ConGeo: Robust Cross-view Geo-localization across Ground View Variations

Li Mi[1*], Chang Xu[2,1*†], Javiera Castillo-Navarro[1], Syrielle Montariol[1], Wen Yang[2], Antoine Bosselut[1], and Devis Tuia[1]

[1] EPFL
[2] Wuhan University
https://eceo-epfl.github.io/ConGeo/

Abstract. Cross-view geo-localization aims at localizing a ground-level query image by matching it to its corresponding geo-referenced aerial view. In real-world scenarios, the task requires accommodating diverse ground images captured by users with varying orientations and reduced field of views (FoVs). However, existing learning pipelines are orientation-specific or FoV-specific, demanding separate model training for different ground view variations. Such models heavily depend on the North-aligned spatial correspondence and predefined FoVs in the training data, compromising their robustness across different settings. To tackle this challenge, we propose **ConGeo**, a single- and cross-view **Con**trastive method for **Geo**-localization: it enhances robustness and consistency in feature representations to improve a model's invariance to orientation and its resilience to FoV variations, by enforcing proximity between ground view variations of the same location. As a generic learning objective for cross-view geo-localization, when integrated into state-of-the-art pipelines, ConGeo significantly boosts the performance of three base models on four geo-localization benchmarks for diverse ground view variations and outperforms competing methods that train separate models for each ground view variation.

1 Introduction

Given an image captured at ground level, cross-view geo-localization (CVGL) aims to determine the location of the image by referring to its corresponding aerial view [4,12,20,21,34]. As an important auxiliary positioning technique, the task enables noisy GPS correction [30] and offers practical applications in fields such as navigation [10] and autonomous driving [5].

Cross-view geo-localization is often addressed as a retrieval task, where a ground-level image acts as the query and a geo-tagged overhead image as the reference. In real-world applications, it often requires a high generalization ability, to handle diverse ground view image variations, including varying orientations and reduced field of view (FoV). In the past, models were trained and evaluated under idealised settings, where both views are aligned to the North [13,26,36]

* Equal contribution † Corresponding author (xuchangeis@whu.edu.cn)

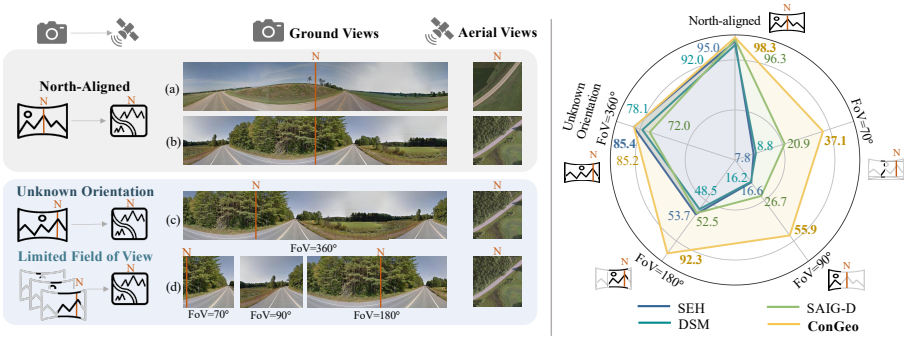

Fig. 1: ConGeo boosts the robustness across ground view variations: North-aligned, unknown orientation (FoV=360°) and limited field of views (FoV=70°, 90°, and 180°). We compare with SEH [6], DSM [21] and SAIG-D [37] and report Top-1 Recall on the CVUSA [26] dataset, one of the geo-localization benchmarks.

which only covers limited scenarios in the real-world challenges (see **North-aligned** setting in the left panel of Fig. 1, rows (a) and (b)). Recent works [21,37] extend existing North-aligned datasets to more challenging settings, cyclically shifting a ground view panorama by a random angle to obtain orientation variations, or further reducing the FoV from 360° to 70°, 90°, or 180° to simulate limited ground view information (see left panel of Fig. 1, **Unknown Orientation** setting in rows (c) and **Limited FoV** setting in row (d)). However, they train and evaluate models separately for each setting. This limits the generalization ability of the model in real-world applications, where the FoV and orientation of the query image are often unknown. Moreover, the orientation-specific or FoV-specific training prevents the model from representing robust features across different settings. Instead, the models are often biased toward training data to capture specific spatial correspondence that is not generalizable across ground view variations. For example, in the left panel of Fig. 1 (rows (a) and (b)), the two aerial views share similar road directions, which can also be found in the corresponding ground views. Our analysis shows such North-aligned spatial correspondence in the training data can serve as a shortcut for the model to improve the specific setting but may sacrifice the orientation invariance and feature consistency towards ground view variations.

To address this challenge, we introduce **ConGeo**, a model-agnostic **Con**trastive learning pipeline for cross-view **Geo**-localization. ConGeo follows the intuition that a robust model should capture consistent features for the same location regardless of orientations or FoVs and identify the same reference image across ground view variations. To achieve that, we design two contrastive objectives to enforce the proximity between ground view variations and their original representation: a single-view contrastive objective and a cross-view one. The ground view contrastive loss works to align ground view variants with the original images (North-aligned, full panorama), and the aerial view contrastive

loss minimizes the representation disparity between the aerial image and its augmented counterpart. The cross-view (ground view and aerial view) contrastive loss further aligns the query image variants with reference aerial images. By disrupting the initial geometric correspondence and mitigating potential shortcuts found in the training data, ConGeo compels the model to focus on learning coherent features across different modalities and view variations.

We run extensive experiments on four geo-localization benchmarks, demonstrating that ConGeo outperforms existing methods across ground view variations. Moreover, our results highlight three insights: **(1)** ConGeo empowers a single model to handle various ground view variations, using a model-agnostic learning pipeline. ConGeo consistently outperforms comparison methods under orientation-specific or FoV-specific training by a large margin when facing ground view orientation and FoV shifts (The right panel of Fig. 1, Table 1). Meanwhile, ConGeo demonstrates strong versatility, as it can be plugged into different geo-localization pipelines (Table 5). **(2)** We demonstrate the advantages of the proposed contrastive learning pipeline over targeted data augmentation when facing ground view variations (Table 3), especially unseen ones (Table 8). **(3)** Our analysis reveals that the trade-off between the model's focus on geometric or semantic cues for matching images from different views affects its robustness among variations (Fig. 5 and 6).

2 Related Works

2.1 Cross-view Geo-localization

Cross-view geo-localization has been an active field of research over the last decade [4, 12, 20, 34]. We can distinguish two categories of works: first is **geo-localization with North alignment**, the classical evaluation setting for cross-view image geo-localization. The standard approach uses a siamese network [2] to encode the ground view and the aerial images and optimizes a well-designed loss to align the feature embeddings. A series of works pointed out the importance of spatial correspondence in paired images and incorporated such kind of prior into the pipeline; SAFA [20] designed a polar transformation that explicitly aligns two domains and used spatial attention to facilitate network learning; Liu et al. [13] encoded the orientation correspondence between views as additional input for the network, allowing the model to be orientation-discriminative. Recent studies [4, 34, 37] use more advanced architectures to model global relationships between views, TransGeo [34] uses learnable position encoding that implicitly learns the cross-view correspondence, Sample4Geo [4] uses the ConvNeXt [14] as feature extractor and designs GPS-based and similarity-based samplers for hard negative sample mining.

Second is **geo-localization with variations in orientation and FoV**: compared to strictly North-aligned panoramas, ground images with unknown orientations and limited FoV (*e.g.*, captured by a smartphone or car camera) are more readily available, but make the task more challenging. Recent studies [19, 21, 35] gradually draw attention to this more realistic scenario. Specifically, DSM [21]

proposed to crop out the query-activated region in the reference image after orientation estimation for accurate matching. Rodrigues et al. [19] customized a pipeline for the retrieval under limited FoVs, where the limited FoV images are cropped from the original image as a form of data augmentation for both views. Despite the steady progress, these methods still face limitations: First, specific FoV information is required for aerial image/feature cropping during both training and testing, while in real-world data this strong prior is unknown, especially during testing. Second, models trained on a specific setting cannot generalize and do not perform satisfactorily on all FoV scenarios. Thus several specialised models are trained for each setting.

Instead, we propose a method that does not rely on the exact FoV information during training nor testing, and only needs to be trained end-to-end once to achieve competitive results in all settings (see the right panel of Fig. 1).

2.2 Contrastive Learning for Geo-localization

Contrastive learning has shown impressive performance for self-supervised and supervised learning in computer vision [3, 7, 9, 23], including image classification [24], object detection [28], and vision-language tasks [11]. Existing methods use either cross-modal objectives to align different modalities (*e.g.*, CLIP [18]) or single-modal ones to enhance single-modal representation (*e.g.*, SLIP [15]). In CVGL, the triplet loss [8, 13, 17, 21, 34] has been a standard choice to solve the cross-view image retrieval problem in a contrastive way. More recently, InfoNCE [4, 16, 37] has shown to be an efficient objective for the task. However, contrastive learning in these works is limited to aligning feature embeddings between the vanilla ground view and aerial images to match image pairs; while it has shown a powerful ability for building view-invariant and robust representations [3, 7, 23]. In this work, we further leverage contrastive learning by using both single- and cross-view contrastive objectives, achieving more robust feature representations shared by ground view variants.

3 ConGeo

To obtain robust representation across diverse ground view variations, we introduce ConGeo, a contrastive learning objective that aims to enhance the robustness of geo-localization models by enforcing proximity between ground view variations and their original representations. The proposed ConGeo is model-agnostic and can be a learning objective for different base models. Here we use Sample4Geo [4] as the base model to illustrate the proposed method (Fig. 2).

3.1 Overview

Problem Statement. Let $\{I_q\}$ be a set of query images (North-aligned ground view), and let $\{I_r\}$ be a set of reference images (aerial view). For geo-localization

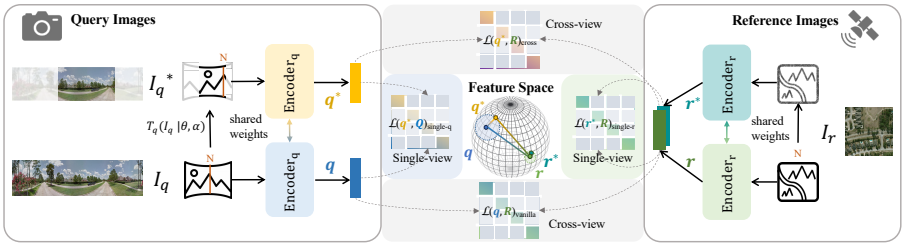

Fig. 2: ConGeo's learning pipeline. For feature representation in the left and right boxes, the North-aligned ground image (I_q), the transformed ground image (I_q^*), and the aerial view (I_r) are sent to their respective encoders. Then in the feature space, the single- and cross-view contrastive learning losses are applied to enforce the proximity of the paired images.

under arbitrary orientation and limited FoVs, the query image I_q can be cyclically shifted with a random angle and restricted by a specific FoV (*e.g.*, 70°, 90°, 180°, and 360°), resulting in its variation I_q^* that still shows the same location. For each I_q or I_q^*, the cross-view geo-localization aims to retrieve the corresponding reference image in $\{I_r\}$.

Architecture. The classic architecture for geo-localization consists of a siamese network, based on Convolutional Neural Network (CNN) or Vision Transformer (ViT) for feature encoding [4,20,21,34]. More specifically, a set of query features $Q := \{q\}$ is obtained by passing query images into the query encoder: $q = \text{Encoder}_\text{q}(I_q)$. Similarly, the set of reference features $R := \{r\}$ is obtained by $r = \text{Encoder}_\text{r}(I_r)$. To boost the model's robustness, ConGeo includes ground image variations $\{I_q{}^*\}$. Such variations are obtained by applying a transformation T to each ground view I_q. A random orientation angle θ is first used to horizontally shift the ground image, followed by the application of a FoV angle α to crop the panorama query to a FoV query $I_q{}^* = T_q(I_q|\theta, \alpha)$.

Model Training and Inference. During training, the dual encoders are learned according to the ConGeo learning objectives (see below). During inference, depending on the setting (North-aligned, unknown orientation, and limited FoV), the specific query image is processed by the ground view encoder. Then, a set of aerial reference images is ranked as retrieval results based on the cosine similarity between the features of the query and reference images.

3.2 Learning Objectives

Single-view Contrastive Learning. To enhance the consistency in feature representations between views across orientations and FoV variations, a single-view contrastive learning objective is designed to produce similar representations between the original images (North-aligned, full panorama) and their transformed counterpart. The loss for the ground view is computed as:

$$\mathcal{L}(q^*, Q)_{\text{single-q}} = -\log \frac{\exp\left(q^* \cdot q_+/\tau_q\right)}{\sum_{q_i \in Q} \exp\left(q^* \cdot q_i/\tau_q\right)}, \tag{1}$$

where q_i, q^* are the feature embeddings of the original and transformed query image respectively, q_+ denotes the positive one corresponding to q^*.

A similar single-view contrastive loss is also applied to the aerial views. However, unlike the ground view contrastive loss that is applied between ground view variations, the aerial view contrastive loss is designed to enforce the similarity between two distorted versions of the same reference image using random data augmentations, aiming to obtain more robust feature representations [3]. The aerial view contrastive objective is:

$$\mathcal{L}(r^*, R)_{\text{single-r}} = -\log \frac{\exp\left(r^* \cdot r_+/\tau_r\right)}{\sum_{r_i \in R} \exp\left(r^* \cdot r_i/\tau_r\right)}, \tag{2}$$

where r_i is the feature embedding of the original aerial image, r_+ denotes the positive one corresponding to r^*, which is the feature of another possible augmentation of the same aerial image. τ_r and τ_q are learnable temperature parameters [4].

Cross-view Contrastive Learning. The cross-view contrastive learning in ConGeo comprises two alignment objectives: First, the loss of the base method is used[3], which we refer to as the vanilla loss; it is a cross-view alignment loss, where the model learns to match the embeddings of a query image q with the embeddings of its corresponding aerial image $r \in R$. In Sample4Geo [4], for example, the vanilla loss is:

$$\mathcal{L}(q, R)_{\text{vanilla}} = -\log \frac{\exp\left(q \cdot r_+/\tau_v\right)}{\sum_{r_i \in R} \exp\left(q \cdot r_i/\tau_v\right)}, \tag{3}$$

where R is a set of reference images, and r_+ is the reference feature embedding that matches the query embedding.

Second, to further reinforce the model's robustness to ground view variations, we enforce the alignment between transformed ground images and aerial images with a cross-view contrastive objective:

$$\mathcal{L}(q^*, R)_{\text{cross}} = -\log \frac{\exp\left(q^* \cdot r_+/\tau_c\right)}{\sum_{r_i \in R} \exp\left(q^* \cdot r_i/\tau_c\right)}, \tag{4}$$

with τ_v and τ_c as learnable temperature parameters.

Final Loss. Combining all the terms previously described, the total loss of ConGeo can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{vanilla}} + w_1 \mathcal{L}_{\text{single-q}} + w_2 \mathcal{L}_{\text{single-r}} + w_3 \mathcal{L}_{\text{cross}}, \tag{5}$$

where w_1, w_2 and w_3 are factors balancing the contribution of the different learning objectives.

[3] Thanks to its flexible learning objective, ConGeo can be plugged into other existing geo-localization models. In this section, we use Sample4Geo [4] as an example. In the experiments, we also plug ConGeo into TransGeo [34], and SAIG-D [37].

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. Experiments are performed on four CVGL datasets. **CVUSA** [31] contains 35,532 view pairs for training and 8,884 for evaluation. **CVACT** [13] is split into training, validation, and test sets, where the first two are of the same size as CVUSA while the test set has 92,802 image pairs. **VIGOR** [36] contains 90,618 satellite images and 105,214 street-view images and can be divided into two train-test splits: same-area and cross-area. The training and test data in the cross-area subset were collected from different cities. **University-1652** [33] contains drone, satellite and street view images. The street view images are commonly with unknown orientation and limited FoV, enabling a real-world evaluation under this challenging setting. Detailed dataset descriptions can be found in the supplementary material.

Evaluation Metrics. Following the literature [4,21,37], we use Top-k recall, denoted as R@k, to measure the retrieval performance. For each ground view query image, the aerial images are ranked based on cosine similarity with ground view representation. If the ground-truth aerial image is among the top k retrieved images, the retrieval is considered a success under R@k. As in previous methods [4,33], we also use Average Precision (AP) for one-to-many and many-to-one matching in the University-1652 dataset.

4.2 Implementation Details

Base Models and Data Preprocessing. Unless specified, ConGeo uses Sample4Geo [4] as the base model, with ConvNeXt-B [14] as the backbone. In Section 5.3, we also plug ConGeo into TransGeo [34] and SAIG-D [37]. We follow the default settings and data augmentation methods used in those models.

Hyper-parameters and Environment. Loss weights w_1, w_2, and w_3 are empirically set to 0.5, 0.5, and 0.25, corresponding to a higher weight on the single-view contrastive loss and lower weight on the cross-view one. For training, the orientation angle θ is randomly drawn between 0° and 360° while the FoV angle α is set to 180°. For evaluation, the orientation angle and FoV angle depend on different settings (see Section 4.3). Experiments comparing different hyper-parameters are provided in the supplementary materials. Besides, following the default setting, for those who used Sample4Geo as the base model, the weights of the query encoder and the reference encoder are shared. For each experiment, the model is trained for 60 epochs with a batch size of 16. We use the AdamW optimizer with an initial learning rate of 0.0001 and a cosine learning rate scheduler. A single NVIDIA GeForce RTX 4090 is used for all the experiments.

4.3 Experimental settings

The evaluation is performed for three different settings.

Set	Methods	FoV=360°				FoV=180°				FoV=90°				FoV=70°				Avg.
		R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1
CVUSA	CVFT [22]	23.4	44.4	55.2	86.6	8.1	24.3	34.5	75.2	4.8	14.8	23.2	61.2	3.8	12.4	19.3	55.6	10.0
	SEH [6]	85.4	93.5	95.8	-	53.7	72.3	79.0	-	16.6	32.2	40.3	-	7.8	18.8	25.6	-	40.9
	DSM [21]	78.1	89.5	92.9	98.5	48.5	68.5	75.6	93.0	16.2	31.4	39.9	71.1	8.8	19.9	27.3	61.2	37.9
	SAIG-D [37]	72.0	90.2	94.0	99.1	52.5	78.1	85.8	97.7	26.7	50.2	59.8	86.6	20.9	41.4	51.2	80.4	43.0
	Sample4Geo [4]	93.3	97.5	98.0	99.1	84.6	95.9	97.6	99.5	55.1	**78.3**	**85.0**	**96.6**	40.9	65.4	74.1	93.0	68.5
	ConGeo	**96.6**	**98.9**	**99.2**	**99.7**	**92.3**	**97.9**	**98.7**	**99.7**	**55.5**	75.4	81.5	93.9	**49.1**	**70.8**	**78.0**	**93.1**	**73.4**
	Sample4Geo [4]	16.3	26.1	31.4	51.7	4.1	8.4	11.3	30.4	2.5	6.7	9.8	26.7	1.5	4.6	6.7	20.4	6.1
	Sample4Geo† [4]	**93.2**	**98.2**	**99.0**	**99.8**	84.6	95.9	97.6	99.5	45.1	64.8	71.3	86.5	28.4	47.1	54.9	75.8	62.8
	ConGeo	85.2	95.1	96.9	98.9	**92.3**	**97.9**	**98.7**	**99.7**	**55.9**	**73.2**	**79.0**	**90.9**	**37.1**	**55.7**	**62.8**	**81.4**	**67.6**
CVACT	CVFT [22]	26.8	46.9	55.1	81.0	7.1	18.5	26.8	63.9	1.9	6.3	10.5	39.3	1.5	5.1	8.2	34.6	9.3
	SEH [6]	77.4	88.6	90.9	-	47.7	67.9	74.3	-	13.9	28.4	36.2	-	6.9	16.5	22.3	-	36.5
	DSM [21]	72.9	85.7	88.9	95.3	49.1	67.8	74.2	89.9	18.1	33.3	40.9	68.7	8.3	20.7	27.1	57.1	37.1
	Sample4Geo [4]	82.4	**90.6**	92.3	95.4	58.9	79.8	85.3	95.0	27.9	52.0	62.3	**87.0**	18.8	40.4	51.0	**81.3**	47.0
	ConGeo	**83.0**	**90.6**	**92.4**	**96.3**	**70.3**	**85.2**	**88.6**	**95.1**	**40.6**	**62.6**	**69.8**	86.6	**24.6**	**45.3**	**54.3**	80.6	**54.6**
	Sample4Geo [4]	12.6	17.5	19.9	31.2	3.4	8.0	10.6	24.5	1.9	5.5	7.9	23.9	1.0	3.2	5.0	16.8	4.7
	Sample4Geo† [4]	**77.1**	**89.8**	**92.6**	**97.2**	58.9	79.8	85.3	95.0	22.4	44.0	52.7	78.3	12.5	28.5	37.5	67.4	42.7
	ConGeo	62.6	79.9	84.7	93.9	**70.3**	**85.2**	**88.6**	**95.1**	**34.8**	**56.9**	**64.8**	**83.6**	**18.5**	**37.4**	**46.7**	**72.5**	**46.6**

Table 1: Comparison on the unknown orientation setting and limited FoV setting on CVUSA [31] and CVACT Val [21] datasets. "-" means the score is not provided in the original paper. The best performance is in **bold**. Results of models trained with FoV-specific images are reported in the white background, while results from a single model without FoV-specialized training are reported with a grey background. † denotes using the same training FoV as ConGeo.

- North-aligned. This is the common configuration of the task: the full-view North-aligned ground image is used to retrieve the aerial view.
- Unknown Orientation. When testing, we shift the full-view ground image by a random angle and use it as the query. This setting is also noted as FoV=360°.
- Limited FoV. We first apply a random cyclical shift to the ground image (same as unknown orientation) and then reduce the FoV to 70°, 90°, or 180°.

For the CVUSA and CVACT datasets, we evaluate the model's performance under these 3 settings. For the VIGOR dataset, we use the unknown orientation and limited FoV settings to verify the method's cross-location robustness. Furthermore, we use the University-1652 dataset for Street-to-Satellite (St2S, many-to-one) and Satellite-to-Street (S2St, one-to-many) retrieval, showing the robustness for street view images. We also test the models' generalization ability on four unseen ground view variations on the CVUSA dataset.

5 Results

5.1 Performance under Different Settings

As in previous works [21], we assess the effectiveness of ConGeo in the unknown orientation setting and the limited FoV settings on the CVUSA and CVACT validation datasets in Table 1. Unlike previous methods [21, 29, 37], which train different models for each setting to improve performance, the single model trained with ConGeo excels in all settings. As shown in the rows highlighted in grey in Table 1, ConGeo improves the base model's (Sample4Geo) performance by a large margin (68.9% and 53.4% R@1 improvement for FoV=360°

Methods	CVUSA		CVACT Val		CVACT Test	
	R@1	R@1%	R@1	R@1%	R@1	R@1%
CVFT [22]	61.4	99.0	61.1	95.9	-	-
DSM [21]	92.0	99.7	82.5	97.3	-	-
TransGeo [34]	94.1	99.8	85.0	98.4	-	-
GeoDTR [32]	95.4	**99.9**	86.2	98.8	64.5	**98.7**
SAIG-D [37]	96.3	**99.9**	89.1	**98.9**	67.5	96.8
Sample4Geo [4]	**98.7**	**99.9**	**90.8**	98.8	71.5	**98.7**
ConGeo	98.3	**99.9**	90.1	98.2	**71.7**	98.3

Table 2: Comparison of the North-aligned setting on CVUSA and CVACT datasets. The second-best performance is underlined. "-" means the score is not provided in the original paper.

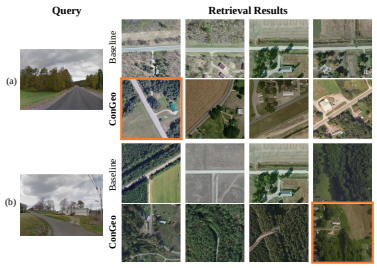

Fig. 3: Examples of the top-4 retrieved images from ConGeo and the baseline when FoV=90°. Images in the orange box denote the correct results.

and FoV=90°, respectively) and significantly outperforms the state-of-the-art methods on most of the challenging settings on the CVUSA dataset, *without training models on separate FoVs*. Besides, when training ConGeo on each FoV separately, the model's performance can be further enhanced in different settings, significantly surpassing all competitors on all FoVs.

Similarly, ConGeo outperforms the state-of-the-art model's R@1 under 360°, 180°, 90°, and 70° settings on the CVACT validation set by 0.6%, 11.4%, 12.7%, and 5.8% respectively. ConGeo also shines on this dataset with a single model, surpassing the previous approaches in most settings by a substantial margin. In short, these results show the consistent superiority of ConGeo over previous state-of-the-art methods and its robustness in handling ground view variations.

Moreover, ConGeo maintains competitive performance in the North-aligned setting. As shown in Table 2, ConGeo always ranks within the top two for R@1 both on the CVUSA and CVACT datasets. Moreover, by comparing Table 2 and Table 1, we observe that when confronting challenging settings, baselines experience a considerable performance drop (as seen from the results with grey background in both tables). For example, for a query image with FoV=180°, R@1 of the base model drops by 82.4%; while shifting the North-aligned ground view with unknown orientation leads to an R@1 decrease of 94.6%. On the contrary, ConGeo maintains robustness to various ground view shifts with R@1 dropping by only 13.1% for 360° and by 6.0% for 180°.

Fig. 3 shows two examples of top-4 retrieved images when the query image's FoV=90°. Limited FoV images offer restricted semantic and spatial information and thus are more challenging. In Fig. 3 (a), the baseline model only retrieves aerial images with a horizontal road, indicating that the model heavily relies on spatial correspondence; ConGeo's results are more diverse among road directions. Fig. 3 (b) shows a building next to a half-circle side road. Several images retrieved by ConGeo include similar visual features, showing that ConGeo elicits the model's focus on information that is consistent across ground view variations.

Aug. type			North-aligned				FoV=360°				FoV=180°				FoV=90°				Avg.
Shift	FoV	Rotate	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1
			98.7	**99.7**	**99.8**	**99.9**	16.3	26.1	31.4	51.7	4.1	8.4	11.3	30.4	2.5	6.7	9.8	26.7	30.4
✓			94.4	98.1	99.6	93.2	**93.1**	**97.6**	**98.2**	99.1	73.8	88.0	90.9	95.8	35.1	54.2	61.2	77.6	74.1
✓	✓		90.3	97.3	98.4	99.6	84.1	95.0	96.9	**99.4**	63.6	84.6	90.1	98.2	32.2	55.1	64.5	87.4	67.6
✓	✓	✓	88.9	96.0	97.1	99.0	89.0	96.1	97.3	98.9	71.8	87.9	91.3	97.2	39.3	59.9	67.9	85.7	72.3
ConGeo			98.3	99.6	99.7	**99.9**	85.2	95.1	96.9	98.9	**92.3**	**97.9**	**98.7**	**99.7**	**55.9**	**73.2**	**79.0**	**90.9**	**82.9**

Table 3: The comparison of results between ConGeo and task-specific data augmentations. We incorporate augmentations into Sample4Geo, "Shift" denotes using shifted query images and "FoV" denotes using query images of limited FoVs, "Rotate" randomly rotating aerial images with an angle in {90°, 180°, 270°} as data augmentation. The second-best performance is underlined.

5.2 Ablations

We perform two ablation studies to show the advantages of ConGeo over alternative data augmentations and the impact of each component of ConGeo.

Comparison with Data Augmentations (Table 3). One possible way to improve robustness under view variations is through data augmentation. We compare ConGeo with three methods: "Shift" (random cyclical shift), "FoV" (random query image FoV cropping, from 70° to 360°), and "Rotate" (random aerial images rotation by 90°, 180° or 270°). Results show that these methods can improve the base model under unknown orientation and limited FoV. However, their overall performance stays far lower than ConGeo, especially since their performance in the North-aligned setting notably decreases, showing the superiority of contrastive objectives over targeted data augmentation. Compared to ConGeo, all the data augmentation approaches lack a crucial single-view contrastive objective, which explicitly enforces *discovering the joint features shared by different ground view variations* and therefore being robust to them.

Ablations of Loss Components (Table 4). The aerial view contrastive objective ($\mathcal{L}_{\text{single-r}}$) slightly enhances the performance (Row 1 and 2), which serves as the basis for the following ablations. The ground view contrastive loss ($\mathcal{L}_{\text{single-q}}$) gradually boosts performance under FoVs as contrastive targets between ground views are added (Row 1, 3, and 4). Finally, adding the cross-view contrastive loss on top of single-view losses yields a notable improvement (Row 4 and 7). This indicates that the contrastive learning between the original and shifted ground images plays an essential role in assisting cross-view alignment.

5.3 Adaptability to Different Base Models

ConGeo is model-agnostic: it can be plugged into different CVGL systems and boosts their robustness to ground view variations. To demonstrate this, we choose two representative methods, TransGeo [34] and SAIG-D [37] as base models, besides Sample4Geo. For each base model, we follow the default configurations — including backbone and data augmentations — and use the vanilla loss in the base model instead of Eq.(5). We train the ConGeo-augmented models on all FoVs jointly (same as the results with grey background in Table 1). We

No.	$\mathcal{L}_{\text{single-r}}$	$\mathcal{L}_{\text{single-q}}$		$\mathcal{L}_{\text{cross}}$		FoV=180°		FoV=90°	
		Shift	FoV	Shift	FoV	R@1	R@1%	R@1	R@1%
1						4.1	30.4	2.5	26.7
2	✓					15.7	60.0	7.7	43.3
3	✓	✓				44.1	93.4	17.8	73.0
4	✓	✓	✓			37.9	81.2	20.5	59.9
5	✓			✓	✓	91.5	**99.7**	40.2	89.0
6	✓	✓		✓		81.7	98.8	35.8	81.6
7	✓	✓	✓	✓	✓	**92.3**	**99.7**	**55.9**	**90.9**

Table 4: Ablation studies on FoV=180° and FoV=90° on the CVUSA dataset. "Shift" and "FoV" mean cyclic shift and FoV cropping for ground view images.

Methods	FoV=360°		FoV=180°		FoV=90°	
	R@1	R@1%	R@1	R@1%	R@1	R@1%
TransGeo	13.5	59.4	4.5	42.2	0.4	13.9
TransGeo + DA	**75.9**	**99.2**	47.8	94.9	18.7	74.7
ConGeo[TransGeo]	52.7	97.2	**54.8**	**97.4**	**26.9**	**83.8**
SAIG-D	12.5	69.4	3.3	40.1	0.3	10.2
SAIG-D + DA	64.8	98.6	49.3	96.9	**29.7**	**90.0**
ConGeo[SAIG-D]	**70.7**	**98.9**	**54.9**	**97.3**	24.0	80.4
Sample4Geo	16.3	51.7	4.1	30.4	2.5	26.7
Sample4Geo + DA	84.1	**99.4**	63.6	98.2	32.2	87.4
ConGeo[Sample4Geo]	**85.2**	98.9	**92.3**	**99.7**	**55.9**	**90.9**

Table 5: ConGeo with three base models on the CVUSA dataset including TransGeo [34], SAIG-D [37] and Sample4Geo [4]. Note that "DA" means data augmentation.

compare each model with two baselines: (1) the base model trained on North-aligned data, (2) the base model trained with strong data augmentations, including shifting each ground image with a random angle and cropping it with a FoV between 70° and 360° (referred to as "DA"). As shown in Table 5, adding ConGeo improves ViT-based TranGeo's R@1% performance by 37.8% and 69.9% for FoV=360° and FoV=90°, respectively. SAIG-D combines convolutional stem and self-attention layers in the encoder and the improvement with ConGeo is also remarkable. ConGeo-augmented models also outperform strong data augmentations under most FoVs. In summary, ConGeo can be plugged into three different baselines with both CNN-based and ViT-based backbones, demonstrating its versatility and effectiveness across model architectures.

5.4 Robustness across Other Ground View Variations

Robustness across Locations (Table 6). In cross-view geo-localization, the VIGOR dataset is particularly challenging, because the two views are not center-aligned. In particular, VIGOR's cross-area subset is regarded as a standard benchmark to test the model's robustness to data across different locations, since training and test images are taken from different areas [4, 27, 36]. As shown in Table 6, ConGeo outperforms the baselines on both the cross-area and same-area subsets. Note that we report the performance of a single model for all FoVs, for ConGeo and Sample4Geo (grey background). The results indicate that ConGeo consistently improves the model's robustness on cross-area data.

Robustness to Street Images (Table 7). Street images in University-1652 are typical examples of real-world limited FoV images. The St2S and S2St settings pose significant challenges as the limited FoV images can be captured from diverse locations with unknown orientations. Different from the experiments considering North-aligned panoramas, here we randomly sample two street view images from one location as the inputs to conduct single-view contrastive loss. Detailed descriptions can be found in the supplementary materials. Results in Table 7 show that the performance on both settings is improved.

Robustness to Unseen Ground View Variations (Table 8). In real-world scenarios, the variations of the ground view image are more diverse than orien-

Methods	Cross-Area				Same-Area			
	FoV=360°		FoV=90°		FoV=360°		FoV=90°	
	R@1	R@1%	R@1	R@1%	R@1	R@1%	R@1	R@1%
VIGOR [36]	1.4	44.6	-	-	19.1	95.1	-	-
TransGeo [34]	5.5	66.9	-	-	47.7	**99.3**	-	-
Sample4Geo [4]	9.0	43.7	0.5	21.6	14.2	54.9	1.1	30.6
ConGeo	**16.2**	**72.9**	**3.9**	**54.3**	**61.9**	98.4	**8.5**	**68.7**

Table 6: Comparison on the VIGOR dataset [36]. "Cross-Area" and "Same-Area" mean its cross-area subset and same-area subset, respectively.

Methods	St2S		S2St	
	R@1	AP	R@1	AP
University-1652 [33]	0.6	1.6	0.9	1.0
LPN [25]	0.7	1.8	1.4	1.3
Sample4Geo* [4]	4.9	8.1	6.6	6.1
ConGeo	**5.9**	**9.2**	**6.8**	**6.4**

Table 7: Comparison on University-1652 [33]. * denotes the reproduced model since the pre-trained weights are not provided.

Methods	Random FoVs				Random Zooming				Gaussian Noise				Motion Blur				Avg.
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1
Sample4Geo	15.7	25.3	31.2	51.7	48.2	60.8	65.8	80.1	42.4	60.9	67.0	81.6	31.4	38.6	42.0	52.9	34.4
Sample4Geo + DA	**85.5**	**95.1**	96.6	98.8	44.5	60.6	66.7	82.7	0.2	1.0	1.7	8.8	16.8	23.2	25.9	36.0	36.8
ConGeo	84.2	**95.1**	**97.0**	**99.4**	**68.7**	**80.3**	**83.8**	**92.1**	**45.8**	**64.4**	**70.3**	**83.4**	**32.5**	**40.7**	**43.5**	**53.5**	**57.8**

Table 8: Comparison on unseen ground view variations between ConGeo and baselines on the CVUSA dataset. Random FoVs are between 0° to 360°, Random Zooming is performed with a ratio between 0.5 and 2.0, Gaussian Noise [1], and Motion Blur [1] are added with severity: 5.

tation and FoV variations. In Table 8, we evaluate the model across four unseen ground view variations on the CVUSA dataset: Random FoVs, Random Zooming, Gaussian Noise, and Motion Blur. The ConGeo-augmented model exhibits comparable performance to the model with data augmentation on Random FoVs test, while demonstrating significant advantages over the two baselines when tested on other unseen ground view variations. According to Table 3 and 8, the improvements that task-oriented data augmentation brings are *not transferable nor generalizable*, on the contrary, ConGeo represents a way of unleashing the potential of data augmentation by enforcing the learning of invariances.

6 Analysis: How does ConGeo achieve robustness?

In this section, we analyze the behavior of the base model and ConGeo under ground view variations, to diagnose what led to the base model's collapse and explain the superior performance of ConGeo. We perform orientation invariance analysis to showcase models' vulnerabilities to orientation shifts and visualize activation maps to investigate the models' focus.

Orientation Invariance Analysis. We introduce the concept of orientation invariance to dissect the siamese network's behavior. Let Φ be a function that maps the input images to predictions, $G := \{g\}$ be a group of cyclic shift transformations. I_q and I_r are the query and reference input images, respectively. Here, the ground view orientation invariance is defined as:

$$\Phi\left[g(I_q), I_r\right] = \Phi(I_q, I_r) \quad \forall (I_q, I_r, g) \in (Q, R, G). \tag{6}$$

In Fig. 4, we experimentally investigate the orientation invariance of different models by evaluating the retrieval performance when using shifted ground views

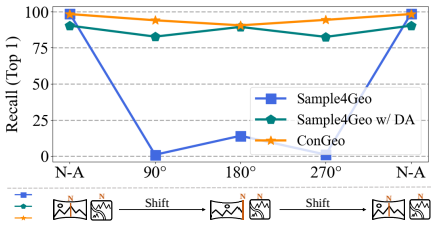

Fig. 4: ConGeo shows better orientation invariance. We cyclically shift the ground view with an angle (x-axis) as the model's input to test its retrieval performance. Note that "N-A" denotes the North-aligned setting and "DA" means data augmentation.

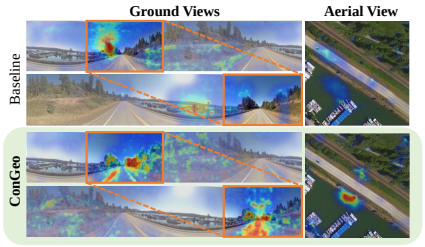

Fig. 5: ConGeo's activation areas are more consistent across ground view variants. The Grad-CAM activation maps of the base model and ConGeo on the North-aligned and the unknown orientation setting. The orange box indicates the same area in different ground view variants.

as input. For the base model trained under North alignment (blue line), the model's recall drops significantly on different orientation angles, indicating a lack of orientation invariance. Although the model with ground view variants data augmentation (green line) shows an improved orientation invariance compared to the base model with the orientation-specific learning pipeline, the performance of the North-aligned setting drops considerably. In contrast, ConGeo (orange line) yields consistently high performance both under the cyclically shifted query input and with the North-aligned image pairs. This indicates that the model trained with the proposed contrastive objectives shows better orientation invariance while simultaneously maintaining a strong ability to leverage the spatial correspondence, allowing it to maintain robustness to ground view variations.

Activation Map Visualization. We analyze the activation map of ConGeo and its baseline in Fig. 5 (North-aligned and unknown orientation settings) and Fig. 6 (North-aligned and limited FoV settings). We make two key observations. First, the focus of the base model is vulnerable to orientations and FoV variations, while ConGeo's representation is more *robust across view variations*. In Fig. 5, the ground view shifts make the base model's attention drift from the roadside to random regions that might not carry geospatial information (*e.g.*, sky), while ConGeo consistently highlights regions (*e.g.*, trees) with similar contents under view variations. Second, the base model focuses more on spatial correspondence cues (*e.g.*, road), while ConGeo focuses more on the *semantically consistent objects* in both views (*e.g.*, trees), as shown in Fig. 6. This further demonstrates that enhancing consistency and mitigating shortcuts of spatial correspondence makes the model more robust to view variations. More examples can be found in supplementary materials.

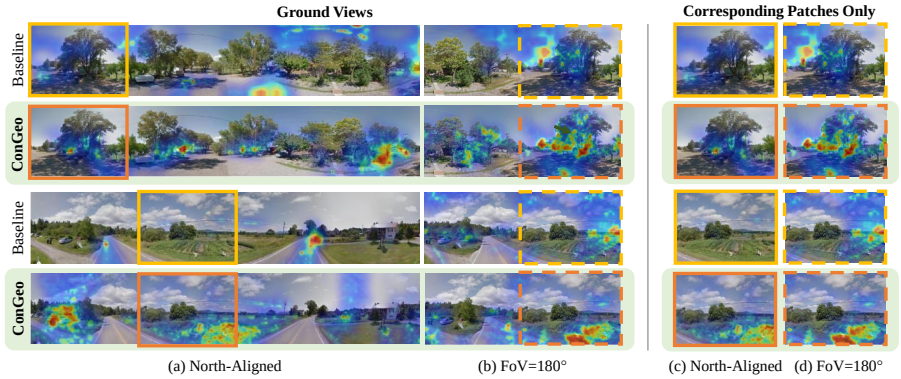

Fig. 6: ConGeo focuses on semantically consistent features for the task and its activated regions are consistent across view variation. Left: GradCam activations of the base model and ConGeo (green background) on the North-aligned and FoV = 180° settings. Right: focus on the regions highlighted by the coloured boxes.

7 Limitations

We showed that ConGeo leads to significant improvements when facing arbitrarily oriented ground images and diverse FoVs, except in the North-aligned setting. This performance drop is nearly unavoidable, as ConGeo tends to disrupt the over-reliance on spatial correspondence shortcuts. However, by keeping the original learning objectives of the base model (vanilla loss), ConGeo achieves competitive performance when orientation information is available (Table 2), and significantly improves the robustness when it is unknown. Additionally, we focus on ground view orientation and FoV variations, but other variations can be envisaged (*e.g.*, zoom, color intensity, blur). We show ConGeo's robustness to some of these variations unseen during training (Table 8). Finally, besides contrastive learning, other ways of aligning modalities (*e.g.*, redundancy reduction) could also be considered. We discuss this in the supplementary materials.

8 Conclusion

Tackling situations where the ground view image orientation is unknown, or the FoV is limited, is crucial for real-world applications of cross-view image geo-localization. We propose ConGeo, a single- and cross-view contrastive method that enhances a model's robustness to ground view variations by aligning image variations with their original representations. Experiments on four datasets and activation map analysis demonstrate that ConGeo consistently boosts the performance of state-of-the-art methods by a large margin when facing diverse challenging settings. With its adaptability to different base models and versatility to accommodate diverse ground view variations, ConGeo strives to be a step towards widening the applicability of geo-localization methods to the real world.

Acknowledgments

We thank the anonymous reviewers for their constructive and thoughtful comments. We thank Haoyuan Li, Zimin Xia, Gencer Sümbül, Silin Gao, Valérie Zermatten, Zeming Chen, Tianqing Fang, Gaston Lenczner, Kuangyi Chen, Robin Zbinden, Giacomo May, Riccardo Ricci, Emanuele Dalsasso, and Sepideh Mamooler for providing helpful feedback on earlier versions of this work. We acknowledge the support from the CSC and EPFL Science Seed Fund and the support in part by the National Natural Science Foundation of China (NSFC) under Grant 62271355. AB gratefully acknowledges the support of the Swiss National Science Foundation (No. 215390), Innosuisse (PFFS-21-29), the EPFL Center for Imaging, Sony Group Corporation, and the Allen Institute for AI.

References

1. Arularasu, A., Kulkarni, P.P., Nayak, G.K., Shah, M.: Robust image geolocalization. Tech. rep. (2023)
2. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "siamese" time delay neural network. Advances in Neural Information Processing Systems **6** (1993)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning (2020)
4. Deuser, F., Habel, K., Oswald, N.: Sample4Geo: Hard negative sampling for cross-view geo-localisation. In: IEEE International Conference on Computer Vision. pp. 16847–16856 (2023)
5. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R.: Uncertainty-aware vision-based metric cross-view geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 21621–21631 (2023)
6. Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M.: Soft exemplar highlighting for cross-view image-based geo-localization. IEEE Transactions on Image Processing **31**, 2094–2105 (2022)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
8. Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 7258–7267 (2018)
9. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems (2020)
10. Li, A., Hu, H., Mirowski, P., Farajtabar, M.: Cross-view policy learning for street navigation. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8100–8109 (2019)
11. Liang, Z., Jiang, W., Hu, H., Zhu, J.: Learning to contrast the counterfactual samples for robust visual question answering. In: EMNLP. pp. 3285–3292 (2020)
12. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2013)

13. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5624–5633 (2019)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
15. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. In: European conference on computer vision. pp. 529–544. Springer (2022)
16. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
17. Peyré, G., Cuturi, M., et al.: Computational optimal transport: With applications to data science. Foundations and Trends in Machine Learning **11**(5-6), 355–607 (2019)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
19. Rodrigues, R., Tani, M.: Global assists local: Effective aerial representations for field of view constrained image geo-localization. In: IEEE Workshops on Applications of Computer Vision. pp. 3871–3879 (2022)
20. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. Advances in Neural Information Processing Systems **32** (2019)
21. Shi, Y., Yu, X., Campbell, D., Li, H.: Where am I looking at? joint location and orientation estimation by cross-view matching. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4064–4072 (2020)
22. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: AAAI Conference on Artificial Intelligence. vol. 34, pp. 11990–11997 (2020)
23. Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., Isola, P.: What makes for good views for contrastive learning? Advances in Neural Information Processing Systems **33**, 6827–6839 (2020)
24. Wang, P., Han, K., Wei, X.S., Zhang, L., Wang, L.: Contrastive learning based hybrid networks for long-tailed image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 943–952 (2021)
25. Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2021)
26. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3961–3969 (2015)
27. Xia, Z., Booij, O., Manfredi, M., Kooij, J.F.: Visual cross-view metric localization with dense uncertainty estimates. In: European Conference on Computer Vision. pp. 90–106 (2022)
28. Xie, E., Ding, J., Wang, W., Zhan, X., Xu, H., Sun, P., Li, Z., Luo, P.: Detco: Unsupervised contrastive learning for object detection. In: IEEE International Conference on Computer Vision. pp. 8392–8401 (2021)
29. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. Advances in Neural Information Processing Systems **34**, 29009–29020 (2021)

30. Zamir, A.R., Shah, M.: Accurate image localization based on google maps street view. In: European Conference on Computer Vision. pp. 255–268 (2010)
31. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 867–875 (2017)
32. Zhang, X., Li, X., Sultani, W., Zhou, Y., Wshah, S.: Cross-view geo-localization via learning disentangled geometric layout correspondence. In: AAAI Conference on Artificial Intelligence. vol. 37, pp. 3480–3488 (2023)
33. Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: ACM International Conference on Multimedia. pp. 1395–1403 (2020)
34. Zhu, S., Shah, M., Chen, C.: TransGeo: Transformer is all you need for cross-view image geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1162–1171 (2022)
35. Zhu, S., Yang, T., Chen, C.: Revisiting street-to-aerial view image geo-localization and orientation estimation. In: IEEE Workshops on Applications of Computer Vision. pp. 756–765 (2021)
36. Zhu, S., Yang, T., Chen, C.: VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3640–3649 (2021)
37. 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)