CRAD: CONFUSED RELATION-AWARE DISTILLATION FOR EFFICIENT CROSS-VIEW GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-view geo-localization aims to match ground-level images with satellite-view counterparts and is a core capability for autonomous navigation. However, state-of-the-art models are often computationally expensive, motivating the use of knowledge distillation for efficient deployment. Similarity-distribution distillation has been explored in deep metric learning, but compact cross-view geo-localization requires transferring relations between satellite queries and ground candidates while learning from informative negatives. We propose **Confused Relation-Aware Distillation (CRAD)**, which distills a pretrained, frozen teacher into a compact student. CRAD aligns cross-view candidate distributions and jointly trains on student–student and teacher–student satellite–ground correspondences. To further strengthen supervision, we adopt Teacher-Guided confused Sampling, which constructs a shared hard-sample pool by jointly considering samples that are challenging for both the teacher and the student, allowing the teacher to actively select confused examples to guide the student's training. Extensive experiments on multiple benchmarks demonstrate the effectiveness of CRAD. On the challenging VIGOR-Same benchmark, CRAD-Tiny (28.6M) achieves 75.16% Recall@1, while the lightweight CRAD-Nano (15.0M) still delivers competitive performance. These results indicate that, for metric learning–based geo-localization, distilling how to compare can be more effective than distilling what to represent.

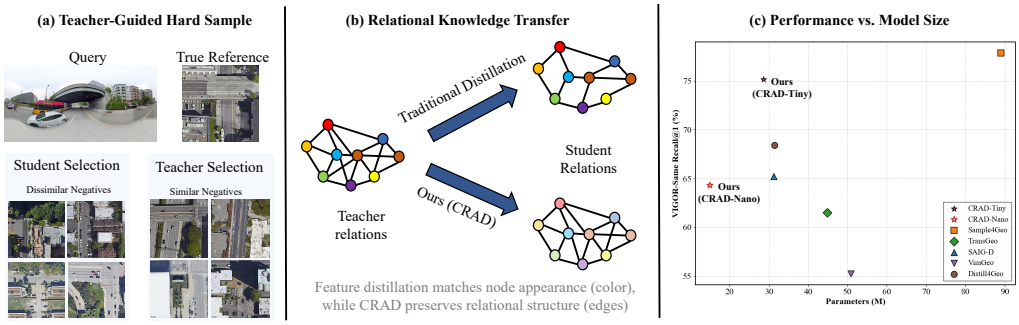


Figure 1: **Motivation and overview of CRAD.** (a) Teacher-guided mining exposes the student to truly confusing negatives missed by student-only sampling. (b) CRAD distills relational similarity structures rather than point-wise features. (c) This yields better Recall@1–parameter trade-offs compared to conventional distillation.

1 INTRODUCTION

Cross-view geo-localization retrieves a location across street-level and satellite viewpoints, supporting navigation and robotics (Viswanathan et al., 2014; Betke & Gurvits, 1997; Mithun et al., 2023). The viewpoint gap makes accurate retrieval difficult, while high-capacity contrastive models remain costly for edge deployment (Deuser et al., 2023; Zhang et al., 2024; Yang et al., 2021).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Knowledge distillation (KD) can compress these models (Hinton et al., 2015), but feature regression and embedding-distribution matching reduce Recall@1 in our CVACT experiments. Retrieval depends on relative candidate similarities: a teacher must separate a true match from visually plausible negatives, not reproduce a particular feature coordinate system. Figure 1 illustrates this distinction.

Teacher rankings also identify hard negatives that the student may miss early in training. These candidates are close to the true reference in the teacher's metric space, so distinguishing them supplies more targeted supervision than repeatedly sampling easy negatives. This motivates transferring both the teacher's candidate ordering and its choice of confusing examples.

We propose **Confused Relation-Aware Distillation (CRAD)** to transfer the teacher's ranking behavior. A cross-view contrastive term aligns matching locations, while relational structure distillation matches teacher and student similarity distributions over candidates. Teacher-Guided confused Sample mining (TGS) combines teacher- and student-mined negatives, emphasizing cases that challenge the student. Together, these components preserve discriminative relations in a smaller model.

Across CVUSA, CVACT, and VIGOR, CRAD improves compact students over the evaluated KD baselines. Its contributions are a cross-view teacher–student objective combining candidate-relation distillation with same-location correspondence supervision, a hard-negative sampling strategy that fuses teacher and student rankings, and an evaluation of these components across same- and cross-area settings.

2 Related Works

2.1 Cross-View Geo-Localization

Cross-view methods match ground images with geo-referenced overhead views (Zhu et al., 2021a; Li et al., 2024a; Xia et al., 2024; Zhang et al., 2024). CVUSA, CVACT, and VIGOR provide increasingly demanding benchmarks (Workman et al., 2015; Liu & Li, 2019; Zhu et al., 2021b). Recent learned embeddings improve recall, but larger backbones increase inference cost (Deuser et al., 2023; Mi et al., 2024; Zahid & Taj, 2025).

2.2 Knowledge Distillation

Logit and feature distillation transfer predictions or representations (Hinton et al., 2015; Romero et al., 2015; Chen et al., 2021). RKD transfers distances and angles (Park et al., 2019), while CRD uses contrastive supervision (Tian et al., 2020). Most relevant to ours, S2SD matches row-wise batch similarity distributions between jointly trained high-dimensional auxiliary embeddings and a lower-dimensional embedding on a shared backbone (Roth et al., 2021). CRAD uses this form of relational supervision for a different relation: satellite queries versus ground candidates, computed by a frozen teacher and a compact student. It further couples the distillation loss with cross-view teacher–student positives and negatives mined from both networks.

2.3 Hard Sample Mining in Metric Learning

Hard negatives sharpen metric learning but can destabilize optimization when selected from immature student embeddings (Kim et al., 2020; Song et al., 2016; Wu et al., 2017; Chen et al., 2018). Cross-view systems such as Sample4Geo mine nearest negatives dynamically (Deuser et al., 2023; Wu et al., 2024; Zahid & Taj, 2025). TGS adds the teacher's rankings to this pool, so the student trains on informative negatives even before its own rankings become reliable.

3 Method

3.1 Teacher and Baseline: Sample4Geo Framework

Our approach builds upon the Sample4Geo framework (Deuser et al., 2023), a strong baseline for cross-view geo-localization. We summarize its key components, which also serve as the foundation of our distillation framework.

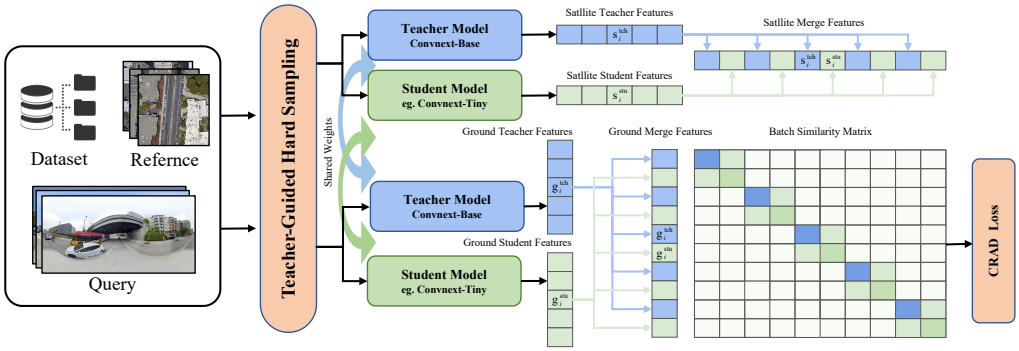


Figure 2: **Overview of CRAD.** A batch of satellite and street-view image pairs is processed by a frozen teacher and a trainable student, which extract cross-view features in parallel. Teacher-guided hard sample mining selects confusing negatives using teacher rankings; these are injected into the batch to strengthen supervision. For each view, features from both networks are used to construct a batch-wise similarity matrix. The proposed loss $\mathcal{L}_{\text{CRAD}}$ then transfers relational knowledge from teacher to student by combining (i) a multi-positive contrastive term that aligns cross-view correspondences across networks, and (ii) a relational structure distillation term that matches row-wise similarity distributions.

Architecture. Sample4Geo adopts a Siamese architecture with shared weights across satellite and street-view branches. A single ConvNeXt-Base encoder extracts feature maps from both modalities, followed by global average pooling and LayerNorm to produce a normalized global descriptor $\mathbf{f} \in \mathbb{R}^d$.

Training Objective. Given a batch of N paired samples $\{(S_i, G_i)\}_{i=1}^{N}$, the model is optimized using InfoNCE loss over both retrieval directions:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(\mathbf{f}_{S_i},\mathbf{f}_{G_i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{f}_{S_i},\mathbf{f}_{G_j})/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{f}_{G_i},\mathbf{f}_{S_i})/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{f}_{G_i},\mathbf{f}_{S_j})/\tau)}\right]. \tag{1}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and τ is a temperature parameter.

Dynamic Similarity Sampling (DSS). To emphasize hard negatives, Sample4Geo employs (DSS). At regular intervals, the current model is used to retrieve top-K most similar negative samples for each query, which are stored in a candidate pool. Subsequent training batches combine these hard negatives with randomly sampled ones, enabling continual refinement of the decision boundary.

3.2 Framework of Confused Relation-Aware Distillation (CRAD)

Retrieval optimizes candidate rankings rather than absolute descriptor coordinates. In particular, orthogonal transformations preserve cosine similarities while changing feature values; point-wise regression can therefore penalize equally useful embeddings. CRAD transfers the teacher's comparative knowledge through the batch-wise similarity distributions that govern InfoNCE training and retrieval.

For a query, the softmax-normalized similarity row determines how the contrastive objective weighs its positive and competing candidates. Matching this row allows the student to learn the teacher's relative preferences without forcing identical feature values. The distinction matters most when several negatives resemble the correct match and small ranking changes determine retrieval success.

For each batch of N matched satellite–ground pairs, the frozen teacher and trainable student extract descriptors in both views. A linear student projection matches the teacher descriptor dimension. Their pairwise similarities define the relations used by two complementary terms: $\mathcal{L}_{\text{CR-InfoNCE}}$ aligns cross-view positives across networks, and $\mathcal{L}_{\text{RSD}}$ matches teacher and student candidate distributions.

The overall loss is

$$\mathcal{L}_{\text{CRAD}} = \lambda \mathcal{L}_{\text{CR-InfoNCE}} + (1-\lambda)\mathcal{L}_{\text{RSD}}, \tag{2}$$

with $\lambda = 0.5$. TGS supplies a shared pool of teacher- and student-mined hard negatives for these objectives; further settings are in the supplementary material.

3.3 Confused Relation-Aware Distillation Loss (CRAD Loss)

3.3.1 Cross-View Multi-Positive Contrastive Learning (CR-InfoNCE Loss).

To robustly align representations across both views and networks, we design a cross-view multi-positive contrastive objective, denoted as $\mathcal{L}_{\text{CR-InfoNCE}}$. Different from conventional InfoNCE formulations that rely on a single positive pair, our loss introduces multiple positive associations for each anchor, covering student–teacher, student–student, and cross-view correspondences.

Specifically, given a mini-batch of N paired satellite and ground samples, we optimize three bidirectional cross-view relations at the same geographic location. For a student satellite anchor $\mathbf{s}_i^{\text{stu}}$, the corresponding student and teacher ground descriptors, $\mathbf{g}_i^{\text{stu}}$ and $\mathbf{g}_i^{\text{tch}}$, form positives in their respective loss terms. Symmetrically, a student ground anchor $\mathbf{g}_i^{\text{stu}}$ is paired with $\mathbf{s}_i^{\text{stu}}$ or $\mathbf{s}_i^{\text{tch}}$. Descriptors indexed by $j \neq i$ are negatives. Thus, positives are defined by geographic correspondence across views and networks rather than by an ambiguous notion of a "counterpart."

Let $\mathbf{g}_i^{\text{tch}}$ and $\mathbf{g}_i^{\text{stu}}$ denote the teacher and student global descriptors of the i-th ground image, and let $\mathbf{s}_i^{\text{tch}}$ and $\mathbf{s}_i^{\text{stu}}$ denote the corresponding teacher and student global descriptors of the satellite image. The temperature is τ, and the batch size is N.

For compactness, let $\ell(\mathbf{a}, \mathbf{b})$ denote the bidirectional InfoNCE operator in Eq. (1), applied to two descriptor sets. The three cross-view terms and their combination are

$$\begin{aligned} \mathcal{L}_{\text{SS}} &= \ell(\mathbf{s}^{\text{stu}}, \mathbf{g}^{\text{stu}}), \\ \mathcal{L}_{\text{TS}} &= \ell(\mathbf{s}^{\text{tch}}, \mathbf{g}^{\text{stu}}), \\ \mathcal{L}_{\text{ST}} &= \ell(\mathbf{s}^{\text{stu}}, \mathbf{g}^{\text{tch}}), \\ \mathcal{L}_{\text{CR-InfoNCE}} &= w_{\text{ss}}\mathcal{L}_{\text{SS}} + w_{\text{ts}}\mathcal{L}_{\text{TS}} + w_{\text{st}}\mathcal{L}_{\text{ST}}. \end{aligned} \tag{3}$$

The loss consists of three bidirectional InfoNCE terms. $\mathcal{L}_{\text{SS}}$ is the student–student cross-view term between satellite and ground descriptors. $\mathcal{L}_{\text{TS}}$ aligns teacher-satellite descriptors with student-ground descriptors, while $\mathcal{L}_{\text{ST}}$ aligns student-satellite descriptors with teacher-ground descriptors. Together, these terms transfer cross-view correspondence across both student–student and teacher–student relations.

The weights w_{ss}, w_{ts}, and w_{st} balance the contribution of each loss term. By adjusting these hyperparameters, one can control the relative importance of, student-student and teacher-student in different views. In our experiments we set all weights to 0.33 to give equal emphasis to each component, but they can be tuned according to the specific task or data characteristics. Detailed hyperparameter experiments are provided in the supplementary material.

3.3.2 Relational Structure Distillation (RSD Loss).

Following similarity-distribution distillation in deep metric learning (Roth et al., 2021), we align softmax-normalized similarity rows. Unlike the within-batch image relations used by S2SD, each row here represents a satellite query's distribution over ground-view candidates, including its paired location and competing locations. The frozen teacher supplies this cross-view distribution, while CR-InfoNCE explicitly supervises the paired correspondences across the teacher and student networks.

Based on the same batch, we construct two cross-view similarity matrices:

$$\mathbf{A}_{ij}^{T} = \text{sim}\big(\mathbf{s}_i^{\text{tch}}, \mathbf{g}_j^{\text{tch}}\big), \qquad \mathbf{A}_{ij}^{S} = \text{sim}\big(\mathbf{s}_i^{\text{stu}}, \mathbf{g}_j^{\text{stu}}\big), \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes a similarity function (cosine similarity).

Each row $\mathbf{A}^{(\cdot)}i$: defines the relative ordering of all in-batch ground candidates for the i-th satellite query. To compare teacher and student behaviors in a scale-controlled manner, we convert logits into row-wise similarity distributions using a distillation temperature τ_d:

$$\mathbf{p}_i^T = \mathrm{softmax}\big(\mathbf{A}_{i:}^T/\tau_d\big)\,, \quad \mathbf{p}_i^S = \mathrm{softmax}\big(\mathbf{A}_{i:}^S/\tau_d\big)\,. \tag{5}$$

Here, τ_d controls distribution sharpness: smaller τ_d emphasizes the highest-scoring (hard) candidates, while larger τ_d retains more global ranking information across the batch. We align the teacher and student distributions by minimizing a symmetric KL divergence averaged over all queries:

$$\mathcal{L}_{\mathrm{RSD}} = \frac{1}{2N}\sum_{i=1}^{N}\Big[\mathrm{KL}(\mathbf{p}_i^T\|\mathbf{p}_i^S) + \mathrm{KL}(\mathbf{p}_i^S\|\mathbf{p}_i^T)\Big]. \tag{6}$$

Expanding $\mathrm{KL}(\mathbf{p}_i^T\|\mathbf{p}_i^S) = \sum_{j=1}^{N} p_{ij}^T \log \frac{p_{ij}^T}{p_{ij}^S}$ shows that the loss penalizes mismatches in teacher-weighted comparisons: street candidates that the teacher considers highly similar (large p_{ij}^T) contribute more to the gradient, thus concentrating supervision near the decision boundary. Importantly, this objective supervises the student at the *relation level*—matching the teacher's induced ranking distribution for each query—and is therefore directly compatible with retrieval training, where optimization and evaluation are both driven by relative similarities rather than absolute feature coordinates.

Overall, CRAD integrates a cross-view multi-positive contrastive objective that aligns satellite–ground representations across student and teacher networks, together with a relational structure distillation term that transfers the similarity geometry learned by the teacher to the student.

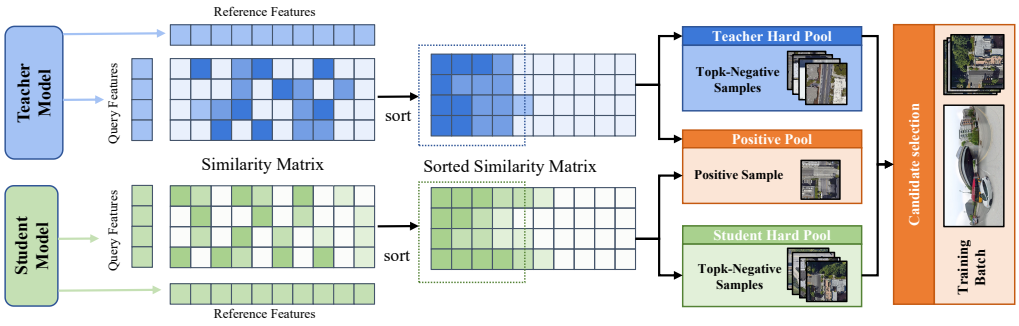


Figure 3: **Overview of Teacher-Guided hard Sample mining (TGS) in CRAD.** For each query, teacher and student models independently compute similarity matrices with respect to reference features and produce ranked candidate lists. Top-ranked negatives from the teacher and student perspectives are collected into separate hard pools, while the ground-truth match is retained in a positive pool. These candidate pools are subsequently fused under a controlled selection rule to form batch-level training samples, enabling structured exposure to informative hard negatives during optimization.

3.4 Teacher-Guided Hard Sample Mining (TGS)

Motivation. Relation-level distillation is highly sensitive to the quality of negative samples. Existing hard negative mining methods (e.g., DSS) rely solely on the student's similarity structure, which can be unreliable in early training and may amplify immature representations. We therefore introduce a teacher-guided mining strategy that injects *teacher-verified confusing samples*—negatives ranked highly by the teacher—into training, forcing the student to learn from the teacher's more discriminative geometry.

Periodic Dual-Model Inference. Following the efficiency principle of DSS, we refresh the mining cache every $e = 4$ epochs. At each refresh, we run one inference pass to extract global descriptors for *both* teacher and student, compute cosine similarities, and build two neighbor lists for each query:

$$\mathcal{P}_i^S = \text{Top-}K\big(\cos(\mathbf{f}_i^S, \mathbf{f}_j^S)\big), \quad \mathcal{P}_i^T = \text{Top-}K\big(\cos(\mathbf{f}_i^T, \mathbf{f}_j^T)\big), \tag{7}$$

where ground-truth matches are excluded from the negative candidates.

Dual-Perspective Hard Negative Construction. For each query i, we first extract student- and teacher-specific hard-negative lists:

$$\mathcal{N}_i^S = \text{Top-}K_s(\mathcal{P}_i^S) \setminus \text{pos}(i), \quad \mathcal{N}_i^T = \text{Top-}K_t(\mathcal{P}_i^T) \setminus \text{pos}(i). \tag{8}$$

Instead of directly taking their union, we adopt a *ratio-controlled fusion strategy* to form the final hard-negative set $\mathcal{N}_i$. Specifically, given a target neighbor budget K and a mixing ratio $\alpha \in (0, 1)$, we first select the top $\lfloor \alpha K \rfloor$ samples from the teacher-mined list $\mathcal{N}_i^T$, prioritizing teacher-verified confusing cases. The remaining slots are then filled by traversing $\mathcal{N}_i^S$ in descending similarity order, while explicitly removing duplicates. If the budget is still not met, additional samples are taken from the remaining teacher list. This process yields an ordered, duplicate-free hard-negative set:

$$\mathcal{N}_i = \text{Fuse}\big(\mathcal{N}_i^T, \mathcal{N}_i^S; K, \alpha\big), \tag{9}$$

where α controls the relative contribution of teacher- versus student-mined negatives. The resulting hard negatives $\mathcal{N}_i$ are cached and reused to construct training batches for the next e epochs, serving as supervision for both contrastive learning and relation-aware distillation.

By prioritizing teacher-mined hard negatives while retaining complementary student-induced ones, our strategy stabilizes early training and remains adaptive as the student evolves. As a result, relational distillation focuses on the most confusing teacher-identified samples, sharpening the student's decision boundary where supervision is most effective.

4 Experiments

4.1 Experimental Details

We detail our model configurations, training procedures, and implementation setup as follows. Our hyperparameter experiments and related visualization experiments can be found in the supplementary materials.

Models: We utilize the 88M-parameter ConvNeXt-B architecture from Sample4Geo (Deuser et al., 2023) as our teacher model. The primary student model is a 28M-parameter ConvNeXt-Tiny, chosen for its architectural homology and shared pre-training with the teacher, indicating strong performance potential.

Training Recipe: Models are trained for 40 epochs using the AdamW optimizer (Loshchilov & Hutter, 2019) with an initial learning rate of 0.001, a batch size of 128, and a cosine learning rate scheduler with a 1-epoch warm-up. For the InfoNCE loss, we use a label smoothing factor of 0.1 and treat the temperature parameter τ as learnable. Select up to 64 neighbors from a pool of 128 candidates.

Data Augmentation: Following Sample4Geo (Deuser et al., 2023), we apply synchronized horizontal flipping and rotation to satellite images, with corresponding shifts to street-view images to preserve orientation. To improve generalization and prevent over-reliance on specific regions, we further employ grid/coarse dropout and color jittering.

Implementation: All experiments are conducted on NVIDIA RTX 4090 GPUs, using 2-4 GPUs depending on model size and input resolution.

4.2 Datasets and Evaluation Metrics

4.2.1 CVUSA and CVACT

CVUSA (Workman et al., 2015) is one of the first cross-view datasets, with 35,532/8,884 image pairs for training/evaluation. Images are aligned such that geographic north is at the top-center of satellite views (750×750) and street views are warped accordingly (224×1232). CVACT (Liu & Li, 2019) shares the same training/validation splits but adds a test set of 92,802 images from Canberra, with higher resolutions (1200×1200 for satellite, 832×1664 for street views). Both datasets use Recall@k (k=1,5,10) and R@1% as evaluation metrics. To assess model efficiency in our distillation work, we also report the number of parameters. We evaluate on the more challenging CVACT test set

Table 1: Comparison of state-of-the-art methods across five benchmarks. **Params** indicates the number of parameters (in millions). Each dataset reports Recall@1 (R@1) and Recall@5 (R@5). Best results are highlighted in **bold**, and second-best results are underlined. Sample4Geo (marked with †) is used as our teacher model.

Model	Params	CVUSA		CVACT-Val		CVACT-Test		Vigor-Same		Vigor-Cross	
	(M)	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
L2LTR (Yang et al., 2021)	195.9	94.05	98.27	84.89	94.59	-	-	-	-	-	-
TransGeo (Zhu et al., 2022)	44.8	94.08	98.36	84.95	94.14	-	-	61.48	87.54	18.99	38.24
GeoDTR (Zhang et al., 2023)	48.5	95.43	98.86	86.21	96.44	64.52	88.59	-	-	-	-
SAIG-D (Zhu et al., 2023)	31.2	96.34	99.10	89.06	96.11	67.49	89.39	65.23	88.08	33.05	55.94
PaSS-KD (Li et al., 2024b)	401.0	94.09	98.42	87.20	94.30	66.81	88.03	52.90	76.60	21.00	39.70
ConGeo (Mi et al., 2024)	89.0	98.27	99.59	90.12	95.69	71.67	91.61	-	-	-	-
VimGeo (Huang et al., 2025)	50.9	96.19	98.62	87.62	94.88	**81.69**	**92.42**	55.24	80.75	19.31	37.50
Distill4Geo (Zahid & Taj, 2025)	31.4	97.92	99.58	90.04	96.40	71.07	91.66	68.39	89.28	59.93	83.09
Sample4Geo (Deuser et al., 2023)†	89.0	**98.68**	**99.68**	**90.81**	**97.64**	71.51	**92.42**	**77.86**	**95.66**	**61.70**	**83.50**
Ours CRAD(Tiny)	**28.6**	98.24	99.50	89.99	96.43	70.17	91.95	75.16	94.30	58.60	81.11
Ours CRAD(Nano)	**15.0**	96.13	98.99	85.54	94.18	62.93	87.35	64.30	86.48	48.13	72.65

Table 2: Ablation study on the core components of CRAD. We compare the following configurations: **Teacher** (ConvNeXt-Base, same as Sample4Geo (Deuser et al., 2023)); **Baseline** (student trained with InfoNCE loss and student-selected hard negatives, using ConvNeXt-Tiny); **+TGS** (teacher-guided hard sample mining, mixing ratio=0.5, replaces student's hard negative mining); **+$\mathcal{L}_{\text{CR-InfoNCE}}$** (Eq. (3) without TGS); **+$\mathcal{L}_{\text{RSD}}$** (Eq. (6) without TGS); **+$\mathcal{L}_{\text{CRAD}}$** (Eq. (3) and Eq. (6) combined as $0.5\mathcal{L}_{\text{CR-InfoNCE}} + 0.5\mathcal{L}_{\text{RSD}}$, without TGS); **Ours** (full CRAD combining both components). Recall@1 and Recall@5 are reported on four datasets.

Method	CVACT-Val		CVACT-Test		VIGOR-Same		VIGOR-Cross	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Teacher	90.81	97.64	71.51	92.42	77.86	95.66	61.70	83.50
Baseline	89.30	96.04	61.85	88.65	71.90	92.70	52.32	75.76
+ TGS	89.30	96.16	68.38	91.38	73.58	93.62	54.05	77.76
+ $\mathcal{L}_{CR-InfoNCE}$	**90.20**	96.30	70.03	91.86	73.85	93.88	55.05	78.68
+ $\mathcal{L}_{RSD}$	89.55	96.05	68.68	91.52	74.24	94.15	57.60	78.24
+ $\mathcal{L}_{CRAD}$	90.09	96.21	69.28	91.77	74.30	94.03	57.35	80.67
Full CRAD (Ours)	89.99	**96.43**	**70.17**	**91.95**	**75.16**	**94.30**	**58.60**	**81.11**

to demonstrate generalization, without polar transformation, using input sizes of 140×768 (street) and 384×384 (satellite).

4.2.2 VIGOR

VIGOR (Zhu et al., 2021b) contains 90,618 satellite and 105,214 street-view images from four US cities. It introduces two evaluation settings: same-area (all cities for training/validation) and cross-area (training on New York/Seattle, testing on San Francisco/Chicago) to assess generalization. A key feature is the inclusion of semi-positive satellite images that partially cover the street-view region, making high R@1 scores difficult. Inputs are resized to 384×768 (street) and 384×384 (satellite) from original resolutions of 1024×2048 and 640×640, respectively.

We evaluate our method on five benchmarks: CVUSA, CVACT (val/test), and VIGOR (same-area/cross-area). Table 1 reports Recall@1 and Recall@5 along with model parameters (in millions). CRAD-Tiny (28.6M, ConvNeXt-Tiny) achieves strong performance on challenging benchmarks, including 70.17% R@1 on CVACT-Test and 75.16% / 58.60% R@1 on VIGOR Same/Cross, while using significantly fewer parameters than some prior methods. CRAD-Nano (15.0M, Convnext-nano) further reduces the model size with moderate performance drop.

4.3 Ablation Study

The ablation study is based on Convnext-tiny as student, and the teacher model is Convnext-Base trained by Sample4Geo (Deuser et al., 2023).

Table 3: Comparison of different knowledge distillation methods (with teacher-guided sampling) on the same student model. We report Recall@1 and Recall@5 on 4 datasets. **MSE**: L2 loss between teacher and student global features $\mathcal{L}_{\text{MSE}} = \|\mathbf{f}_T - \mathbf{f}_S\|_2^2$. **KL**: KL divergence between feature distributions after softmax $\mathcal{L}_{\text{KL}} = D_{\text{KL}}(P_T \| P_S)$. **CRD**: Contrastive Representation Distillation (Tian et al., 2020) $\mathcal{L}_{\text{CRD}} = -\log \frac{\exp(\mathbf{f}_S \cdot \mathbf{f}_T/\tau)}{\sum_{k=1}^{N} \exp(\mathbf{f}_S \cdot \mathbf{f}_k/\tau)}$. **SP**: Similarity-Preserving distillation (Tung & Mori, 2019) aligning pairwise similarity matrices $\mathcal{L}_{\text{SP}} = \|\mathbf{A}_T - \mathbf{A}_S\|_F^2$. **Ours**: our proposed CRAD loss.

Method	CVACT-Val		CVACT-Test		Vigor-Same	
	R@1	R@5	R@1	R@5	R@1	R@5
MSE	85.64	94.83	61.05	87.49	61.26	85.52
KL	85.75	94.69	61.19	87.54	59.71	83.71
CRD	85.65	95.00	60.75	87.39	59.28	83.46
SP	85.88	95.05	62.15	87.73	58.61	83.06
Ours	**89.99**	**96.43**	**70.17**	**91.95**	**75.16**	**94.30**

Ablation Study on Core Components. From Table 2, we conduct ablation studies to evaluate the contribution of each component in the proposed framework. Overall, each module consistently improves the baseline across all benchmarks, demonstrating the effectiveness of the proposed training strategy. First, introducing TGS already brings noticeable improvements, particularly on CVACT-Test (R@1 increases from 61.85 to 68.38) and VIGOR-Cross (R@1 increases from 52.32 to 54.05), indicating that the teacher-guided supervision can effectively enhance the representation capability of the student model. More importantly, the two proposed supervision objectives exhibit different advantages across datasets with distinct geographical settings. On CVACT, which is a single-city dataset where training and testing share the same urban distribution, the model with $\mathcal{L}_{CR\text{-}InfoNCE}$ achieves the strongest performance among the individual losses (e.g., 70.03 R@1 on CVACT-Test). This suggests that CR-InfoNCE excels in city-specific settings by enforcing strong instance-level alignment, allowing the model to precisely match ground–satellite pairs within a consistent geographic domain.

In contrast, on VIGOR, which involves multi-city training and cross-city evaluation, $\mathcal{L}_{RSD}$ shows stronger generalization ability, especially on the VIGOR-Cross split (57.60 R@1 vs. 55.05 with CR-InfoNCE). This indicates that RSD is more robust under cross-city scenarios, as its relation-level, geometry-preserving supervision focuses on maintaining structural relationships among samples rather than enforcing strict pairwise alignment, making it less sensitive to city-specific appearance variations. Finally, combining both objectives in the Full CRAD framework yields the best overall performance across all benchmarks, achieving 70.17 / 91.95 on CVACT-Test and 58.60 / 81.11 on VIGOR-Cross. This demonstrates that the two losses are complementary: CR-InfoNCE strengthens instance discrimination in homogeneous environments, while RSD enhances structural consistency for cross-domain generalization.

Effectiveness of Relational Knowledge Distillation. To verify the effectiveness of relational knowledge distillation for cross-view retrieval, we compare our method with commonly used distillation strategies, including direct feature regression (MSE), distribution matching (KL), contrastive distillation (CRD), and similarity-preserving distillation (SP). Results are summarized in Table 3. Conventional feature-level distillation methods (MSE/KL) and existing representation distillation approaches (CRD/SP) provide limited improvements over the student model and in some cases even degrade performance on challenging datasets such as VIGOR. This suggests that directly aligning global features or distributions does not effectively transfer the discriminative knowledge required for cross-view retrieval. In contrast, the complete CRAD loss performs best among the evaluated distillation objectives. Relative to the baseline in Table 2, it improves Recall@1 from 61.85% to 70.17% on CVACT-Test and from 71.90% to 75.16% on VIGOR-Same. These results support combining cross-view correspondence with relational supervision rather than relying on point-wise feature alignment alone.

All methods in Table 3 use the same teacher, student backbone, teacher-guided sampling, data processing, and training schedule; only the distillation objective changes. The KL baseline aligns softmax-normalized feature distributions, whereas RSD aligns distributions over cross-view candidates. SP also transfers pairwise similarities, but matches their absolute values rather than each query's relative candidate distribution. This controlled setting helps attribute the measured differ-

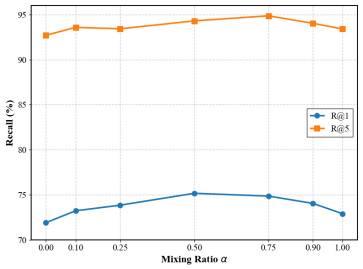


Figure 4: Impact of Teacher-Guided Mixing Ratio α on R@1 and R@5 for VIGOR-Same Dataset.

ences to how relational supervision is combined with cross-view correspondence, instead of to a different model or sampling procedure.

The component ablation and the controlled distillation comparison answer complementary questions: the former isolates the contributions of CR-InfoNCE and RSD, while the latter holds the teacher, student, and sampling procedure fixed. Together, they support transferring cross-view candidate relations alongside explicit same-location correspondence.

Analysis of Teacher-Guided Hard Sample Mining

As shown in Fig. 4, using only student-mined negatives ($\alpha = 0$) leads to inferior performance because the student cannot reliably identify highly confusing samples during the early stages of training. In contrast, relying solely on teacher-mined negatives ($\alpha = 1$) introduces overly difficult samples that may destabilize optimization. By prioritizing teacher-mined hard negatives while retaining complementary student-induced ones, our strategy stabilizes early training and remains adaptive as the student evolves. As a result, relational distillation focuses on the most confusing teacher-identified samples, sharpening the student's decision boundary where supervision is most effective. Consequently, a balanced combination achieves the best results, validating the effectiveness of our ratio-controlled fusion strategy.

Qualitative Analysis. Figure 5 in the appendix provides complementary views of this behavior. Its Grad-CAM maps show that the student trained with CRAD attends to regions similar to those used by the teacher for a correctly retrieved ground–satellite pair, whereas the baseline and MSE-distilled student fail on this example. For a CVACT validation query, the similarity plot orders candidates by the teacher's ranking: CRAD retains a clearer distinction between the positive match and hard negatives than either comparison model. Finally, the mini-batch query–reference similarity matrices show that CRAD better preserves the teacher's cross-view structure, including the diagonal corresponding to correct pairs. These visualizations illustrate the same candidate-level behavior measured by the retrieval metrics, rather than relying solely on feature-level agreement.

5 Conclusion

We propose a lightweight distillation framework that integrates with existing mainstream training paradigms, enabling cross-view geo-location to maintain competitive accuracy with significantly fewer parameters. Our experiments validate the effectiveness of transferring both positive samples and confused negative samples. The two loss functions we design complement each other: CR-InfoNCE excels in city-specific settings by enforcing strong instance-level alignment, while RSD shows greater robustness in cross-city evaluations through its relation-level, geometry-preserving supervision. Furthermore, the proposed TGS effectively assists the student model in mining hard negative samples. Together, these components enable strong performance even with a compact model. The proposed training framework provides a practical and general solution for CVGL distillation and also shows potential for cross-modal tasks such as image–text retrieval.

AI USE STATEMENT

In this work, we used generative AI tools to a limited extent to aid language polishing, identify potentially relevant literature, and assist with preliminary research ideation and execution. We did not use generative AI tools for other tasks requiring disclosure; the remaining required disclosure categories are not applicable to this work. We did not use generative AI tools for tasks with recommended disclosure beyond those stated above. We reviewed all AI-assisted work, checked identified references against their original sources, and independently evaluated any AI-assisted ideas and research outputs. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Margrit Betke and Leonid Gurvits. Mobile robot localization using landmarks. *IEEE transactions on robotics and automation*, 13:251–263, 1997.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 7028–7036, 2021.

YunTao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2852–2859. AAAI Press, 2018. doi: 10.1609/aaai.v32i1.11783. URL https://ojs.aaai.org/index.php/AAAI/article/view/11783.

Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4Geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16847–16856, October 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Jinglin Huang, Maoqiang Wu, Peichun Li, Wen Wu, and Rong Yu. VimGeo: Efficient cross-view geo-localization with vision mamba architecture. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1188–1196, 2025. doi: 10.24963/ijcai.2025/133. URL https://www.ijcai.org/proceedings/2025/133.

Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3238–3247, June 2020.

Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16719–16729, 2024a.

Songlian Li, Min Hu, Xiongwu Xiao, and Zhigang Tu. Patch similarity self-knowledge distillation for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5091–5103, 2024b. doi: 10.1109/TCSVT.2023.3336844.

Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5624–5633, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net, 2019.

Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. ConGeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/2187_ECCV_2024_paper.php.

Niluthpol Chowdhury Mithun, Kshitij S Minhas, Han-Pang Chiu, Taragay Oskiper, Mikhail Sizintsev, Supun Samarasekera, and Rakesh Kumar. Cross-view visual geo-localization for outdoor augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 493–502. IEEE, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *CVPR*, pp. 3967–3976, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Karsten Roth, Timo Milbich, Bjorn Ommer, Joseph Paul Cohen, and Marzyeh Ghassemi. Simultaneous similarity-based self-distillation for deep metric learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9095–9106. PMLR, 2021. URL https://proceedings.mlr.press/v139/roth21a.html.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4004–4012, 2016. doi: 10.1109/CVPR.2016.434.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations*, 2020.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Anirudh Viswanathan, Bernardo R Pires, and Daniel Huber. Vision based robot localization by ground to satellite matching in gps-denied situations. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 192–198. IEEE, 2014.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3961–3969, 2015.

Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2840–2848, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Wu_Sampling_Matters_in_ICCV_2017_paper.html.

Qiong Wu, Yi Wan, Zhi Zheng, Yongjun Zhang, Guangshuai Wang, and Zhenyang Zhao. CAMP: Across-view geo-localization method using contrastive attributes mining and position-aware partitioning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Zimin Xia, Olaf Booij, and Julian F. P. Kooij. Convolutional cross-view pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3813–3831, 2024. doi: 10.1109/TPAMI.2023.3346924.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021.

Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. In *European Conference on Computer Vision*, pp. 74–90, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/5379_ECCV_2024_paper.php.

Muhammad Haad Zahid and Murtaza Taj. Distill4Geo: Streamlined knowledge transfer from contrastive weight-sharing teachers to independent, lightweight view experts. In *Neural Information Processing: 32nd International Conference, ICONIP 2025*, pp. 546–560, 2025. doi: 10.1007/978-981-95-4097-6_37.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3480–3488, 2023. doi: 10.1609/aaai.v37i3.25457. URL https://ojs.aaai.org/index.php/AAAI/article/view/25457.

Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. GeoDTR+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10419–10433, 2024. doi: 10.1109/TPAMI.2024.3443652.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 756–765, 2021a.

Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3640–3649, 2021b.

Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1162–1171, New Orleans, LA, USA, 2022. IEEE.

Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization, 2023. URL https://arxiv.org/abs/2302.01572.

A Supplementary Material

A.1 Additional Benchmark Results

As shown in Tables 4 and 5, due to space limitations we only report Recall@1 and Recall@5 in the main paper. In this section, we additionally provide the remaining important metric, Recall@10, for completeness.

For the VIGOR dataset, each satellite image contains three semi-positive ground images. As discussed in the original benchmark, these semi-positive samples make it difficult to achieve a high Recall@1 score because they are visually similar yet treated as negatives during evaluation. To better reflect the true retrieval performance, the benchmark also reports the hit rate, which can be interpreted as Recall@1 with the semi-positive samples masked out, since these samples may otherwise act as distractors.

For consistency with the complete benchmark results below, the CRAD-Tiny entry in the main comparison reports 58.60 R@1 and 81.11 R@5 on VIGOR-Cross.

Table 4: Comparison of methods on CVUSA and CVACT benchmarks. Params are in millions (M).

Model	Params (M)	CVUSA			CVACT-Val			CVACT-Test		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
L2LTR (Yang et al., 2021)	195.9	94.05	98.27	98.99	84.89	94.59	95.96	-	-	-
TransGeo (Zhu et al., 2022)	44.88	94.08	98.36	99.02	84.95	94.14	95.78	-	-	-
GeoDTR (Zhang et al., 2023)	48.5	95.43	98.86	99.34	86.21	96.44	96.72	64.52	88.59	91.96
SAIG-D (Zhu et al., 2023)	31.2	96.34	99.10	99.42	89.06	96.11	97.08	67.49	89.39	92.30
PaSS-KD (Li et al., 2024b)	401.0	94.09	98.42	99.13	87.20	94.30	95.67	66.81	88.03	90.87
ConGeo (Mi et al., 2024)	89.0	98.27	99.59	99.70	90.12	95.69	96.56	71.67	91.61	93.50
VimGeo (Huang et al., 2025)	50.9	96.19	98.62	99.00	87.62	94.88	96.06	**81.69**	**92.42**	94.32
Distill4Geo (Zahid & Taj, 2025)	31.4	97.92	99.58	99.76	90.04	96.40	97.22	71.07	91.66	93.83
Sample4Geo (Deuser et al., 2023)	89.0	**98.68**	**99.68**	**99.78**	**90.81**	**97.64**	**97.48**	71.51	**92.42**	**94.45**
Ours CRAD(Tiny)	28.6	98.24	99.50	99.76	89.99	96.43	97.32	70.17	91.95	94.08
Ours CRAD(Nano)	15.0	96.13	98.99	99.17	85.54	94.18	95.73	62.93	87.35	90.64

Table 5: Comparison of methods on VIGOR benchmarks (Same and Cross splits). Metrics include Recall@K and Hit Rate. Params are in millions (M). Note that for VimGeo reported R@10 is slightly lower than R@5. This phenomenon comes from the original reported results in the corresponding papers, and we keep them unchanged for faithful reproduction.

Model	Params (M)	Vigor-Same				Vigor-Cross			
		R@1	R@5	R@10	Hit Rate	R@1	R@5	R@10	Hit Rate
TransGeo (Zhu et al., 2022)	44.88	61.48	87.54	91.88	73.09	18.99	38.24	46.91	21.21
SAIG-D (Zhu et al., 2023)	31.2	65.23	88.08	-	74.11	33.05	55.94	-	36.71
PaSS-KD (Li et al., 2024b)	401.0	52.90	76.60	-	57.00	21.00	39.70	-	22.20
VimGeo (Huang et al., 2025)	50.9	55.24	80.75	76.12	57.43	19.31	37.50	46.03	20.72
Distill4Geo (Zahid & Taj, 2025)	31.4	68.39	89.28	92.70	77.84	59.93	83.09	87.66	68.14
Sample4Geo (Deuser et al., 2023)	89.0	**77.86**	**95.66**	**97.21**	**89.82**	**61.70**	**83.50**	**88.00**	**69.87**
Ours CRAD(Tiny)	28.6	75.16	94.30	96.37	86.31	58.60	81.11	86.15	66.31
Ours CRAD(Nano)	15.0	64.30	86.48	90.87	71.66	48.13	72.65	80.12	52.42

A.2 Hyperparameter Studies

We conduct hyperparameter studies on three aspects: the balancing coefficient λ in the overall loss, the weights of different components in $\mathcal{L}_{\text{CR-InfoNCE}}$, and the batch size.

First, we study the effect of the balancing parameter λ in the overall objective:

$$\mathcal{L}_{\text{CRAD}} = \lambda \mathcal{L}_{\text{CR-InfoNCE}} + (1-\lambda)\mathcal{L}_{\text{RSD}}. \tag{10}$$

Here, λ controls the contribution of $\mathcal{L}_{\text{CR-InfoNCE}}$ and $\mathcal{L}_{\text{RSD}}$. As shown in Table 6, using only one loss (i.e., $\lambda = 0$ or $\lambda = 1$) results in inferior performance. Combining the two losses consistently improves the results, indicating that they provide complementary supervision. The best performance is achieved when $\lambda = 0.5$, which balances the two objectives. Therefore, we adopt $\lambda = 0.5$ in all experiments (The w parameter in CR-InfoNCE is fixed to 0.33).

Table 6: Comparison of different λ values for the loss function $\mathcal{L}_{\text{CRAD}} = \lambda\mathcal{L}_{\text{CR-InfoNCE}} + (1-\lambda)\mathcal{L}_{\text{RSD}}$. The maximum is in **bold** and the second maximum is underlined.

λ	CVACT-Val			CVACT-Test		
	R@1	R@5	R@10	R@1	R@5	R@10
0.0	89.55	96.05	96.96	68.68	91.52	93.73
0.1	89.62	96.36	97.24	69.13	91.62	93.83
0.25	89.54	96.16	97.17	69.03	91.73	93.90
0.5	**89.99**	**96.43**	**97.32**	**70.17**	**91.95**	**94.08**
0.75	89.68	96.38	97.21	70.04	91.76	93.96
0.9	89.93	**96.43**	97.29	70.08	91.92	93.97
1.0	89.30	96.04	96.97	68.61	91.17	93.57

Table 7: Hyperparameter study on the weighting coefficients of $\mathcal{L}_{\text{SS}}$, $\mathcal{L}_{\text{TS}}$, and $\mathcal{L}_{\text{ST}}$ in $\mathcal{L}_{\text{CR-InfoNCE}}$ on CVACT-Val and CVACT-Test. The maximum is in **bold** and the second maximum is underlined.

w1:w_{ss},w2:w_{ts},w3:w_{st}	CVACT-Val			CVACT-Test		
	R@1	R@5	R@10	R@1	R@5	R@10
w1=1.0,w2=0.0,w3=0.0	88.96	95.88	96.97	68.39	91.20	93.52
w1=0.0,w2=1.0,w3=0.0	88.70	96.03	97.05	68.97	91.55	93.75
w1=0.0,w2=0.0,w3=1.0	89.24	96.29	**97.38**	69.37	91.77	93.94
w1=0.5,w2=0.5,w3=0.0	89.52	96.27	97.28	69.29	91.63	93.88
w1=0.5,w2=0.0,w3=0.5	89.50	96.20	97.16	70.03	91.86	94.01
w1=0.0,w2=0.5,w3=0.5	89.85	96.42	97.31	70.01	**91.96**	**94.12**
w1=0.33,w2=0.33,w3=0.33	89.99	96.43	97.32	**70.17**	91.95	94.08
w1=0.66,w2=0.167,w3=0.167	89.63	96.31	97.21	69.04	91.52	93.77
w1=0.167,w2=0.66,w3=0.167	**90.06**	**96.47**	97.29	70.05	91.90	94.10
w1=0.167,w2=0.167,w3=0.66	89.77	96.38	97.33	69.72	91.87	94.03

Next, we investigate the weights of the three components in $\mathcal{L}_{\text{CR-InfoNCE}}$:

$$\mathcal{L}_{\text{CR-InfoNCE}} = w_{\text{ss}}\mathcal{L}_{\text{SS}} + w_{\text{ts}}\mathcal{L}_{\text{TS}} + w_{\text{st}}\mathcal{L}_{\text{ST}}.$$

Table 7 reports the results under different weight combinations. When only a single component is used, the performance is relatively lower. In contrast, combining multiple components generally yields better results, suggesting that they capture complementary cross-view relations. Among all settings, the balanced configuration $(w_{\text{ss}}, w_{\text{ts}}, w_{\text{st}}) = (0.33, 0.33, 0.33)$ achieves competitive and stable performance across both validation and test sets. Therefore, we adopt this balanced setting in the final model. Moreover, the hyperparameter studies indicate that the framework is relatively robust across different non-zero parameter settings, suggesting that the combined objective provides stable and effective supervision.

Finally, we study the effect of batch size, as shown in Table 8. When the batch size is small (e.g., 32 or 64), the performance drops significantly. This is mainly because the number of negative samples within a batch is limited, which restricts the model's ability to learn effective discrimination, especially for confused negatives. As the batch size increases, the performance improves substantially. Although a batch size of 256 achieves slightly higher values on some metrics, the improvement over 128 is marginal. This indicates that a batch size of 128 already provides sufficient negative samples for effective learning, while larger batch sizes introduce redundancy and increase GPU memory consumption. Therefore, we use a batch size of 128 in our experiments.

A.3 Additional Visualizations

Figure 5 provides qualitative and quantitative visualizations to compare different distillation strategies.

(a) Grad-CAM visualizations. We visualize Grad-CAM (Selvaraju et al., 2017) responses from the last convolutional layer to highlight spatial regions contributing to retrieval. The selected ground–satellite pair is correctly retrieved by both the teacher and our method, while the Baseline and MSE-distilled models fail. As shown in Figure 5(a), our method focuses on regions highly consistent with the teacher, especially on structurally discriminative areas in both ground and satellite views. In contrast, the Baseline and MSE models exhibit attention misalignment, particularly on

Table 8: Hyperparameter study on the batch size. Results are reported on CVACT-Val and CVACT-Test using Recall@1, Recall@5, and Recall@10. The maximum is in **bold** and the second maximum is underlined.

Batch size	CVACT-Val			CVACT-Test		
	R@1	R@5	R@10	R@1	R@5	R@10
32	81.26	92.30	93.72	57.02	82.86	87.11
64	87.56	95.10	96.25	66.04	89.63	92.25
128	**89.54**	96.16	97.17	**69.03**	**91.73**	**93.90**
256	89.34	**96.35**	**97.25**	68.14	91.51	93.85

satellite images, indicating limited transfer of region-level discriminative knowledge. More results are shown in Figure 6.

(b) Similarity distribution over ranked candidates. Figure 5(b) plots similarity scores following the teacher's ranking order for a query from the CVACT validation set. Our method produces a similarity distribution closely matching the teacher, with clear separation between positive samples and hard negatives (high responses at top ranks and rapid decay afterward). The Baseline fails to sufficiently distinguish positives from hard negatives, while MSE distillation further degrades relative similarity modeling due to overly strict feature alignment. This explains why MSE distillation yields lower retrieval accuracy than the Baseline, as also reflected quantitatively in Table 3, highlighting the inherent limitation of naive feature matching for ranking-based retrieval tasks.

(c) Mini-batch similarity matrices. Figure 5(c) visualizes query–reference similarity matrices computed on the CVACT training set with a mini-batch size of 16. Our method preserves a clear diagonal structure consistent with the teacher, indicating effective relational structure learning. In contrast, MSE distillation disrupts the similarity structure, confirming that point-wise feature alignment alone is insufficient for ranking-based retrieval tasks.

Key Insights from Visualization. The visualizations reveal three important patterns:

- **Teacher's superior discrimination**: In early training stages, the teacher consistently identifies negatives that are geographically closer to the query but exhibit significant visual viewpoint differences—precisely the cases that confuse novice students.
- **Student's learning trajectory**: As training progresses, the overlap between teacher-selected and student-selected hard negatives substantially increases, indicating effective knowledge transfer.
- **Boundary refinement evidence**: Negatives selected by *both* teacher and student (intersection set) predominantly lie near the decision boundary, confirming their crucial role in metric refinement.

A.4 Dynamic Quality of Teacher-Guided Sampling

We evaluate the quality of the negative samples selected throughout training on CVACT-train. At each epoch, the corresponding student checkpoint and the fixed teacher extract query and reference descriptors. Their similarity matrices are then used to construct student-only, teacher-only, and TGS-fused hard-negative pools. We measure (i) the average teacher similarity of each selected pool, which indicates how confusing the negatives are in the teacher metric space, and (ii) the Jaccard overlap between the teacher and student hard pools, which tracks agreement between their mining decisions.

Figure 7 shows that TGS consistently selects negatives with high teacher-view similarity and maintains strong teacher–student hard-pool consistency during training. The candidates are updated as the student evolves; the mixing ratio controls only the contribution of teacher- and student-mined candidates. These dynamics show that TGS supplies informative negatives even when early student-driven mining is unreliable, while progressively incorporating the student's improving similarity estimates.

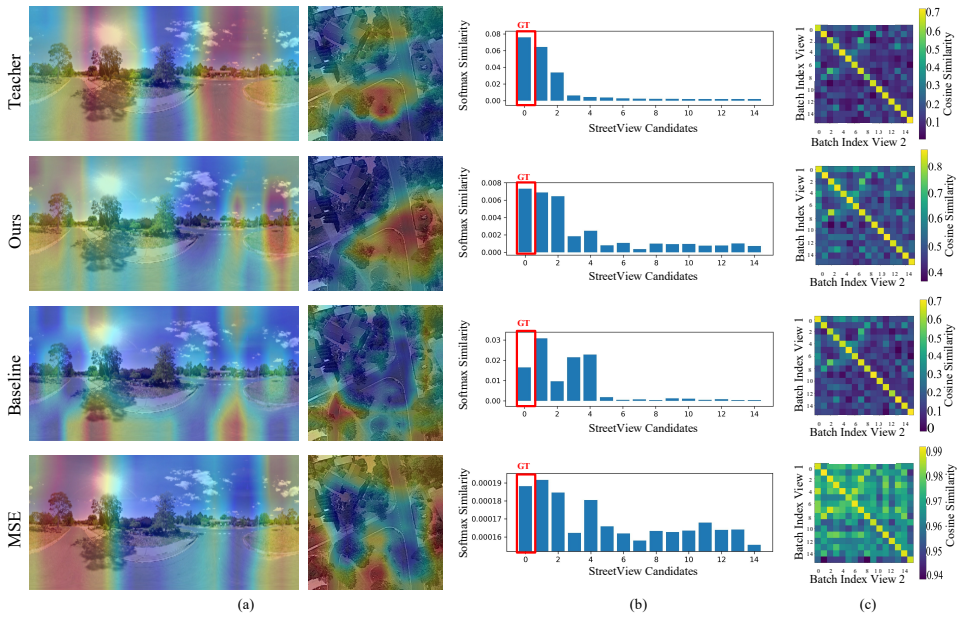


Figure 5: **Qualitative analysis of different distillation strategies.** (a) Grad-CAM visualizations of the last convolutional layer for a correctly retrieved ground–satellite pair. Our method exhibits attention patterns highly consistent with the teacher, while the Baseline and MSE models show noticeable misalignment, especially in the satellite view. (b) Similarity distributions over ranked street-view candidates (ordered by the teacher) on the CVACT validation set, where our method closely matches the teacher's discrimination between positives and hard negatives. (c) Query–reference similarity matrices computed on the CVACT training set with a mini-batch size of 16, showing that our method preserves the relational structure learned by the teacher, whereas MSE distillation disrupts similarity ordering.

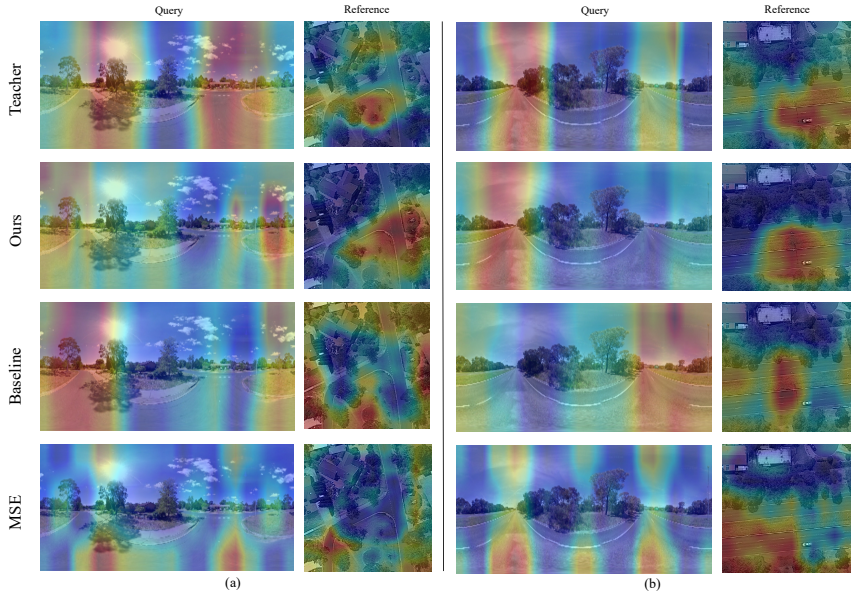
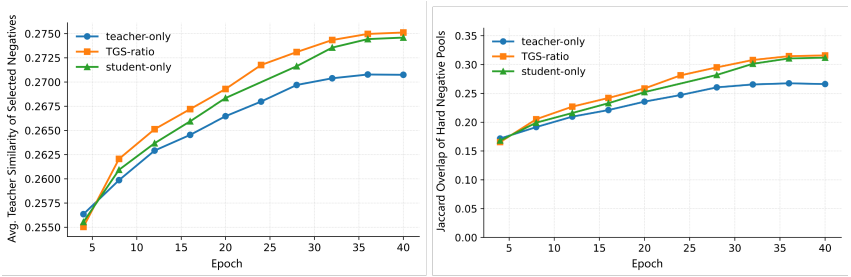


Figure 6: **More Grad-CAM visualizations** Grad-CAM visualizations of the last convolutional layer for a correctly retrieved ground–satellite pair.



(a) Quality of Mined Nagatives over Training Epochs (b) Evolution of Hard Negative Pool Overlap

Figure 7: Dynamic evaluation of hard-negative quality during training. TGS is compared with teacher-only and student-only mining using teacher-view similarity and teacher–student hard-pool overlap.

Table 9: Ablation study on the TGS mixing ratio α on VIGOR-Same.

Mixing Ratio α	**R@1 / R@5 / R@10 / Hit Rate**
0 (teacher-only)	71.90 / 92.70 / 95.31 / 82.96
0.1	73.22 / 93.58 / 95.79 / 84.23
0.25	73.84 / 93.42 / 95.77 / 85.04
0.5 (TGS)	**75.16 / 94.30 / 96.37 / 86.31**
0.75	74.85 / 94.86 / 96.35 / 86.12
0.9	74.04 / 94.03 / 96.24 / 85.45
1 (student-only)	72.89 / 93.40 / 95.85 / 83.96

A.5 Mixing Ratio in TGS

The mixing coefficient α specifies the fraction of the K hard negatives contributed by teacher-guided mining, with the remaining candidates selected by the student. Table 9 shows that combining both sources is consistently stronger than either endpoint. The balanced setting $\alpha = 0.5$ provides the best overall result and is therefore used in the main experiments.

A.6 Controlled Distillation Comparison

The distillation comparison in Table 3 of the main paper is controlled: every method uses the same teacher, student backbone, TGS strategy, data processing, training schedule, and evaluation protocol; only the distillation objective changes. The comparison includes not only direct MSE and KL matching but also CRD and similarity-preserving (SP) distillation. CRAD remains stronger under this common setup. Together with the component and hyperparameter ablations, these results attribute the improvement to retrieval-oriented relational supervision rather than to the number of loss terms. In particular, $\mathcal{L}_{\text{CR-InfoNCE}}$ anchors same-location cross-view correspondence, whereas $\mathcal{L}_{\text{RSD}}$ transfers candidate-wise ranking geometry; their combination improves upon either objective alone.

For reproducibility, our ConvNeXt-Tiny baseline uses the `fb_in22k_ft_in1k_384` checkpoint. The exact Tiny checkpoint in the original Sample4Geo configuration is not specified, so absolute differences from its reported number may reflect pretraining or implementation details. All ablations in this work nevertheless share the same reproduced baseline and training pipeline. Thus, their relative gains are directly controlled. CRAD changes only training-time objectives and sampling, adding no inference-time parameters or computation.

A.7 Heterogeneous Lightweight Students

We further evaluate whether CRAD benefits compact architectures outside the ConvNeXt family. MobileNetV3-Large and EfficientNet-B0 use two non-shared view encoders because weight sharing produces weak baselines for these heterogeneous lightweight models. Both students are distilled from the Sample4Geo ConvNeXt-Base teacher. As shown in Table 10, CRAD consistently improves both architectures on CVACT, supporting its applicability to compact students with different inductive biases.

Table 10: CRAD with heterogeneous lightweight student architectures.

Model	Params (M)	CVACT-Val R@1/5/10	CVACT-Test R@1/5/10
MobileNet baseline	4.2×2	80.57 / 93.29 / 95.21	52.93 / 81.75 / 86.94
MobileNet CRAD	4.2×2	84.57 / 95.02 / 96.57	59.37 / 86.55 / 90.50
EfficientNet baseline	4.0×2	87.20 / 95.16 / 96.47	66.10 / 89.22 / 91.97
EfficientNet CRAD	4.0×2	**88.94 / 96.32 / 97.32**	**68.26 / 90.97 / 93.43**

Table 11: CRAD on other InfoNCE-compatible cross-view geo-localization pipelines.

SAIG setting	VIGOR-Same R@1	EP-BEV setting	CVACT-Val R@1
Teacher	62.92	Teacher (ConvNeXt-Base)	91.90
Student baseline	59.21	Student (ConvNeXt-Tiny)	89.30
Student + CRAD	**61.82**	Student + CRAD	**91.10**

A.8 Applicability to Other Retrieval Pipelines

CRAD is designed for InfoNCE-based retrieval, so we test it on two additional compatible pipelines. SAIG provides an InfoNCE-compatible model setting, while EP-BEV (Ye et al., 2024) transforms ground panoramas into bird's-eye-view representations within a Sample4Geo-style pipeline. Table 11 shows consistent gains over the corresponding students. These results support applicability across multiple InfoNCE-based cross-view geo-localization frameworks, while we do not claim universality for substantially different objectives such as triplet-loss training.