First Learn, Then Review : Human-like Continual Learning for Cross-View Geo-localization with Limited Field of View

Lei Cheng, Daikun Liu, Zhikun Chen, Teng Wang *

Southeast University
{leicheng, dkliu, zhikunchen, wangteng}@seu.edu.cn

Abstract

This paper addresses cross-view geo-localization in real-world scenarios, where the field-of-view (FoV) is restricted and the orientation is unknown for ground-view images. This task is extremely challenging due to the huge domain gap. Existing methods typically treat tasks with different FoVs as independent tasks. These approaches not only require separate retraining for each FoV, but also neglect the strong correlations between different FoVs, leading to poor performance under extremely limited FoV. To overcome these limitations, we propose **HCL-Geo**, a framework that follows **h**uman-like **c**ontinual **l**earning paradigm of "first learn, then review" for **geo**-localization: in the first "learn" stage, tasks are presented to the model in an easy-to-hard sequence to enable gradual learning and knowledge retention, so that their natural correlations could be exploited to facilitate knowledge transfer. In the second "review" stage, expert modules are incorporated to efficiently handle tasks with varying FoVs. This approach eliminates the need for retraining separate models and demonstrates state-of-the-art performance across different FoVs with strong generalization capabilities. Remarkably, the recall rate@top-1 improves from 49.1% to 68.3% and from 24.6% to 34.3% respectively on CVUSA and CVACT benchmarks with 70° FoV.

Code — https://github.com/Orange3stone/HCL-Geo.git

Introduction

Cross-view geo-localization aims to determine the location of a ground-view query image by matching it with reference geo-tagged satellite images. It has attracted wide attention for its potential in GPS-denied environments, including autonomous driving (Li et al. 2019), navigation (Shetty and Gao 2019), and augmented reality (Chiu et al. 2018). Despite advancements, this task remains challenging due to the significant appearance difference between the views. Most approaches (Shi et al. 2019; Yang, Lu, and Zhu 2021) frame this task as an image retrieval task, using siamese networks with metric learning to align features from different viewpoints. However, these methods assume panoramic images with known orientations, limiting their applicability to real-world scenarios where the field of view (FoV) of ground-view images is restricted and orientation is unknown.

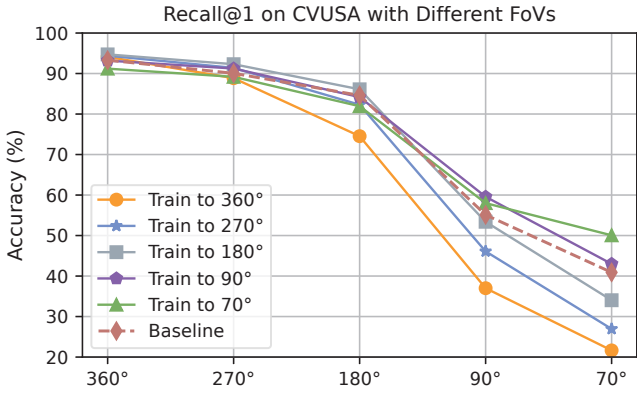

Figure 1: The ConvNext-Base network is sequentially trained on FoV ranging from easy to hard: 360° with north-aligned, 360° with unknown orientation, 270°, 180°, 90° and 70°. The "Baseline" indicates the best R@1 accuracy of the ConvNext-Base network when trained separately on each FoV.

To augment real-world geo-localization, DSM (Shi et al. 2020a) estimates the relative orientation of the ground images relative to aerial and performs orientation alignment in feature space. Subsequent improvement (Hu et al. 2022) refines orientation prediction methods and evaluation metrics. Although these techniques have successfully improved retrieval accuracy, they suffer from poor generalization across different FoVs. Furthermore, these methods generally treat cross-view geo-localization with different FoVs as independent tasks, thus requiring separate models or retraining to adapt to different FoVs. This does not align with practical applications, where the FoV of ground-view images varies dynamically and unpredictably. ConGeo (Mi et al. 2025) represents a recent effort to improve the robustness of the model to different FoVs and orientation variations through contrastive learning. By training a single model on panoramic with north-aligned and 180° street-view images with known orientation, ConGeo achieves significant improvements across different FoVs. However, it still struggles with cross-view geo-localization under narrow FoVs (e.g.,

*Corresponding author: Teng Wang.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

90°and 70°), where performance is still far from satisfactory for practical deployment.

We conjecture that the poor performances of existing methods under narrow FoVs are attributed to their learning paradigm, where cross-view geo-localization with different FoVs is considered as independent tasks. Actually, these tasks naturally exhibit strong correlations, allowing knowledge transfer among them. To validate this conjecture, we build a simple continual learning framework by implementing a vanilla progressive curriculum learning strategy (Bengio et al. 2009), as illustrated in fig. 1. The results demonstrate that curriculum learning effectively facilitates the knowledge transfer from prior tasks to subsequent ones while revealing a critical limitation of knowledge forgetting during the learning process. This highlights the challenges in the field of continual learning: "*Catastrophic Forgetting*". Researchers have long theorized that the principal human learning process starts with the acquisition of basic knowledge, progressively advances to more complex concepts by a teacher, and finally concludes with a review phase to reinforce and consolidate the learned knowledge. This motivates us to propose **HCL-Geo**, a two-stage **h**uman-like **c**ontinual **l**earning strategy of "first learn, then review" based on curriculum learning and the mixture of experts (MoE) architecture for cross-view **geo**-localization. In the first "learn" stage, tasks are presented to the model in an easy-to-hard sequence, enabling the model to effectively learn and inherit knowledge from the data flow. This process is further enhanced by the integration of knowledge distillation, which alleviates the issue of knowledge forgetting and facilitates the model in learning common knowledge across different tasks. In the second "review" stage, expert modules are integrated to learn discriminative knowledge across different tasks, thus enabling model to handle varying FoVs. The results of the experiment on benchmark datasets show that the single model achieves state-of-the-art performance across various FoVs. We summarize our contributions as follows:

- By leveraging the strong correlations between ground-view images with different FoVs, we are the first to leverage continual learning to enhance cross-view geo-localization under varying FoVs.
- We propose a two-stage human-like continual learning framework. The first knowledge distillation-augmented curriculum learning stage is proposed to effectively learn diverse knowledge, while the second stage, through "review", further strengthens the model's ability to distinguish between different tasks, enhancing its multitask processing capability.
- Through extensive experimentation, we demonstrate the effectiveness of our two-stage human-like learning framework. Under a range of commonly encountered FoV conditions, our model consistently achieves state-of-the-art performance.

Raleted Work

Cross-View Geo-localization

Cross-view geo-localization, which matches ground-view images to satellite images to determine location, is framed as a retrieval task (Lin et al. 2015; Hu et al. 2018; Cai et al. 2019). Early methods (Shi et al. 2019; Zhu et al. 2023) enhance these models by incorporating dual-branch CNNs and feature fusion modules. More recently, transformer-based networks have been adopted to enhance global feature extraction (Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022; Wang, Li, and Sun 2023). Other approaches involve using generative adversarial networks (GANs) for viewpoint translation (Lu et al. 2020; Regmi and Shah 2019). However, a key limitation of these methods is their assumption that ground-view images are panoramas with north-aligned. This restricts their real-world applicability, as real-world scenarios often involve ground-view images with limited FoVs and unknown orientations.

To address scenarios with limited FoVs and unknown orientations, several works (Shi et al. 2020a; Hu et al. 2022; Fervers et al. 2023) have focused on orientation prediction. DSM (Shi et al. 2020a) proposes a dynamic similarity matching module to predict relative orientation. Work (Hu et al. 2022) refines the orientation prediction task and evaluation metrics, achieving improved accuracy through fine-grained orientation prediction. These methods introduce orientation alignment into the retrieval task, successfully improving model performance in scenarios with limited FoVs and unknown orientations. However, in dynamic real-world environments, the generalization ability of these models and their robustness to variations in both FoV and orientation remain insufficient, requiring retraining to adapt to different FoVs. To the best of our knowledge, ConGeo (Mi et al. 2025) is the first work to consider training a single model to address the challenges of different FoVs. It uses a contrastive learning strategy to capture similarities and differences between panoramic (north-aligned) and limited FoV (unknown orientation) ground-view images, showing robustness to variations in both FoV and orientation. However, ConGeo's performance still remains insufficient for practical use when FoV is reduced to 90°or even 70°. Therefore, we propose a novel two-stage human-like continual learning training framework, which allows us to train a single model that demonstrates superior performance compared to state-of-the-art methods across commonly used FoVs, particularly under smaller FoVs (e.g., 90° and 70°), while also exhibiting improved generalization capabilities on different FoVs with unknown orientation.

Continual Learning

Continual Learning refers to the ability of the model to retain the capacity to learn new tasks while avoiding the forgetting of previously acquired knowledge. Storing exemplars from previous tasks is a simple and effective solution (Hou et al. 2019; Zhao et al. 2020), but it introduces concerns related to privacy and memory management. Therefore, some methods (Smith et al. 2021; Yu et al. 2020) employ regularization techniques to achieve exemplar-free continual learning. Another mainstream approach to enhancing continual learning is the expansion of the model architecture. For instance, Expert Gate (Aljundi, Chakravarty, and Tuytelaars 2017) proposes assigning a specialized " *expert*" to each task, thereby enabling multi-task learning capabilities.

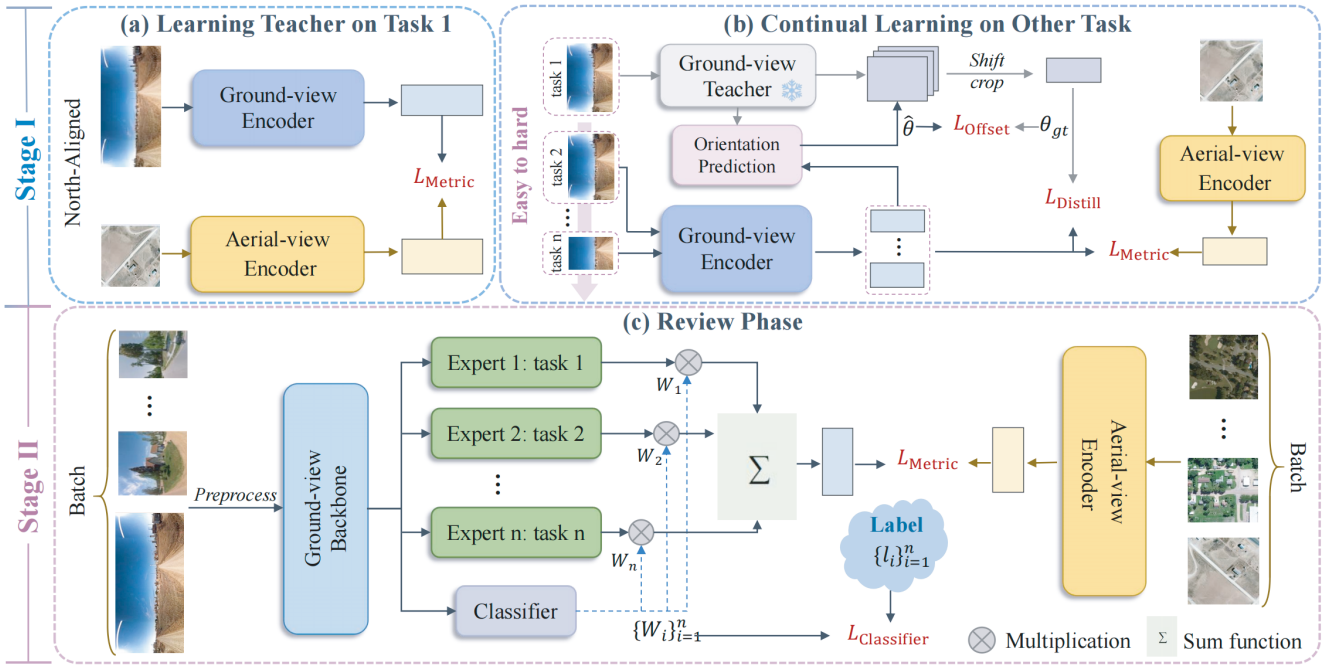

Figure 2: Our two-stage "first learn, then review" human-like learning framework is illustrated as follows: (a) shows the training process during the first stage on the simplest task, where the training architecture is similar to Sample4Geo (Deuser, Habel, and Oswald 2023). (b) depicts the subsequent tasks in the first stage, where knowledge distillation and orientation prediction are integrated into the training framework. The backbone module inherits the parameters from the previous task, while the teacher model remains frozen. (c) presents the training framework for the second "review" stage.

SEED (Rypeść et al. 2024) improves model stability and plasticity by selectively fine-tuning expert models and using multivariate Gaussian distributions. However, in cross-view geo-localization, the high correlation between scene images from different FoVs provides a natural advantage for leveraging continual learning without additional inter-task processing. To fully exploit this potential, combining the "class-incremental" knowledge transfer ability and the "ensemble expert" capability for handling multiple tasks, we propose a human-like learning strategy "first learn, then review" to enhance the model's efficiency in learning and inheriting knowledge across tasks with different FoVs.

Methodology

With regard to the cross-view geo-localization task, the smaller the FoV, the more challenging the localization task becomes. Actually, the geo-localization tasks with different FoVs naturally exhibit strong correlations, as a narrow FoV image can be regarded as a subimage of a panoramic view. To leverage this characteristic, we for the first time introduce continual learning into cross-view geo-localization to address the challenges posed by limited FoVs . We propose the HCL-Geo training framework, an easy-to-hard learning procedure followed by a subsequent review phase as shown in fig. 2. Below, we will provide a detailed introduction to our continual learning framework.

Backbone Model

In this work, we follow previous works (Deuser, Habel, and Oswald 2023; Mi et al. 2025) to employ ConvNext-Base (Liu et al. 2022) as the backbone network for feature extraction. It follows a hierarchical multi-stage design, comprising four sequential stages that progressively reduce spatial resolution while expanding channel dimensions, and the final output is generated by a head block.

Stage I: Easy-to-Hard Learning Strategy

Inspired by curriculum learning (Bengio et al. 2009), we sequentially present tasks from easy to hard to the model without repetition, leveraging easy tasks to enhance the understanding of hard tasks and perceiving commonalities between different tasks, thereby improving the model's ability to handle difficult tasks. During training, we divide the tasks into four levels, ranked from easiest to hardest: panoramic images with north-aligned, panoramic images with unknown orientation, 180° images with unknown orientation, and 70° images with unknown orientation.

In this stage, to facilitate the model's learning of difficult tasks, we leverage knowledge distillation from a teacher model. As shown in fig. 2 (b), we view the model trained under the first task characterized by the panoramic image with north-aligned as the teacher model, since it could provide full semantic information of the location while simultaneously mitigating the forgetting of initially learned tasks. It should be noted that the teacher supervision is applied only

to the ground-view image encoder, since FoV variations occur solely in ground-view images while the satellite database remains invariant. When training subsequent tasks, the student model inherits the trained parameters of the previous task and takes the current task's FoV images as input, while receiving supervision from the teacher model.

However, the panoramic images from the teacher model and the limited FoV images from the student models are not perfectly orientation-aligned, posing a challenge to implement teacher supervision. To enhance the model's orientation perception, we introduce an orientation prediction module to the student model. The output of the student model, after being fused with the teacher model's output through cross-attention, is passed through three fully connected layers, each followed by a GELU activation layer, and finally processed by a Sigmoid activation function to predict the orientation $\hat{\theta}$. The predicted $\hat{\theta}$ is used to offset the teacher model's pre-pooling features. The features are then cropped to match the FoV of the student model and used for teacher supervision. To be specific, for the first task, given the input ground-view image I_g and satellite image I_a, the loss function can be represented as follows:

$$z_g = f_g(I_g), z_a = f_a(I_a), \tag{1}$$

$$\mathcal{L}(z_g, z_s)_{\text{InfoNCE}} = -\log \frac{\exp\left(z_g \cdot z_a^+ / \tau\right))}{\sum_{i=1}^{B} \exp\left(z_g \cdot z_a^i / \tau\right)\Big)}, \tag{2}$$

where f_g and f_a represent the ground image encoder and satellite image encoder, respectively. z_g and z_a represent the features of ground-view and satellite images, and z_a^+ denotes the positive sample for ground-view image. $(\cdot)$ indicates the dot product operation. The temperature parameter τ is a hyperparameter. The loss for subsequent tasks can be expressed as:

$$z_g^s, \hat{\theta} = f_g^s(I_g); z_g^t = f_g^t(I_g^{pano}; \hat{\theta}); z_a = f_a(I_a), \tag{3}$$

$$\mathcal{L}_{\text{Metric}} = \mathcal{L}(z_g^s, z_a)_{\text{InfoNCE}}, \tag{4}$$

$$\mathcal{L}_{\text{Distillation}} = \mathcal{L}(z_g^s, z_g^t)_{\text{InfoNCE}}, \tag{5}$$

$$\mathcal{L}_{\text{Offset}} = \begin{cases} \left(\hat{\theta} - \theta_{gt}\right)^2 / 2 & \text{if } \left|\hat{\theta} - \theta_{gt}\right| < 1 \\ \left|\hat{\theta} - \theta_{gt}\right| - 0.5 & \text{otherwise} \end{cases}, \tag{6}$$

$$\mathcal{L}_{\text{first-stage}} = \mathcal{L}_{\text{Metric}} + \alpha_1 \mathcal{L}_{\text{Distillation}} + \alpha_2 \mathcal{L}_{\text{Offset}}, \tag{7}$$

where f_g^s and f_g^t represent the student model and teacher model for encoding ground-level images, respectively. I_g^{pano} denotes the panoramic image with north-aligned, and θ_{gt} represents the ground truth value for the orientation offset. The hyperparameters α_1 and α_2 are used to control the weights of different loss.

Stage II: Mixture-of-Experts-Based Review

For human skill learning, the "review" is an essential stage for consolidating and summarizing knowledge. This inspires us to propose a review phase to further enhance the model's ability to inherit and differentiate multi-task knowledge. We implement the review procedure by adopting the MoE structure (Aljundi, Chakravarty, and Tuytelaars 2017). To fully exploit the knowledge learned from stage I, we remove the orientation prediction module from the student model, and split the ConvNeXt backbone model into two parts. As shown in fig. 2 (c), the model's final block and head block serve as the expert structure, while all preceding blocks constitute the shared backbone network. This expert structure is duplicated into four copies, each corresponding to one of the four predefined tasks. To avoid introducing restrictions on deployment scenarios, we assume no prior knowledge of the FoV or orientation of the query street-view image. This requires a single model to infer effectively across different scenarios. Therefore, we add a fully connected module on top of the expert structure as a classifier to identify the category of the task, and the output logit is treated as weight for the corresponding expert. To effectively supervise the classification, we define a one-shot label for each task and employ cross-entropy loss for supervision. Specifically, given an input I_g and I_a, the model in the second stage outputs the following:

$$w_i = \text{Classfier}(\phi(I_g)) \tag{8}$$

$$z_g = \sum_{i=1}^{n=4} w_i E_i(\phi(I_g)); z_a = f_a(I_a) \tag{9}$$

where ϕ represents the backbone of the ground-view encoder, Classifier denotes the classifier module, and E_i and w_i represent the expert for the i-th task and the weight of its feature, respectively. The loss supervision can be formulated as follows:

$$\mathcal{L}_{\text{Classifier}} = \sum_{i=1}^{n=4} (l_i \log w_i), \tag{10}$$

$$\mathcal{L}_{\text{Metric}} = \mathcal{L}(z_g, z_a)_{\text{InfoNCE}}, \tag{11}$$

$$\mathcal{L}_{\text{second-stage}} = \mathcal{L}_{\text{Metric}} + \beta \mathcal{L}_{\text{Classifier}}, \tag{12}$$

where l_i is the ground truth label of input I_g for the i-th task, and β is a hyperparameter used to balance the loss.

Experiment

Datasets

CVUSA (Zhai et al. 2017) and CVACT (Liu and Li 2019) are widely used benchmarks for cross-view geo-localization, featuring one-to-one matching between panoramic street-view images and directionally aligned aerial images. And the VIGOR (Zhu, Yang, and Chen 2021) introduces a more challenging setting, where each query image is associated with one positive reference and three semi-positive references. The dataset is divided into training and testing sets based on two configurations: same-area and cross-area. Performance is evaluated using recall accuracy at recall-k (R@k), with metrics reported for top-1, top-5,and top 1%.

Experiment Details

Our models are trained with the AdamW optimizer with an initial learning rate of 0.0001 and a cosine learning rate schedule. During the "first learn" stage, we train for 40 epochs on each predefined task. The loss weight α_1 and α_2 are set to 0.01 and 0.01, respectively. In the "review" stage, a total of 80 epochs are performed and loss weight

Set	Method	FoV=360°		FoV=270°		FoV=180°		FoV=90°		FoV=70°	
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CVSUA	CVFT	23.4	44.4	-	-	8.1	24.3	4.8	14.8	3.8	12.4
	DSM	78.1	89.5	-	-	48.5	68.5	16.2	31.4	8.8	19.9
	SEH	85.4	93.5	-	-	53.7	72.3	16.6	32.2	7.8	18.8
	L2LTR	-	-	-	-	56.7	80.9	26.9	50.5	14.0	33.1
	SAIG-D	72.0	90.2	-	-	52.5	78.1	26.7	50.2	20.9	41.4
	Sample4Geo	93.3	97.5	-	-	84.6	95.9	55.1	**78.3**	40.9	65.4
	ConGeo	**96.6**	**98.9**	-	-	**92.3**	**97.9**	**55.5**	75.4	**49.1**	**70.8**
	ConGeo	85.2	95.1	86.0	95.1	92.3	**97.9**	55.9	73.2	37.1	55.7
	HCL-Geo	**97.6**	**99.1**	**95.9**	**98.6**	**92.4**	97.4	**75.2**	**89.5**	**68.3**	**85.5**
CVACT_val	CVFT	26.8	46.9	-	-	7.1	18.5	1.9	6.3	1.5	5.1
	SEH	77.4	88.6	-	-	47.7	67.9	13.9	28.4	6.9	16.5
	DSM	72.9	85.7	-	-	49.1	67.8	18.1	33.3	8.3	20.7
	Sample4Geo	82.4	**90.6**	-	-	58.9	79.8	27.9	52.0	18.8	40.4
	ConGeo	**83.0**	**90.6**	-	-	**70.3**	**85.2**	**40.6**	**62.6**	**24.6**	**45.3**
	ConGeo	62.6	79.9	67.5	80.1	**70.3**	**85.2**	34.8	56.9	18.5	37.4
	HCL-Geo	**80.3**	**88.3**	**73.8**	**84.4**	66.4	79.7	**40.6**	**60.9**	**34.3**	**55.6**

Table 1: Comparison on CVUSA and CVACT datasets with the unknown orientation and limited FoV setting. The performance of the single model is indicated with a gray background, while the performance of models trained on specific FoVs is shown with a white background. **Bold** denotes the best performance.

Set	Method	FoV=360°		FoV=90°	
		R@1	R@1%	R@1	R@1%
Same-Area	TransGeo	47.7	99.3	-	-
	Sample4Geo	14.2	54.9	1.1	30.6
	ConGeo	61.9	**98.4**	8.5	68.7
	HCL-Geo	**69.3**	97.8	**30.4**	**93.0**
Cross-Area	TransGeo	5.5	66.9	-	
	Sample4Geo	9.0	43.7	0.5	21.6
	ConGeo	16.2	72.9	3.9	54.3
	HCL-Geo	**29.8**	**85.8**	**4.7**	**58.2**

Table 2: Comparison on the unknown orientations and limited FoV setting on VIGOR dataset.

β is set to 0.01. To reduce training parameters and accelerate convergence, the ground and aerial-view encoders share the identical architecture with parameter sharing. In the "review" stage, to enhance feature diversity, the final output of the aerial-view encoder is obtained by averaging all expert outputs without classifier weights.

Comparison with State-of-the-art Methods

We evaluate HCL-Geo on three cross-view geo-localization datasets: CVUSA, CVACT, and VIGOR, compared with prior state-of-the-art methods (Shi et al. 2020b,a; Yang, Lu, and Zhu 2021; Guo et al. 2022; Zhu et al. 2023; Deuser, Habel, and Oswald 2023; Mi et al. 2025) under conditions of unknown orientation and limited FoV. It should be noted our method and ConGeo (Mi et al. 2025) only require training one single model for different FoVs, whereas other methods require retraining models. As shown in table 1, our method achieves the best results for the key metric R@1 and R@5. Remarkably, significant improvements are observed under narrow FoVs of 90° and 70°. Specifically, on the CVUSA dataset, our approach improves R@1 from 55.9% to 75.2% (+19.3%) and from 37.1% to 68.3% (+31.2%) respectively for FoVs of 90° and 70° compared to the runner-up ConGeo. Similarly, on the CVACT dataset, it improves at R@1 by a margin of 5.8% and 15.8%, respectively, for FoVs of 90° and 70°. As shown in table 2, on the VIGOR dataset, we improve R@1 by a margin of 21.9% for 90° FoV under the "Same-Area" setting. Furthermore, we surpass ConGeo by 13.6% at R@1 for the 360° FoV under the "Cross-Area" setting. In summary, HCL-Geo not only eliminates the need for retraining under different settings but also exhibits superior performance under limited FoVs. This superiority can be attributed to our two-stage learning paradigm—first learn, then review, which encourages the model to inherit the shared characteristics across different FoVs while simultaneously recognizing their distinctive features.

Generalization of the Model

In table 1, HCL-Geo is trained on north-aligned, 360°, 180°, and 70° FoVs and tested on 360°, 270°, 180°, 90°, and 70° FoVs. It not only demonstrates superior performance on small FoVs but also exhibits robust generalization capabilities at 270° and 90°. Even on the unseen 90° FoV, it still achieves state-of-the-art performance. To further systematically evaluate the model's generalization ability, we test HCL-Geo across a series of FoV ranges: [360°-270°), [270°-180°), [180°-90°), [90°-70°), [70°-45°). During testing, the ground-level images are randomly cropped to an arbitrary FoV within their corresponding range. As illustrated in fig. 3, HCL-Geo consistently outperforms ConGeo (Mi

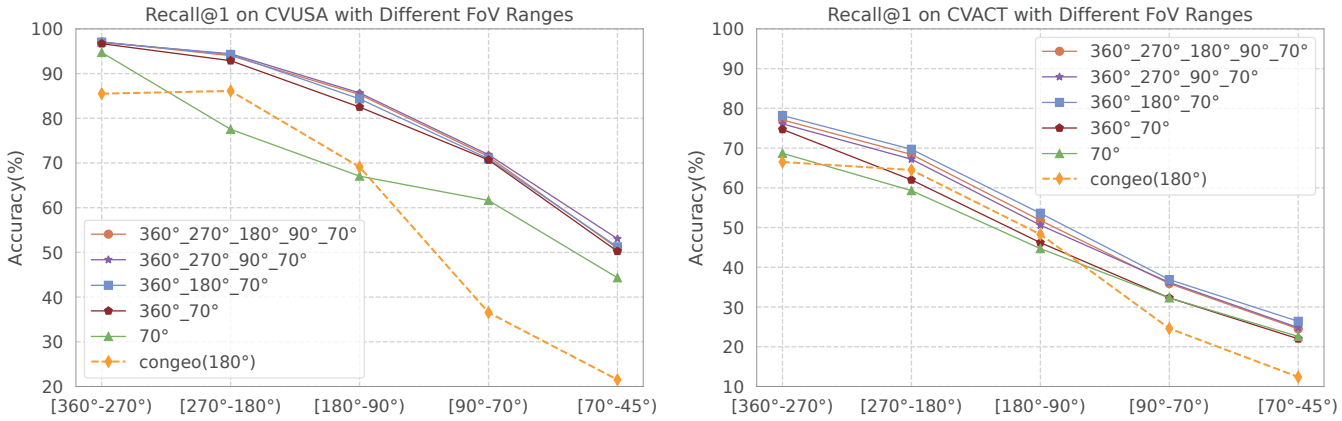

Figure 3: Comparison of generalization capabilities across different FoV ranges under varying numbers of training FoVs.

Method	FoV=360°			FoV=180°			FoV=90°			FoV=70°			Avg.
	R@1	R@5	R@1%	R@1	R@5	R@1%	R@1	R@5	R@1%	R@1	R@5	R@1%	R@1
ConGeo (360°)	96.6	98.9	**99.7**	83.8	94.2	98.8	38.2	58.2	82.0	19.5	35.9	66.5	59.5
ConGeo (270°)	70.4	86.6	97.4	77.9	91.7	98.2	44.9	66.1	88.1	28.1	48.9	79.6	55.3
ConGeo (180°)	85.2	95.1	98.9	**92.3**	97.9	**99.7**	55.9	73.2	90.9	37.1	55.7	81.4	67.6
ConGeo (90°)	81.4	93.0	98.6	92.2	**98.1**	**99.7**	55.5	75.4	93.9	35.5	54.8	81.0	66.2
HCL-Geo (360°)	**97.4**	99.0	99.5	89.0	96.2	98.7	60.1	77.4	92.1	32.4	50.4	76.7	69.7
HCL-Geo (270°)	**97.4**	**99.2**	**99.7**	88.4	96.0	98.6	52.9	71.5	88.0	24.7	42.7	70.0	65.9
HCL-Geo (180°)	97.2	99.1	99.6	92.2	97.5	99.2	47.7	68.6	89.5	22.2	40.6	72.1	64.8
HCL-Geo (90°)	88.1	94.4	97.8	86.4	94.9	99.2	**71.8**	**87.7**	**97.0**	48.9	70.9	91.0	**73.8**
HCL-Geo (70°)	80.1	89.4	94.9	85.3	93.6	97.6	67.4	85.3	96.2	**61.5**	**81.1**	**95.5**	73.4

Table 3: Comparison of the model's performance under different training tasks for orientation unknown images with FoV angle of 360°, 180°, 90°, and 70°on the CVUSA dataset. "Avg" denotes the average performance across these four test tasks.

et al. 2025) in different FoV ranges. Notably, while ConGeo exhibits significant performance degradation at 90° and 70° FoVs, HCL-Geo demonstrates a more gradual decline. This observation highlights HCL-Geo's superior robustness and generalization capability under varying FoV ranges. Both HCL-Geo and ConGeo aim to address FoV robustness by single model. The table 3 provides a more fair comparison of their performance when trained under the same FoV. It can be observed that HCL-Geo utilizes small-angle datasets more effectively for training compared to ConGeo, thereby improving performance within small-angle ranges (HCL-Geo (90°) vs. ConGeo (90°)). We speculate that this is because, in the more challenging small-FoV scenarios, contrastive learning forces the features of limited FoV and panoramic images to be similar, making it difficult to learn effective features. In contrast, HCL-Geo, through its step-by-step learning and review stage, is more conducive to learning discriminative features.

Ablation of Training Tasks

While more tasks intuitively boost performance, HCL-Geo achieves optimal generalization with just a few well-designed tasks, thanks to the strong correlation between FoVs and its continual learning strategy. The fig. 3 shows HCL-Geo's generalization ability with varying training task numbers (optimal FoV combinations). Notably, performance improves notably from two to four training tasks, but plateaus at six tasks, suggesting that HCL-Geo does not require an excessive number of training tasks. This demonstrates HCL-Geo's efficiency - its training scheme leverages strong FoV correlations to achieve broad generalization with a limited set of tasks. Considering the trade-off between efficiency and performance, we train HCL-Geo with four different FoVs. In fig. 4, we statistically analyze the one-shot classification results and the average weight obtained by each expert when testing models trained on 360° with north-aligned, 360°, 180°, and 70°. This confirms that HCL-Geo can effectively utilize experts with the closest FoV during training for processing with larger weights during testing.

Ablation of Model's Component

We conduct an ablation study to validate the effectiveness of our proposed framework. First, we establish two baselines: Baseline-1 employs only the easy-to-hard data stream for one-stage continual learning. Baseline-2 represents our MoE structure framework. Notably, as shown in table 4, results in the first and second rows show that both baselines outperform existing methods on the smaller FoVs, highlighting the effectiveness of continual learning for this task. We further combine Baseline 1 and Baseline 2 to simplify our

Ablation	Stage I			Stage II	North-aligned	FoV = 360°	FoV = 180°	FoV = 90°	FoV = 70°
	Strat.	KD	Ori.		R@1(%)	R@1(%)	R@1(%)	R@1(%)	R@1(%)
Baseline-1	✓	×	×	×	91.5	91.2	81.8	58.7	49.1
Baseline-2	×	×	×	✓	92.4	92.1	82.6	62.2	54.5
HCL-Geo	✓	×	×	✓	94.1	94.1	84.2	63.0	55.7
HCL-Geo	✓	✓	×	✓	97.6	97.3	91.6	73.8	67.2
HCL-Geo	✓	✓	✓	✓	**97.9**	**97.6**	**92.4**	**75.2**	**68.3**

Table 4: In various settings of the CVUSA dataset, it demonstrates the effectiveness of the main components of our model. "Strat." denotes our easy-to-hard curriculum learning strategy, "KD" represents knowledge distillation, and "Ori." indicates the orientation prediction module.

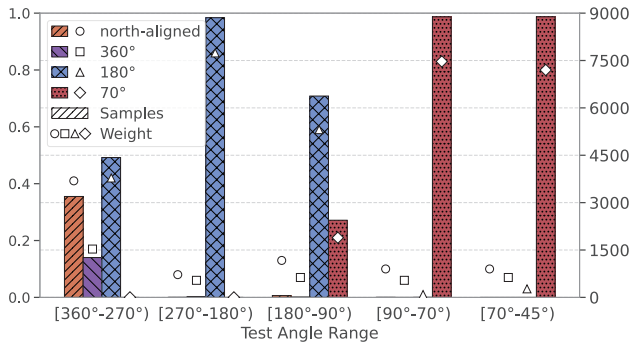

Figure 4: Results of expert classification across different test ranges are presented. Bars denote the average classification weight for each expert over varying test angle ranges (left-Y), whereas markers signify the corresponding number of classified samples (right-Y).

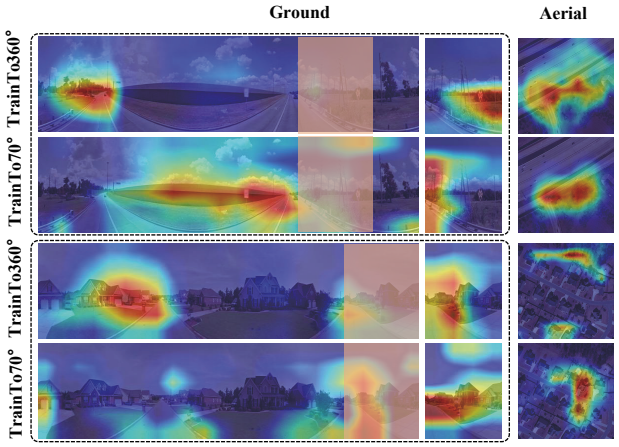

Figure 5: Visualization of regions of interest for the "TrainedTo_360°" and "TrainTo_70°" models across 360° ground-level panoramas with known orientation, 70° ground-level image, and aerial image. Pink shaded portion highlight cropped areas corresponding to the 70° FoV.

two-stage training implementation. The results (third row) reveal significant performance improvements across all settings. Finally, we progressively introduce knowledge distillation and orientation prediction into the above framework. The last two rows confirm their effectiveness. Particularly, we gain substantial improvement by adding knowledge distillation. This indicates that the introduction of knowledge distillation not only effectively mitigates knowledge forgetting but also provides meaningful guidance for new tasks due to the strong correlation between tasks.

Analysis

To explain the performance enhancement of cross-view geo-localization through continual learning, we visualize the model's attention regions. The fig. 5 shows visualization results of two models: the "TrainedTo_360°" model trained until 360° FoV and the "TrainedTo_70" model trained until reaching 70° FoV, evaluated under 360° FoV with known orientation and 70° FoV. The "TrainedTo_70" model demonstrates expanded attention coverage on 360° ground image during continual learning, with consistent feature alignment in satellite image, while maintaining precise focus on geolocation-critical patterns under 70° conditions. This improvement is attributed to the continual learning paradigm where ground images with different FoVs from same locations are used during training, enabling extraction of fine-grained discriminative features. These results confirm that continual learning effectively exploits the correlation between varying FoV tasks to enhance performance under limited FoV conditions.

Conclusion

Training a single model to handle real-world scenarios with limited FoV and unknown orientation is crucial for advancing the practical application of cross-view geo-localization. we propose HCL-Geo, a two-stage human-like continual learning framework that incorporates knowledge distillation and orientation prediction. HCL-Geo significantly boost performance across different FoVs and enhances robustness to FoV and orientation variations. However, HCL-Geo has two main limitations: its computational cost scales with the number of training FoVs, necessitating a cost-performance trade-off; and while knowledge distillation effectively mitigates catastrophic forgetting, further improvements are needed to fully resolve this issue.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62273093, 62236002).

References

Aljundi, R.; Chakravarty, P.; and Tuytelaars, T. 2017. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3366–3375.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Cai, S.; Guo, Y.; Khan, S.; Hu, J.; and Wen, G. 2019. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8391–8400.

Chiu, H.-P.; Murali, V.; Villamil, R.; Kessler, G. D.; Samarasekera, S.; and Kumar, R. 2018. Augmented reality driving using semantic geo-registration. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 423–430. IEEE.

Deuser, F.; Habel, K.; and Oswald, N. 2023. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16847–16856.

Fervers, F.; Bullinger, S.; Bodensteiner, C.; Arens, M.; and Stiefelhagen, R. 2023. C-BEV: Contrastive Bird's Eye View Training for Cross-View Image Retrieval and 3-DoF Pose Estimation. *arXiv preprint arXiv:2312.08060*.

Guo, Y.; Choi, M.; Li, K.; Boussaid, F.; and Bennamoun, M. 2022. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE transactions on image processing*, 31: 2094–2105.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 831–839.

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

Hu, W.; Zhang, Y.; Liang, Y.; Yin, Y.; Georgescu, A.; Tran, A.; Kruppa, H.; Ng, S.-K.; and Zimmermann, R. 2022. Beyond geo-localization: Fine-grained orientation of street-view images by cross-view matching with satellite imagery. In *Proceedings of the 30th ACM International Conference on Multimedia*, 6155–6164.

Li, A.; Hu, H.; Mirowski, P.; and Farajtabar, M. 2019. Cross-view policy learning for street navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8100–8109.

Lin, T.-Y.; Cui, Y.; Belongie, S.; and Hays, J. 2015. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5007–5015.

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

Liu, Z.; Mao, H.; Wu, C.-Y.; Feichtenhofer, C.; Darrell, T.; and Xie, S. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986.

Lu, X.; Li, Z.; Cui, Z.; Oswald, M. R.; Pollefeys, M.; and Qin, R. 2020. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 859–867.

Mi, L.; Xu, C.; Castillo-Navarro, J.; Montariol, S.; Yang, W.; Bosselut, A.; and Tuia, D. 2025. Congeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, 214–230. Springer.

Regmi, K.; and Shah, M. 2019. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 470–479.

Rypeść, G.; Cygert, S.; Khan, V.; Trzciński, T.; Zieliński, B.; and Twardowski, B. 2024. Divide and not forget: Ensemble of selectively trained experts in Continual Learning. *arXiv preprint arXiv:2401.10191*.

Shetty, A.; and Gao, G. X. 2019. Uav pose estimation using cross-view geolocalization with satellite imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, 1827–1833. IEEE.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32.

Shi, Y.; Yu, X.; Campbell, D.; and Li, H. 2020a. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4064–4072.

Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; and Li, H. 2020b. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11990–11997.

Smith, J.; Hsu, Y.-C.; Balloch, J.; Shen, Y.; Jin, H.; and Kira, Z. 2021. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9374–9384.

Wang, T.; Li, J.; and Sun, C. 2023. Dehi: A decoupled hierarchical architecture for unaligned ground-to-aerial geolocalization. *IEEE Transactions on Circuits and Systems for Video Technology*.

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

Yu, L.; Twardowski, B.; Liu, X.; Herranz, L.; Wang, K.; Cheng, Y.; Jui, S.; and Weijer, J. v. d. 2020. Semantic drift

compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6982–6991.

Zhai, M.; Bessinger, Z.; Workman, S.; and Jacobs, N. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 867–875.

Zhao, B.; Xiao, X.; Gan, G.; Zhang, B.; and Xia, S.-T. 2020. Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13208–13217.

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

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

Zhu, Y.; Yang, H.; Lu, Y.; and Huang, Q. 2023. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*.