RelaVPR: Relation-Based Knowledge Distillation for Efficient Visual Place Recognition

Zheyuan Zhang[1,2] **Jiwei Zhang**[1,2] **Rongjing Cai**[1,2]
Nihua Shrestha[1,2] **Zilong Wang**[1,2] **Boyu Zhou**[1,2] **Hong Chen**[1,2*]
[1]School of Computer Science, Beijing University of Posts and Telecommunications
[2]Key Laboratory of Interactive Technology and Experience System,
Ministry of Culture and Tourism, Beijing University of Posts and Telecommunications
zzy1998@bupt.edu.cn, chenhong76@bupt.edu.cn

Abstract

Visual Place Recognition (VPR) aims to identify and retrieve geographic locations from visual observations. With the advent of Visual Foundation Models (VFMs), recent VPR methods have primarily adopted them as backbones to exploit their strong generalization capabilities, effectively mitigating challenges such as perceptual aliasing and long-term appearance variations. However, their large parameter sizes and high inference latency impose substantial hardware demands, significantly hindering practical deployment. To retain the generalization strength of VFMs while enabling efficient VPR, we revisit this problem from the perspective of Knowledge Distillation (KD). Specifically, we conduct a systematic investigation of various KD losses for VPR from both theoretical and empirical standpoints. Our analysis reveals that relation-based KD methods generally achieve superior performance and efficiency, which we attribute to their larger feasible solution space and better alignment with the VPR objective. Building upon this insight, we propose a relation-based KD framework, termed RelaVPR, which is further enhanced along three key dimensions: 1) dynamically filtering noisy knowledge, 2) improving hard-sample learning, and 3) mitigating gradient interference in multi-teacher distillation. Compared with state-of-the-art (SOTA) KD-based VPR methods, RelaVPR requires no additional fine-tuning while delivering higher performance, faster inference, and a more compact model. Moreover, RelaVPR establishes a stronger performance–efficiency trade-off than existing SOTA VPR approaches across multiple popular benchmarks. The code is available at https://github.com/tony19980810/RelaVPR.

1 Introduction

Visual Place Recognition (VPR) aims to determine the coarse geographical location of a query image by retrieving the most visually similar images from a pre-collected, geo-tagged database [4, 17]. It serves as a fundamental component in many computer- and robotic-vision applications, including visual relocalization, SLAM, and augmented reality [21]. Owing to the low cost of image sensors and the wide availability of visual data compared to modalities such as LiDAR, VPR has attracted extensive research interest [13, 41, 14]. Nevertheless, despite significant progress, VPR remains challenged by perceptual aliasing, long-term appearance changes, viewpoint variations, and illumination differences [26].

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The recent rise of Visual Foundation Models (VFMs) has reshaped the landscape of VPR research [9, 23, 16, 17]. Pre-trained on internet-scale data, VFMs exhibit strong generalization capabilities that effectively mitigate the aforementioned challenges, ushering VPR into the VFM era. Yet this progress comes at a cost that is often underappreciated: substantially increased model size and inference latency. Earlier VPR methods [5, 7] typically relied on lightweight Convolutional Neural Network (CNN) backbones such as ResNet-50 [15] (23.5M parameters), whereas widely adopted VFMs like DINOv2-B (base) and DINOv2-L (large) [30] contain roughly 86M and 300M parameters, respectively—a significant and unignorable increase that imposes prohibitive demands on hardware resources and storage, severely limiting real-world deployment. Existing remedies fall into three categories: 1) adopting smaller VFMs (e.g., DINOv2-S (small)) as backbones; 2) applying model compression techniques such as quantization or pruning [13]; and 3) transferring knowledge from large VFM-based teachers to lightweight students via Knowledge Distillation (KD) [47]. Among these, KD stands out for two reasons: it offers a more stable optimization process than training small models from scratch [11]—indeed, DINOv2-S, B, and L are themselves all distilled from DINOv2-G (giant)—and it mitigates the evident representational degradation that often accompanies aggressive quantization or pruning. Therefore, we focus on KD as the promising path toward efficient VPR in the VFM era.

However, a critical question has been overlooked by recent KD methods for VPR: *which type of KD loss should be used, and what is the principled basis for this choice?* Existing efficient VPR approaches [47, 13] predominantly default to Mean Squared Error (MSE) for distillation, without thoroughly justifying this choice or comparing it against alternative formulations. We argue that this oversight is far from innocuous: the choice of KD loss is not merely an implementation detail, but fundamentally shapes what the student can—and cannot—learn during distillation. A poorly chosen loss may over-constrain the student, forcing it to approximate teacher features in ways that are both unnecessary for retrieval and beyond its expressive capacity, thereby limiting the benefit obtainable from subsequent architectural refinements.

To this end, we conduct a systematic investigation of various KD losses for VPR from both theoretical and empirical standpoints. Our analysis reveals that relation-based methods generally achieve superior performance–efficiency trade-offs, which we attribute to their larger feasible solution space and better alignment with the VPR objective. Building upon this insight, we propose RelaVPR, a relation-based KD framework with two core loss-level designs, each addressing a distinct challenge that emerges once the solution space is appropriately relaxed. First, teacher models are imperfect: particularly near decision boundaries, they may produce unreliable or even contradictory supervisory signals, and treating all relational knowledge equally risks amplifying these errors. To this end, Uncertainty-aware Relational Soft Filtering (URSF) estimates the reliability of each relational supervision signal and uses it as a sampling distribution to softly filter noisy teacher guidance, ensuring that reliable knowledge dominates student learning. Second, hard samples, despite carrying rich information, exhibit large representational discrepancies between student and teacher, making them disproportionately difficult to optimize under relational constraints alone. Hardness-aware Anchor Alignment (HAA) addresses this by providing additional alignment on hard anchors as a complementary optimization shortcut, making their distillation learning more tractable. Furthermore, we extend RelaVPR to the multi-teacher setting via a decoupled distillation strategy that assigns each teacher's supervision to a dedicated representational subspace of the student, mitigating gradient conflicts across heterogeneous teachers. Extensive experiments suggest that RelaVPR compares favorably against existing methods in terms of the performance–efficiency trade-off across popular VPR benchmarks (see Fig. 1).

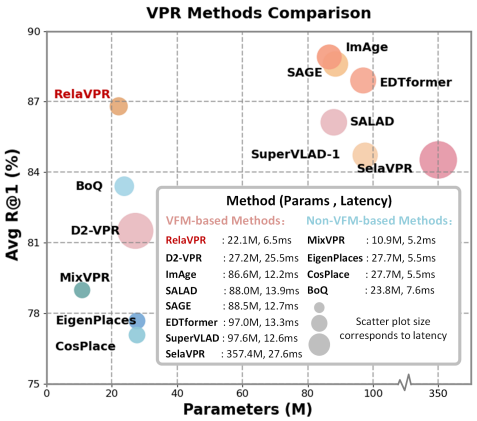
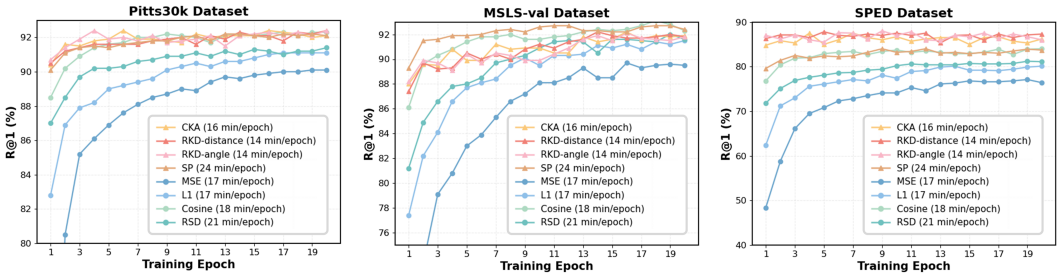


Figure 1: Comparison of average R@1 vs. parameters and latency across Pitts30k, Pitts250k [37], MSLS-val [42], SPED [45], Nordland [35], and AmsterTime [44]. VFM-based methods are shown in warm colors, while non-VFM-based methods are shown in cool colors. Our RelaVPR achieves an excellent performance–efficiency trade-off.

In summary, the main contributions are as follows:



(a) KD results on Pitts30k (b) KD results on MSLS (c) KD results on SPED

Figure 2: Comparison of distillation performance (R@1) across three datasets under identical learning rates and batch sizes for different KD methods, along with training time analysis. Relation-based loss is represented by warm-color triangles, and value-based loss is represented by cold-color dots.

- We analyze different KD losses for VPR from both theoretical and empirical perspectives, showing that relation-based methods outperform others due to their larger solution space and better alignment with VPR objectives, providing practical insights for KD in VPR.
- We propose RelaVPR, a relation-based KD framework that dynamically filters noisy knowledge, improves hard sample optimization, and mitigates gradient interference in multi-teacher KD.
- Extensive experiments show that RelaVPR establishes a competitive performance–efficiency trade-off on popular VPR benchmarks.

2 Related Works

Classical VPR methods primarily relied on handcrafted features (e.g., Bag-of-Words [33]) to construct global descriptors. With the advent of deep learning, CNNs (e.g., NetVLAD [4], MixVPR [2], CosPlace [5], and EigenPlaces [7]) and Transformers (e.g., R2Former [48]) have been introduced to extract local features, which are subsequently aggregated into global representations.

VFM-based VPR. The recent rise of VFMs has redefined VPR through their strong generalization capabilities from large-scale pre-training. Most existing methods adopt DINOv2-B or DINOv2-L [30] as backbones, differing mainly in aggregation design. SALAD [16] fine-tunes DINOv2-B with optimal transport-based aggregation; ImAge [21] replaces explicit aggregators with implicit aggregation tokens injected into DINOv2-B; SelaVPR [23] and FoL [41] adopt DINOv2-L for global retrieval followed by local feature matching to refine the top-K candidates.

Efficient VFM-based VPR. While VFM-based methods deliver strong accuracy, their heavy backbones hinder deployment on resource-constrained devices. TeTRA-VPR [13] improves the efficiency–performance trade-off by quantizing VFMs with a two-stage KD-then-fine-tuning strategy. Most closely related to ours, D^2-VPR [47] distills from DINOv2-B-based teachers into a DINOv2-S student, yet still relies on post-distillation fine-tuning to reach competitive performance.

Although these approaches represent meaningful steps toward efficient VFM-based VPR, two key limitations persist. First, none of them systematically examines which KD loss is suited for VPR or justifies their loss choice on principled grounds. Second, all of them rely on additional fine-tuning after distillation to reach satisfactory performance, which undermines the simplicity and efficiency that motivated KD in the first place. Our work fills both gaps: we provide an in-depth theoretical and empirical analysis of KD loss selection for VPR, and we demonstrate that a principled relation-based framework, without any post-hoc fine-tuning, achieves stronger KD performance than existing efficient VPR methods.

3 Analysis

3.1 Experiment Setting

We revisit KD for VPR by comparing different categories of KD methods on several benchmarks, including MSLS-val [42], SPED [45], and Pitts30k [37]. KD methods are generally categorized into logit-based, value-based (feature-level), and relation-based (sample-wise) approaches [12]. Since VPR is a metric learning task without classification logits, we focus on the latter two. Specifically, we evaluate value-based losses (MSE, L1, cosine similarity and RSD [46]) and relation-based losses

(RKD-distance, RKD-angle [31], SP [39] and CKA [18]). All experiments are conducted under a unified setting using an RTX 4090: the teacher is the SOTA DINOv2-B-based ImAge [21], and the student is a lightweight DINOv2-S-based variant. For value-based KD, feature dimensionality alignment is required; thus, we introduce a linear projection layer to upsample student features to the teacher dimension. This layer is removed during evaluation, and the original student features are used as final descriptors. We avoid projecting teacher features to the student dimension to prevent information loss and degenerate solutions. Results are shown in Fig. 2.

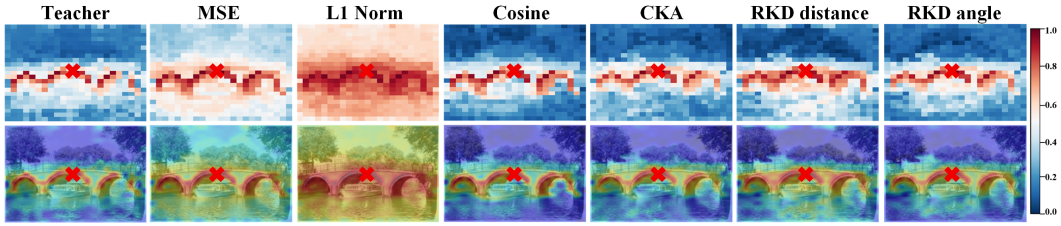


Figure 3: Visualization comparison of cosine similarity between anchor patch tokens at red cross positions and other patch tokens in teacher models and student models distilled via six representative KD schemes. The anchor patches correspond to bridge arches. See more visualizations in Sec. A.5.2.

3.2 Phenomenon Analysis

We observe two main findings from the results: 1) Relation-based distillation generally outperforms value-based methods in both efficiency and final performance. In terms of efficiency, most relation-based methods achieve strong performance from the first epoch and consistently surpass value-based losses across all three datasets. For performance, relation-based methods are generally superior to value-based methods. Cosine loss achieves comparable results on Pitts30k and MSLS-val, while RSD and L1 can approach relation-based methods on MSLS-val but show significant disadvantages on all other datasets. 2) Among value-based methods, cosine loss significantly outperforms others in both efficiency and performance. To further illustrate the two findings, Fig. 3 visualizes local semantic relationships via cosine similarity between anchor patch tokens and all other tokens in teacher and student models, where the student is expected to align with the teacher. It can be observed that relation-based methods and cosine loss produce semantic structures much closer to the teacher than other value-based methods.

We attribute both findings to a unified principle: the feasible solution space of a KD loss and how well that space aligns with the VPR objective. Value-based losses like MSE and L1 enforce strict point-wise alignment between projected student and teacher features, imposing constraints unnecessary for retrieval. Relation-based distillation, by contrast, only requires the student to preserve inter-sample relational structures, relaxing the requirement for direct coordinate alignment and focusing supervision on relational structure. Since VPR retrieval depends mainly on the relative ordering of pairwise similarities rather than absolute feature values, this focus on relational structure may reduce unnecessary constraints and facilitate optimization, providing a possible explanation for Finding 1. Finding 2 can be interpreted from the same perspective: cosine loss, while still value-based, discards the magnitude constraint present in MSE and L1, thereby enlarging the solution space, and directly optimizes directional alignment, which is consistent with the cosine similarity metric used in VPR.

3.3 Theoretical Analysis

We theoretically analyze the constraint structures of value-based and relation-based distillation to help interpret their empirical performance differences (see Sec. A.2.2 for specific instantiations of MSE vs. CKA and RKD).

Problem Definition. VPR aims to learn an embedding function $f(\cdot)$ that maps an input image x to a feature vector $X = f(x)$ such that images depicting the same place have similar representations while images from different places are far apart in the feature space. Given a query image x_q and a database $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$, VPR retrieves the image whose feature embedding is most similar to that of the query:

$$x^* = \arg\max_{x_i \in \mathcal{D}} s(f(x_q), f(x_i)), \tag{1}$$

where $s(\cdot, \cdot)$ denotes a similarity function. We analyze different loss functions from the perspective of the constraints they impose on the student representation. Let $X \in \mathbb{R}^{B\times C}$ and $Y \in \mathbb{R}^{B\times C}$ denote the student and teacher features, respectively, where B is the batch size and C is the feature dimension. For simplicity, we assume the student and teacher share the same feature dimension, which can be achieved in practice via a learnable projection layer. The analysis concerns the representations directly compared by the losses. A learnable projection may absorb transformations of the original student descriptors; thus, the comparison does not characterize the joint solution space of the student and projection parameters. All losses and feasible sets below are understood to be restricted to a common domain on which they are well-defined, and this domain is assumed to be invariant under orthogonal transformations.

Definition of Value-based Distillation. We consider value-based losses that enforce direct numerical alignment between student and teacher features:

$$\mathcal{L}_{\mathrm{val}}(X, Y) = d(X, Y), \tag{2}$$

where $d(\cdot, \cdot)$ is a nonnegative, continuous divergence satisfying $d(A, B) = 0 \Leftrightarrow A = B$. We additionally assume that, for fixed Y, $d(X_n, Y) \to 0$ implies $\|X_n - Y\|_F \to 0$, as holds for MSE and L1 distance. The feasible solution set is:

$$\mathcal{S}_{\mathrm{val}}(\epsilon) = \left\{ X \in \mathbb{R}^{B\times C} \;\middle|\; d(X, Y) < \epsilon \right\}, \tag{3}$$

defining a bounded neighborhood of Y for sufficiently small $\epsilon > 0$ under the above assumptions.

Definition of Relation-based Distillation. $\mathcal{L}_{\mathrm{rel}}$ depends on X and Y only through their pairwise relation matrices $\mathcal{R}(X), \mathcal{R}(Y) \in \mathbb{R}^{B\times B}$:

$$\mathcal{L}_{\mathrm{rel}}(X, Y) = \delta\big(\mathcal{R}(X),\, \mathcal{R}(Y)\big), \tag{4}$$

where $\mathcal{R}(\cdot)$ maps a set of B feature vectors to a $B \times B$ matrix of pairwise relations (e.g., Euclidean distances, cosine similarities, or inner products), and $\delta(\cdot, \cdot)$ is a nonnegative discrepancy satisfying $\delta(A, A) = 0$. The feasible solution set is

$$\mathcal{S}_{\mathrm{rel}}(\epsilon) = \left\{ X \in \mathbb{R}^{B\times C} \;\middle|\; \delta\big(\mathcal{R}(X),\, \mathcal{R}(Y)\big) < \epsilon \right\}. \tag{5}$$

Proposition. *Suppose d satisfies the above assumptions, $Y \neq \mathbf{0}$, and $C \geq 2$. Assume that $\mathcal{R}$ and δ are continuous on their relevant domains and that $\mathcal{R}$ is orthogonally invariant, i.e., $\mathcal{R}(XQ) = \mathcal{R}(X)$ for any orthogonal matrix $Q \in \mathbb{R}^{C\times C}$ with $Q^\top Q = I$; this holds for pairwise Euclidean distance matrices, Gram matrices, and centered Gram matrices used in linear CKA. Then, for any tolerance $\epsilon > 0$, there exists a tolerance $\epsilon' > 0$ such that $\mathcal{S}_{\mathrm{val}}(\epsilon') \subseteq \mathcal{S}_{\mathrm{rel}}(\epsilon)$. Moreover, ϵ' can be chosen such that there exists $X^* \in \mathcal{S}_{\mathrm{rel}}(\epsilon)$ with $X^* \notin \mathcal{S}_{\mathrm{val}}(\epsilon')$, making the inclusion strict.*

Proof sketch. For containment, the assumption on d gives $d(X_n, Y) \to 0 \Rightarrow X_n \to Y$. By the continuity of $\mathcal{R}$ and δ and the identity $\delta(A, A) = 0$, this implies $\delta(\mathcal{R}(X_n), \mathcal{R}(Y)) \to 0$. Hence, for any $\epsilon > 0$, a sufficiently small $\epsilon' > 0$ ensures $\mathcal{S}_{\mathrm{val}}(\epsilon') \subseteq \mathcal{S}_{\mathrm{rel}}(\epsilon)$. For strict inclusion, choose an orthogonal matrix Q such that $YQ \neq Y$, which exists for nonzero Y, and define $X^* = YQ$. By orthogonal invariance, $\mathcal{R}(X^*) = \mathcal{R}(Y)$, so $X^* \in \mathcal{S}_{\mathrm{rel}}(\epsilon)$ for all $\epsilon > 0$. Since $X^* \neq Y$, we have $d(X^*, Y) > 0$. If necessary, further reduce ϵ' so that $0 < \epsilon' < d(X^*, Y)$; containment is preserved, while $X^* \notin \mathcal{S}_{\mathrm{val}}(\epsilon')$. The full proof is in Sec. A.2. □

Comparison and Implications. Compared with $\mathcal{S}_{\mathrm{val}}$, the feasible set $\mathcal{S}_{\mathrm{rel}}$ admits a broader family of solutions under the tolerances specified above. However, a larger solution space is not unconditionally beneficial: if the constraints are too loose, optimization may converge to task-irrelevant solutions. The key insight is that the structure of the solution space should align with the task objective. For VPR, the retrieval result $x^* = \arg\max_{x_i \in \mathcal{D}} s(f(x_q), f(x_i))$ depends on the relative ordering of query–database similarities rather than absolute feature coordinates. Consequently, two embedding functions that preserve this ordering yield the same retrieval results on the considered queries and database. In particular, common orthogonal transformations preserve both cosine similarities and Euclidean distances, while positive global scaling preserves retrieval rankings under either metric. These transformations form a class of retrieval-equivalent solutions and are admitted by relation-based losses with the corresponding invariances, including CKA and normalized RKD-distance. By contrast, under direct numerical alignment, $\mathcal{S}_{\mathrm{val}}$ excludes many such solutions (e.g., rotated or scaled versions of the teacher features that differ from Y) for sufficiently small tolerances by imposing absolute-value constraints. Therefore, these relation-based objectives provide additional flexibility while retaining relational structure relevant to VPR, which may reduce unnecessary constraints and facilitate optimization. This offers a possible explanation for the advantages observed in our empirical analysis.

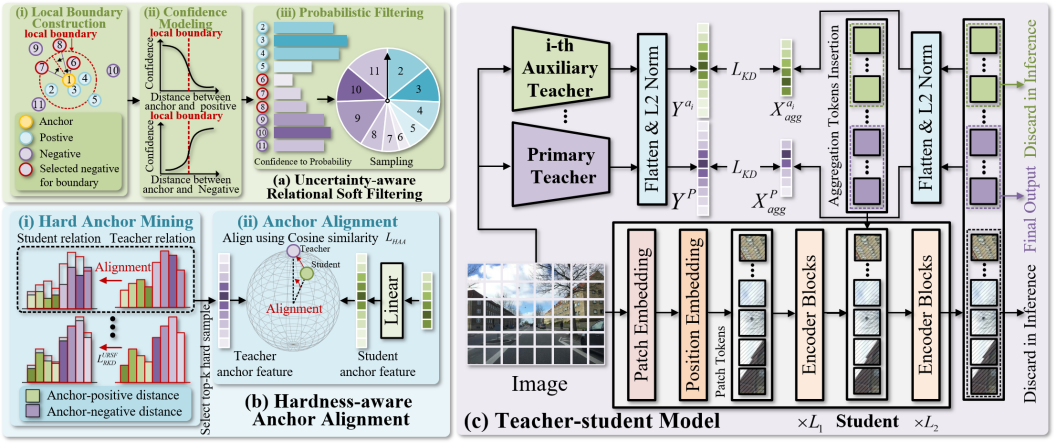


Figure 4: Overview of RelaVPR. At the KD loss level, (a) Uncertainty-aware Relational Soft Filtering: based on the triplet formulation, we estimate the reliability of each pairwise relation and convert the reliability scores into a sampling distribution, enabling soft probabilistic filtering of noisy teacher knowledge. (b) Hardness-aware Anchor Alignment: for hard samples (dashed boxes), we impose additional alignment on anchors as a potential optimization shortcut to enhance learning of challenging cases. At the model level, (c) Teacher–student Model: the student adopts an implicit aggregation paradigm, replacing conventional aggregators with aggregation tokens. For multi-teacher KD, a decoupled distillation strategy assigns separate subspaces to each teacher (Different aggregation tokens are assigned to different teachers, shown in different colors), mitigating multi-source gradient interference.

4 Methods

Based on the previous findings, we propose RelaVPR, a relation-based KD framework (Fig. 4). We adopt RKD-distance as the base KD loss, as it achieves comparable performance to other relation-based methods while requiring the shortest KD time per epoch, as illustrated in Fig. 2. The following section presents the design of RelaVPR from the KD loss and KD model perspectives.

4.1 KD Loss Design

4.1.1 Uncertainty-aware Relational Soft Filtering

In KD, transferred knowledge is not always reliable [8], as teachers may produce ambiguous or incorrect signals near decision boundaries, which can harm student optimization. To address this, we propose Uncertainty-aware Relational Soft Filtering (URSF) (Fig. 4(a)), which models relation reliability and stochastically suppresses noisy supervision via confidence-guided sampling.

Built on relation-based distillation and the triplet formulation in VPR, our method follows the principle that positives should be closer to the anchor than negatives. We estimate pairwise relation reliability by measuring consistency with the local ranking structure: for each anchor, we construct a local boundary and assess relations based on their positions relative to it.

Local Boundary Construction. For each anchor A_i, we define a local boundary r_i as the mean distance to its k nearest negatives:

$$N_i^{(k)} = \mathrm{TopK}_{j \neq i}\left(\tilde{D}_Y(A_i, N_j)\right), \quad r_i = \frac{1}{k} \sum_{N_j \in N_i^{(k)}} \tilde{D}_Y(A_i, N_j), \tag{6}$$

where $\tilde{D}_Y(\cdot,\cdot)$ denotes the normalized teacher distance and r_i separates positive and negative relations.

Confidence Modeling. For each anchor-positive pair (A_i, P_j) and anchor-negative pair (A_i, N_j), we define a confidence score based on its relative position to r_i:

$$w_{ij} = \begin{cases} \sigma\Big((r_i - \tilde{D}_Y(A_i, P_j))/\tau\Big), \\ \sigma\Big((\tilde{D}_Y(A_i, N_j) - r_i)/\tau\Big), \end{cases} \tag{7}$$

where $\sigma(\cdot)$ is the sigmoid and τ controls transition smoothness. Pairs well aligned with the teacher's ranking receive higher weights, while those near or violating the boundary are down-weighted.

Probabilistic Filtering. Rather than using w_{ij} as deterministic weights, we convert them into a per-anchor sampling distribution $p_{ij} = w_{ij} / \sum_{l \neq i} w_{il}$, from which we draw a subset $\mathcal{S}_i$ of $m = \lfloor \rho(B-1) \rfloor$ pairs without replacement (ρ: sampling ratio, B: batch size). The filtered RKD loss is:

$$\mathcal{L}_{\mathrm{RKD}}^{\mathrm{URSF}}(X, Y) = \frac{1}{\sum_i |\mathcal{S}_i|} \sum_i \sum_{j \in \mathcal{S}_i} \ell\big(\tilde{D}_X(A_i, X_j),\ \tilde{D}_Y(A_i, X_j)\big), \tag{8}$$

where $\ell(\cdot, \cdot)$ is the smooth L1 loss and X_j is P_j or N_j depending on pair type.

This formulation performs a soft probabilistic filtering over the relational graph: high-confidence pairs are sampled more frequently, while noisy ones are naturally suppressed. We adopt stochastic sampling rather than hard thresholding because low-confidence pairs are not necessarily erroneous—they may correspond to ambiguous samples near the decision boundary that still carry meaningful structural information. Retaining a diminished but non-zero selection probability preserves their contribution while also introducing beneficial stochasticity that regularizes training.

4.1.2 Hardness-aware Anchor Alignment

Hard samples dominate gradient updates during training and exhibit higher information density, making them the primary targets for effective learning [32]. Existing methods [19, 20] typically emphasize hard samples by assigning them larger weights, thereby forcing the model to focus on these challenging cases. However, such strategies often suffer from optimization difficulties.

In this work, we focus on hard anchor samples with large relational discrepancies between student and teacher models. To better leverage these challenging samples, we introduce Hardness-aware Anchor Alignment (HAA) (Fig. 4(b)) with a new supervision perspective: rather than merely aligning inter-sample relations, we additionally align the anchor features themselves, providing a potential optimization shortcut that makes the originally difficult process more stable and tractable.

Hard Anchor Mining. We define the hardness of an anchor A_i as the average discrepancy between student and teacher relational structures:

$$h_i = \frac{1}{B} \sum_j \left| \tilde{D}_X(A_i, X_j) - \tilde{D}_Y(A_i, X_j) \right|. \tag{9}$$

This quantity measures how well the relational structure centered at node A_i is preserved. Then, we select the k_{hard} hardest anchors $\mathcal{H} = \mathrm{TopK}_i^{k_{\mathrm{hard}}}(h_i)$, corresponding to samples with the largest relational inconsistency.

Anchor Alignment. For the selected anchors X_i, we enforce feature alignment via cosine similarity:

$$\mathcal{L}_{\mathrm{HAA}} = 1 - \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{\langle X_i P,\ Y_i \rangle}{\|X_i P\| \cdot \|Y_i\|}, \tag{10}$$

where P is a learnable projection matrix. The use of the cosine loss is motivated by the analysis in Sec. 3: It imposes fewer constraints than other value-based losses while achieving better results.

Overall KD loss. The final loss combines relational filtering and hardness-aware alignment:

$$\mathcal{L}_{\mathrm{KD}} = \mathcal{L}_{\mathrm{RKD}}^{\mathrm{URSF}} + \lambda \mathcal{L}_{\mathrm{HAA}}. \tag{11}$$

We present the complete KD loss in algorithmic form for clarity in Alg. 1 in the Appendix.

4.2 Teacher-student Model Design

KD Model Design. VPR models generally consist of a backbone and an aggregator. For backbone selection, following [47], we adopt the lightweight VFM DINOv2-S as the student backbone. For aggregation, efficiency is the primary consideration. We adopt the implicit aggregation paradigm [21], which removes explicit post-backbone aggregators and inserts aggregation tokens into intermediate backbone layers. These tokens progressively fuse patch-level features, and only they are retained as

Method	Pitts30k			MSLS-val			SPED			Pitts250k			AmsterTime			Nordland			Avg.
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1
CKA* ICML' 19	92.0	96.3	97.4	91.9	95.9	96.5	87.3	93.1	94.9	94.6	98.4	99.1	50.4	72.5	77.8	85.1	94.0	96.0	83.6
RKD* CVPR' 19	92.1	96.1	97.4	91.8	95.8	96.5	87.5	93.2	94.9	94.7	98.4	99.1	51.7	72.6	79.0	83.1	92.8	95.3	83.5
RSD* ICCV' 25	91.3	95.3	96.5	91.6	95.9	96.5	80.7	89.6	91.3	93.9	97.5	98.4	46.5	66.6	72.1	82.8	91.2	93.6	81.1
D^2-VPR AAAI' 26	91.7	95.8	96.8	90.7	95.4	96.4	86.0	92.9	94.1	94.4	98.3	98.9	49.1	70.7	76.1	77.1	88.6	91.9	81.5
RelaVPR Single-teacher	92.5	**96.6**	**97.7**	**93.0**	96.1	96.6	89.3	93.9	95.1	94.9	98.5	99.2	52.8	73.8	79.4	**89.4**	**96.0**	**97.5**	**85.3**
RelaVPR Multi-teacher	**92.6**	96.4	97.4	92.2	**96.4**	**96.8**	**89.8**	**94.6**	**95.4**	**95.0**	**98.7**	**99.3**	**53.2**	**74.3**	**79.9**	86.1	94.3	96.3	84.8
CosPlace CVPR' 22	90.9	95.7	96.7	87.4	94.1	94.9	72.2	84.8	89.1	92.3	97.4	98.4	47.7	69.8	75.8	71.9	83.8	88.1	77.1
MixVPR WACV' 23	91.5	95.5	96.3	87.2	93.1	94.3	84.7	92.3	94.4	94.1	98.2	98.9	40.2	59.1	64.6	76.2	86.9	90.3	79.0
EigenPlaces ICCV' 23	92.5	96.8	97.6	89.1	93.8	95.0	70.2	83.5	87.5	94.1	97.9	98.7	48.9	69.5	76.0	71.2	83.8	88.1	77.7
LPS-VPR RAL' 2024	91.8	96.0	96.8	89.9	94.2	95.0	84.8	93.9	95.7	94.1	98.0	98.9	-	-	-	-	-	-	-
EPSA-VPR JVCI' 25	-	-	-	89.3	93.8	95.4	84.5	93.7	**96.0**	93.9	97.9	98.7	-	-	-	-	-	-	-
Clus-VPR TAI' 25	90.8	95.2	96.6	82.7	88.5	92.4	-	-	-	92.4	96.9	97.6	-	-	-	-	-	-	-
BoQ CVPR' 24	92.4	96.5	97.4	91.2	95.3	96.1	86.5	93.4	95.7	95.0	98.5	99.1	52.2	72.5	78.4	83.1	91.0	93.5	83.4
RelaVPR Single-teacher	**92.6**	**96.9**	**97.8**	**92.8**	**96.8**	**97.6**	**89.8**	**94.2**	94.6	**95.1**	**98.8**	**99.4**	**56.9**	**77.2**	**82.4**	**93.7**	**97.7**	**98.6**	**86.8**
SelaVPR ICLR' 24	92.8	96.8	97.7	90.8	96.4	97.2	89.0	94.6	96.4	95.7	98.8	99.2	53.6	72.8	78.1	85.2	95.5	98.5	84.5
SALAD CVPR' 24	92.5	96.4	97.5	92.2	96.4	97.0	92.1	**96.2**	96.5	95.1	98.5	99.1	58.5	79.3	84.6	86.3	93.7	95.8	86.1
SuperVLAD† NeurIPS' 24	95.0	97.4	98.2	92.2	96.6	97.4	93.2	97.0	98.0	97.2	99.4	99.7	63.9	83.2	87.2	91.0	96.4	97.7	88.8
SuperVLAD-1 NeurIPS' 24	92.3	96.7	97.8	91.2	95.9	96.9	90.0	95.1	95.9	94.9	98.8	99.4	54.3	76.8	82.3	85.6	94.0	96.3	84.7
CricaVPR† CVPR' 24	94.9	97.3	98.2	90.0	95.4	96.4	91.4	95.6	96.7	97.5	99.4	99.7	64.7	82.8	87.5	90.7	96.3	97.6	88.2
CricaVPR-1 CVPR' 24	91.6	95.7	96.9	88.5	95.1	95.7	87.3	92.9	94.7	94.5	98.3	99.0	49.4	70.3	76.7	79.4	90.1	93.3	81.8
SAGE* ICLR' 26	93.4	97.0	97.9	93.4	**97.3**	97.6	91.8	95.7	96.5	95.7	98.6	99.2	63.1	82.4	86.0	94.1	98.0	98.8	88.6
ImAge NeurIPS' 25	**94.1**	**97.3**	**98.1**	**94.5**	**97.3**	**98.0**	90.9	95.6	96.2	**96.4**	**99.1**	**99.5**	59.5	80.1	84.6	**97.7**	**99.3**	**99.6**	**88.9**
EDTformer TCSVT' 25	93.4	97.0	97.9	92.0	96.6	97.2	**92.4**	95.9	**96.9**	96.0	98.8	99.4	**65.2**	**85.0**	**89.0**	88.3	95.3	97.0	87.9
RelaVPR Single-teacher	92.6	96.9	97.8	92.8	96.8	97.6	89.8	94.2	94.6	95.1	98.8	99.4	56.9	77.2	82.4	93.7	97.7	98.6	86.8

Table 1: Comparison with SOTA methods on six benchmarks. Best and second-best results are marked in **bold** and underlined, respectively. The comparison includes three groups: 1) KD-based VPR methods; 2) similar-scale methods; 3) larger-scale methods including the teacher model ImAge. † denotes methods excluded from the comparison due to cross-image interaction (see Sec. A.4.1 for reasons); for fairness, SuperVLAD and CricaVPR are reported with single-query inference results from EDTformer, and SAGE is reproduced using its official weights without such interaction. * indicates methods reproduced with official weights and codes. "–" indicates unreported results. Group 1 methods are uniformly evaluated at 224×224 for a fair comparison. For Groups 2 and 3, except for SAGE (which is reproduced), all results are taken from the original or published papers; therefore, our RelaVPR is evaluated at 322×322 to match the most common setting in this configuration.

final representations. Compared with conventional methods with explicit aggregators, this approach introduces only negligible additional parameters (0.006M, see Tab. 5)

Decoupled Distillation for Multi-Teacher KD. For a single teacher, the student follows conventional KD. To extend RelaVPR to the multi-teacher setting where heterogeneous teachers may produce conflicting supervisory signals, we introduce a decoupled distillation strategy. For multi-teacher KD, to reduce gradient interference from non-collinear teacher gradients, teachers are categorized into a primary teacher (focusing on generality and robustness) and auxiliary teachers (emphasizing specific attributes or local cues). In the student, eight aggregation tokens are allocated per teacher, and each teacher's outputs are aligned with its corresponding tokens via distillation loss. This decoupling assigns separate representational subspaces to each teacher, increasing usable representation space, mitigating gradient conflicts, and improving training stability.

During inference, no gating or explicit fusion of aggregation tokens is performed. Only the aggregation token aligned with the primary teacher is retained as the final 3072-dimensional representation, consistent with the single-teacher KD output. This design is motivated by following factors: 1) auxiliary knowledge is implicitly absorbed into the shared backbone during training; 2) the primary teacher emphasizes global consistency and robustness, yielding stable features for inference. See Sec. A.4.4 for ablation study.

The overall decoupled distillation loss $\mathcal{L}_{\text{decoupled}}$ is defined as:

$$\mathcal{L}_{\text{decoupled}} = \mathcal{L}_{\text{KD}}(X^{p}_{\text{agg}}, Y^{p}) + \sum_{i=1}^{N_a} \mathcal{L}_{\text{KD}}\left(X^{a_i}_{\text{agg}}, Y^{a_i}\right), \tag{12}$$

where X^{p}_{agg} and Y^{p} are the aggregation tokens of the student and the primary teacher's output features, respectively, and $X^{a_i}_{\text{agg}}$ and Y^{a_i} are those aligned with the i-th auxiliary teacher; N_a is the number of auxiliary teachers.

5 Experiments

5.1 Benchmarks and Implementation Details

Benchmarks. We evaluate our method on popular VPR benchmarks covering key challenges such as viewpoint variations, environmental changes, and perceptual aliasing. Specifically, Pitts30k and Pitts250k [37] focus on large viewpoint changes; MSLS [42] includes urban, suburban, and natural scenes collected over multiple years, introducing significant appearance variations; and SPED [45] consists of imagery from fixed surveillance cameras. We further consider more challenging datasets: Nordland [35], emphasizing seasonal changes, AmsterTime [44], targeting long-term appearance variations and SVOX [27], which tests robustness to diverse weather and lighting conditions. See Sec. A.4.1 for evaluation protocols and A.6 for dataset details.

Implementation Details. For single-teacher KD, we use 8 aggregation tokens aligned with the teacher ImAge [21]; for multi-teacher KD (here we implement it with two teachers), a total of 16 tokens are employed to align with the primary teacher ImAge [21] and the auxiliary teacher EDTformer [17], respectively. The student adopts DINOv2-S as the backbone and outputs 3072-D descriptors. Training is performed on GSV-Cities [1] using Adam with a learning rate of 2.5×10^{-5}, batch size 24, where each place contains 4 images of resolution 224×224. The total training time for single-teacher KD is less than 7 hours using an RTX 4090. See Sec. A.4.7 for hyperparameter details.

5.2 Comparisons with State-of-the-art Methods

This section compares our method with SOTA approaches. Tab. 1 reports performance results, and Tab. 2 lists method configurations. The comparison is divided into three parts, separated by lines in Tab. 1: 1) Comparison with the SOTA KD-based VPR method (D^2-VPR [47]) and other KD method (CKA [18],RKD [31],RSD [46]). 2) Comparison with other VPR methods of similar scale: LPS-VPR [28], Clus-VPR [43], EPSA-VPR [29], BoQ [3], CosPlace [5], MixVPR [2], and EigenPlaces [7]. 3) Comparison with larger-scale VFM-based VPR methods: CricaVPR [22], SALAD [16], SuperVLAD [24], SelaVPR [23], teacher model ImAge [21], EDTformer [17] (auxiliary teacher model in multi-teacher KD) and SAGE [9].

In the first comparison among KD-based VPR methods, RelaVPR achieves comprehensive superiority, with single-teacher and multi-teacher variants ranking first and second across nearly all benchmarks. In the second comparison with similar-scale methods, RelaVPR likewise dominates all six benchmarks. In terms of efficiency, it uses a 3072-D descriptor, has the second fewest parameters, and is only 1.3 ms slower than the fastest. In the third comparison with larger-scale methods including the teacher model ImAge, RelaVPR delivers highly competitive performance while having the smallest parameter count—just one-fourth that of the second lightweight method, SALAD (except for teacher ImAge)—and the fastest inference speed; additional efficiency comparisons in Sec. A.4.2, covering single-image inference time, batched inference time, and retrieval time, further demonstrate RelaVPR's strong trade-off between accuracy and efficiency.

Method	Dim.	Backbone	Evaluated Image Size	Param. (M)	Latency (ms)
D^2-VPR AAAI' 26	4096	DINOv2-S	224×224	27.2	25.5
RelaVPR Multi-teacher	3072	DINOv2-S	224×224	22.1	6.5
RelaVPR Single-teacher	3072	DINOv2-S	224×224	22.1	6.5
CosPlace CVPR' 22	2048	ResNet50	No Resize	27.7	5.5
MixVPR WACV' 23	4096	ResNet50	320×320	10.9	5.2
EigenPlaces ICCV' 23	2048	ResNet50	No Resize	27.7	5.5
LPS-VPR RAL' 24	2048	ResNet50	640×480	29.7	-
EPSA-VPR JVCI' 25	1024	ResNet50	-	27.7	-
Clus-VPR TAI' 25	4096	CWTNet	640×480	53.1	-
BoQ CVPR' 24	16384	ResNet50	384×384	23.8	7.6
RelaVPR Single-teacher	3072	DINOv2-S	322×322	22.1	6.5
SelaVPR ICLR' 24	1024	DINOv2-L	224×224	357.4	27.6
SALAD CVPR' 24	8448	DINOv2-B	322×322	88.0	13.9
SuperVLAD NeurIPS' 24	3072	DINOv2-B	322×322	97.6	12.6
CricaVPR CVPR' 24	4096	DINOv2-B	224×224	106.8	23.1
SAGE ICLR' 26	8448	DINOv2-B	322×322	88.5	12.7
ImAge NeurIPS' 25	6144	DINOv2-B	322×322	86.6	12.2
EDTformer TCSVT' 25	4096	DINOv2-B	322×322	97.0	13.3
RelaVPR Single-teacher	3072	DINOv2-S	322×322	22.1	6.5

Table 2: Details of the compared SOTA methods in Tab. 1. Latency is measured with a 322×322 input, except for MixVPR and SelaVPR, which only support 320×320 and 224×224, respectively. RKD, RSD, and CKA are not listed because the same model is used for the remaining RelaVPR. See Section A.4.1 for more information.

To further investigate robustness under severe illumination and weather changes, we additionally evaluated RelaVPR on SVOX-Night/ Rain/ Sun [27] in Tab. 3. RelaVPR achieves R@1 scores of

Method	Dim.	Param. (M)	SVOX-Night			SVOX-Rain			SVOX-Sun			R@1
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
MixVPR WACV' 23	4096	10.9	64.4	79.2	83.1	91.5	97.2	98.1	84.8	93.2	94.7	80.2
EigenPlaces ICCV' 23	2048	27.7	58.9	76.9	82.6	90.0	96.4	98.0	86.4	95.0	96.4	78.4
CricaVPR-1 CVPR' 24	4096	106.8	76.8	88.0	92.3	93.5	98.5	99.0	87.8	97.2	97.9	86.0
SuperVLAD-1 NeurIPS' 24	3072	97.6	91.9	97.8	98.4	96.7	98.8	99.4	97.3	99.2	99.3	95.3
SelaVPR ICLR' 24	1024	357.4	88.9	95.7	97.3	93.8	98.4	98.9	90.9	96.0	96.8	91.2
EDTformer TCSVT' 25	4096	97.0	96.2	98.7	99.4	98.1	**99.8**	**99.8**	**98.1**	**99.5**	**99.8**	97.5
ImAge NeurIPS' 25	6144	86.6	**97.7**	**99.5**	**99.5**	**99.0**	99.6	99.7	97.3	99.2	99.6	**98.0**
RelaVPR Single-teacher	3072	22.1	91.5	97.1	98.3	98.0	99.7	99.7	96.4	98.4	98.9	95.3

Table 3: Comparison of different methods on the SVOX benchmark datasets. The average score is calculated over R@1 of SVOX-Night, SVOX-Rain, and SVOX-Sun. RelaVPR is evaluated at a resolution of 322×322.

91.5/98.0/96.4, with an average of 95.3. Despite having only 22.1M parameters, it matches the much larger SuperVLAD-1 (95.3, 97.6M) and substantially outperforms SelaVPR (91.2, 357.4M) and CricaVPR-1 (86.0, 106.8M). It also significantly surpasses lightweight methods of a comparable scale, including MixVPR (80.2, 10.9M) and EigenPlaces (78.4, 27.7M).

5.3 Ablation Study

Loss Design. We compare the performance of RKD baseline with that of our method incorporating URSF and HAA. The results show that URSF and HAA effectively filter out noisy knowledge and enhance the learning of hard samples, thereby improving the overall KD performance. See Sec. A.4.4 for more ablation study.

Method	Pitts30k		MSLS-val		SPED	
	R@1	R@5	R@1	R@5	R@1	R@5
ST	92.1	96.1	91.8	95.8	87.5	93.2
ST + U	92.4	96.3	92.8	95.9	88.8	93.6
ST + U + H	**92.5**	**96.6**	**93.0**	**96.1**	**89.3**	**93.9**
MT	92.2	96.3	91.5	96.1	88.5	94.2
MT + DD	**92.6**	**96.4**	**92.2**	**96.4**	**89.8**	**94.6**

Table 4: Ablation study on multi-teacher (MT) and single-teacher (ST) frameworks. DD: decoupled distillation; U: URSF; H: HAA.

Model Design. In Tab. 4, we compare the multi-teacher KD approach—where multiple teachers simultaneously constrain the student model's output—with our decoupled distillation method. The results demonstrate that our approach effectively mitigates conflicts among multiple-teacher gradient signals. See Sec. A.4.4 for more ablation study.

6 Conclusion

In this work, we propose RelaVPR, a relation-based multi-teacher KD framework for efficient VPR. Built on a systematic analysis of KD loss designs, RelaVPR achieves faster inference, a smaller model size, and stronger performance than existing KD and efficient VPR methods, all without requiring fine-tuning, while remaining competitive with top-performing VPR methods. We believe our analysis and framework will offer valuable insights for efficient VPR in the VFM era.

Acknowledgments and Disclosure of Funding

This work was supported by Key Research and Development Program of Ningxia Hui Autonomous Region (2026BEG02081).

References

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

[2] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Mixvpr: Feature mixing for visual place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2998–3007, January 2023.

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

[4] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016.

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

[6] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022.

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

[8] Byungchul Chae and Seonyeong Heo. What to distill? fast knowledge distillation with adaptive sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2407–2416, 2025.

[9] Shunpeng Chen, Changwei Wang, Rongtao Xu, Xingtian Pei, Yukun Song, Jinzhou Lin, Wenhao Xu, Jingyi Zhang, Li Guo, and Shibiao Xu. SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *The Fourteenth International Conference on Learning Representations*, 2026.

[10] Mark Cummins and Paul Newman. Highly scalable appearance-only slam–fab-map 2.0. 2010.

[11] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. {SEED}: Self-supervised distillation for visual representation. In *International Conference on Learning Representations*, 2021.

[12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819, 2021.

[13] Oliver Grainge, Michael J. Milford, Indu Bodala, Sarvapali D. Ramchurn, and Shoaib Ehsan. Tetra-vpr: A ternary transformer approach for compact visual place recognition. *IEEE Robotics and Automation Letters*, 10(8):8396–8403, 2025.

[14] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*, pages 14141–14152, 2021.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[16] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[17] Tong Jin, Feng Lu, Shuyu Hu, Chun Yuan, and Yunpeng Liu. Edtformer: An efficient decoder transformer for visual place recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 3519–3529. PMLR, 09–15 Jun 2019.

[19] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in neural information processing systems*, 33:21002–21012, 2020.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[21] Feng Lu, Tong Jin, Canming Ye, Xiangyuan Lan, Yunpeng Liu, and Chun Yuan. Towards implicit aggregation: Robust image representation for place recognition in the transformer era. In *The Annual Conference on Neural Information Processing Systems*, 2025.

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

[23] Feng Lu, Lijun Zhang, Xiangyuan Lan, Shuting Dong, Yaowei Wang, and Chun Yuan. Towards seamless adaptation of pre-trained models for visual place recognition. In *ICLR*, 2024.

[24] Feng Lu, Xinyao Zhang, Canming Ye, Shuting Dong, Lijun Zhang, Xiangyuan Lan, and Chun Yuan. Supervlad: Compact and robust image descriptors for visual place recognition. In *Advances in Neural Information Processing Systems*, volume 37, pages 5789–5816, 2024.

[25] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. 1 year, 1000 km: The oxford robotcar dataset. *The International Journal of Robotics Research*, 36(1):3–15, 2017.

[26] Jinyu Miao, Kun Jiang, Tuopu Wen, Yunlong Wang, Peijin Jia, Benny Wijaya, Xuhe Zhao, Qian Cheng, Zhongyang Xiao, Jin Huang, Zhihua Zhong, and Diange Yang. A survey on monocular re-localization: From the perspective of scene map representation. *IEEE Transactions on Intelligent Vehicles*, pages 1–33, 2024.

[27] Gabriele Moreno Berton, Valerio Paolicelli, Carlo Masone, and Barbara Caputo. Adaptive-attentive geolocalization from few queries: a hybrid approach. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2917–2926, 2021.

[28] Jiwei Nie, Dingyu Xue, Feng Pan, Zuotao Ning, Wei Liu, Jun Hu, and Shuai Cheng. Efficient saliency encoding for visual place recognition: Introducing the lightweight pooling-centric saliency-aware vpr method. *IEEE Robotics and Automation Letters*, 9(7):6035–6042, 2024.

[29] Jiwei Nie, Qixi Zhao, Dingyu Xue, Feng Pan, and Wei Liu. Epsa-vpr: A lightweight visual place recognition method with an efficient patch saliency-weighted aggregator. *Journal of Visual Communication and Image Representation*, 110:104440, 2025.

[30] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[31] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3962–3971, 2019.

[32] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[33] Sivic and Zisserman. Video google: a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, 2003.

[34] Andreas Peter Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *Transactions on Machine Learning Research*, 2022.

[35] Niko Sünderhauf, Peer Neubert, and Peter 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.

[36] Akihiko Torii, Relja Arandjelović, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(2):257–271, 2018.

[37] Akihiko Torii, Josef Sivic, Tomas Pajdla, and Masatoshi Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890, 2013.

[38] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

[39] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[40] Shanshan Wan, Yingmei Wei, Lai Kang, Tianrui Shen, Haixuan Wang, and Yee-Hong Yang. Scicevpr: Stable cross-image correlation enhanced model for visual place recognition. *arXiv preprint arXiv:2502.20676*, 2025.

[41] Changwei Wang, Shunpeng Chen, Yukun Song, Rongtao Xu, Zherui Zhang, Jiguang Zhang, Haoran Yang, Yu Zhang, Kexue Fu, Shide Du, et al. Focus on local: Finding reliable discriminative regions for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7536–7544, 2025.

[42] Frederik Warburg, Soren Hauberg, Manuel Lopez-Antequera, Pau Gargallo, Yubin Kuang, and Javier Civera. Mapillary street-level sequences: A dataset for lifelong place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2626–2635, 2020.

[43] Yifan Xu, Pourya Shamsolmoali, Masoume Zareapoor, and Jie Yang. Clusvpr: Efficient visual place recognition with clustering-based weighted transformer. *IEEE Transactions on Artificial Intelligence*, 2024.

[44] Burak Yildiz, Seyran Khademi, Ronald Maria Siebes, and Jan Van Gemert. Amstertime: A visual place recognition benchmark dataset for severe domain shift. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 2749–2755. IEEE, 2022.

[45] Mubariz Zaffar, Sourav Garg, Michael Milford, Julian Kooij, David Flynn, Klaus McDonald-Maier, and Shoaib Ehsan. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *International Journal of Computer Vision*, 2021.

[46] Weijia Zhang, Yuehao Liu, Wu Ran, and Chao Ma. Cross-architecture distillation made simple with redundancy suppression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23256–23266, October 2025.

[47] Zheyuan Zhang, Jiwei Zhang, Boyu Zhou, Linzhimeng Duan, and Hong Chen. D²-vpr: A parameter-efficient visual-foundation-model-based visual place recognition method via knowledge distillation and deformable aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):13034–13042, Mar. 2026.

[48] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2former: Unified retrieval and reranking transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023.

A Technical appendices and supplementary material

A.1 Overview

This supplementary material provides the following additional content:

A.2 Additional Theoretical Analysis of Relation-based and Value-based Loss
- **A.2.1** Full Proof of the Proposition
- **A.2.2** Specific Instantiations

A.3 Details of Method
- **A.3.1** Student Model
- **A.3.2** Proposed KD Loss

A.4 Additional Experimental Results
- **A.4.1** Evaluation Protocols
- **A.4.2** Trade-offs Comparison in Practical VPR Tasks
- **A.4.3** Additional Comparisons on Challenge Datasets
- **A.4.4** Additional Ablation Study Results
- **A.4.5** Comparison with More Lightweight Backbones
- **A.4.6** Batch Size Analysis of Relation-based KD Loss
- **A.4.7** Hyperparameter Analysis

A.5 Additional Visualizations
- **A.5.1** Qualitative Visual Comparison
- **A.5.2** Visualizations of Patch-feature-relation
- **A.5.3** Visualizations of Feature t-SNE

A.6 Details of the Datasets

A.7 Limitations

A.8 Broader Impacts

A.2 Additional Theoretical Analysis of Relation-based and Value-based Losses

A.2.1 Full Proof of the Proposition

Proof.

(i) Containment. Fix the teacher feature $Y \in \mathbb{R}^{B\times C}$. By assumption, $d(X_n, Y) \to 0$ implies $\|X_n - Y\|_F \to 0$. The continuity of $\mathcal{R}$ and δ, together with $\delta(A, A) = 0$, therefore gives

$$d(X_n, Y) \to 0 \quad \Longrightarrow \quad \delta\big(\mathcal{R}(X_n), \mathcal{R}(Y)\big) \to 0. \tag{13}$$

Consequently, for any tolerance $\epsilon > 0$, there exists $\epsilon_0 > 0$ such that

$$d(X, Y) < \epsilon_0 \implies \delta\big(\mathcal{R}(X), \mathcal{R}(Y)\big) < \epsilon, \tag{14}$$

which directly gives $\mathcal{S}_{\text{val}}(\epsilon_0) \subseteq \mathcal{S}_{\text{rel}}(\epsilon)$.

(ii) Strict inclusion. By assumption, the teacher feature Y is non-trivial, i.e., $Y \neq \mathbf{0}$, and $C \geq 2$. We construct an orthogonal matrix $Q \in \mathbb{R}^{C\times C}$ such that $YQ \neq Y$ as follows. Since $Y \neq \mathbf{0}$, there exists a non-zero row y_k of Y. Pick any unit vector $u \in \mathbb{R}^C$ that is not parallel to y_k, and let $\text{span}(y_k, u)$ denote the two-dimensional subspace spanned by y_k and u. Define $Q \in \mathbb{R}^{C\times C}$ to act as a non-trivial rotation (e.g., by an angle $\theta \in (0, \pi)$) within $\text{span}(y_k, u)$ and as the identity on its orthogonal complement. Then Q is orthogonal, $Q \neq I$, and $y_kQ \neq y_k$ by construction. Define the candidate student feature $X^* = YQ \in \mathbb{R}^{B\times C}$. By the orthogonal invariance assumption on $\mathcal{R}$,

$$\mathcal{R}(X^*) = \mathcal{R}(YQ) = \mathcal{R}(Y), \tag{15}$$

so $\delta(\mathcal{R}(X^*), \mathcal{R}(Y)) = 0 < \epsilon$ for all $\epsilon > 0$, and hence $X^* \in \mathcal{S}_{\text{rel}}(\epsilon)$. On the other hand, since the k-th row of X^* is $y_kQ \neq y_k$, we have $X^* \neq Y$, and thus $d(X^*, Y) > 0$. Choose

$$0 < \epsilon' < \min\{\epsilon_0, d(X^*, Y)\}. \tag{16}$$

Then

$$\mathcal{S}_{\mathrm{val}}(\epsilon') \subseteq \mathcal{S}_{\mathrm{val}}(\epsilon_0) \subseteq \mathcal{S}_{\mathrm{rel}}(\epsilon), \tag{17}$$

while $X^* \notin \mathcal{S}_{\mathrm{val}}(\epsilon')$. This establishes the strict inclusion $\mathcal{S}_{\mathrm{val}}(\epsilon') \subsetneq \mathcal{S}_{\mathrm{rel}}(\epsilon)$. □

Remark. The continuity and convergence conditions are satisfied by standard choices such as MSE and L1 distance for d, pairwise Euclidean distance or Gram matrices for $\mathcal{R}$, and the Frobenius norm of the matrix difference for δ. Orthogonal invariance also holds for centered Gram matrices used in linear CKA and normalized pairwise distance matrices used in RKD. For these normalized losses, all normalization denominators are assumed to be nonzero. The assumption $Y \neq \mathbf{0}$ is mild for learned teacher representations. Intuitively, the construction above produces a continuous family of embeddings $\{YQ_\theta : \theta \in (0, \pi)\}$ that are indistinguishable to $\mathcal{R}$ (and hence to the relation-based loss $\mathcal{L}_{\mathrm{rel}}$) but distinguishable to direct value-based alignment. This illustrates that relation-based distillation admits richer solution structures, providing a possible explanation for its empirical advantages in VPR. The comparison uses loss-dependent tolerances and does not assert containment at the same numerical threshold. The proposition applies to strict numerical alignment losses such as MSE and L1; cosine alignment, which permits additional scale freedom, is not covered by the assumption $d(A, B) = 0 \Leftrightarrow A = B$.

A.2.2 Specific Instantiations

To provide concrete intuition for the proposition in Sec. 3.3, we illustrate its implications through three representative instantiations: MSE as a value-based loss, and CKA and RKD as relation-based losses. These examples help visualize *what kinds of student features* are considered feasible under each type of loss, thereby supporting the observation that relation-based losses admit a broader family of feasible representations.

Problem Definition. Following the setup in Sec. 3.3, VPR aims to learn an embedding function $f(\cdot)$ that maps an input image x to a feature vector $z = f(x) \in \mathbb{R}^C$ such that images depicting the same place have similar representations while images from different places are far apart in the feature space. Given a query image x_q and a database $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$, VPR retrieves the image whose feature embedding is most similar to that of the query:

$$x^* = \arg\max_{x_i \in \mathcal{D}} s(f(x_q), f(x_i)), \tag{18}$$

where $s(\cdot, \cdot)$ denotes a similarity function. Let $X, Y \in \mathbb{R}^{B \times C}$ be the student and teacher feature matrices, where B is the batch size and C is the feature dimension. We adopt the standard row-vector convention, i.e., the i-th sample feature $y_i \in \mathbb{R}^{1 \times C}$ is the i-th row of Y (and analogously for x_i in X). We next examine what feature transformations are tolerated by each loss.

MSE Loss. The MSE loss is defined as

$$\mathcal{L}_{\mathrm{MSE}}(X, Y) = \frac{1}{BC}\|X - Y\|_F^2, \tag{19}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. MSE requires the student feature X to approximate the teacher feature Y in numerical values. For a given tolerance $\epsilon > 0$, the feasible set is

$$\mathcal{S}_{\mathrm{MSE}}(\epsilon) = \left\{ X \in \mathbb{R}^{B \times C} \;\middle|\; \|X - Y\|_F^2 < BC\,\epsilon \right\},$$

which corresponds to a bounded region centered at Y. Rotations or rescalings that change Y incur positive MSE and are excluded for sufficiently small tolerances, reflecting the strict point-wise alignment required by this loss.

CKA Loss. The CKA loss measures the similarity of relational structures between student and teacher features. For the following linear CKA expression, X and Y denote features centered across the batch dimension, and both centered feature matrices are assumed to be nonzero:

$$\mathrm{CKA}(X, Y) = \frac{\|X^\top Y\|_F^2}{\|X^\top X\|_F \, \|Y^\top Y\|_F}. \tag{20}$$

We define $\mathcal{L}_{\mathrm{CKA}}(X, Y) = 1 - \mathrm{CKA}(X, Y)$. Relation-based losses such as CKA do not require the student and teacher to share the same feature dimension, since the Gram matrix $XX^\top \in \mathbb{R}^{B\times B}$ depends only on the batch size B. CKA can be equivalently rewritten via Gram matrices as

$$\mathrm{CKA}(X, Y) = \frac{\langle XX^\top, YY^\top\rangle_F}{\|XX^\top\|_F \, \|YY^\top\|_F}. \tag{21}$$

Consider a linear transformation $X = YA$ with $A \in \mathbb{R}^{C\times C}$. If $AA^\top = cI$ for some $c > 0$ (equivalently, $A/\sqrt{c}$ is orthogonal), then

$$XX^\top = YAA^\top Y^\top = cYY^\top,$$

which yields $\mathrm{CKA}(X, Y) = 1$ and thus zero CKA loss. Hence CKA admits *at least* all such orthogonally related and uniformly scaled variants of Y as feasible solutions—a structurally richer family than the small neighborhood permitted by direct MSE alignment. The same orthogonal and scaling invariance argument applies to uncentered kernel alignment when it is computed on nonzero features.

RKD Distance Loss. RKD-distance focuses on preserving pairwise distances between samples. Let $D_X(i, j) = \|x_i - x_j\|_2$ and $D_Y(i, j) = \|y_i - y_j\|_2$ denote pairwise Euclidean distances. To remove dependence on the overall feature scale, let $\tilde{D}_X(i, j)$ and $\tilde{D}_Y(i, j)$ denote the distances normalized by their respective means over all non-diagonal pairs. We assume $B \geq 2$ and that both mean distances are strictly positive. The RKD distance loss is defined as

$$\mathcal{L}_{\mathrm{RKD}}(X, Y) = \sum_{i\neq j} \ell\big(\tilde{D}_X(i, j), \tilde{D}_Y(i, j)\big), \tag{22}$$

where $\ell(\cdot, \cdot)$ denotes the smooth L_1 loss. Consider a linear transformation $X = YA$ with $A \in \mathbb{R}^{C\times C}$. Under the row-vector convention, $x_i = y_i A$, so for any pair (i, j),

$$\|x_i - x_j\|_2^2 = (y_i - y_j)AA^\top(y_i - y_j)^\top.$$

If $AA^\top = cI$ with $c > 0$, this reduces to $\|x_i - x_j\|_2^2 = c\|y_i - y_j\|_2^2$, i.e., all pairwise distances are uniformly rescaled by $\sqrt{c}$. After mean normalization, $\tilde{D}_X(i, j) = \tilde{D}_Y(i, j)$ for every pair, yielding zero RKD loss. Hence, as for CKA, RKD admits all orthogonally related and uniformly scaled variants of Y as feasible solutions.

Summary. These three instantiations concretize the proposition in Sec. 3.3. For small tolerances, the feasible set of direct value alignment (MSE) is restricted to a neighborhood of Y, while relation-based losses (CKA, RKD) admit at least the family of transformations $X = YA$ satisfying $AA^\top = cI$, $c > 0$—i.e., orthogonally related and uniformly scaled variants of Y. When descriptors are constrained to have unit norms, admissible transformations must additionally respect that constraint. The examples identify sufficient transformations for zero relation loss; they do not characterize all retrieval-equivalent representations or guarantee ranking preservation at nonzero loss. This structural difference illustrates how relation-based distillation can provide more flexible supervision, consistent with the observations in our empirical analysis.

A.3 Details of Method

A.3.1 Details of Student Model

Component	Param. (M)
aggregation tokens	0.006
Backbone	22.058
All	22.064

Table 5: Analysis of model parameter quantity.

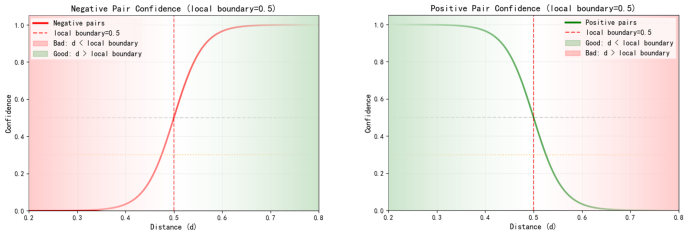
Tab. 5 presents a detailed breakdown of the model's parameter composition (dual-teacher version). Owing to our implicit aggregation design, the model completely eliminates traditional explicit

aggregation modules and instead relies on a small set of learnable aggregation tokens. As shown in the table, the aggregation tokens introduce only 0.006M parameters, which is negligible compared to the backbone network (22.058M parameters). Consequently, the total parameter count of the entire model remains almost identical to that of the backbone alone (22.064M parameters). This lightweight design significantly reduces architectural overhead while preserving strong representation capability, making the proposed framework particularly suitable for efficient and scalable VPR deployment.

A.3.2 Details of Proposed KD loss

Algorithm 1 KD Loss with URSF and HAA

Input: Student features X, teacher features Y, batch size B, temperature τ, sampling ratio ρ, number of nearest negatives k, number of hard anchors k_{hard}, loss weight λ
Output: KD loss $\mathcal{L}_{\text{KD}}$
for each anchor A_i **do**
  Compute Local Boundary:
  $N_i^{(k)} \leftarrow \text{TopK}_{j\neq i}^{k}\left(\tilde{D}_Y(A_i, N_j)\right)$ ▷ k nearest negatives
  $r_i \leftarrow \frac{1}{k}\sum_{N_j \in N_i^{(k)}} \tilde{D}_Y(A_i, N_j)$
  Compute Confidence Weights:
  for each positive pair (A_i, P_j) **do**
    $w_{ij} \leftarrow \sigma\left(\frac{r_i - \tilde{D}_Y(A_i, P_j)}{\tau}\right)$
  end for
  for each negative pair (A_i, N_j) **do**
    $w_{ij} \leftarrow \sigma\left(\frac{\tilde{D}_Y(A_i, N_j) - r_i}{\tau}\right)$
  end for
  Confidence-guided Probabilistic Filtering:
  $p_{ij} \leftarrow w_{ij} \big/ \sum_{l\neq i} w_{il}$ ▷ normalize into sampling distribution
  $m \leftarrow \lfloor \rho(B-1) \rfloor$
  $\mathcal{S}_i \leftarrow \text{Multinomial}\left(\{p_{ij}\}_{j\neq i},\ m,\ \text{without replacement}\right)$
end for
Compute Filtered RKD Loss:
$\mathcal{L}_{\text{RKD}}^{\text{URSF}}(X, Y) \leftarrow \frac{1}{\sum_i |\mathcal{S}_i|}\sum_i \sum_{j\in\mathcal{S}_i} \ell\left(\tilde{D}_X(A_i, X_j),\ \tilde{D}_Y(A_i, X_j)\right)$
Hard Anchor Mining:
for each anchor A_i **do**
  $h_i \leftarrow \frac{1}{B}\sum_j \left|\tilde{D}_X(A_i, X_j) - \tilde{D}_Y(A_i, X_j)\right|$
end for
$\mathcal{H} \leftarrow \text{TopK}_i^{k_{\text{hard}}}(h_i)$ ▷ select k_{hard} hardest anchors
Hardness-aware Anchor Alignment:
$\mathcal{L}_{\text{HAA}} \leftarrow 1 - \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \frac{\langle X_i P,\ Y_i\rangle}{\|X_i P\| \cdot \|Y_i\|}$
Combine Losses:
$\mathcal{L}_{\text{KD}} \leftarrow \mathcal{L}_{\text{RKD}}^{\text{URSF}} + \lambda\mathcal{L}_{\text{HAA}}$
return $\mathcal{L}_{\text{KD}}$



(a) Confidence for Negative (b) Confidence for Positive

Figure 5: The curve graph of the confidence function

We provide the proposed URSF and HAA in algorithmic form for readers to understand more clearly, as shown in Alg. 1. Meanwhile, we also provide the confidence function curve as shown in Fig. 5.

A.4 Additional Experimental Results

A.4.1 Evaluation Details

Evaluation Protocols and Details. We adopt Recall@N (R@N) as the evaluation metric, which reports the proportion of queries for which at least one of the top-N retrieved database images falls within the predefined ground-truth tolerance. Following established evaluation protocols [42, 37], the matching thresholds are set to 25 m and 40° for MSLS; 25 m for Pitts30k, Pitts250k, Tokyo24/7, Eynsham, SVOX and SPED; ±10 frames for Nordland, and unique counterpart matching for AmsterTime. In Tab. 2, the Clus-VPR and EPSA-VPR papers only report aggregator parameters, so we use their full ResNet50 versions, resulting in 23.5M backbone parameters. MixVPR and BoQ employ a cropped ResNet50.

Cross-image Encoder in Comparison. In Tab. 1, to ensure fairness of comparison, we do not directly compare methods that employ a cross-image encoder. The reason is that a cross-image encoder correlates multiple images from the same place during testing to achieve better performance, which would be unfair to other baselines. Therefore, existing methods such as ImAge [21], EDTformer [17], a follow-up work by the author of the cross-image encoder mechanism, and SciceVPR [40] all make a clear distinction in their comparisons. For methods that do use a cross-image encoder, including SuperVLAD [24], SAGE [9], and D^2-VPR [47], official weights without the cross-image encoder are also provided to enable fair evaluation.

A.4.2 Trade-offs Comparison in Practical VPR Tasks

Method	Avg. R@1	Param. (M)	Extract Database (s) / batchsize=32	Extract Queries (s) / batchsize=1	Calculation Time (s)	All Time (s)
MixVPR	79.0	10.9	20.7	23.6	10.5	54.8
EigenPlaces	77.7	27.7	20.9	25.3	5.3	51.5
SALAD	86.1	88.0	28.7	30.9	21.6	81.2
RelaVPR	86.8	22.1	21.1	26.4	7.8	55.3

Table 6: Trade-offs comparison on AmsterTime. Experiments are conducted on an RTX 3090 with 1231 query and database images. Input resolution is fixed to 322 × 322 for fairness, except for MixVPR (only supports 320 × 320).

To clearly illustrate that our method achieves a new state-of-the-art trade-off, we compare it with the two most efficient similar-scale methods (EigenPlaces and MixVPR) and a representative efficient large-VFM-based method (SALAD) in Tab. 6. Image extraction time reflects inference latency, while retrieval time depends on descriptor dimensionality. Compared to SALAD, our method uses only a quarter of the parameters and reduces total computation time by 32%, while achieving better average R@1 (86.8 vs. 86.1) across six datasets. Compared to similar-scale methods, RelaVPR achieves comparable runtime while improving average R@1 by 9.1 and 7.8 over EigenPlaces and MixVPR, respectively.

A.4.3 Additional Comparisons on Challenge Datasets

Method	Dim.	Backbone	Evaluated Image Size	Param. (M)	Latency (ms)	Tokyo24/7			Eynsham		
						R@1	R@5	R@10	R@1	R@5	R@10
SALAD CVPR' 24	8448	DINOv2-B	322×322	88.0	13.9	94.6	97.5	97.8	91.6	95.0	95.9
SAGE ICLR' 26	8448	DINOv2-B	322×322	88.5	12.7	97.5	98.7	99.4	92.3	95.8	96.5
RelaVPR Single-teacher	3072	DINOv2-S	322×322	22.1	6.5	95.2	97.1	97.8	92.1	95.4	96.3

Table 7: Additional Comparisons on Tokyo24/7 [36] and Eynsham [10, 6].

As shown in Tab. 7, we further evaluate our method on the additional benchmarks Tokyo24/7 [36] and Eynsham [10, 6]. The results demonstrate that our method significantly outperforms SALAD on Tokyo24/7, achieving +0.6 improvement on R@1 while using substantially fewer parameters (22.1M vs. 88.0M) and lower latency (6.5ms vs. 13.9ms). Compared with SAGE, our method achieves competitive performance despite using a much lighter backbone (DINOv2-S vs. DINOv2-B) and significantly reduced model complexity. Specifically, our approach attains comparable results on both Tokyo24/7 (95.2 vs. 97.5 R@1) and Eynsham (92.1 vs. 92.3 R@1), while reducing parameters by

approximately 75% and latency by about 49%. These results highlight the effectiveness and efficiency of our method, demonstrating a strong trade-off between performance and computational cost.

A.4.4 Additional Ablation Study Results

Method	Pitts30k			MSLS-val			SPED			Pitts250k			AmsterTime			Nordland			Avg.
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1
ImAge* NeurIPS' 25	**93.8**	**97.0**	**97.9**	**94.5**	**97.3**	**97.7**	90.0	**95.4**	96.0	**96.3**	**99.0**	**99.4**	54.6	75.8	80.7	**95.9**	**98.6**	**99.3**	**87.5**
EDTformer* TCSVT' 25	92.9	96.8	97.8	91.5	96.4	96.6	**90.8**	**95.4**	**96.9**	95.5	98.7	99.3	**58.4**	**80.8**	**84.8**	81.0	91.2	94.1	85.0
RelaVPR Single-teacher	92.5	**96.6**	**97.7**	**93.0**	96.1	96.6	89.3	93.9	95.1	94.9	98.5	99.2	52.8	73.8	79.4	**89.4**	**96.0**	**97.5**	**85.3**
RelaVPR Multi-teacher	**92.6**	96.4	97.4	92.2	**96.4**	**96.8**	**89.8**	**94.6**	**95.4**	**95.0**	**98.7**	**99.3**	**53.2**	**74.3**	**79.9**	86.1	94.3	96.3	84.8
RelaVPR Multi-teacher (Primary)	**92.6**	**96.4**	**97.4**	**92.2**	**96.4**	**96.8**	89.8	**94.6**	**95.4**	95.0	**98.7**	**99.3**	53.2	74.3	79.9	86.1	94.3	96.3	84.8
RelaVPR Multi-teacher (Auxiliary)	91.9	96.3	97.3	90.0	95.4	96.2	88.5	93.9	95.1	94.7	98.4	99.0	52.5	73.4	79.5	83.0	92.4	94.9	83.4
RelaVPR Multi-teacher (Concat)	92.5	96.5	97.4	91.4	96.2	**96.8**	**90.0**	94.2	95.1	**95.2**	98.6	99.2	**54.2**	**76.2**	**81.2**	**87.9**	**95.1**	**96.9**	**85.2**

Table 8: Ablation study on multi-teacher KD. we compare three aspects: 1) a comparison between the two teachers in the multi-teacher KD setting, 2) a comparison between multi-teacher KD and single-teacher KD, and 3) an ablation on the use of only the aggregation token aligned with the primary teacher in multi-teacher KD. For the last three rows, we compare three inference strategies: using only the aggregation token aligned with the primary teacher (our default), using only the token aligned with the auxiliary teacher, or concatenating both tokens. * indicates that ImAge and EDTformer are reproduced under resolution of 224 to match our evaluation resolution.

Model Design. As shown in Table 8, we analyze whether RelaVPR can effectively integrate the advantages of multiple teachers. In RelaVPR, the single-teacher KD version uses ImAge as the teacher, while the multi-teacher version employs both ImAge and EDTformer as teachers. We observe that compared to ImAge, EDTformer exhibits clear advantages on SPED and AmsterTime. This is also reflected in the multi-teacher version of RelaVPR, which achieves better results than the single-teacher version on SPED and AmsterTime, while maintaining comparable performance on other datasets. This demonstrates that the integration is effective.

Meanwhile, we observed that multi-teacher performance degrades on some datasets (e.g., Nordland, MSLS). This is because the auxiliary teacher EDTformer performs significantly worse than the main teacher ImAge on these datasets (ImAge achieves 95.9 on Nordland, while EDTformer achieves only 81.0), introducing negative transfer.

As shown in third part of Table 8, we validate our inference design choice by comparing three strategies within the multi-teacher setting: using only the primary teacher's aggregation token (our default), using only the auxiliary teacher's token, or concatenating both.

Two key observations emerge. First, the auxiliary token alone yields the weakest performance across all six benchmarks (Avg. R@1: 83.4), consistently lagging behind both primary-only and concat by a large margin. This indicates that the auxiliary subspace, by design, captures attribute-specific or localized cues rather than the globally consistent representations required for retrieval—supporting our motivation that the primary teacher yields more stable and discriminative features for inference. Second, concatenating both tokens (Avg. R@1: 85.2) achieves only a marginal improvement over primary-only (Avg. R@1: 84.8), despite doubling the feature dimension and introducing additional retrieval computation. Moreover, the improvement of concat mainly comes from AmsterTime and Nordland; on the remaining four benchmarks, primary-only is roughly on par with or exceeds concat. This confirms that auxiliary knowledge is largely absorbed into the shared backbone during training, so the primary token alone already captures most of the complementary information, and fusion offers diminishing returns at the cost of efficiency. These results collectively justify our design: retaining only the primary teacher's token preserves inference simplicity and efficiency while delivering competitive—and in several cases superior—performance.

Method	Pitts30k		MSLS-val		SPED	
	R@1	R@5	R@1	R@5	R@1	R@5
deterministic weighting	92.3	96.4	92.2	95.9	89.0	**94.2**
probabilistic sampling	**92.5**	**96.6**	**93.0**	**96.1**	**89.3**	93.9

Table 9: Ablation study on probabilistic sampling in URSF.

KD Loss. We further ablate the filtering strategy in URSF. As shown in Tab. 9, probabilistic sampling consistently outperforms deterministic weighting on R@1 across all three datasets (+0.2 on Pitts30k,

+0.8 on MSLS-val, +0.3 on SPED), confirming the benefit of confidence-guided stochastic sampling over direct weighted averaging. We attribute this improvement to two factors: (i) sampling injects stochasticity into the relational supervision, acting as an implicit regularizer that prevents the student from overfitting to any particular pair configuration; (ii) deterministic weighting propagates gradients from every pair in every step, so low-confidence pairs still consistently bias the optimization, whereas sampling only exposes them occasionally in proportion to their reliability, limiting their aggregate influence while retaining their structural information.

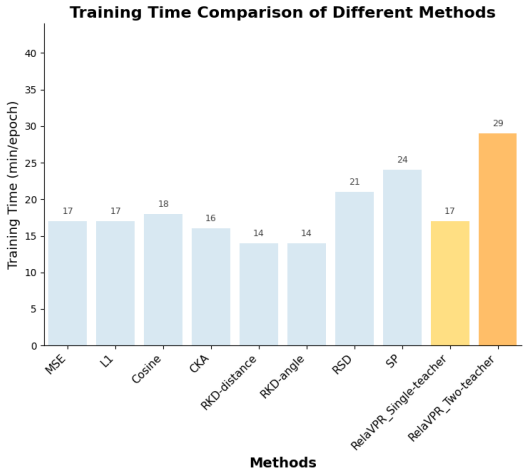


Figure 6: Comparison of training time between existing KD methods and our RelaVPR under identical batch size using an RTX 4090.

As shown in Fig. 6, when using the single-teacher version, our method only slightly increases the time compared to the baseline RKD-distance.

A.4.5 Comparison with More Lightweight Backbones

Method	Param. (M)	Pitts30k		MSLS-val		SPED		Avg. R@1
		R@1	R@5	R@1	R@5	R@1	R@5	
ViT-Small	21.67	90.2	95.4	85.3	93.5	71.0	81.2	82.2
DeiT-Small	21.67	89.9	95.7	85.4	93.2	73.5	86.2	82.9
DINOv2-Small	22.10	92.5	96.6	93.0	96.1	89.3	93.9	91.6

Table 10: Comparison of different visual backbones on three VPR benchmarks.

RelaVPR is backbone-agnostic at the loss level, as URSF and HAA operate on feature relations without relying on DINOv2-specific components. We further replaced DINOv2-Small with two lightweight Transformer backbones (ViT-Small [34] and DeiT-Small [38]) under identical settings. ViT-Small and DeiT-Small achieve 82.2 and 82.9 Avg. R@1, respectively, confirming that RelaVPR can generalize to different lightweight Transformer architectures. DINOv2-Small achieves the best result of 91.6 Avg. R@1 with a comparable parameter count. We attribute this advantage to its large-scale self-supervised pre-training and the more homogeneous DINOv2-B teacher–DINOv2-S student architecture, which reduces the representation gap and facilitates relational knowledge transfer. These results show that pre-training quality and teacher–student compatibility substantially affect the performance–efficiency frontier.

A.4.6 Batch Size Analysis of Relation-based KD Loss

In Tab. 11, we analyze the impact of batch size on distillation performance. The results show that while larger batch sizes lead to more stable performance—consistent with the known sensitivity of relation-based losses to batch size and the benefit of reducing estimation noise—we ultimately choose a batch size of 24.

batch size	Pitts30k			MSLS-val			SPED		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
8	92.2	95.8	97.0	92.3	95.5	96.4	88.1	93.6	94.2
14	92.4	96.1	97.1	92.4	95.7	96.5	88.3	93.6	94.1
24	92.5	96.6	97.7	93.0	96.1	96.6	89.3	93.9	95.1

Table 11: Batch size analysis.

A.4.7 Hyperparameter Analysis

Param.	Value	Pitts30k		MSLS-val	
		R@1	R@10	R@1	R@10
ρ	0.7	92.3	97.5	92.6	96.5
	0.8	92.2	97.3	92.3	96.2
	0.9	**92.5**	**97.7**	**93.0**	**96.6**
	1.0	92.4	97.6	92.8	96.5
k_{neg}	0.05	92.3	97.5	92.7	96.4
	0.10	**92.5**	**97.7**	**93.0**	**96.6**
	0.20	92.4	97.6	92.8	96.5
k_{hard}	0.1	92.3	97.5	92.7	96.5
	0.2	**92.5**	**97.7**	**93.0**	**96.6**
	0.3	92.4	**97.7**	92.8	**96.6**
multi-teacher ratio	2:1	92.1	97.3	91.6	96.6
	1:1	**92.6**	97.4	**92.2**	**96.8**
	1:2	92.3	**97.5**	92.0	96.6

Table 12: Sensitivity analysis of key hyperparameters in RelaVPR. ρ: URSF sampling ratio; k_{neg}: fraction of batch used as nearest negatives for local boundary construction; k_{hard}: fraction of batch selected as hard anchors in HAA. Bold denotes the chosen configuration.

We conduct sensitivity experiments on key hyperparameters, including ρ, k_{neg}, k_{hard}, and the multi-teacher subspace allocation ratio. As shown in Tab. 12, RelaVPR maintains stable performance under different settings. For multi-teacher distillation, the allocation ratio is achieved by adjusting the number of aggregation tokens assigned to each teacher. Specifically, the primary teacher consistently uses 8 tokens, while the auxiliary teacher uses 4, 8, and 16 tokens, corresponding to ratios of 2:1, 1:1, and 1:2, respectively. The balanced allocation (8:8) achieves the best performance, while other configurations only cause marginal variations, demonstrating the robustness of our framework to hyperparameter selection.

A.5 Additional Visualizations

A.5.1 Qualitative Visual Comparison

In Fig. 7, we present qualitative comparison results from the evaluation, which demonstrate the robustness of our method in handling perceptual aliasing, long-term changes, and variations in illumination and viewpoint.

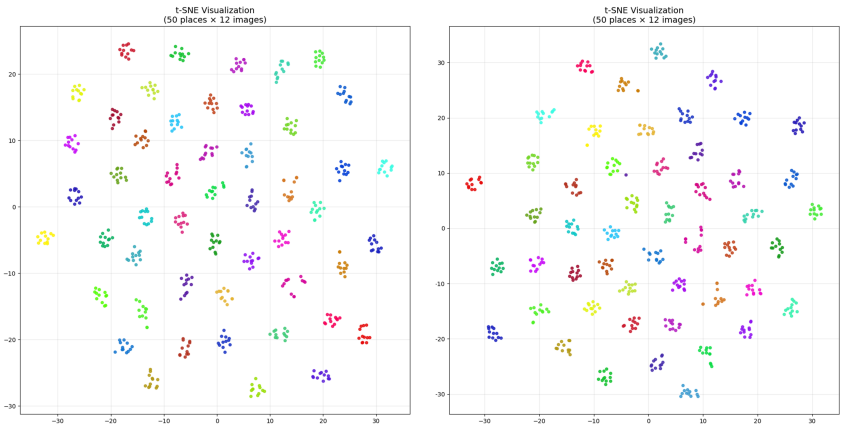
A.5.2 Visualizations of Patch-feature-relation

As shown in Fig. 8, we offer more visualizations of cosine similarity between patch tokens at specific positions for the different KD methods. Student models trained with CKA exhibit patch token relationships that are more consistent with those of the teacher, while MSE-only distillation results in less clearly structured semantic relationships. This comparison suggests that incorporating relational knowledge can provide more informative supervision during the distillation process.



Figure 7: Qualitative comparison results of VPR. Our method exhibits competitive performance against other SOTA models under various challenging scenarios: long-term appearance variations (first row), perceptual aliasing (second row), lighting changes (third row), and viewpoint variations (fourth row). Correct matches are marked in green, whereas incorrect ones are marked in red.

A.5.3 Visualizations of Feature t-SNE



(a) RelaVPR AID=0.7624 (b) ImAge AID=0.7853

Figure 9: t-SNE visualization comparing feature embeddings from RelaVPR and ImAge (Teacher) on 600 GSV Cities images spanning 50 locations. Cluster quality is measured by Average Intra-class Distance (AID).

As shown in Fig. 9, to evaluate the discriminative power of the learned representations, we visualize the t-SNE projections of features extracted by RelaVPR and its teacher model, ImAge. We observe that our method successfully distills the knowledge from ImAge, achieving comparably compact clusters.

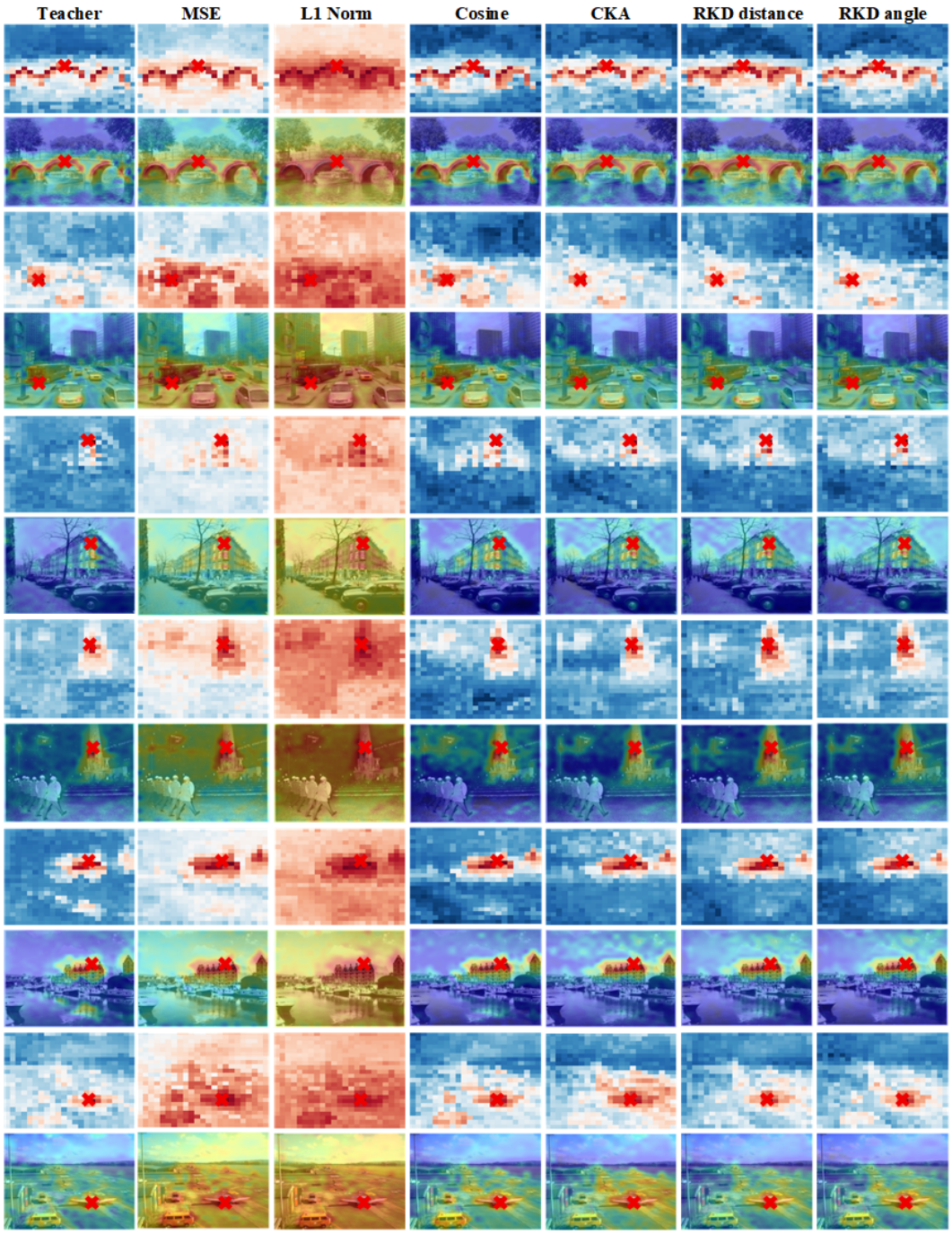


Figure 8: Visualization comparison of cosine similarity between patch tokens at red cross positions and other patch tokens in teacher models and student models distilled via different schemes.

A.6 Details of the Datasets

Dataset	Description	Number	
		Database	Queries
Pitts30k	urban	10,000	6,816
Pitts250k	urban	83952	8280
MSLS-val	urban, suburban	18,871	740
Nordland	natural, seasonal	27,592	27,592
AmsterTime	long-term	1,231	1,231
SPED	various scenes	607	607
Tokyo24/7	urban, time-varying	75984	315
Eynsham	rural, historical	23935	23935
SVOX	cross-domain	17166	4356

Table 13: Summary of the test datasets in experiments.

In this section, we detail the datasets used in this paper. An overview is provided in Tab. 13.

Pitts30k and Pitts250k [37] are large-scale image retrieval datasets commonly used for evaluating VPR systems. Constructed from Google Street View panoramas in the city of Pittsburgh, these datasets consist of cropped perspective images with precise GPS coordinates. Pitts30k contains around 30,000 images and is often used for training or quick evaluation, while Pitts250k includes approximately 250,000 images and offers a more challenging benchmark. Positive and negative samples are defined based on geographic distance, typically considering images within 25 meters as positives and those farther away as negatives. These datasets have become standard benchmarks for testing the robustness and generalization ability of VPR models in real-world urban environments.

MSLS [42] is a large-scale dataset for evaluating visual place recognition and long-term localization under real-world conditions. It contains over 1.6 million images collected from 30 cities across six continents, covering diverse seasons, weather conditions, times of day, and camera devices. Images are organized into sequences with GPS and compass data, enabling robust benchmarking for sequence-based and single-image VPR methods. MSLS is particularly challenging due to its extreme variations in viewpoint and appearance, making it a key benchmark for testing model generalization in global-scale applications.

Nordland [35] is a VPR dataset constructed from a train journey recorded along the same 728 km route across Norway during the four seasons: spring, summer, autumn, and winter. Each journey was recorded using a front-facing camera mounted on a train, resulting in four time-synchronized video sequences. The dataset is especially challenging due to the drastic seasonal appearance changes, while the viewpoint and motion remain nearly identical. It serves as a benchmark for evaluating the robustness of VPR models under extreme environmental changes.

SPED (test) [45] is a curated subset of the original SPED dataset, introduced in the VPR-Bench framework for standardized evaluation of VPR algorithms. It comprises images captured from fixed-position outdoor surveillance cameras under extreme appearance variations such as day/night, seasonal, and weather changes. The dataset emphasizes long-term appearance change while maintaining fixed viewpoints, making it ideal for testing a model's robustness to illumination and temporal variation in place recognition tasks.

AmsterTime [44] is a recently introduced dataset focused on evaluating long-term VPR under extreme appearance changes over years. It contains images spanning over a decade, covering changes in lighting, weather, seasons, and even structural modifications in the environment. Unlike many other datasets, AmsterTime highlights the challenges of long-term autonomy, making it valuable for benchmarking models that aim to operate reliably over extended periods in the same environment.

Eynsham [10, 6] comprises a long-range driving sequence captured in Oxfordshire, UK, spanning roughly 35 km. The data are recorded with a vehicle-mounted camera and aligned with GPS measurements, providing accurate location annotations for each frame. Thanks to its continuous temporal coverage and reliable ground truth, it is widely used for evaluating sequence-based place recognition and loop closure detection.

Tokyo24/7 [36] consists of street-level imagery captured in Tokyo, Japan, with images largely derived from Google Street View panoramas and associated GPS annotations. It includes both daytime and nighttime scenes, capturing significant long-term appearance changes. The dataset covers densely

built urban regions with high-rise buildings, along with occasional open areas, resulting in substantial variations in viewpoint, lighting conditions, and dynamic elements. It is commonly used to assess the robustness of VPR methods in complex urban settings.

SVOX [27] is a cross-domain VPR benchmark covering diverse weather and illumination conditions. Its large-scale reference database consists of Google Street View images from across Oxford, while its queries come from the Oxford RobotCar dataset [25] and are grouped by weather and lighting condition.

A.7 Limitations

Although our study analyzes knowledge distillation for VPR and provides a novel solution for achieving robust and lightweight VPR, we acknowledge the following limitations of our work: 1) Our designed loss for filtering teacher noise relies on the anchor-based partitioning of positive and negative samples as the basis for filtering. This means that if the partitioning itself contains noise, our filtering algorithm will also be affected. 2) Although this work focuses on the VPR task, we believe that the proposed framework is broadly applicable to a wide range of image (or other modalities) retrieval tasks. This potential needs to be further explored through more experiments in future work.

A.8 Broader Impacts

VPR serves as a fundamental component in many computer- and robotic-vision applications, including visual relocalization, SLAM, and augmented reality. Our work analyzes knowledge distillation methods for VPR and proposes an improved KD framework based on our conclusions, delivering an efficient VPR method. We believe this provides insights and guidance for future work on KD for VPR in terms of KD selection and KD framework design. However, we recognize that current VPR research raises certain privacy concerns. It is essential to take proactive measures to prevent the application of VPR research in harmful or unethical contexts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the introduction and abstract, we present our analysis of knowledge distillation in the VPR field, along with the resulting conclusions. Based on these findings, we propose a knowledge distillation framework and corresponding improvements.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the appendix, we discuss the limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: In our analysis of different KD losses for VPR, we validate our proposed claims both theoretically and experimentally.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the implementation details for both training and testing of our experiments, and we commit to open-sourcing the relevant code and weights.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The full codes and model checkpoints for reproducing our methods will be publicly available upon paper publication.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the implementation details for both training and testing of our experiments, and we commit to open-sourcing the relevant code and weights.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not commonly used in the VPR research.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: During testing, we clearly specify the model and configuration of the computing resources used.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, it is.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss this part in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our model achieves a new state-of-the-art trade-off between performance and efficiency in the VPR task. Although there is a small possibility of unintended or malicious use, our project is not at high risk since we do not release new datasets. We will include a reminder of the associated risks in the README.md file of our upcoming open-source project.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In the paper, we have accurately cited the original sources of the datasets in the experimental section and the appendix. We respect the licenses of the referenced code and data and will properly acknowledge them in the project's README.md file.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Codes and model checkpoints for reproducing our methods will be publicly available upon paper publication.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper neither involves crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: In this research, the development of the core methods is entirely based on our own original thinking, without any involvement of LLMs as important, original, or non-standard components. LLMs were only utilized for paper polishing purposes, which falls under the category of writing, editing, or formatting.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.