Distill4Geo: Streamlined Knowledge Transfer from Contrastive Weight-Sharing Teachers to Independent, Lightweight View Experts

Muhammad Haad Zahid and Murtaza Taj

Computer Vision and Graphics Lab, Lahore University of Management Sciences
{24100134, murtaza.taj}@lums.edu.pk

Abstract. Cross-View Geo-Localization (CVGL) aims to align images from different perspectives (e.g., satellite and street views) to a shared geographic location—a complex task due to variations in viewpoint, intricate scene geometry, and visual discrepancies across views. Current methods commonly employ contrastive loss, which requires matching and non-matching (negative) pairs and often demands large batch sizes, leading to significant training overhead. This challenge is compounded in weight-sharing models which—while typically achieving better accuracy—incur high parameter and computational costs. We introduce a novel knowledge distillation approach that trains lightweight, view-specific student models without weight sharing. Optimized with a cosine embedding-based dual distillation loss, our method eliminates the need for large batch sizes. We also introduce augmentation noise to improve the student models' pairwise generalization. Our approach reduces parameters by 3× and GFLOPs by over 13.5×, achieving state-of-the-art (SOTA) accuracy on leading cross-view datasets, including CVUSA, CVACT, and VIGOR.

Keywords: Cross-View Geo-localization · Knowledge Distillation

1 Introduction

Cross-View Geo-Localization (CVGL) aims to determine the geographical location of objects or scenes from visual data captured from various viewpoints with a shared geospatial reference [8]. Precise geolocalization is essential for autonomous vehicle navigation [5, 14]. It also supports augmented reality by aligning virtual objects with real-world settings [21, 35] and assists in environmental monitoring by linking visual data to specific geolocations [33]. Additionally, CVGL plays a crucial role in enabling location-based search and analysis [29].

Early CVGL approaches attempted to assign precise latitude and longitude coordinates to each pixel in a query image through geodetic alignment [3].

With the advent of deep neural networks, geo-localization was reformulated as a cross-view retrieval problem, where contrastive loss—often used with Siamese networks—improved performance by minimizing the distance between similar pairs and maximizing the distance between dissimilar ones [22, 24, 48]. Multi-stage architectures, particularly those based on transformers [37, 2], further enhanced localization by leveraging self-attention [45] and feature aggregation.

Recent advances, such as hard negative mining [7], help disambiguate geographically close or visually similar pairs to learn fine-grained features. Self-distillation approaches [15], leveraging multiple scales of aerial views, learn location-specific features. However, these architectures often introduce substantial computational costs, with model sizes averaging 80 million parameters [7, 20]. While some methods have adopted smaller local batch sizes in contrastive loss to address computational challenges, they suffer from the log-K curse.

In contrast to recent methods relying on weight sharing [7], self-distillation [15], self-supervision [38, 39], and hard negative mining [7], our proposed approach leverages knowledge distillation [12] to achieve competitive retrieval performance with lightweight, non-weight-sharing networks (see Fig. 1). The key aspects of our work are:

- We propose a novel architecture that performs knowledge distillation (KD) for cross-view geo-localization. Our model conducts dual distillation by transferring knowledge from a contrastive, weight-shared teacher network [7] to each view-specific student.
- Whereas Hinton et. al.[12] proposed KD using softmax with temperature, we instead employ a simple, sample-wise, cosine embedding-based dual distillation loss to align each student's representation with that of the teacher.
- We eliminate the need for weight sharing between aerial and ground-view encoders, resulting in independent, identical encoders that are significantly smaller in size and can be trained independently—further reducing resource requirements during both training and inference.
- Our strategy achieves a 3× reduction in parameters and over a 13.5× reduction in GFLOPs, resulting in an efficient and lightweight model.
- We achieve accuracy on par with current SOTA approaches when evaluated on the popular CVUSA [42], CVACT [18], and VIGOR [46] datasets.

2 Related Works

Feature-based Geodetic Alignment: Early methods focused on pixel-wise techniques such as geodetic alignment. These approaches used geodetically accurate reference imagery to assign specific latitude and longitude coordinates to each pixel in a query image, but faced limitations in scalability and in coverage of less-documented areas [3]. They relied on handcrafted feature descriptors that captured both local image properties and geometric invariants, ensuring robustness against changes in scale, rotation, translation, and illumination [17, 36]. However, these methods were limited in adaptability to novel contexts or diverse, real-world environments.

CNN-Based Retrieval: The advent of deep learning technologies revolutionized the field of geo-localization. Deep models, particularly those based on Siamese architectures, have been successfully applied to cross-view geo-localization tasks [18, 23, 46, 31]. These networks excel at learning similarity metrics between image pairs, making them suitable for handling viewpoint variations. They have

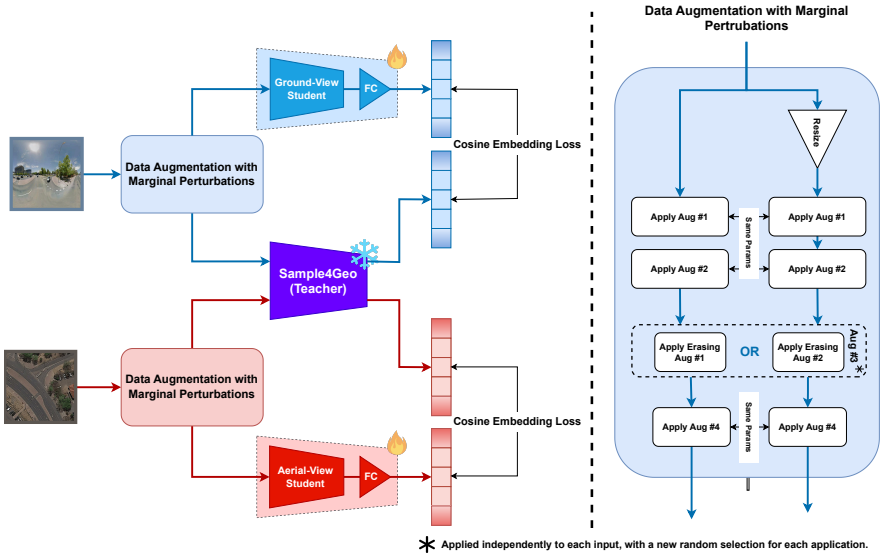

Fig. 1: An illustration of our training pipeline: inputs from each view are first augmented, each view-specific student is then independently trained with a cosine-based dual distillation loss, aligning the student's embeddings with those from the teacher model.

been used to compare ground-level query images with aerial reference images and estimate their geographic coordinates [18, 25, 4]. Subsequent research explored Capsule Networks [28, 47], in which groups of neurons—called capsules—encode both the probability and pose of objects. Unlike traditional CNNs, Capsule Networks use dynamic routing to preserve spatial hierarchies, enabling improved handling of viewpoint changes and object relationships. In addition, the inherent misalignment between aerial and street imagery has been addressed by incorporating orientation information [1] through polar representations [25, 26, 30, 16].

Attention-Based Multi-Stage Architectures: More recently, multi-stage architectures have demonstrated strong performance [27, 9, 45, 13]. One notable example is TransGeo [45], which leverages transformers for feature extraction and matching between ground and aerial images. TransGeo employs a dual-branch architecture with shared weights, incorporating self-attention mechanisms to capture long-range dependencies in both views. It also integrates a polar transform module to address orientation discrepancies, along with Adaptive Sharpness-Aware Minimization [45]. Another significant approach is the Multi-Scale Aggregation Network (MSA-Net) [13], which captures multi-scale features from both ground and aerial images. MSA-Net uses a hierarchical feature pyramid network for multi-scale feature extraction, followed by an attention-based aggre-

gation module to combine features effectively. This design improves performance in handling variations in scale and perspective between views.

Models with Weight Sharing: The use of contrastive losses has led to the development of networks such as Sample4Geo [7], which implements the InfoNCE loss—a contrastive objective commonly used in multi-modal tasks—for cross-view localization. Sample4Geo treats different views as separate modalities while employing a weight-sharing Siamese network. Further works, such as ConGeo [20], enhance generalization across varying fields of view and orientation conditions.

Knowledge Distillation in Cross-View Geo-Localization: Knowledge distillation has been widely explored in both cross-view and cross-modal retrieval [8]. Many existing studies address visual place recognition across heterogeneous sensors, such as cameras and LiDAR [44, 34]. PaSS-KD [15] introduced distillation specifically for cross-view geolocation. Unlike traditional teacher–student frameworks, PaSS-KD employs self-distillation using multiple scaled copies of the same view to capture fine-grained, location-specific features.

3 Methodology

3.1 Overview

The CVGL problem is usually formulated as the task of finding an embedding space in which corresponding aerial and street-view image pairs are close to each other, whereas non-matching pairs are far apart. This is achieved through contrastive methods such as the Triplet loss and the InfoNCE loss.

We instead propose a cosine-based *dual distillation loss* that attempts to learn the embedding space of a pre-trained teacher model using a student model. The goal is to find an embedding space in which the street-view image features of the student model are close to those of the teacher model, and the aerial/satellite-view image features are similarly close to those of the teacher model. This approach also eliminates the need for comparison with ground truth. The loss for a single view is defined as:

$$\mathcal{L}v = \frac{1}{B} \sum i = 1^B \left(1 - \frac{\mathbf{z}_i \cdot \mathbf{t}_i}{|\mathbf{z}_i||\mathbf{t}_i|}\right), \tag{1}$$

where $\mathbf{t}_i$ and $\mathbf{z}_i$ are the teacher and student embeddings, respectively, and B is the batch size. This loss can be used when training a single student independently. For training both student encoders simultaneously, the loss for both views can be calculated and averaged as $\mathcal{L}_{total} = (\mathcal{L}_{v_1} + \mathcal{L}_{v_2})/2$. This loss is effective because the teacher model has already been pre-trained using a contrastive loss, which ensures discriminative representations. This allows the student model to mimic them efficiently with low computational and memory costs. In addition, our student models can be trained separately, which is ideal for memory-bound, processing-constrained GPU clusters. Our strategy also eliminates the need for computationally expensive hard-negative mining involved in contrastive distillation losses by treating every pair as a positive pair.

3.2 Teacher/Student Models

Since our goal is to learn a lightweight model with low training and inference cost, we propose using knowledge distillation for the CVGL problem. Many large-scale models have been proposed for CVGL, such as TransGeo [45], SAIG-D [48], Sample4Geo [7], and the Panorama-BEV Co-Retrieval Network [41]. Any of these models can serve as a teacher model; however, in our implementation, we chose Sample4Geo [7] as our teacher due to its superior performance compared to its non-weight-sharing counterparts.

Recently, many lightweight vision transformers have been proposed for various computer vision tasks [37, 2]. Among these models, **FasterViT** [11] was chosen for our student encoder due to its hybrid architecture, which combines fast local representation learning with the global modeling capabilities of Vision Transformers (ViTs). It also introduces Hierarchical Attention (HAT), which reduces computational complexity while capturing long-range dependencies with minimal additional cost. FasterViT has only 31 million parameters and can operate with low-resolution images, further reducing computational requirements; hence, it is adopted as the student model in our architecture.

As illustrated in Fig. 1, we employ two distinct FasterViT-based student encoders, denoted by f_{s_i}, where $i \in \{1, 2\}$ corresponds to the street view ($i = 1$) and aerial view ($i = 2$), respectively. These learn from a single Sample4Geo teacher f_t. Each encoder is trained independently using a novel training strategy characterized by two principal components:

1. Improving the embedding space of input samples via augmentation noise and achieving *pair-wise generalization for retrieval*.
2. Performing *dual distillation via cosine similarity* instead of using contrastive distillation losses.

This approach introduces *inter-student proximity*, thereby improving inference performance.

3.3 Pair-wise Generalization via Augmentation Noise for Retrieval

We found that there is an inherent dissonance between teacher embeddings generated for the same location from different views. This means that, to improve convergence, it is beneficial to train a student model within its respective view-related embedding space. In addition, data augmentation helps the student better estimate the variance boundaries of the respective teacher embeddings and explore the space near a given embedding. Previous knowledge distillation literature shows that the standard deviation of a teacher's mean probability has been used for data augmentation [32]; in contrast, we perform noise-based augmentation.

For simplicity, we describe the method for a single view. Let v_i represent the view fed to both the teacher and the student. The data augmentation for this view consists of a set of noise vectors $\eta_{v_i} = \{\eta_t, \eta_{s_i}\}$, where η_t is the augmentation applied to the teacher model and η_{s_i} is the augmentation applied to the student

model, with $\eta_{s_i} \sim \eta_t$, i.e., there is a slight variation between the augmentations applied to the teacher and student inputs. This transformation can be expressed as $\tilde{v}_i = v_i + \eta_{v_i}$, and the resulting augmentation set is $\tilde{v}_i = \{\tilde{v}_t, \tilde{v}_{s_i}\}$.

The teacher network f_t and the student network f_{s_i} process this noisy input $\tilde{v}_i$, generating embeddings $\tilde{t}_{v_i}$ and $\tilde{z}_{v_i}$, respectively:

$$\tilde{t}_{v_i} = f_t(\tilde{v}_t), \quad \tilde{z}_{v_i} = f_{s_i}(\tilde{v}_{s_i}).$$

Without noise, each input v_i would produce a fixed embedding $t_{v_i} = f_t(v_i)$. However, with noise, the embedding $\tilde{t}_{v_i} = f_t(v_i + \eta_i)$ varies with η_i, leading to a distribution of embeddings around t_{v_i}.

Let $\mathbb{V}(\tilde{t}_{v_i})$ represent the variance of these embeddings due to noise. For sufficiently small noise σ, this distribution has a variance

$$\mathbb{V}(\tilde{t}_{v_i}) \approx \sigma^2 \|\nabla f_t(v_i)\|^2,$$

where $\nabla f_t(v_i)$ is the gradient of the teacher's embedding function with respect to the input. This variance encourages the student to explore a larger embedding space around t_{v_i} during training, improving generalization by learning to approximate a distribution of embeddings rather than a single point.

Without noise, the expected generalization error would have a weaker bound

$$\mathbb{E}_{v_i} \left[\mathcal{L}_{\cos}(f_{s_i})\right] \leq \epsilon,$$

where ϵ is a much smaller constant, indicating less robustness and higher sensitivity to input perturbations. With noise, the expected loss over the noisy input distribution $v + \eta$ is minimized, leading to a generalization bound of the form

$$\mathbb{E}_{v_i, \eta_i} \left[\mathcal{L}_{\cos}(f_{s_i})\right] \leq \sigma^2,$$

where σ^2 is the variance of the noise. This bound implies that, as σ increases, the model's generalization improves due to better exploration of the embedding space.

We can now state that there is a variation difference of ϵ between η_t and η_{s_i}, which can be modeled for the student as

$$\mathbb{V}(z_{v_i}) \approx (\sigma^2 + \epsilon^2) \|\nabla f_s(v)\|^2.$$

Adding noise implicitly regularizes the objective to minimize the worst-case deviation within this distribution, leading to the following $\mathcal{L}_{\cos}(f_{s_i})$ loss:

$$\min_{f_{s_i}} \mathbb{E}_{\eta_t, \eta_{s_i}} \left[1 - \frac{f_t(\tilde{v}_t) \cdot f_s(\tilde{v}_s)}{\|f_t(\tilde{v}_t)\| \|f_s(\tilde{v}_s)\|}\right].$$

This can be rewritten as

$$\lim_{f_{s_i} \to f_t} \mathcal{L}_{\cos}(f_{s_i}) \Rightarrow \mathbb{E}_{\eta_t, \eta_{s_i}} \left[1 - \frac{f_t(v_i + \eta_i) \cdot f_{s_i}(v_i + \eta_i)}{\|f_t(v_i + \eta_i)\| \|f_{s_i}(v_i + \eta_i)\|}\right].$$

With the variance under consideration, the expectation of the loss function can now be expressed as

$$\mathbb{E}_{v,\eta_t,\eta_s}\left[\mathcal{L}_{\cos}(f_s)\right] \leq \sigma^2 + \epsilon^2.$$

This means that the student embedding z_{s_i} will approximate the teacher embedding t_i for the given view v_i under various noisy conditions η. Consequently, this optimization ensures that f_{s_i} learns to represent the embedding space similarly to f_t, leading each student to converge toward the teacher's embedding distribution.

3.4 Inter-student proximity during inference

Let the respective student embeddings be $z_{s_1} = f_{s_1}(v_1)$ and $z_{s_2} = f_{s_2}(v_2)$, where each student independently approximates the teacher's output. Since each student has learned to approximate the teacher's embeddings, we expect inter-student proximity to satisfy:

$$1 - \frac{z_{s_1} \cdot z_{s_2}}{\|z_{s_1}\|\|z_{s_2}\|} \leq 2\delta,$$

where δ is a small positive constant reflecting the teacher's expected similarity between views. By the transitivity of embedding proximity, we expect that the student embeddings z_{s_1} and z_{s_2} will also be close, leading to:

$$\mathbb{P}\left(1 - \frac{z_{s_1} \cdot z_{s_2}}{\|z_{s_1}\|\|z_{s_2}\|} < \delta\right) \to 1, \tag{2}$$

which implies that the student embeddings align closely despite the absence of a comparison loss during training. Thus, applying identical noise during training enables the student embeddings to remain close during inference, even without explicitly enforcing alignment between the students.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

We evaluate our results on three popular benchmark datasets, namely CVUSA [42], CVACT [18], and VIGOR [46].

CVUSA consists of $35,532$ view pairs for training and $8,884$ for evaluation. Both splits are north-aligned. The satellite images have a resolution of 750×750, and the street-view images have a resolution of $224 \times 1,232$.

CVACT: Similar to CVUSA, CVACT also consists of $35,532$ view pairs for training; however, it includes an extended test split with $92,802$ images. Both splits are north-aligned. Higher-resolution images are provided, with $1,200 \times$

$1,200$ pixels for satellite views and $832 \times 1,664$ pixels for street views.

VIGOR is a relatively larger dataset comprising $90,618$ aerial images and $238,696$ street panoramas from four cities. The image resolutions are 640×640 and $2,048 \times 1,024$ pixels, respectively, for aerial and ground views. Unlike CVUSA and CVACT, the images in VIGOR are not north-aligned. It also offers two distinct evaluation protocols: *same-area* (training and testing within the same cities) and *cross-area* (testing transferability across different cities). In this dataset, the IoU (Intersection over Union) for perfectly aligned view pairs is only 0.39, which illustrates its difficulty.

Evaluation Metrics: We use **Recall@K** and **Hit Rate@K** to evaluate the performance of our proposed strategy. **Recall@K** measures the proportion of queries for which the model successfully retrieves at least one relevant image within the top K results. Similarly, **Recall@K%** measures the proportion of queries for which at least one relevant image is retrieved within the top $K\%$ of results. **Hit Rate@K** evaluates the proportion of queries in which at least one relevant image is found within the top K results.

Table 1: Results on CVUSA, CVACT datasets. Top-1, Top-2, and Top-3 results are highlighted in magenta, blue, and teal, respectively.

	CVUSA				CVACT Val				CVACT Test			
Method	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CDE [30]	92.56	97.55	98.50	99.57	83.28	93.57	95.42	98.22	60.72	85.85	89.88	96.12
L2LTR [40]	94.05	98.27	98.99	99.76	84.89	94.59	95.96	98.37	-	-	-	-
SEH [10]	95.04	98.31	98.43	99.76	85.13	95.24	96.97	97.97	-	-	-	-
TransGeo [45]	94.08	98.36	99.02	99.77	84.95	94.14	95.78	97.37	-	-	-	-
GeoDTR [43]	95.43	98.86	99.34	99.86	86.21	96.44	96.72	98.77	64.52	88.59	91.96	98.74
SAIG-D [48]	96.34	99.10	99.42	99.86	89.06	96.11	97.08	98.89	67.49	89.39	92.30	96.80
PaSS-KD [15]	94.09	98.42	99.13	99.77	87.20	94.30	95.67	97.85	66.81	88.03	90.87	98.02
GSRA [9]	97.25	99.39	-	-	88.35	95.32	-	-	-	-	-	-
ConGeo [20]	98.27	99.59	99.70	99.86	90.12	95.69	96.56	98.24	71.67	91.61	93.50	98.30
Sample4Geo [7]	98.68	99.68	99.78	99.87	90.81	97.64	97.48	98.77	71.51	92.42	94.45	97.90
UnifyGeo [27]	98.91	99.62	99.72	99.84	91.38	96.27	97.04	98.45	73.42	92.46	94.27	98.43
Ours (Distill4Geo)	97.92	99.58	99.76	99.86	90.04	96.40	97.22	98.83	71.07	91.66	93.83	98.76

4.2 Implementation Details

We adopt Sample4Geo [7] as our teacher model; it uses ConvNext-B [19] with shared weights and a default image size of 384. Our students, however, use the FasterVit-T0 [11] architecture, pretrained on ImageNet-1K, followed by a fully connected layer of size 1000×1024. This architecture is chosen for its high throughput (5802 images per second compared to ConvNext-B's 95.7 images per

second) and strong ImageNet-1K top-1 accuracy of 82.1% (compared to 85.1% for ConvNext-B). All inputs to the student models are resized to 224×224.

We train our student models on 4× RTX 2080 Ti GPUs with a global batch size of 48 for CVUSA and CVACT, and 36 for VIGOR (due to the larger image size input to the frozen teacher model). Training uses Distributed Data Parallel in the PyTorch Lightning framework with the AdamW optimizer. We employ a cosine annealing scheduler with a 1-epoch linear warm-up. All models are trained for 300 epochs with a learning rate of $1e^{-4}$ and an embedding size of 1024. **We use a separate flow of augmentation for street and satellite views; however, parity of augmentations to both the teacher and student is kept the same, with minor differences arising due to some augmentations having different effects on different image sizes. Augmentations referred to in Fig 1 include Random Image Compression, Random Color Jitter, Random Selection b/w Advanced Blur or Sharpening, Random Selection b/w Grid and Coarse Dropout, Flipping, and Rotation (for Satellite-View)/Horizontal Roll (i.e. Cyclic Shift for Street View).**

5 Results & Evaluation

5.1 Comparative Analysis of Retrieval

We compare our proposed approach with 12 recent SOTA methods: CDE [30], L2LTR [40], SEH [10], TransGeo [45], GeoDTR [43], SAFA [25], SAIG-D [48], PaSS-KD [15], GSRA [9], Sample4Geo [7], UnifyGeo [27], and ConGeo [20]. As shown in Table 1, our proposed strategy consistently achieves competitive results on both the CVUSA and CVACT datasets despite having 3× fewer parameters and requiring at least 13× fewer GFLOPs.

Compared to the teacher model [7], our proposed Distill4Geo achieves higher R@1 on both CVACT Validation and Test. It also outperforms all SOTA methods and achieves the highest R@1 on CVACT Test. In most other measures, our proposed Distill4Geo ranks second best, just behind the best-performing Sample4Geo teacher model. Moreover, it outperforms previous self-distillation baselines such as PaSS-KD [15] by a large margin on all measures. For example, our Distill4Geo achieves 4.26% higher R@1 on CVACT Test compared to PaSS-KD. This demonstrates that our student models successfully learn the knowledge transferred from the teacher model without the need for contrastive learning or hard-negative mining.

Benchmarking on the more challenging VIGOR dataset is shown in Table 2. Our proposed Distill4Geo outperforms all SOTA methods except the Sample4Geo teacher model on both the same-area and cross-area splits. When compared to the teacher model, Distill4Geo demonstrates satisfactory performance on the same-area split, which evaluates the model's ability to memorize the distribution within the area of interest. When compared to the newer UnifyGeo [27] model, which outperforms the teacher, our lightweight alternative

Table 2: Quantitative comparison with state-of-the-art on VIGOR [46]. † indicates use of polar transformation for satellite input. Top-1/2/3 in magenta, blue, teal.

Method	R@1	R@5	R@10	R@1%	Hit Rate	R@1	R@5	R@10	R@1%	Hit Rate
	Same Split					Cross Split				
SAFA† [25]	33.93	58.42	68.12	98.24	36.87	8.20	19.59	26.36	77.61	8.85
TransGeo [45]	61.48	87.54	91.88	99.56	73.09	18.99	38.24	46.91	88.94	21.21
SAIG-D [48]	65.23	88.08	-	99.68	74.11	33.05	55.94	-	94.64	36.71
PaSS-KD [15]	52.90	76.6	-	-	57.00	21.00	39.7	-	-	22.20
Sample4Geo [7]	77.86	95.66	97.21	99.61	89.82	61.70	83.50	88.00	98.17	69.87
UnifyGeo [27]	82.80	94.92	96.57	99.47	88.72	67.58	81.60	86.02	97.26	68.24
Ours (Distill4Geo)	68.39	89.28	92.70	99.16	77.84	59.93	83.09	87.66	97.92	68.14

consistently outperforms it on R@1% benchmarks across all datasets, while also achieving highly competitive R@5 and R@10 results on CVACT Validation and Test. A similar trend can be seen in Table 2, where we outperform UnifyGeo in the cross-area split on R@5 and higher benchmarks.

Our approach achieves competitive results on the cross-area split, where knowledge distillation was performed only on New York and Seattle, while testing was conducted on San Francisco and Chicago. This indicates that our proposed Distill4Geo strategy closely mimics the teacher's output, demonstrating excellent generalization capabilities to new areas, including large-scale inference scenarios such as CVACT Test and VIGOR Cross-Split.

5.2 Comparison of Loss Functions and Design Choices

We evaluated the performance of our proposed KD method using three different loss functions, namely MSE, contrastive, and cosine (see Table 3). Since both MSE and cosine embedding loss are batch-agnostic, they outperformed contrastive loss on CVUSA. Between MSE and cosine embedding loss, the latter performed marginally better due to its inherent similarity to the teacher's loss function. Therefore, in our proposed Distill4Geo approach, we adopt cosine embedding loss. Table 3 compares the proposed knowledge distillation approach with and without weight sharing using all three loss functions, namely MSE, contrastive, and cosine. It can be observed that for smaller architectures, weight sharing yields lower accuracy than its non-weight-sharing counterparts. Furthermore, weight-separated models are preferable in resource-constrained environments. This experiment also demonstrates that the model trained with the proposed approach outperforms all other configurations.

5.3 Variants of the Proposed Approach

Table 4 compares our proposed knowledge distillation-based Distill4Geo approach, trained using two different strategies, with a baseline method that does not use knowledge distillation and is directly trained on the dataset using contrastive loss. The comparison is conducted for both the CVUSA and CVACT

Table 3: We compare results with and without weight sharing between aerial and ground view encoders/backbone.

(a) Loss-wise comparison across CVACT Val.

Loss	R@1	R@5	R@10	R@1%
MSE	89.32	96.04	96.92	98.68
Contrastive	84.67	93.37	94.84	97.69
Cosine	**89.70**	**96.25**	**97.16**	**98.84**

(b) Loss-wise comparison on CVUSA dataset with or w/o weight-sharing.

Loss	WS	R@1	R@5	R@10	R@1%
MSE	✓	93.94	98.47	99.09	99.70
MSE	✗	97.16	99.40	99.65	**99.86**
Contrastive	✓	89.22	96.62	97.71	99.52
Contrastive	✗	89.85	97.82	98.93	99.85
Cosine	✓	94.01	98.75	99.27	99.75
Cosine	✗	**97.92**	**99.58**	**99.76**	**99.86**

(c) Contrastive loss only comparison on CVACT Val and Test with or w/o weight-sharing.

	WS	R@1	R@5	R@10	R@1%
Val	✗	84.67	93.37	94.84	97.69
	✓	81.16	92.71	94.51	97.83
Test	✗	60.32	85.74	89.43	97.88
	✓	55.56	82.72	87.14	97.98

(validation and test) benchmark datasets. In this experiment, we use FasterViT [11], pretrained on ImageNet-1K, as the architecture for both the street-view and aerial-view encoders. Hence, we present three scenarios:

- **Baseline Only:** Training our ImageNet-1K-pretrained student encoders directly on the dataset using contrastive loss, with no teacher involved.
- **KD Only:** Using our ImageNet-1K-pretrained student encoders and distilling knowledge from the teacher model using cosine embedding loss.
- **Baseline + Fine-tuning via KD:** Taking our baseline model and then fine-tuning it via knowledge distillation from the teacher model on the respective dataset.

It can be clearly seen that the proposed Distill4Geo (**KD Only**) approach outperforms the **Baseline**. Furthermore, it is on par with the **Baseline + Fine-tuning via KD** while requiring only half the computation time during training, as it avoids the baseline pretraining phase on the respective dataset. In addition, for CVACT (test), the proposed Distill4Geo achieved almost 21% higher R@1 compared to the baseline.

5.4 Computational Efficiency

One of the reasons behind proposing Distill4Geo is to reduce memory (i.e. parameters) and compute cost (i.e. FLOPS) without compromising retrieval performance. Table 5 shows the cost-effectiveness of our chosen model, where it

Table 4: Results demonstrating that the proposed Distill4Geo method outperforms training without a distillation loss, regardless of pre-training. The results with and without pre-training on the respective datasets are highly comparable, indicating that the primary benefit stems from the dual distillation-based cosine embedding loss.

Training Method	Fine-tuned	Evaluated	R@1	R@5	R@10	R@1%
Baseline Only	CVUSA Train	CVUSA Test	92.43	98.32	99.22	99.86
KD Only (Ours)			97.88	**99.62**	99.74	**99.89**
Baseline + KD			**97.92**	99.58	**99.76**	99.86
Baseline Only	CVACT Train	CVACT Val	75.13	93.37	95.52	98.67
KD Only (Ours)			89.87	**96.42**	**97.24**	98.80
Baseline + KD			**90.04**	96.40	97.22	**98.83**
Baseline Only	CVACT Train	CVACT Test	40.46	75.62	83.91	98.64
KD Only (Ours)			**71.15**	**91.67**	93.81	98.73
Baseline + KD			71.07	91.66	**93.83**	**98.76**

Table 5: Computational efficiency of our model on CVUSA and CVACT. Although our model employs two separate students, both training and inference can be performed with a single student by loading different weights; therefore, we treat it as a single model. Higher R@1 per parameter (M) and R@1 per GFLOP values indicate better performance. The top-1/2/3 results are shown in magenta, blue, and teal, respectively. *M* denotes mixed.

Method	Weight Sharing	Params (M)	FLOPS	R@1 Per Million Params			R@1 Per GFLOPS		
				CVUSA	CVACT Val	CVACT Test	CVUSA	CVACT Val	CVACT Test
L2LTR [40]	✗	195.9	44.06	0.48	0.43	-	2.13	1.93	-
TransGeo [45]	✗	2×22.44	11.32	2.10	1.89	-	8.31	7.50	-
GeoDTR [43]	✗	48.5	39.89	1.97	1.78	1.33	2.40	2.16	1.62
SAIG-D [48]	✗	2×15.6	13.3	3.09	2.85	2.16	7.24	6.70	5.07
PaSS-KD [15]	✓	≈ 401.0	≈ 46.5	0.23	0.22	0.17	2.02	1.86	1.44
ConGeo [20]	✓	89.0	45.0	1.10	1.01	0.81	2.18	2.00	1.59
Sample4Geo [7]	✓	89.0	45.0	1.11	1.02	0.80	2.19	2.02	1.59
UnifyGeo [27]	M	57.7	-	1.70	1.56	1.27	-	-	-
Ours (Distill4Geo)	✗	31.4	3.3	3.12	2.87	2.26	29.68	27.28	21.54

maintains the second-best position for the R@1 per million parameters and the best position for R@1 per GFLOPS with over 13.5x lower compute cost (FLOPS) and 3x fewer parameters as compared to our Sample4Geo [7] teacher and and 1.8x lower parameters against the SOTA UnifyGeo while being hugely cost effective for the performance gains (3.12 vs 1.70) in the R@1 metric.

The primary reason behind this reduction in cost is the choice of lightweight student models. In addition, our method uses an input size of 224×224, which is one of the smallest inputs in the recent Geo-localization literature.

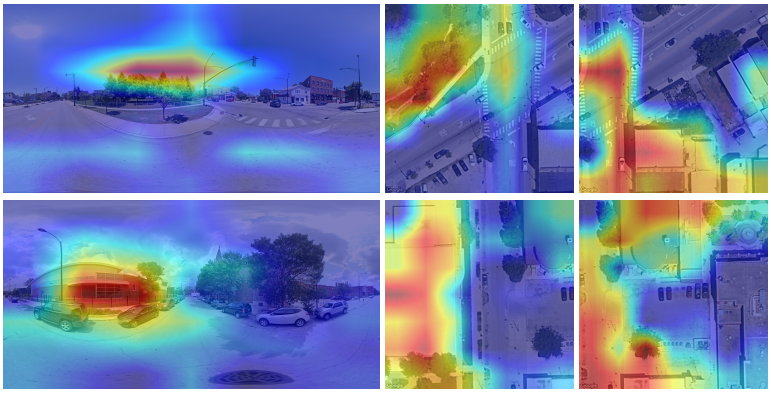

Fig. 2: Heatmap visualization of our approach on the VIGOR cross split, using AblationCAM [6]. (1^{st} col.) Ground truth street view, (2^{nd}, 3^{rd} col.) visualization on positive and semi-positive hit respectively.

5.5 Visualization

Figure 2 shows a heatmap visualization generated using AblationCAM [6] on the VIGOR cross split. It can be seen that student models trained via the proposed Distill4Geo framework identify common regions across the two distinct views. VIGOR presents a unique challenge, as each location has multiple viable matches between aerial and ground perspectives. This setting forces the model to move beyond simple road alignments or direct spatial correspondences, focusing instead on broader, context-rich features such as vegetation, terrain, and other landscape elements. As a result, VIGOR encourages our model to capture and utilize semantically meaningful regions that remain stable despite variations in orientation and field of view. As shown in Fig. 2, the model learns to maintain robust feature representations, demonstrating excellent heatmap alignment on positive matches while capturing overlaps across semi-aligned cross-views.

6 Conclusion

In conclusion, this study introduces Distill4Geo, a knowledge distillation framework optimized for efficient cross-view geo-localization. By leveraging lightweight, non-weight-sharing student models that learn from a contrastively trained teacher via a cosine embedding loss, Distill4Geo eliminates the need for large batch sizes and hard-negative mining, addressing common inefficiencies in contrastive-based approaches. Our method achieves competitive accuracy across standard datasets such as CVUSA, CVACT, and VIGOR while requiring 3× fewer parameters and over 13.5× lower computational cost, demonstrating a strong balance between performance and efficiency. Although the results are competitive on established benchmarks, future work should explore ways to further reduce training complexity and adapt the approach to a wider range of real-world scenarios.

References

1. Alom, M.Z., et al.: A state-of-the-art survey on deep learning theory and architectures. Electronics **8**(3), 292 (2019)
2. Anasosalu Vasu, P.K., Gabriel, J., Zhu, J., Tuzel, O., Ranjan, A.: FastViT: A fast hybrid vision transformer using structural reparameterization. In: ICCV (Oct 2023)
3. Bansal, M., Daniilidis, K., Sawhney, H.: Ultrawide baseline facade matching for geo-localization. In: Large-Scale Visual Geo-Localization, pp. 77–98. Springer (2016)
4. Cao, R., Zhu, J., Li, Q., Zhang, Q., Li, Q., Liu, B., Qiu, G.: Learning spatial-aware cross-view embeddings for ground-to-aerial geolocalization. In: Int. Conf. on Image and Graphics. pp. 57–67. Springer (Aug 2019)
5. Charroud, A., El Moutaouakil, K., Palade, V., Yahyaouy, A., Onyekpe, U., Eyo, E.U.: Localization and mapping for self-driving vehicles: A survey. Machines **12**(2) (2024)
6. Desai, S., Ramaswamy, H.G.: Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization. In: WACV. pp. 972–980 (2020)
7. Deuser, F., Habel, K., Oswald, N.: Sample4geo: Hard negative sampling for cross-view geo-localisation. In: ICCV (2023)
8. Durgam, A., Paheding, S., Dhiman, V., Devabhaktuni, V.: Cross-view geo-localization: a survey. In: IEEE Access (2024)
9. Gan, W., Zhou, Y., Hu, X., Zhao, L., Huang, G., Hou, M.: Learning robust feature representation for cross-view image geo-localization. IEEE Trans. Geoscience and Remote Sensing Letters **22** (2025)
10. Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M.: Soft exemplar highlighting for cross-view image-based geo-localization. IEEE TIP **31**, 2094–2105 (2022)
11. Hatamizadeh, A., Heinrich, G., Yin, H., Tao, A., Alvarez, J.M., Kautz, J., Molchanov, P.: FasterViT: Fast vision transformers with hierarchical attention. In: ICLR (2024)
12. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: Neural Information Processing Systems Workshop (2014)
13. Iqbal, E., Safarov, S., Bang, S.: MSANet: Multi-similarity and attention guidance for boosting few-shot segmentation (2022), https://arxiv.org/abs/2206.09667
14. Laconte, J., Kasmi, A., Aufrère, R., Vaidis, M., Chapuis, R.: A survey of localization methods for autonomous vehicles in highway scenarios. Sensors **22**(1) (2022)
15. Li, S., Hu, M., Xiao, X., Tu, Z.: Patch similarity self-knowledge distillation for cross-view geo-localization. IEEE TCSVT **34**(6), 5091–5103 (2024)
16. Li, S., Tu, Z., Chen, Y., Yu, T.: Multi-scale attention encoder for street-to-aerial image geo-localization. CAAI Trans. on Intelligence Tech. **8**(1), 166–176 (2023)
17. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: CVPR. pp. 891–898 (2013)
18. Liu, L., Li, H.: Lending Orientation to Neural Networks for Cross-View Geo-Localization . In: CVPR (Jun 2019)
19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s . In: CVPR (Jun 2022)
20. Mi, L., Xu, C., Castillo-Navarro, J., Montariol, S., Yang, W., Bosselut, A., Tuia, D.: ConGeo: Robust cross-view geo-localization across ground view variations. In: ECCV (2024)
21. Mithun, N.C., Minhas, K.S., Chiu, H.P., Oskiper, T., Sizintsev, M., Samarasekera, S., Kumar, R.: Cross-view visual geo-localization for outdoor augmented reality. In: IEEE Conf. Virtual Reality and 3D User Interfaces. pp. 493–502 (2023)

22. Rasna, A.A., Mohan, C.K.: Geodesic based image matching network for the multi-scale ground to aerial geo-localization. In: IEEE Aerospace Conf. (2023)
23. Rodrigues, R., Tani, M.: Are these from the same place? seeing the unseen in cross-view image geo-localization. In: WACV. pp. 3753–3761 (2021)
24. Rodrigues, R., Tani, M.: Semgeo: Semantic keywords for cross-view image geo-localization. In: ICASSP. pp. 1–5. IEEE (2023)
25. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. NeurIPS **32** (2019)
26. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am i looking at? joint location and orientation estimation by cross-view matching. In: CVPR. pp. 4064–4072 (2020)
27. Song, Z., Zhang, Y., Li, K., Wang, L., Guo, Y.: A unified hierarchical framework for fine-grained cross-view geo-localization over large-scale scenarios (2025), https://arxiv.org/abs/2505.07622
28. Sun, B., Chen, C., Zhu, Y., Jiang, J.: GEOCAPSNET: Ground to aerial view image geo-localization using capsule network. In: ICME. pp. 742–747 (2019)
29. Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geo-localization in urban environments. In: CVPR (2017)
30. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: CVPR. pp. 6488–6497 (2021)
31. Vyas, S., Chen, C., Shah, M.: Gama: Cross-view video geo-localization. In: ECCV. pp. 440–456. Springer (2022)
32. Wang, H., Wu, Y., Huang, T., You, S., Xu, C., Qian, C.: What makes a good data augmentation in knowledge distillation – a statistical perspective. In: NeurIPS. vol. 35 (2022)
33. Wang, S., Zhang, Y., Perincherry, A., Vora, A., Li, H.: View Consistent Purification for Accurate Cross-View Localization . In: ICCV (Oct 2023)
34. Wang, S., She, R., Kang, Q., Jian, X., Zhao, K., Song, Y., Tay, W.P.: DistilVPR: cross-modal knowledge distillation for visual place recognition. In: AAAI (2024)
35. Wilson, D., Zhang, X., Sultani, W., Wshah, S.: Visual and object geo-localization: A comprehensive survey (2024)
36. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: ICCV. pp. 1–9 (2015), acceptance rate: 30.3%
37. Wu, K., et al.: TinyViT: Fast pretraining distillation for small vision transformers. In: ECCV (2022)
38. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. ECCV **12354**, 588–604 (2020)
39. Yang, C., An, Z., Cai, L., Xu, Y.: Hierarchical self-supervised augmented knowledge distillation. In: Int. Joint Conf. on Artificial Intelligence. pp. 1217–1223 (2022)
40. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. In: NeurIPS (2021)
41. Ye, J., Lv, Z., Li, W., Yu, J., Yang, H., Zhong, H., He, C.: Cross-view image geo-localization with panorama-bev co-retrieval network. In: ECCV (2024)
42. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: CVPR (2017)
43. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence (2023)
44. Zhang, X., Wang, L., Su, Y.: Visual place recognition: A survey from deep learning perspective. Pattern Recognition **113**, 107760 (2021)
45. Zhu, S., Shah, M., Chen, C.: Transgeo: Transformer is all you need for cross-view image geo-localization. In: CVPR. pp. 1162–1171 (2022)

46. Zhu, S., Yang, T., Chen, C.: VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In: CVPR (Jun 2021)
47. Zhu, Y., Sun, B., Lu, X., Jia, S.: Geographic semantic network for cross-view image geo-localization. IEEE Trans. Geoscience and Remote Sensing **60**, 1–15 (2021)
48. Zhu, Y., Yang, H., Lu, Y., Huang, Q.: Simple, effective and general: A new backbone for cross-view image geo-localization. arXiv preprint arXiv:2302.01572 (2023)