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

Cross-Distill: Multi-Manifold and Viewpoint-Decoupled Distillation for Cross-View Geo-Localization

Jiaxu Gao[1], Shuying Zhao[1*], Yunzhou Zhang[1], Hongyu Zhou[1], Man Qi[1], Jiabo Shen[1], Yu Zhang[1]

***Abstract*— Cross-View Geo-Localization (CVGL) localizes a query image via retrieval from georeferenced satellite imagery, yet severe viewpoint variation remains a central challenge. Recent advances often rely on heavy backbones or add-on modules that achieve high accuracy but are impractical on resource-constrained UAVs. To balance accuracy and efficiency, we introduce Cross-Distill, a knowledge-distillation framework for CVGL. Cross-Distill performs Cross-Similarity Ranking Distillation by constructing a teacher–student interaction matrix to enforce ranking consistency and enhance discrimination. Building on this, it introduces Viewpoint Decoupling, which partitions ranking relations into intra-view, intra-to-cross-view, and cross-to-cross-view, enabling precise modeling of cross-view dependencies and improving class compactness and separability. Cross-Distill further employs Multi-Manifold Feature Distillation that jointly enforces angular consistency on the spherical manifold, preserves local distances in Euclidean space, and leverages hyperbolic distance as a negatively curved metric to strengthen teacher–student alignment. Experiments on University-1652 and SUES-200 show that the distilled student achieves significant gains with low complexity (31.43M parameters, 13.09 GFLOPs), and an inference time of only 62.02 ms per image on an RK3588. For instance, on University-1652 UAV→SAT retrieval, R@1 improves from 75.97% to 94.43% and AP from 79.24% to 95.33%.**

I. INTRODUCTION

CVGL aims to determine the location of an image within large-scale environments by establishing correspondences between a query image and georeferenced satellite images. This task has applications in autonomous driving and robotic navigation [1]. Initially, CVGL was developed as a complementary approach to GPS-based localization, particularly for ground-to-satellite matching in urban areas where tall buildings often degrade Global Navigation Satellite System (GNSS) signals [2]. In recent years, with the rapid proliferation of low-altitude unmanned aerial vehicles (UAVs), CVGL research based on UAV imagery from near-ground perspectives has attracted growing attention in both academia and industry.

Prior work on CVGL has largely pursued stronger feature representations. Early approaches learned global descriptors for holistic structure [3] and local/part features for fine-grained cues [4]; globals may miss discriminative detail, whereas locals are brittle under viewpoint/scale changes.

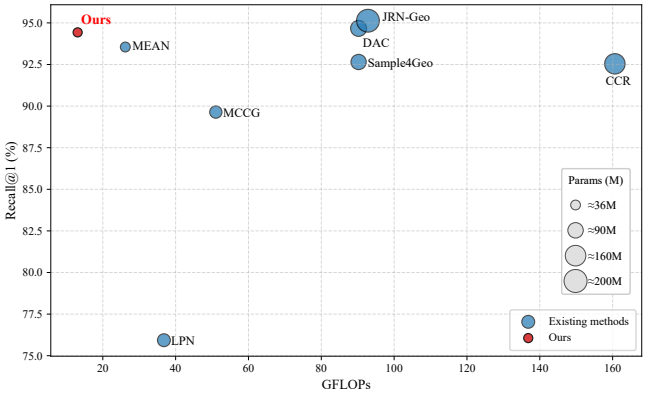

Fig. 1: Scatter plot of UAV→SAT retrieval performance on University-1652. Recall@1 (vertical axis), computational cost in GFLOPs (horizontal axis), and parameter size (circle radius) are jointly illustrated.

Subsequent methods integrated complementary global–local cues with stronger architectures. Transformer-based models [5] [6] (primarily for ground–satellite) capture long-range dependencies and cross-view context. Despite these advances, many pipelines still hinge on heavy backbones [7] or costly features fusion [8], hindering real-time deployment on resource-constrained UAVs. Lightweight designs [9] [10] narrow the gap but typically introduce auxiliary modules to recover accuracy, increasing system complexity.

However, Knowledge Distillation (KD) [11] has shown strong potential in related vision tasks by transferring knowledge from large teacher models to lightweight students, offering a promising direction for balancing accuracy and efficiency. Existing KD methods in image retrieval mainly target person re-identification [12] [13], where intra-class and inter-class relations are relatively constrained. In contrast, CVGL-oriented distillation remains underexplored. A key challenge is recovering the teacher's fine-grained relational distribution under severe viewpoint shifts while bridging the student–teacher gap via geometry-aware alignment beyond a single manifold. Addressing these challenges is crucial for advancing both accuracy and robustness of lightweight CVGL models.

To address these challenges, we propose **Cross-Distill**, a knowledge distillation framework tailored for cross-view geo-localization. Our method jointly models teacher–student feature relationships and cross-view geometric dependencies, while introducing a cross-manifold consistency constraint that further contracts features in multiple manifold spaces. As illustrated in Fig. 1, **Cross-Distill** significantly enhances student performance, achieving high accuracy while maintaining low parameter count and computational cost. The

*The corresponding author of this paper
[1]Jiaxu Gao, Shuying Zhao, Yunzhou Zhang, Hongyu Zhou, Man Qi, Jiabo Shen, Yu Zhang are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. zhaoshuying@ise.neu.edu.cn
This project is funded by National Natural Science Foundation of China (No. 61973066), Major Science and Technology Projects of Liaoning Province(2021JH1/10400049).

main contributions of this work are summarized as follows:

- **Cross-Similarity Ranking Distillation.** We propose cross-similarity ranking distillation, where the student learns the teacher's self-similarity structures via cross-rankings, providing more discriminative supervision and enhancing feature representations.
- **Viewpoint Decoupling.** Building on cross-similarity ranking distillation, we introduce viewpoint decoupling, which splits ranking relations into intra-view, intra-to-cross-view, and cross-to-cross-view, enabling more precise modeling of cross-view dependencies and improving class compactness and separability.
- **Multi-Manifold Feature Distillation.** We distill features across spherical, Euclidean, and hyperbolic manifolds, collaboratively capturing directional, local, and global geometric properties to improve training stability and retrieval performance.
- **Extensive Experimental Validation.** Extensive experiments on University-1652 and SUES-200 demonstrate that our distilled student models significantly improve retrieval accuracy while maintaining low parameter counts and FLOPs, outperforming existing methods and substantially narrowing the gap to large teacher models.

II. Related Work

A. Cross-View Geo-Localization

The central challenge of CVGL is the drastic appearance variation caused by viewpoint differences, particularly in UAV-to-satellite (UAV–SAT) scenarios. The University-1652 dataset [1] first established a benchmark spanning UAV, satellite, and ground views, accelerating UAV–SAT research. Later datasets such as SUES-200 [14] and Game4Loc [15] introduced altitude, pose, and resolution variations, making the task closer to real-world conditions.

Early CVGL methods relied on CNNs to extract global or local features [16] [17]. Global descriptors miss fine details, while locals are sensitive to scale and viewpoint changes. More recently, Transformer architectures have shown strong ability in modeling long-range dependencies and cross-view alignment. For instance, L2LTR [5] leverages layer-to-layer attention for cross-view relations, while TransGeo [6] aligns ground and satellite images through a unified Transformer with attention-guided cropping. Despite accuracy gains, these methods often rely on heavy backbones (e.g., Swin [7]) or costly fusion strategies [8], hindering UAV deployment. Lightweight backbones (e.g., MobileNetV3 [18]) reduce cost but still require auxiliary modules to retain accuracy, increasing complexity. These trade-offs motivate knowledge distillation to transfer knowledge from large teachers to compact students.

B. Knowledge Distillation

Knowledge Distillation (KD) [11] was introduced to transfer teacher knowledge to lightweight students via soft labels. It has since evolved into feature-level distillation (FitNets [19]), structural distillation (FSP [20]), and relational distillation (RKD [12]), which preserves inter-sample geometry.

For retrieval and metric learning, ranking-based KD is dominant. DarkRank [21] transfers similarities via listwise rank matching, enabling students to inherit ranking behavior. In asymmetric retrieval, Contextual Similarity Distillation (CSD) [22] improves lightweight queries while retaining gallery encoders. In CVGL, GeoDistill [23] applies FoV-based self-distillation, while cross-modal tasks such as Distill-lVPR [24] highlight KD's effectiveness. Nevertheless, most retrieval KD methods do not explicitly model fine-grained teacher–student relations under large viewpoint shifts, nor disentangle intra- and cross-view dependencies. They typically operate in a single Euclidean manifold, leaving angular and curvature-aware discrepancies unresolved.

C. Multi-Manifold Spaces

Representation learning traditionally relies on Euclidean geometry, which preserves local distances. Recent work highlights the value of leveraging multiple manifolds to capture complementary properties. Euclidean space emphasizes local distances [25], spherical space enforces angular consistency [26], and hyperbolic space, with its negative curvature, provides a natural embedding for hierarchical or exponentially growing structures [27], while also better preserving global geometry. SPD manifolds [28], in contrast, utilize covariance features and Riemannian metrics for statistical alignment.

In cross-view retrieval and VPR, multi-manifold geometry has been incorporated into distillation and ranking. Distill-lVPR [24], for example, integrates Euclidean, spherical, and hyperbolic constraints for cross-modal distillation, demonstrating the benefit of cross-geometry consistency. However, these methods mainly focus on ranking alignment and often overlook explicit feature coupling with geometric regularization. In this context, our **Cross-Distill** jointly leverages spherical (angular consistency), Euclidean (local distance) , and hyperbolic (negative curvature) manifolds to align teacher–student features in UAV–SAT scenarios, ensuring cross-space consistency.

III. Method

A. Method Overview

We first build a teacher feature database, where global representations from the teacher network are stored offline. For each image in the current batch, teacher and student features are extracted (or teacher features retrieved from the database) and aligned via **Multi-Manifold Feature Distillation** across spherical, Euclidean, and hyperbolic spaces. Each teacher feature then serves as a query to retrieve top-k_1 UAV and top-k_2 satellite features, enabling computation of teacher–teacher and teacher–student similarities. We further apply **Viewpoint Decoupling** to split these similarities into intra-view, intra-to-cross-view, and cross-to-cross-view relations, from which a similarity difference matrix is derived capturing ranking deviations. Finally, **Cross-Similarity Ranking Distillation** is applied in the decoupled space to enforce local ranking consistency and global structural alignment for the student model, improving feature discriminability and cross-view retrieval. The overall is shown in Fig. 2.

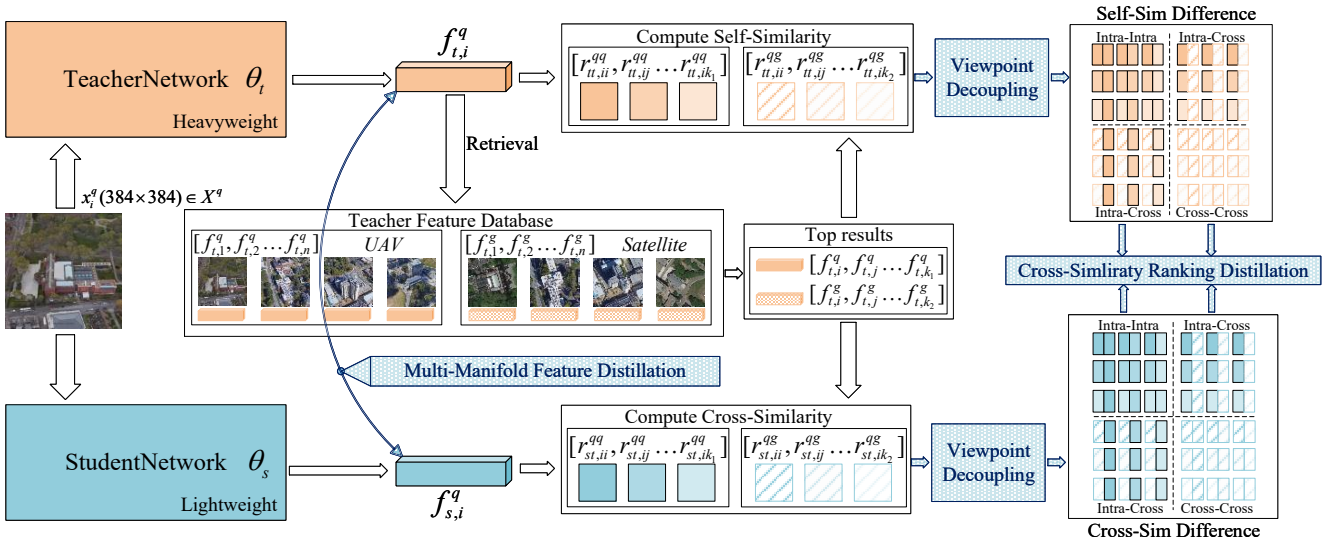

Fig. 2: Overview of the **Cross-Distill** framework. Multi-Manifold Feature Distillation aligns student features with the teacher, while Viewpoint Decoupling and Cross-Similarity Ranking Distillation enhance the preservation of discriminative structures.

B. Problem Formulation

Let $\theta_s(\cdot)$ and $\theta_t(\cdot)$ denote the student and teacher networks that map images into normalized d-dimensional embeddings. We denote the extracted student features as f_s and teacher features as f_t. Given a batch of query images $X^q = [x^q_1, \ldots, x^q_n]$ and gallery images $X^g = [x^g_1, \ldots, x^g_n]$, their representations are obtained as

$$\begin{aligned} f^q_{s,i} &= \theta_s(x^q_i), \quad f^q_{t,i} = \theta_t(x^q_i), \quad i = 1, 2, \ldots, n, \\ f^g_{s,j} &= \theta_s(x^g_j), \quad f^g_{t,j} = \theta_t(x^g_j), \quad j = 1, 2, \ldots, n, \end{aligned} \tag{1}$$

where f^q_s and f^g_s denote student features for query and gallery images, while f^q_t and f^g_t are the corresponding teacher features.

Let $\langle \cdot, \cdot \rangle$ denote cosine similarity. We define $r_{ss,ij}$, $r_{tt,ij}$, and $r_{st,ij}$ as the pairwise similarities for the student–student, teacher–teacher, and student–teacher embeddings, respectively, formulated as:

$$\begin{aligned} r_{ss,ij} &= \langle f_{s,i}, f_{s,j} \rangle, \\ r_{tt,ij} &= \langle f_{t,i}, f_{t,j} \rangle, \\ r_{st,ij} &= \langle f_{s,i}, f_{t,j} \rangle, \quad i, j = 1, 2, \ldots, n. \end{aligned} \tag{2}$$

C. Cross-Simliraty Ranking Distillation

In cross-view geo-localization, global feature similarity r^{qg}_{ij} is used to judge whether two images correspond to the same location. If the relation $r^{qg}_{ij} > r^{qg}_{il} > \cdots$ holds, the reference image x^g_j is considered the best match for query x^q_i. We thus expect the student to generate features and similarity relations consistent with the teacher. However, due to limited capacity, the student struggles to replicate the teacher's feature structure. We therefore reformulate the objective as consistency in similarity ranking:

$$[r_{tt,ij} > r_{tt,ik} > r_{tt,il} > \cdots] = [r_{ss,ij} > r_{ss,ik} > r_{ss,il} > \cdots]. \tag{3}$$

As shown in Fig. 3(a), while the student partially preserves the teacher's ranking, it fails to capture intra-class compactness and inter-class separability. Moreover, its early-stage features are weak and unstable, making self-ranking prone to noise and error amplification. To overcome this, we propose a cross-similarity ranking strategy, where the student imitates student-teacher similarity ordering instead of relying solely on student–student ordering (Fig. 3(b)):

$$[r_{tt,ij} > r_{tt,ik} > r_{tt,il} > \cdots] = [r_{st,ij} > r_{st,ik} > r_{st,il} > \cdots]. \tag{4}$$

As long as a student feature produces the same similarity ranking to all teacher features as its corresponding teacher feature, the student naturally inherits the structure of the teacher space. To formalize this intuition, we introduce teacher prototype $f_{t,c}$ for class c as the normalized sum of all teacher features in that class, and the angular separation $\Theta_{cc'}$ between two prototypes c and c' is given by

$$f_{t,c} = \frac{\sum_{j:y_j=c} f_{t,j}}{\left\| \sum_{j:y_j=c} f_{t,j} \right\|}, \quad \Theta_{cc'} = \arccos\langle f_{t,c}, f_{t,c'} \rangle. \tag{5}$$

If two student features $f_{s,i}, f_{s,j} \in c$ satisfy $r_{st,ic} \geq \tau^+$ and $r_{st,jc} \geq \tau^+$, then

$$\|f_{s,i} - f_{s,j}\| \leq \|f_{s,i} - f_{t,c}\| + \|f_{s,j} - f_{t,c}\| \leq 2\sqrt{2(1-\tau^+)}. \tag{6}$$

This result indicates that as the cross-similarity $\langle f_s, p_c \rangle$ increases, the upper bound of intra-class distances contracts significantly.

If $f_{s,i} \in c$ and $f_{s,j} \in c'$ $(c \neq c')$ both satisfy $r_{st,ic} \geq \tau^+$ and $r_{st,jc'} \geq \tau^+$, then

$$\angle(f_{s,i}, f_{s,j}) \geq \Theta_{cc'} - 2\arccos(\tau^+), \tag{7}$$

consequently, we obtain

$$\|f_{s,i} - f_{s,j}\| \geq 2\sin\left(\frac{\Theta_{cc'} - 2\arccos(\tau^+)}{2}\right). \tag{8}$$

This result indicates that as the cross-similarity $\langle f_s, p_c \rangle$ increases, the inter-class separability gradually improves.

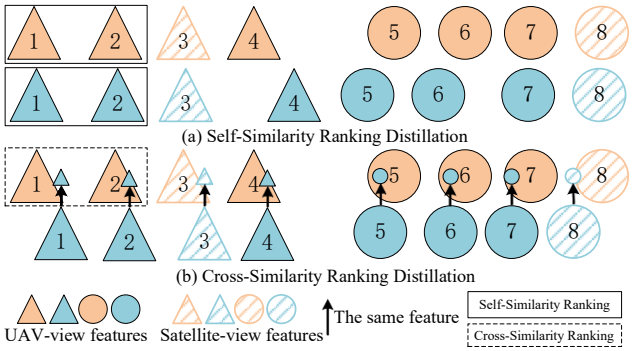

Fig. 3: Comparison of self-similarity and cross-similarity ranking distillation. Light Orange denotes teacher features, Light Blue denotes student features, and different shapes indicate different locations (categories). Shorter distances imply higher similarity.

Essentially, we realize cross-similarity ranking (CSR) distillation by constructing a semantically consistent reference space from teacher prototypes. CSR imposes explicit ranking constraints that stabilize training and improve cross-view discrimination. To achieve this, the student is trained to match the teacher's pairwise ordering so that retrieval rankings remain consistent at inference. We therefore introduce the strict ranking distillation loss L_{rank}:

$$L_{\text{rank}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\left(\mathcal{H}(r_{tt,ij}-r_{tt,ik})-\mathcal{H}(r_{st,ij}-r_{st,ik})\right)^2, \tag{9}$$

where $\mathcal{H}(\cdot)$ denotes the Heaviside step function.

However, directly optimizing the non-differentiable Heaviside step function is difficult. The constraint holds when $r_{tt,ij}-r_{tt,ik}$ and $r_{st,ij}-r_{st,ik}$ share the same sign. Thus, we rewrite it as Eq. (10) and further decompose the expression, where the objective can be satisfied once the second term approaches zero:

$$\frac{r_{st,ij}-r_{st,ik}}{r_{tt,ij}-r_{tt,ik}} = 1+\frac{(r_{st,ij}-r_{st,ik})-(r_{tt,ij}-r_{tt,ik})}{r_{tt,ij}-r_{tt,ik}} > 0. \tag{10}$$

To prevent instability caused by small denominators, we introduce a relaxation factor m. The final similarity ranking loss is:

$$L_{\text{rank}} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\left(\frac{(r_{st,ij}-r_{st,ik})-(r_{tt,ij}-r_{tt,ik})}{m+|r_{tt,ij}-r_{tt,ik}|}\right)^2\right)^{\frac{1}{2}}. \tag{11}$$

Building on this, inspired by [29], we further divide ranking consistency into easy and hard pairs, enabling finer-grained decomposition of L_{rank}.

$$L_{\text{rank}} = L_{\text{erank}} + L_{\text{hrank}}. \tag{12}$$

Specifically, the loss for easy pairs is defined as:

$$L_{\text{erank}} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathcal{H}\left(\frac{r_{st,ij}-r_{st,ik}}{r_{tt,ij}-r_{tt,ik}}\right)\times\left(\frac{(r_{st,ij}-r_{st,ik})-(r_{tt,ij}-r_{tt,ik})}{m+|r_{tt,ij}-r_{tt,ik}|}\right)^2\right)^{\frac{1}{2}}, \tag{13}$$

and the loss for hard pairs is defined as:

$$L_{\text{hrank}} = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n}\mathcal{H}\left(-\frac{r_{st,ij}-r_{st,ik}}{r_{tt,ij}-r_{tt,ik}}\right)\times\left(\frac{(r_{st,ij}-r_{st,ik})-(r_{tt,ij}-r_{tt,ik})}{m+|r_{tt,ij}-r_{tt,ik}|}\right)^2\right)^{\frac{1}{2}}. \tag{14}$$

Finally, we combine the two with fixed weighting to form the overall ranking loss:

$$L_{\text{rank}} = aL_{\text{erank}} + bL_{\text{hrank}}. \tag{15}$$

where a and b are tunable hyper-parameters that balance the contributions of easy and hard pairs.

D. Viewpoint Decoupling

As shown in Fig. 2, cross-view geo-localization involves UAV and satellite perspectives. The ranking relations in Eq. (4) can be categorized into intra-view, intra-to-cross-view, and cross-to-cross-view. Properly modeling these relations enhances the student model's intra-class compactness and inter-class separability.

a) Intra-view.

The intra-view consistency can be formulated as:

$$\begin{aligned}\left[r^{qq}_{tt,ij}>r^{qq}_{tt,ik}>r^{qq}_{tt,il}>\cdots\right] &= \left[r^{qq}_{st,ij}>r^{qq}_{st,ik}>r^{qq}_{st,il}>\cdots\right],\\ \left[r^{gg}_{tt,ij}>r^{gg}_{tt,ik}>r^{gg}_{tt,il}>\cdots\right] &= \left[r^{gg}_{st,ij}>r^{gg}_{st,ik}>r^{gg}_{st,il}>\cdots\right],\end{aligned} \tag{16}$$

where r^{qq} and r^{gg} denote similarities between UAV–UAV and SAT–SAT pairs. Formally, we define the intra-view cross-similarity ranking loss as:

$$L_{in-in} = \partial L_{rank}, \tag{17}$$

where ∂ is a weighting coefficient.

b) Intra-to-Cross-view.

The intra-to-cross-view ranking consistency can be expressed as:

$$\begin{aligned}\left[r^{qq}_{tt,ij}>r^{qg}_{tt,ik}>r^{qq}_{tt,il}>\cdots\right] &= \left[r^{qq}_{st,ij}>r^{qg}_{st,ik}>r^{qq}_{st,il}>\cdots\right],\\ \left[r^{gg}_{tt,ij}>r^{gq}_{tt,ik}>r^{gg}_{tt,il}>\cdots\right] &= \left[r^{gg}_{st,ij}>r^{gq}_{st,ik}>r^{gg}_{st,il}>\cdots\right],\end{aligned} \tag{18}$$

where r^{qg} and r^{gq} denote UAV–SAT similarities. Formally, we define the intra-to-cross-view cross-similarity ranking loss as:

$$L_{in-cross} = \beta L_{rank}, \tag{19}$$

where β is a weighting coefficient.

c) Cross-to-Cross-view.

The cross-to-cross-view consistency is formulated as:

$$\begin{aligned}\left[r^{qg}_{tt,ij}>r^{qg}_{tt,ik}>r^{qg}_{tt,il}>\cdots\right] &= \left[r^{qg}_{st,ij}>r^{qg}_{st,ik}>r^{qg}_{st,il}>\cdots\right],\\ \left[r^{gq}_{tt,ij}>r^{gq}_{tt,ik}>r^{gq}_{tt,il}>\cdots\right] &= \left[r^{gq}_{st,ij}>r^{gq}_{st,ik}>r^{gq}_{st,il}>\cdots\right].\end{aligned} \tag{20}$$

Formally, we define the cross-to-cross-view cross-similarity ranking loss as:

$$L_{cross-cross} = \gamma L_{rank}, \tag{21}$$

where γ is a weighting coefficient.

In summary, intra-view, intra-to-cross-view, and cross-to-cross-view ranking relations are disentangled to supervise the student network. This hierarchical decomposition enables the student to inherit intra-class compactness and inter-class separability, improving robustness in UAV–SAT retrieval.

E. Multi-Manifold Feature Distillation

Although cross-similarity ranking distillation enhances compactness and separability, it provides only relative supervision, leaving distribution shifts and scale mismatches. To address this, we introduce multi-manifold feature distillation: embeddings are simultaneously constrained on spherical, Euclidean, and hyperbolic manifolds, encouraging student features to approach the teacher space and exhibit alignment, thereby improving representation and stability. Concretely, we project features f_i onto the three manifolds, define the corresponding metrics, and optimize them jointly to achieve multi-level alignment and error contraction.

a) Spherical Space (Angular Consistency).

The spherical manifold constrains features on a high-dimensional unit hypersphere, focusing on angular information rather than magnitudes. For normalized features, the spherical (cosine) distance is defined as:

$$d_{\cos}(f_{s,i}, f_{t,i}) = \frac{\langle f_{s,i}, f_{t,i} \rangle}{\|f_{s,i}\| \, \|f_{t,i}\|}. \tag{22}$$

The corresponding spherical loss is given by:

$$L_f^{\cos} = \frac{1}{n} \sum_{i=1}^{n} \left(\left(d_{\cos}(f_{s,i}, f_{t,i}) - 1 \right)^2 \right)^{\frac{1}{2}}. \tag{23}$$

b) Euclidean Space (Local Distance).

Euclidean space remains the most common representation space, where the Euclidean distance effectively reflects local linear structures of features. The Euclidean distance is defined as:

$$d_{\text{euc}}(f_{s,i}, f_{t,i}) = \|f_{s,i} - f_{t,i}\|. \tag{24}$$

The corresponding Euclidean loss is:

$$L_f^{ecu} = \frac{1}{n} \sum_{i=1}^{n} d_{\text{euc}}(f_{s,i}, f_{t,i}). \tag{25}$$

c) Hyperbolic Space (Negative Curvature).

Unlike Euclidean space, hyperbolic space with negative curvature provides a more suitable geometry for modeling distances in high-dimensional embeddings, as it better preserves global structures. We adopt the Poincaré ball model for hyperbolic embeddings and alignment.

Given a scaling factor $c > 0$ (commonly set as $c = 1$), a normalized Euclidean feature f_i in the tangent space can be projected into the Poincaré ball as:

$$f_i^P = \tanh\left(\sqrt{c}\,\|f_i\|\right) \frac{f_i}{\sqrt{c}\,\|f_i\|}, \quad (f_i \neq 0). \tag{26}$$

For two feature vectors $f_{s,i}^P, f_{t,i}^P \in \mathbb{B}_c^n = \{x \in \mathbb{R}^n : \|x\| < c\}$, the hyperbolic distance is defined as:

$$d_{\text{hyp}}(f_{s,i}, f_{t,i}) = \frac{2}{\sqrt{c}} \tanh^{-1}\left(\sqrt{c}\,\| -f_{s,i}^P \oplus f_{t,i}^P \| \right), \tag{27}$$

where $\oplus$ denotes the Möbius addition, defined as:

$$f_{s,i}^P \oplus f_{t,i}^P = \frac{(1+2c\langle f_{s,i}^P, f_{t,i}^P \rangle + c\|f_{t,i}^P\|^2) f_{s,i}^P + (1-c\|f_{s,i}^P\|^2) f_{t,i}^P}{1 + 2c\langle f_{s,i}^P, f_{t,i}^P \rangle + c^2 \|f_{s,i}^P\|^2 \|f_{t,i}^P\|^2}. \tag{28}$$

The hyperbolic loss is then defined as:

$$L_f^{hyp} = \frac{1}{n} \sum_{i=1}^{n} \left(\left(d_{hyp}(f_{s,i}, f_{t,i}) \right)^2 \right)^{\frac{1}{2}}. \tag{29}$$

Finally, by integrating the alignment losses from the three manifold spaces with adaptive weights, the overall feature distillation loss is formulated as:

$$L_f = \lambda^{\cos} L_f^{\cos} + \lambda^{ecu} L_f^{ecu} + \lambda^{hyp} L_f^{hyp}, \tag{30}$$

where $\lambda^{\cos}$,λ^{ecu} and λ^{hyp} are balancing coefficients.

In addition, we incorporate the task-specific loss (cross-entropy and triplet loss) together with the three decoupled distillation losses. The final objective is therefore given by:

$$L_{loss} = L_{task} + L_{in-in} + L_{in-cross} + L_{cross-cross} + L_f. \tag{31}$$

IV. Experiments

A. Experimental Setup

Datasets. We evaluate the proposed method on two public cross-view geo-localization benchmarks:

a) University-1652 [1]: A large-scale CVGL benchmark covering 1,652 locations, with over 50k training images and additional drone-, satellite-, and ground-view splits.

b) SUES-200 [14]: A cross-view dataset with UAV images captured at multiple altitudes (150m–300m), designed to simulate altitude variations under real-world flight conditions.

Evaluation Metrics. We report Recall@K (R@K), Average Precision (AP), Params (M), and GFLOPs. R@K reflects retrieval accuracy, AP summarizes the precision–recall curve, Params (M) denotes model size, and GFLOPs indicate computational cost.

Implementation Details. All experiments are conducted in PyTorch on a single NVIDIA RTX 4090 GPU. The student model is a lightweight ConvNeXt-Tiny with an MLP head projecting features to 1024 dimensions. We use Adam with a learning rate of 0.001 on University-1652 and 0.0001 on SUES-200, a batch size of 8, a cosine scheduler, and train for 120 epochs. Following [29], we set $a = 2$, $b = 10$, and other hyperparameters as $\lambda^{\cos} = 170, \lambda^{\text{ecu}} = 10, \lambda^{\text{hyp}} = 10, \partial = 1.10, \beta = 1.20, \gamma = 1.00$. The top-$k_1$ and top-k_2 are determined by the number of UAV and satellite images per batch. JRN-Geo [8] and DAC [34] serve as teacher models, where the teacher feature database is built from all images in the current batch. Both teacher and student adopt the same augmentation factor k, originally proposed in JRN-Geo [8]. All images are resized to 384×384, and cosine similarity is used for UAV–satellite matching.

TABLE I

COMPARISON WITH STATE-OF-THE-ART CVGL METHODS ON THE SUES-200 DATASET. THROUGHOUT ALL TABLES, THE BEST, SECOND-BEST, AND THIRD-BEST RESULTS ARE HIGHLIGHTED IN RED, BLUE, AND PURPLE, RESPECTIVELY.

Method	Publication	Params (M)	GFLOPs	Drone → Satellite								Satellite → Drone							
				150m		200m		250m		300m		150m		200m		250m		300m	
				R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LPN [4]	TCSVT'22	62.39	36.78	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
CCR [30]	TCSVT'24	156.57	160.61	87.08	89.55	93.57	94.90	95.42	96.82	96.82	97.39	92.50	88.54	97.50	95.22	97.50	97.10	97.50	97.49
MCCG [9]	TCSVT'23	56.65	51.04	82.22	85.47	89.30	91.41	93.82	95.04	95.07	96.20	93.75	89.72	93.75	92.21	96.25	96.14	95.00	92.03
MFJR [31]	TGRS'24	>88.0	—	88.95	91.05	93.60	94.72	95.42	96.28	94.45	97.84	95.00	89.31	96.25	94.69	97.50	96.92	98.75	97.14
SRLN [32]	TGRS'24	193.03	—	89.90	91.90	94.32	95.65	95.92	96.79	96.37	97.21	93.75	93.01	97.50	95.08	97.50	96.52	97.50	96.71
Sample4Geo [33]	ICCV'23	87.57	90.24	92.60	94.00	97.38	97.81	98.28	98.64	99.18	99.36	97.50	93.63	98.75	96.70	98.75	98.28	98.75	98.05
DAC [34]	TCSVT'24	96.50	90.24	96.80	97.54	97.48	97.97	98.20	98.62	97.58	98.14	97.50	94.06	98.75	96.66	98.75	98.09	98.75	97.87
JRN(k=0) [8]	ICRA'25	191.83	92.81	86.15	89.02	93.80	97.12	97.12	97.77	96.22	97.10	97.50	91.09	98.75	93.85	98.75	96.77	98.75	97.92
JRN(k=4) [8]				96.47	97.26	98.60	98.92	99.28	99.45	99.10	99.33	98.75	96.05	98.75	98.02	98.75	99.06	98.75	98.97
Ours (k=0, w/o distill.)	—	31.43	13.09	81.82	85.28	85.15	88.03	89.97	92.02	94.45	95.61	92.50	82.81	95.00	84.73	96.25	92.13	97.50	96.22
Ours (k=0, DAC [34])				83.82	87.12	92.28	93.98	95.95	96.86	97.35	97.9	93.75	85.34	95.00	92.94	98.75	95.48	98.75	97.30
Ours (k=0, JRN [8])				88.67	91.00	85.38	88.45	92.75	94.16	96.40	97.19	95.00	87.25	96.25	89.24	97.50	94.54	98.75	96.84
Ours (k=4, JRN [8])				94.62	95.80	96.15	96.99	98.28	98.64	98.80	99.04	98.75	95.52	98.75	98.56	98.75	97.12	98.75	98.03

B. Comparison with State-of-the-Art CVGL Methods

TABLE II

COMPARISON WITH STATE-OF-THE-ART CVGL METHODS ON THE UNIVERSITY-1652 DATASET.

Method	Publication	Params	GFLOPs	Drone → Satellite		Satellite → Drone	
				R@1	AP	R@1	AP
LPN [4]	TCSVT'22	62.39	36.78	75.93	79.14	86.45	74.79
TransFG [10]	TGRS'24	>86.0	—	84.01	86.31	90.16	84.61
MCCG [9]	TCSVT'23	56.65	51.04	89.64	91.32	94.30	89.39
MFJR [31]	TGRS'24	>88.0	—	91.87	93.15	95.29	91.51
CCR [30]	TCSVT'24	156.57	160.61	92.54	93.78	95.15	91.80
Sample4Geo [33]	ICCV'23	87.57	90.24	92.65	93.81	95.14	91.39
SRLN [32]	TGRS'24	193.03	—	92.70	93.77	95.14	91.97
MEAN [35]	TGRS'25	36.50	26.18	93.55	94.53	96.01	92.08
DAC [34]	TCSVT'24	96.50	90.24	94.67	95.50	96.43	93.79
JRN(k=0) [8]	ICRA'25	191.83	92.81	94.32	95.29	96.15	93.81
JRN(k=4) [8]				95.13	95.85	96.72	94.93
Ours (0,w/o distill)	—	31.43	13.09	75.97	79.24	85.31	74.72
Ours (0,DAC [34])				88.69	90.38	92.58	87.21
Ours (0,JRN [8])				91.68	92.93	94.01	90.45
Ours (4,JRN [8])				94.43	95.33	95.72	93.47

Results on the University-1652 Dataset. Results on the University-1652 Dataset. We compare our method to existing approaches on both UAV→SAT and SAT→UAV tasks. As shown in Table II, using JRN-Geo [8] as the teacher model with k=4, our method achieves competitive retrieval accuracy while reducing parameters and computational cost. Compared to MEAN [35] with ConvNeXt-Tiny, our approach outperforms in R@K and AP, validating the effectiveness of our distillation strategy. Even with no augmentation (k=0), our model surpasses larger, more complex models, proving its robustness in improving lightweight models.

Results on the SUES-200 Dataset. We evaluate our model's performance with UAV images at varying altitudes. As shown in Table I, our method consistently improves lightweight models across altitudes from 150m to 300m. With JRN-Geo [8] and k=4, our model performs well at all altitudes, demonstrating robustness to altitude variation. Even without augmentation (k=0), our method outperforms more complex models, showing strong generalization even without additional data.

Inference Time on the RK3588. To further assess the practical performance of our model, we evaluate the inference time for a single image on the RK3588. Our method achieves an inference time of 62.02 ms per image, which is faster than DAC [34] (135.13 ms), Sample4Geo [33] (135.72 ms). These results demonstrate that our approach not only improves retrieval accuracy but also offers competitive inference efficiency on edge devices.

TABLE III

COMPARISON OF DIFFERENT DISTILLATION METHODS ON THE UNIVERSITY-1652 DATASET.

Method	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
Tea: JRN(k=0) [8]	94.32	95.29	96.15	93.81
Student w/o distill.	75.97	79.24	85.31	74.72
RKD [12]	77.55	80.59	88.02	78.69
PKT [13]	70.44	74.52	85.31	71.25
FitNet [19]	87.45	89.39	91.73	84.81
CCKD [36]	83.14	85.60	88.59	80.16
CSD [22]	80.54	82.37	88.73	76.93
RAML [37]	81.53	84.27	88.59	78.66
ROP [38]	76.54	79.88	87.16	76.44
Ours (k=0)	91.68	92.93	94.01	90.45

C. Comparison with Other Knowledge Distillation Methods

For fair comparison, all methods are re-implemented under a symmetric retrieval protocol, where both queries and gallery images are encoded by the same student encoder with identical preprocessing. As shown in Table III, our method achieves the best performance among existing distillation baselines. In particular, it substantially improves the student baseline and significantly narrows the gap to the teacher model, demonstrating the effectiveness and practicality of the proposed framework in cross-view geo-localization.

D. Ablation Studies

In the ablation studies, all models are trained with the same task loss L_{task}, and each experiment introduces only the specific component under investigation on top of this baseline. Unless otherwise specified, JRN-Geo [8] with $k = 0$ is adopted as the teacher.

TABLE IV
ABLATION STUDY ON DIFFERENT SIMILARITY RANKING STRATEGIES ON THE UNIVERSITY-1652 DATASET.

Method	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
w/o distill.	75.97	79.24	85.31	74.72
self-sim	79.73	82.46	88.02	78.48
cross-sim	**81.56**	**84.20**	**89.59**	**79.72**

Cross-Similarity Ranking Distillation. As shown in Table IV, cross-similarity ranking distillation consistently outperforms both the non-distilled baseline and self-similarity distillation.Constructing a cross-similarity interaction matrix provides more discriminative signals, enabling the student to better inherit the teacher's knowledge and achieve superior intra-class compactness and inter-class separability.

To further examine this effect, we remove task losses and compare self- and cross-similarity ranking. We use t-SNE to visualize the image features. As shown in Fig. 4, Student-B (cross-sim) forms tighter clusters and preserves the teacher's geometry, while Student-A (self-sim) shows overlap and centroid shifts. These results demonstrate that cross-similarity ranking distillation better retains the teacher's discriminative structure and improves inter-class separability.

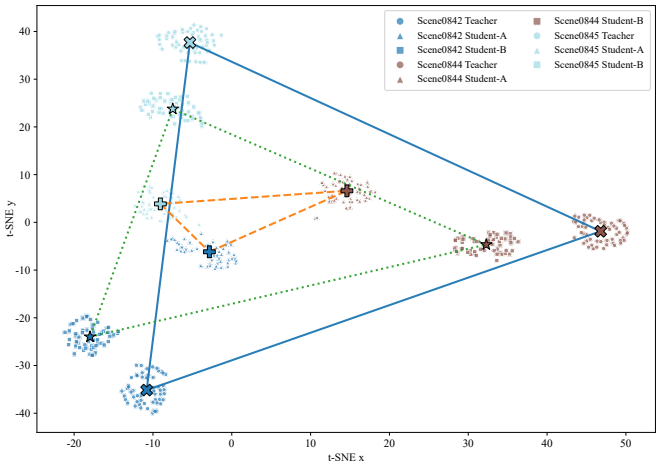

Fig. 4: t-SNE on University-1652: Teacher-JRN [8] vs Student-A (Self-Sim) vs Student-B (Cross-Sim), with colors representing different scenes and shapes distinguishing the models.

TABLE V
ABLATION STUDY ON DIFFERENT WEIGHTS OF $\lambda^{\cos}$, λ^{ecu}, AND λ^{hyp} ON THE UNIVERSITY-1652 DATASET.

$\lambda^{\cos}$	λ^{ecu}	λ^{hyp}	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
—	—	—	75.97	79.24	85.31	74.72
200	—	—	90.45	91.89	93.30	88.89
—	200	—	90.41	91.88	93.72	88.73
—	—	200	90.33	91.82	93.01	88.86
200	200	200	90.44	91.90	92.87	88.81
200	100	100	91.06	92.47	93.87	88.87
170	10	10	**91.18**	**92.55**	**94.01**	**89.66**

Multi-Manifold Feature Distillation. As shown in Table V, spherical manifold distillation outperforms Euclidean and hyperbolic spaces, consistent with cosine-based retrieval. Excessive emphasis on Euclidean or hyperbolic weights degrades performance, whereas balanced weighting across all three manifolds yields further gains, confirming the effectiveness of multi-manifold optimization.

TABLE VI
ABLATION STUDY ON DIFFERENT WEIGHTS OF ∂, β, AND γ ON THE UNIVERSITY-1652 DATASET.

∂	β	γ	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
—	—	—	75.97	79.24	85.31	74.72
1.00	—	—	77.22	80.30	87.73	76.64
—	1.00	—	78.70	81.80	88.73	78.75
—	—	1.00	76.70	79.91	87.30	76.37
1.00	1.00	1.00	81.56	84.20	89.59	79.72
1.10	1.20	1.00	**82.37**	**85.04**	**90.16**	**82.94**

Viewpoint Decoupling. As shown in Table VI, the Intra-view and Intra-to-Cross-view relations play a more prominent role. Assigning them higher weights leads to further improvements in retrieval performance compared with the non-decoupled setting.

E. Impact of Teacher Model Performance on Student Models

TABLE VII
COMPARISON OF TEACHER MODELS ON FEATURE DISTRIBUTION INDICATORS ON THE UNIVERSITY-1652 DATASET.

Teacher	Intra-class Variance	Inter-class Distance	Centroid Gap	ρ	Silhouette (Student)
JRN [8]	0.1821	**0.9690**	**58.49**	0.9623	0.5944
DAC [34]	**0.0858**	0.7456	61.63	**0.9670**	**0.6875**

To further analyze the impact of teacher model selection on student performance, we compare JRN-Geo [8] and DAC [34] on the University-1652 dataset (three randomly sampled categories). As reported in Table VII, we evaluate five indicators: Intra-class Variance, Inter-class Distance, Teacher–Student Centroid Gap, Silhouette Score (student), and Spearman's ρ. Results show DAC [34] achieves lower intra-class variance and higher silhouette scores, guiding students toward more compact features. In contrast, JRN-Geo [8] yields larger inter-class distances, highlighting stronger discriminative power and greater gains. Both student models maintain high ranking consistency ($\rho > 0.96$), confirming supervisory stability.

V. CONCLUSION

In this work, we introduced **Cross-Distill**, a knowledge distillation framework tailored for cross-view geo-localization. Our approach jointly leverages cross-similarity ranking distillation to transfer structural relations, viewpoint decoupling to explicitly model intra- and cross-view dependencies, and multi-manifold feature distillation to align features across spherical, Euclidean, and hyperbolic spaces. Through this integrated design, lightweight student models are endowed with stronger discriminative ability and improved stability. Extensive experiments on University-1652 and SUES-200 demonstrate that **Cross-Distill** achieves state-of-the-art performance among evaluated distillation baselines, while producing student models with substantially lower computational cost and parameter count compared to existing cross-view geo-localization methods, while still maintaining high accuracy.

REFERENCES

[1] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1395–1403. 1, 2, 5

[2] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616. 1

[3] J. Lin, Z. Zheng, Z. Zhong, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Joint representation learning and keypoint detection for cross-view geo-localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 3780–3792, 2022. 1

[4] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 867–879, 2021. 1, 6

[5] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 009–29 020, 2021. 1, 2

[6] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1162–1171. 1, 2

[7] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019. 1, 2

[8] H. Zhou, Y. Zhang, T. Huang, F. Ge, M. Qi, X. Zhang, and Y. Zhang, "Jrn-geo: A joint perception network based on rgb and normal images for cross-view geo-localization," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 3662–3668. 1, 2, 5, 6, 7

[9] T. Shen, Y. Wei, L. Kang, S. Wan, and Y.-H. Yang, "Mccg: A convnext-based multiple-classifier method for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1456–1468, 2023. 1, 6

[10] H. Zhao, K. Ren, T. Yue, C. Zhang, and S. Yuan, "Transfg: A cross-view geo-localization of satellite and uavs imagery pipeline using transformer-based feature aggregation and gradient guidance," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–12, 2024. 1, 6

[11] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015. 1, 2

[12] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976. 1, 2, 6

[13] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Transactions on Neural Networks and learning systems*, vol. 32, no. 5, pp. 2030–2039, 2020. 1, 6

[14] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu, "Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4825–4839, 2023. 2, 5

[15] Y. Ji, B. He, Z. Tan, and L. Wu, "Game4loc: A uav geo-localization benchmark from game data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3913–3921. 2

[16] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *European conference on computer vision*. Springer, 2016, pp. 494–509. 2

[17] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267. 2

[18] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324. 2

[19] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets. arxiv 2014," *arXiv preprint arXiv:1412.6550*, 2014. 2, 6

[20] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4133–4141. 2

[21] Y. Chen, N. Wang, and Z. Zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018. 2

[22] H. Wu, M. Wang, W. Zhou, H. Li, and Q. Tian, "Contextual similarity distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 9489–9498. 2, 6

[23] S. Tong, Z. Xia, A. Alahi, X. He, and Y. Shi, "Geodistill: Geometry-guided self-distillation for weakly supervised cross-view localization," *arXiv preprint arXiv:2507.10935*, 2025. 2

[24] S. Wang, R. She, Q. Kang, X. Jian, K. Zhao, Y. Song, and W. P. Tay, "Distilvpr: Cross-modal knowledge distillation for visual place recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 9, 2024, pp. 10 377–10 385. 2

[25] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742. 2

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699. 2

[27] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," *Advances in neural information processing systems*, vol. 30, 2017. 2

[28] M. T. Harandi, M. Salzmann, and R. Hartley, "From manifold to manifold: Geometry-aware dimensionality reduction for spd matrices," in *European conference on computer vision*. Springer, 2014, pp. 17–32. 2

[29] Y. Xie, Y. Lin, W. Cai, X. Xu, H. Zhang, Y. Du, and S. He, "D3still: Decoupled differential distillation for asymmetric image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 181–17 190. 4, 5

[30] H. Du, J. He, and Y. Zhao, "Ccr: A counterfactual causal reasoning-based method for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6

[31] F. Ge, Y. Zhang, L. Wang, W. Liu, Y. Liu, S. Coleman, and D. Kerr, "Multilevel feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization," *IEEE transactions on geoscience and remote sensing*, vol. 62, pp. 1–15, 2024. 6

[32] H. Lv, H. Zhu, R. Zhu, F. Wu, C. Wang, M. Cai, and K. Zhang, "Direction-guided multiscale feature fusion network for geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–13, 2024. 6

[33] F. Deuser, K. Habel, and N. Oswald, "Sample4geo: Hard negative sampling for cross-view geo-localisation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 847–16 856. 6

[34] P. Xia, Y. Wan, Z. Zheng, Y. Zhang, and J. Deng, "Enhancing cross-view geo-localization with domain alignment and scene consistency," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 5, 6, 7

[35] Z. Chen, Z.-X. Yang, and H.-J. Rong, "Multi-level embedding and alignment network with consistency and invariance learning for cross-view geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 6

[36] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation congruence for knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5007–5016. 6

[37] P. Suma and G. Tolias, "Large-to-small image resolution asymmetry in deep metric learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1451–1460. 6

[38] H. Wu, M. Wang, W. Zhou, and H. Li, "A general rank preserving framework for asymmetric image retrieval," in *The Eleventh International Conference on Learning Representations*, 2023. 6