8**	96.1	98.6	**88.1**	**94.1**	2.619	2.777
Tinyvpr11M_BoQ	**91.4**	95.6	84.4	93.0	**94.2**	**98.4**	**96.9**	**98.7**	**91.6**	**97.4**	77.4	89.4	**97.3**	**98.6**	88.1	93.9	3.804	4.035
Tinyvpr5M_BoQ	90.8	95.4	85.4	90.8	93.0	97.7	94.7	98.6	89.7	96.4	71.0	84.3	96.0	98.4	86.6	93.5	5.051	5.394
Tinyvpr22M_SALAD	89.7	95.2	87.6	94.0	91.9	98.2	95.4	98.6	90.0	96.3	71.2	86.4	96.2	98.1	84.5	93.5	4.108	4.420
Tinyvpr11M_SALAD	89.7	95.1	85.1	93.0	89.5	96.5	94.4	98.6	85.8	94.7	64.3	79.8	95.2	98.1	85.1	92.4	7.691	8.350
Tinyvpr5M_SALAD	89.5	94.6	81.3	91.1	86.9	96.1	93.2	97.5	80.2	90.7	46.5	66.2	94.0	97.4	81.2	90.1	**14.571**	**16.154**

TABLE II

R@1 AND R@5 PERFORMANCE COMPARISON ACROSS DATASETS. WE REPORT RESULTS UNDER THREE STUDENT BACKBONE SIZES (22M / 11M / 5M). THE BASELINE USES ONLY MS LOSS FOR TRAINING; "+ CROSSALIGN" ADDS THE CROSSALIGN MODULE TO THE BASELINE; "+ CMS LOSS" ADDS OUR PROPOSED CONTRASTIVE DISTILLATION LOSS (CMS LOSS) TO THE BASELINE; "OURS" COMBINES BOTH COMPONENTS. ALL RESULTS ARE BASED ON USING BOQ AS THE TEACHER MODEL.

	Method	Pitts30k		Tokyo24/7		SVOX-Night		MSLS	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
22M	Baseline	89.6	95.0	76.8	86.3	54.9	75.3	73.4	84.1
	CrossAlign	88.3	94.4	71.1	83.8	41.6	62.8	69.7	82.0
	CMS loss	90.9	95.4	84.4	92.7	77.4	89.7	88.1	**94.5**
	Ours	**91.2**	**95.8**	**89.8**	**94.0**	**78.4**	**89.8**	**88.1**	93.9
11M	Baseline	90.0	95.1	77.8	88.4	48.8	67.2	70.4	83.4
	CrossAlign	88.8	94.3	70.8	80.6	36.3	57.6	69.2	82.6
	CMS loss	90.6	95.5	83.8	91.1	67.4	81.7	87.1	93.4
	Ours	**91.4**	**95.6**	**84.4**	**93.0**	**77.4**	**89.4**	**88.1**	**93.9**
5M	Baseline	89.1	94.5	73.7	87.0	38.5	58.0	67.6	78.5
	CrossAlign	88.3	94.0	69.8	79.7	30.1	46.8	69.2	80.9
	CMS loss	90.4	95.0	79.7	87.3	57.4	73.9	86.1	93.0
	Ours	**90.8**	**95.4**	**85.4**	**90.8**	**71.0**	**84.3**	**86.6**	**93.5**

TABLE III

COMPARISON BETWEEN OUR CROSSALIGN MODULE AND A STANDARD MLP-BASED ALIGNMENT HEAD ACROSS DIFFERENT STUDENT MODEL SIZES. CMS LOSS IS KEPT FIXED FOR ALL SETTINGS.

	Method	Pitts30k		Tokyo24/7		SVOX-Night		MSLS	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
22M	MLP	91.0	95.6	84.4	91.4	72.9	88.5	81.2	90.4
	Cross	**91.2**	**95.8**	**89.8**	**94.0**	**78.4**	**89.8**	**88.1**	**93.9**
11M	MLP	90.9	95.2	83.8	93.0	65.7	82.9	80.3	89.5
	Cross	**91.4**	**95.6**	**84.4**	**93.0**	**77.4**	**89.4**	**88.1**	**93.9**
5M	MLP	90.8	95.4	80.6	89.2	62.6	80.3	80.7	88.2
	Cross	**90.8**	**95.4**	**85.4**	**90.8**	**71.0**	**84.3**	**86.6**	**93.5**

suggests that our cross-branch attention mechanism is more effective at capturing fine-grained structural correspondences between teacher and student representations, and provides a stronger semantic foundation for robust knowledge transfer under difficult visual scenarios.

(2) Distillation Loss Replacement.

We evaluate several widely-used distillation objectives while keeping CrossAlign fixed: As shown in Table IV, conventional distillation losses (MSE, KL, Reverse-KL) consistently underperform compared to even the finetuned Baseline, especially on challenging datasets like Tokyo247 and SVOX-night. This confirms their limited capacity in transferring discriminative information in retrieval tasks dominated by hard negatives. By contrast, our CMS Loss significantly improves performance across all metrics and model sizes. For instance, on the 22M model, CMS Loss raises the R@1 on Tokyo24/7 from 79.7% (MSE) to 89.8%, and on SVOX-night from 45.1% to 78.4% in Table IV. The gain is particularly notable on benchmarks with high inter-class similarity and real-world ambiguity.

Additionally, disabling online mining leads to noticeable performance drops across all datasets, demonstrating its essential role in exposing hard cases and enabling students to internalize the teacher's discriminative boundaries.

In summary, the CrossAlign module enables effective structural alignment, while the CMS Loss provides relational-level distillation supervision. Combined with dynamic online mining, these components jointly form a robust knowledge transfer framework that enables lightweight student models to generalize well in complex retrieval settings.

G. Hard Negative Similarity Distribution Analysis

To assess whether the student model truly learns the teacher's discriminative ability on confusing samples, we analyze the cosine similarity between queries and their hard negative retrievals. The experiment is conducted on the Tokyo247 dataset, which contains a small number of queries and over 70,000 distractor images, making it a challenging benchmark for evaluating retrieval robustness.

TABLE IV

COMPARISON OF DIFFERENT DISTILLATION LOSS FUNCTIONS WITH CROSSALIGN FIXED. WE EVALUATE STANDARD MSE, KL DIVERGENCE, REVERSE KL, AND A VERSION WITHOUT ONLINE HARD NEGATIVE MINING, AGAINST OUR PROPOSED CMS LOSS. RESULTS ACROSS ALL STUDENT SIZES SHOW THAT CMS LOSS CONSISTENTLY OUTPERFORMS OTHER OBJECTIVES IN R@1 AND R@5, HIGHLIGHTING ITS EFFECTIVENESS IN TRANSFERRING BOTH POSITIVE AND HARD NEGATIVE KNOWLEDGE.

	Method	Pitts30k		Tokyo24/7		SVOX-Night		MSLS	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
22M	MSE	89.6	95.1	79.7	88.6	45.1	64.9	73.4	84.2
	KL	89.5	95.0	80.3	90.8	44.6	64.3	70.8	83.8
	Reverse KL	89.6	94.8	77.8	87.0	49.2	71.0	72.8	83.5
	No Online	88.0	93.9	75.2	86.3	47.0	65.4	68.4	81.6
	Ours	**91.2**	**95.8**	**89.8**	**94.0**	**78.4**	**89.8**	**88.1**	**93.9**
11M	MSE	89.8	94.8	76.2	88.6	54.8	75.3	72.6	84.2
	KL	89.6	94.7	77.1	88.6	48.8	66.6	70.5	82.7
	Reverse KL	90.1	95.1	78.1	88.9	50.9	67.7	73.9	85.1
	No Online	87.1	93.5	71.4	84.8	39.2	55.0	60.5	75.5
	Ours	**91.4**	**95.6**	**84.4**	**93.0**	**77.4**	**89.4**	**88.1**	**93.9**
5M	MSE	88.5	94.0	70.8	85.1	39.2	60.6	67.7	80.4
	KL	88.9	94.2	73.3	85.7	38.0	60.3	68.6	79.3
	Reverse KL	88.9	94.6	74.3	85.4	36.2	55.9	66.3	78.6
	No Online	86.8	93.7	67.6	80.0	34.8	53.0	60.9	74.1
	Ours	**90.8**	**95.4**	**85.4**	**90.8**	**71.0**	**84.3**	**86.6**	**93.5**

We compare three models: 1) a ViT-based teacher with BoQ aggregation, 2) a TinyViT-22M student trained with MS loss, and 3) our TinyViT-22M student trained with the proposed Confusion-aware MS (CMS) distillation. For each model, we extract global descriptors, compute cosine similarities, and collect the top-10 hard negatives—retrievals that are not in the ground-truth set but have high similarity scores.

As shown in Fig. 3, the teacher model yields the lowest mean similarity (0.13), demonstrating clear separation of hard negatives. The baseline model, by contrast, has a higher mean (0.35) and a sharp peak around 0.36, indicating that many false positives are highly similar to the queries, revealing poor discriminative ability. Our student model achieves a lower mean (0.30) and eliminates the high-similarity peak, suggesting improved robustness and effective suppression of confusing negatives. This confirms that our distillation strategy successfully transfers the teacher's ability to separate hard negatives, even under limited model capacity.

V. CONCLUSIONS

We present a lightweight distillation framework for visual place recognition, incorporating a confusion-aware Multi-Similarity loss and a cross-feature attention alignment module. Experiments across multiple benchmarks confirm the effectiveness and synergy of each component. Our method achieves comparable accuracy to the teacher model while significantly reducing parameter count, and adapts well to diverse VPR architectures such as BoQ and SALAD. The proposed CMS loss notably enhances the student's ability to distinguish between positive and hard negative samples, which is essential for retrieval tasks. Overall, our approach offers a practical and generalizable solution for compact VPR deployment and has broader implications for feature-based retrieval systems.

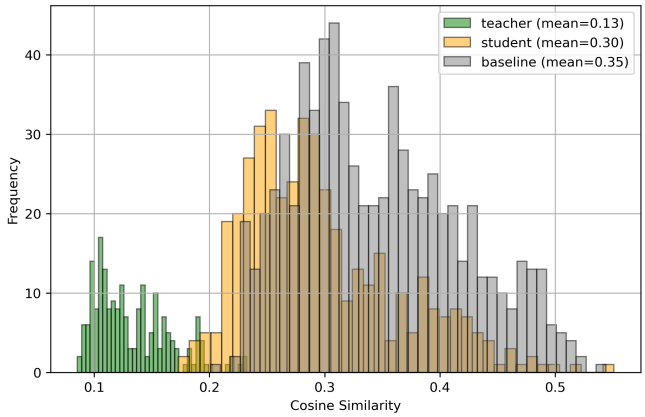

Fig. 3. Cosine similarity distributions between query images and their Top-5 hard negative retrievals on Tokyo24/7.

REFERENCES

[1] S. Middelberg, T. Sattler, O. Untzelmann, and L. Kobbelt, "Scalable 6-dof localization on mobile devices," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*. Springer, 2014, pp. 268–283.

[2] Z. Chen, F. Maffra, I. Sa, and M. Chli, "Only look once, mining distinctive landmarks from convnet for visual place recognition," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 9–16.

[3] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. Mcdonald-Maier, "A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, pp. 561–569, 2020.

[4] S. Hausler, A. Jacobson, and M. Milford, "Multi-process fusion: Visual place recognition using multiple image processing methods," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1924–1931, 2019.

[5] A. Ali-bey, B. Chaib-draa, and P. Giguère, "MixVPR: feature mixing for visual place recognition," in *WACV*, 2023, pp. 2998–3007.

[6] 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*, 2022, pp. 4878–4888.

[7] S. Izquierdo and J. Civera, "Close, but not there: Boosting geographic distance sensitivity in visual place recognition," 2024. [Online]. Available: https://arxiv.org/abs/2407.02422

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2020.

[9] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, and H. Wang, "R^2Former: unified retrieval and reranking transformer for place recognition," in *CVPR*, 2023, pp. 19 370–19 380.

[10] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[11] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.

[12] 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.

[13] 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 *ICLR*, 2024.

[14] F. Lu, X. Lan, L. Zhang, D. Jiang, Y. Wang, and C. Yuan, "Cricavpr: Cross-image correlation-aware representation learning for visual place recognition," in *Accepted to CVPR*, June 2024.

[15] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Accepted to CVPR*, June 2024.

[16] 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*, 2024, pp. 17 794–17 803.

[17] F. Lu, X. Zhang, C. Ye, S. Dong, L. Zhang, X. Lan, and C. Yuan, "Supervlad: Compact and robust image descriptors for visual place recognition," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 5789–5816.

[18] 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," 2025. [Online]. Available: https://arxiv.org/abs/2405.18065

[19] C. Wang, S. Chen, Y. Song, R. Xu, Z. Zhang, J. Zhang, H. Yang, Y. Zhang, K. Fu, S. Du, Z. Xu, L. Gao, L. Guo, and S. Xu, "Focus on local: Finding reliable discriminative regions for visual place recognition," 2025. [Online]. Available: https://arxiv.org/abs/2504.09881

[20] K. Wu, J. Zhang, H. Peng, M. Liu, B. Xiao, J. Fu, and L. Yuan, "Tinyvit: Fast pretraining distillation for small vision transformers," in *European conference on computer vision (ECCV)*, 2022.

[21] K. Cai, C. X. Lu, and X. Huang, "Stun: Self-teaching uncertainty estimation for place recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2022, pp. 6614–6621.

[22] Y. Shen, M. Liu, H. Lu, and X. Chen, "Tscm: A teacher-student model for vision place recognition using cross-metric knowledge distillation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, May 2024, p. 1789–1795.

[23] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 883–890.

[24] A. Ali-bey *et al.*, "Gsv-cities: Toward appropriate supervised visual place recognition," *Neurocomputing*, vol. 513, pp. 194–203, 2022.

[25] 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.

[26] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *Lecture notes in computer science*, vol. 3951, pp. 404–417, 2006.

[27] 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.

[28] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[29] B. Ferrarini, M. J. Milford, K. D. McDonald-Maier, and S. Ehsan, "Binary neural networks for memory-efficient and effective visual place recognition in changing environments," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2617–2631, 2022.

[30] O. Grainge, M. Milford, I. Bodala, S. D. Ramchurn, and S. Ehsan, "Structured pruning for efficient visual place recognition," *IEEE Robotics and Automation Letters*, vol. 10, no. 2, pp. 2024–2031, 2025.

[31] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[32] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *IJCV*, 2021.

[33] S. Yu, T. Chen, J. Shen, H. Yuan, J. Tan, S. Yang, J. Liu, and Z. Wang, "Unified visual transformer compression," in *ICLR*, 2022.

[34] Z. Shen and E. Xing, "A fast knowledge distillation framework for visual recognition," *arXiv*, 2021.

[35] Y. Shen, S. Zhou, J. Fu, R. Wang, S. Chen, and N. Zheng, "Structvpr: Distill structural knowledge with weighting samples for visual place recognition," in *CVPR*, 2023, pp. 11 217–11 226.

[36] G. Peng, Y. Huang, H. Li, Z. Wu, and D. Wang, "Lsdnet: A lightweight self-attentional distillation network for visual place recognition," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 6608–6613.

[37] O. Grainge, M. Milford, I. Bodala, S. D. Ramchurn, and S. Ehsan, "Tetra-vpr: A ternary transformer approach for compact visual place recognition," 2025. [Online]. Available: https://arxiv.org/abs/2503.02511

[38] Z. Li, P. Xu, Z. Dong, R. Zhang, and Z. Deng, "Feature-level knowledge distillation for place recognition based on soft-hard labels teaching paradigm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 2, pp. 2091–2101, 2025.

[39] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," in *CVPR*, 2015, pp. 1808–1817.

[40] 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 *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2626–2635.

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