Supplementary Materials for "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/

The supplementary materials contain the following information:

- Dataset details (Section S1).
- Implementation details (Section S2).
- Detailed method description (Section S3).
- Hyper-parameters analysis: about the training FoV angle, loss weights, and the learning rate (Section S4).
- Additional experiments: architecture analysis (CNN-based and ViT-based encoders), transfer experiments, polar transformation, different paradigms of aligning modalities (Section S5).
- Supplementary results to the main paper: full tables (Section S6).
- Visualization: additional samples and failure case analysis (Section S7).
- Discussions: limitations and future works (Section S8).

S1 Dataset Details

CVUSA. The CVUSA dataset [20] contains 35,532 ground-aerial view pairs for training and 8,884 pairs for evaluation. The satellite images are with the size of 750 × 750 and street-view images are with the size of 224 × 1232. Both types of images are North-aligned to ensure that the geographical North is located in the upper center of the satellite image and the center of the street view images.
CVACT. The CVACT dataset [8] is split into training, validation, and test sets, where training and validation sets are of the same size as CVUSA while the test set has 92,802 image pairs, which is around 10 times larger than the validation set. The image size is larger than the CVUSA dataset, with 1200 × 1200 for satellite images and 832 × 1664 for ground views.
VIGOR. The VIGOR dataset [24] contains 90,618 aerial-view images and 105,214 ground-view images from four cities: New York, Seattle, San Francisco, and Chicago. The raw image sizes for aerial view and ground view are 640 × 640 and 2048 × 1024, respectively. Different from the one-to-one matching in CVUSA

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

Method	Metric Learning Loss	Architecture	Optimizer	Weight-sharing	Zero-padding
Sample4Geo-CNN [2]	InfoNCE	CNN	AdamW [11]	✓	✗
Sample4Geo-ViT [2]	InfoNCE	ViT	AdamW [11]	✗	✓
TransGeo [23]	Soft-triplet	ViT	ASAM [7]	✗	✓
SAIG-D [25]	Soft-triplet	CNN & Attention	AdamW [11]	✗	✓

Table S1: A comparison of the basic settings in different base models.

and CVACT datasets, each query image in VIGOR can be paired with multiple reference images and vice versa. There are two subsets in VIGOR datasets: same-area and cross-area. In the paper, we use the cross-area subset to evaluate the model's robustness across locations. In the training set, the cross-area setting contains 44,055 aerial view images with 51,520 ground view images from New York and Seattle. The cross-area test set is sampled from San Francisco and Chicago, with 46,563 aerial view images and 53,694 ground view images.

University-1652. The University-1652 dataset [22] contains different types of cross-view images, including satellite images, unmanned aerial vehicle (UAV) images, and ground view images. In our experiments, we train and evaluate models based on street-satellite matching to demonstrate the robustness of the model on real-world limited FoV images. For street-to-satellite matching, there are 2,579 street images and 951 satellite images. For satellite-to-street matching, there are 701 satellite images and 2,921 street images. The image size is 512×512.

S2 Implementation Details

Data Preprocessing. We followed the data preprocessing methods used in the base models. For example, when we use Sample4Geo [2] as our base model, we resize the ground view images to 140 × 768, and the aerial view images to 384×384 for CVUSA and CVACT. For the VIGOR dataset, we resize the ground view to 384×768 and the satellite image to 384×384. As for the University-1652 dataset, all images are resized from 512×512 to 384×384. Note that for models utilizing TransGeo, SAIG-D, and Sample4Geo[ViT] architectures, we employ zero-padding in the area after FoV cropping for both training and testing inputs, enabling the learnable positional encoding to adapt to inputs with varying FoVs. Unless specified, we retain the corresponding data augmentation in different base models [2, 23, 25] for fair comparisons. For those using Sample4Geo as the base model, the data augmentation includes dropout, color jitter, flipping on both ground and aerial views, rotation of satellite images, and the corresponding shift on the ground view images.

Model and Training Details. As described in the paper, for different base models, we follow their default settings respectively, including model architecture, weight-sharing operations, and basic loss *etc.* to ensure a fair comparison. A comparison of the detailed settings in different base models is shown in Table. S1. For instance, for experiments based on Sample4Geo[CNN], the weights of the ground view encoder and the aerial view encoder are shared, while for others (Sample4Geo[ViT] and other base models), the weights are not shared between the two encoders. For Sample4Geo[CNN], instead of a batch size of 128 on multiple GPUs, we set the batch size as 16 on a single GPU, which is much smaller. Accordingly, we set the starting learning rate as 0.0001 instead of 0.001 in the default settings.

Experimental Environment. The code was developed using Python version 3.8, PyTorch version 2.0.1, Timm version 0.9.7, and OpenCV-Python version 4.8.1.78. A single NVIDIA GeForce RTX 4090 is utilized for computation.

S3 Detailed Method Description

S3.1 Pseudo Code

The core pseudo code of ConGeo's training process is provided in Alg. 1 in a Pytorch style. Here we use the CNN-based Sample4Geo as an example.

Training Phase: In each mini-batch, we take paired ground-aerial images (I_Q, I_R) as input, and perform defined transformations T_q (*i.e.*, random shift, FoV cropping) to each ground view image I_q ($I_q \in I_Q$), obtaining the transformed ground image ${I_q}^* = T_q(I_q|\theta, \alpha)$. We also perform augmentations in the base model to I_R (described in the data preprocessing of Section S2), obtaining ${I_R}^*$. Then, the image batchs I_Q, ${I_Q}^*$, I_R and ${I_R}^*$ are encoded by two encoders, a query encoder E_q for encoding I_Q, ${I_Q}^*$, and a reference encoder E_r for encoding I_R, ${I_R}^*$, with output feature embeddings Q, Q^*, R, and R^* respectively. The optimization goal contains four targets, two single-view contrastive objectives, one for aligning the feature space between Q and Q^* while another for aligning the feature space between R and R^*, a cross-view contrastive objective aligning the feature space between R and Q^*, and a vanilla contrastive objective aligning the feature space between R and Q. The total loss is a weighted summation of the four loss functions, where the ablation of weights will be studied in the next section.

Testing Phase: Depending on different evaluation settings, the query image is transformed accordingly. We use the original full-view ground images and the transformed ground view images for the North-aligned setting and the other two challenging settings (unknown orientation and limited FoV settings), respectively. For each query input and a set of reference aerial images, the trained E_q and E_r are used for feature representation, and then the reference images are ranked as the retrieval results based on the cosine similarity between the query feature and reference features.

S3.2 Training on the University-1652 Dataset

Street images in the University-1652 dataset are ground view images with unknown orientations and limited FoVs. In order to adapt ConGeo to street images without collecting a full-view panorama, we provide an alternative way of building the ground view single-view contrastive objective (Eq. (1) in the paper). Instead of enforcing the representations of ground view variants to be closer to the

Method(α)	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
Sample4Geo+DA	84.1	95.0	96.9	99.4	63.6	84.6	90.1	98.2	32.2	55.1	64.5	87.4	21.5	42.0	51.6	79.9	50.4
ConGeo(90)	81.4	93.0	95.3	98.6	92.2	**98.1**	**98.9**	**99.7**	55.5	75.4	81.5	93.9	35.5	54.8	61.9	81.0	66.2
ConGeo(180)	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**
ConGeo(360)	**96.6**	**98.9**	**99.2**	**99.7**	83.8	94.2	96.1	98.8	38.2	58.2	65.2	82.0	19.5	35.9	43.8	66.5	59.5
ConGeo(0-360)	63.5	80.0	85.4	95.5	82.8	94.1	96.3	99.3	55.2	**77.3**	**83.0**	**94.7**	**46.4**	**67.1**	**75.2**	**91.6**	62.0

Table S2: Analysis of training ConGeo with different FoV angle α: 90°, 180°, 360°, and random angles between 0°-360°, on the CVUSA dataset. Sample4Geo+DA means training Sample4Geo with targeted data augmentation (random shift and random FoVs 70°-360°).

Loss weights		North-aligned			FoV=90°		
w_1 w_2	w_3	R@1	R@10	R@1%	R@1	R@10	R@1%
0.25	0.25	**98.7**	**99.7**	**99.9**	51.7	75.9	89.6
0.25	0.5	98.5	**99.7**	**99.9**	54.0	78.2	**90.9**
0.5	0.25	98.3	**99.7**	**99.9**	**55.9**	**79.0**	**90.9**
0.5	0.5	98.4	**99.7**	99.8	55.0	77.9	90.3

Table S3: Analysis of loss weights of single- and cross-view contrastive objectives on the CVUSA dataset.

LR	North-aligned			FoV=90°		
	R@1	R@10	R@1%	R@1	R@10	R@1%
0.001	97.6	99.5	99.8	41.4	66.8	83.8
0.0001	**98.3**	**99.7**	**99.9**	**55.9**	**79.0**	**90.9**
0.00001	95.6	99.4	99.8	28.8	59.2	89.4

Table S4: Analysis of the starting learning rate on the CVUSA dataset.

original ones, we construct the single-view contrastive objective by emphasizing the proximity between street images obtained from the same geographic locations but depicting different perspectives or angles. Since the street-to-satellite retrieval is a many-to-one matching, there is more than one street view image located in the region of satellite images. Therefore, for each street view sample, we randomly select another street image with the same location to construct the single-view contrastive objective for training.

S4 Hyper-parameter Analysis

S4.1 Training FoV Angle α

In order to achieve robustness across ground view variations, ConGeo applies a set of transformations on ground view images during training. The transformation includes shifting the ground view images with a random orientation angle θ and cropping the shifted images with a FoV angle α. In the main experiments, we empirically set α to 180°. Here, we provide detailed results by using different training FoV angles. Results are shown in Table S2. In the table, we report models trained with FoV images from 90°, 180°, 360°, and random FoV from 0° to 360° on the unknown orientation and FoV settings. It can be seen that ConGeo's performance is robust under different training FoVs, which consistently surpasses the baseline with targeted data augmentation by a large margin.

S4.2 Loss Weights w_1, w_2 and w_3

We investigate the effect of loss weights on the retrieval performance. In Table S3, model performance with different loss weights on North-aligned setting and FoV=90° are reported. Results suggest that different loss weights may lead to slight performance waves. Specifically, when we gradually increase the w_1, w_2, and w_3, the performance of R@1 under the North-aligned setting will slightly decrease within 1 point, meanwhile, the limited FoV performance will be improved. When the single-view loss weight w_1 and w_2 value are set higher, there is a significant improvement on FoV=90° with around 4 points on R@1.

S4.3 Learning Rate

We also study the effects of the starting learning rate on training results. In Table S4, we take the CVUSA dataset for example, and verify the training performance under 0.001, 0.0001, and 0.00001, respectively. Results show that the best performance is obtained under a learning rate of 0.0001. Moreover, we can observe that the performance under the North-aligned setting is robust to learning rates while the limited FoV setting is sensitive to learning rates.

S5 Supplementary Experiments

S5.1 Architecture Analysis

In the main paper, we demonstrate that ConGeo, as a learning objective, can be plugged into different base models, including the CNN-based model [2] and ViT-based model [23]. To better compare those two mainstream architectures for cross-view geo-localization, we use Sample4Geo as the base model and switch the backbone between CNN and ViT in Table S5 to test the corresponding performance on the North-aligned settings and FoV=90°. In other words, the only difference between these two base models is the encoder architecture: Sample4Geo-CNN is with ConvNeXt [10] and Sample4Geo-ViT is with Swin Transformer [9]. From the results, we can see that ConGeo brings considerable improvement to both architectures, between them, the improvement with the CNN-based backbone is larger. We also compare the effect of training FoV angle (α) on different architectures. The results indicate that CNN-based architecture yields the best overall performance when setting α to 180°. The ViT-based architecture, although shows less competitive performance compared to the CNN-based one, seems to be more robust with different training angles α. Here, we present preliminary findings regarding the analysis of different architectures under unknown orientations and limited FoV settings. A more comprehensive investigation into this direction could serve as an intriguing topic for future research.

S5.2 Polar Transformation Analysis

As we reviewed in the *Related Works* section in the main paper, polar transformation is commonly used to improve the models' performance on the North-aligned

Method	North-aligned			FoV=90°		
	R@1	R@10	R@1%	R@1	R@10	R@1%
Sample4Geo-CNN	**98.7**	**99.8**	**99.9**	2.5	9.8	26.7
ConGeo-CNN[180]	98.3	99.7	**99.9**	**55.9**	**79.0**	**90.9**
ConGeo-CNN[Random]	98.1	99.6	99.8	31.0	58.0	77.2
Sample4Geo-ViT	**97.6**	**99.6**	**99.9**	1.7	6.5	20.8
ConGeo-ViT[180]	96.5	99.3	99.8	**51.0**	**77.8**	**91.9**
ConGeo-ViT[Random]	96.9	99.3	99.8	48.4	74.9	90.7

Table S5: Architecture analysis. The comparison of the base model and ConGeo (training FoV angle α set to 180 and random, respectively) with different backbones: CNN-based and ViT-based.

Method	North-aligned			FoV=90°		
	R@1	R@10	R@1%	R@1	R@10	R@1%
Sample4Geo w/o Polar	**98.7**	**99.8**	**99.9**	2.5	9.8	26.7
ConGeo w/o Polar	98.3	99.7	**99.9**	**55.9**	**79.0**	**90.9**
Sample4Geo w/ Polar	**98.8**	**99.7**	**99.8**	3.9	12.0	27.4
ConGeo w/ Polar	98.4	99.6	**99.8**	**39.0**	**67.7**	**86.6**

Table S6: Comparison between ConGeo and the base model when the aerial images are with or without polar transformation.

Method	CVUSA→CVACT								CVACT→CVUSA							
	FoV=360°				FoV=90°				FoV=360°				FoV=90°			
	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%
Sample4Geo	3.4	6.4	8.1	17.6	0.2	0.8	1.5	7.6	3.1	5.7	7.0	14.5	0.3	0.9	1.7	7.7
ConGeo	**4.2**	**9.8**	**13.3**	**33.6**	**1.0**	**3.4**	**5.2**	**19.8**	**5.7**	**12.2**	**17.1**	**39.6**	**2.2**	**5.3**	**8.1**	**24.9**

Table S7: Transfer results between CVUSA dataset and CVACT dataset on the unknown orientation setting and limited FoV setting.

setting [14, 23]. Polar transformation transfers the aerial images to the polar coordinates and narrows the view gap between aerial images and ground images. We compare the performance between the base model and ConGeo with the original satellite images or with the polar transformed satellite images in Table S6. Results indicate ConGeo can consistently improve the robustness, whether the aerial images are polar-transformed or not. Note that polar transformation enhances the spatial correspondences in the model. Therefore, compared with ConGeo without polar transformation, ConGeo with polar transformation is more likely to maintain strong performance in the North-aligned setting, while being less competitive in the limited FoV settings.

S5.3 Transfer between CVUSA and CVACT Datasets

The transfer analysis between CVUSA and CVACT datasets is often reported to show the generalization ability of the model on different datasets. We also report the transfer results of the base model and ConGeo on the unknown orientation setting and limited FoV setting in Table S7. Results suggest ConGeo can significantly improve the robustness of the base model (R@1% gains on average 20.6% on FoV=360° over the two settings).

S5.4 Different Paradigms of Aligning Modalities

ConGeo requires the alignment of different modalities. Besides contrastive learning (CL), other methods, e.g., direct feature alignment (FA), and redundancy reduction (RR), can also be used to align modalities. To investigate this, we also compare different paradigms of alignment modalities using InfoNCE [12] (CL),

Paradigm	Method	North-Aligned		FoV=90°	
		R@1	R@1%	R@1	R@1%
FA	Cosine Similarity	86.6	99.6	31.2	81.9
RR	Barlow Twins [19]	85.1	99.2	41.4	82.1
CL [**ConGeo**]	InfoNCE [12]	**98.3**	**99.9**	**55.9**	**90.9**

Table S8: Comparison of different paradigms of aligning modalities. "FA", "RR" and "CL" denote direct alignment, redundancy reduction and contrastive learning, respectively.

Method	CVUSA				CVACT Val				CVACT Test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVM-Net [6]	22.47	49.98	63.18	93.62	82.49	92.44	93.99	97.32	-	-	-	-
LPN [17]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
SAFA [13]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
CVFT [15]	61.43	84.69	90.49	99.02	61.05	81.33	86.52	95.93	-	-	-	-
DSM [14]	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	-	-	-	-
CDE [16]	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22	61.29	85.13	89.14	98.32
L2LTR [18]	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37	60.72	85.85	89.88	96.12
SEH [4]	95.04	98.31	98.92	99.76	85.13	93.84	95.24	97.97	-	-	-	-
TransGeo [23]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR [21]	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	98.74
SAIG-D [25]	96.34	99.10	99.50	99.86	89.06	96.11	97.08	98.89	67.49	89.39	92.30	96.80
Sample4Geo [2]	98.68	99.68	99.78	99.87	90.81	96.74	97.48	98.77	71.51	92.42	94.45	98.70
ConGeo	98.27	99.59	99.70	99.86	90.12	95.69	96.56	98.24	71.67	91.61	93.50	98.30

Table S9: The full table of comparison of ConGeo and state-of-the-art methods on the North-aligned setting on the CVUSA and CVACT.

Barlow Twins [19] (RR) and Consine similarity (FA) to represent each paradigm respectively. The results shown in Table S8 suggest that aligning modalities helps to improve robustness, but due to the huge cross-modal view gap and single-view information asymmetry, FA fails to reduce the feature distance, RR struggles to find the shared features, while CL excels in learning by comparing.

S6 Supplementary Results

S6.1 Full Table of the North-aligned Setting

We provide the full comparison of the North-aligned setting on CVUSA and CVACT datasets in Table S9. Among all the methods, ConGeo achieves competitive performance on the North-aligned setting on both datasets meanwhile maintaining robustness across different challenging settings. The results indicate the effectiveness and versatility of the proposed contrastive objectives.

S6.2 Full Tables of Ablation Studies

We provide the results of component ablations and the comparison with different data augmentation methods in Table S10 and Table S11, respectively.

As shown in Table S10, through the full results provided, we can further confirm the individual effectiveness of each component in the proposed ConGeo.

$\mathcal{L}_{\text{single-r}}$	$\mathcal{L}_{\text{single-q}}$		$\mathcal{L}_{\text{cross}}$		FoV=360°				FoV=180°				FoV=90°				FoV=70°			
	Shift	FoV	Shift	FoV	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%
					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
✓					30.1	48.4	57.1	79.1	15.7	28.2	34.7	60.0	7.7	16.2	21.5	43.3	5.8	12.9	17.4	38.9
✓	✓				89.7	96.9	97.9	99.5	44.1	70.3	78.9	93.4	17.8	35.8	44.8	73.0	12.0	26.5	33.7	56.3
✓	✓	✓			28.6	45.9	54.8	83.3	37.9	56.1	62.4	81.2	20.5	34.9	40.8	59.9	14.2	26.3	32.3	53.4
✓			✓	✓	73.8	88.6	91.9	97.3	91.5	97.8	**98.8**	**99.7**	40.2	69.6	75.7	89.0	29.6	48.3	55.9	76.0
✓	✓		✓		**96.5**	**98.9**	**99.4**	**99.7**	81.7	93.4	95.7	98.8	35.8	56.2	63.5	81.6	20.3	37.0	44.3	67.6
✓	✓	✓	✓	✓	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**

Table S10: The full table of ablation studies on FoV=70°, 90°, 180°, and 360° on the CVUSA dataset. "Shift" and "FoV" mean cyclic shift and FoV cropping. "Single" and "Cross" denote single-view contrastive objective cross-view contrastive objective, respectively.

Augmentation			FoV=360°				FoV=180°				FoV=90°				FoV=70°			
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%
			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
	✓		9.6	17.6	22.5	46.8	6.0	10.8	13.7	29.8	3.4	6.8	9.5	24.7	2.6	5.6	7.1	21.1
✓			**93.1**	**97.6**	**98.2**	99.1	73.8	88.0	90.9	95.8	35.1	54.2	61.2	77.6	18.9	34.0	41.5	62.5
✓	✓		84.1	95.0	96.9	**99.4**	63.6	84.6	90.1	98.2	32.2	55.1	64.5	87.4	21.5	42.0	51.6	79.9
✓	✓	✓	89.0	96.1	97.3	98.9	71.8	87.9	91.3	97.2	39.3	59.9	67.9	85.7	29.8	50.0	58.0	79.3
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**

Table S11: The full table of comparison with different data augmentation methods on FoV=70°, 90°, 180°, and 360° on the CVUSA dataset. "Shift" applies random cyclical shift, "FoV" means applying random FoV cropping (from 70° to 360°) to ground images as data augmentation for training, 'Rotate" randomly rotating aerial images with an angle in {90°, 180°, 270°} as data augmentation.

Moreover, the combination of single-view and cross-view contrastive objectives can further boost the model performance indicating that they can complement each other's alignment. Besides, we can see from the results that merely applying "Shift" transformations (FoV=360°) to the ground view images in single-view and cross-view contrastive objectives achieves the best performance, showing that the model performance on specific FoVs can be further improved by using the evaluation FoV for training like in previous methods.

Experimental results in Table S11 confirm the superiority of contrastive objectives over data augmentation methods. Although data augmentation methods can improve the model performance in specific settings, they struggle to obtain robustness across various settings. For example, when randomly shifting the panorama image as data augmentation, the retrieval performance under FoV=360° can be marginally better than ConGeo, while it trails ConGeo by about 20 points at other settings.

S7 Visualization Results

S7.1 Retrieval Results

More visualization results of the North-aligned setting and FoV=90° on the CVUSA dataset are provided in Fig. S1. We make two main observations:

1) Compared to ConGeo, the top candidates retrieved by the baseline method are of similar geometric structure (*e.g.* road direction). When provided with full-

view panoramas, both the baseline and the proposed ConGeo can find the target aerial images. When using the limited FoV setting, the spatial correspondence between the ground view images and the paired aerial images is partially broken, which brings difficulties to the models. As shown in the second and third examples, the baseline model's results are fairly limited to similar road structures (North-south direction), however, with the contrastive objective, the top retrieval candidates exhibit more diverse structures (arbitrarily orientated roads).

2) Compared to the baseline, the top candidates retrieved by ConGeo are of distinguishable similar content. It indicates that, besides the geometric information, ConGeo also focuses on the consistent semantic content shared by both views. This is particularly noticeable when the FoV is limited. For example, in the first example, ConGeo notices the buildings in the ground views are with the same roof color and returns results that are consistent with that. However, even given color clues in the query image, the baselines' retrieval results are with different roof colors. The second and third examples show that ConGeo can retrieve images with the distinguishable semantic cues in images (*e.g.*, water and building), while the baseline fails to capture them.

For better understanding, we plot the rank of ground truth reference images in the retrieval results of the base model and of ConGeo on the left part of Fig. S2, when FoV=90°. For example, $rank = 0$ means the reference image is correctly retrieved by the corresponding model. Compared with the baseline model, the distribution of ConGeo is gathered on the head and is much sparser in the tail, which demonstrates that ConGeo not only achieves better performance among metrics but also improves the feature representation of hard examples.

S7.2 Visualization of Activation Maps

We provide more cases of activation map visualization in Fig. S3 and Fig. S4. As mentioned in the paper, we use Grad-CAM [3] to visualize input regions that contribute the most to the model's predictions, on the CVUSA dataset. The feature maps are extracted from the ConvNeXt blocks of the model and superimposed with the images.

In Fig. S3, we compare the activation maps of different orientations of the panorama. Since the shifted image shares same content with its prototype, the activation areas of a robust model are supposed to be similar across different orientations. We find that the base model's focus is vulnerable to orientations, while ConGeo's attention is more robust. For example, in the first example (the left of the first row), ConGeo's activation areas are distributed in similar regions of the buildings, while the baseline fails to keep consistent across the North-aligned and unknown orientation settings. In the last example (the right of the last row), the ground view variations make the base model's attention drift to the sky which might not carry geospatial information, while ConGeo consistently highlights the road trees.

In Fig. S4, there are comparisons between the baseline and the proposed ConGeo on the North-aligned setting and the limited FoV setting. Compared with the baseline, the activation areas of ConGeo focus more on the coherent semantic objects across views. For example, in the first example (the first two rows), the activation areas of ConGeo (the second row) are mostly on the building along the road while for the baseline model (the first row), the attention is misleading (on the sky). This problem will be severe when FoV=90°, the activation areas of the baseline model are distributed on the sky. In addition, we observe that ConGeo alleviates the model's over-reliance on spatial correspondence. For example, from the third example, we can see that the baseline model mostly relies on the road structure to match the full-view panorama with the aerial view image. When the spatial correspondence is broken as in the FoV images, the model fails to find the cues merely based on the learned geometric structure. However, for ConGeo, with the help of semantic content, the model learns to retrieve the aerial image with the incomplete geometric structure. The activation map analysis further confirms that ConGeo is more robust to view variations.

S7.3 Failure Cases: What Kind of Samples Are More Sensitive?

Our method may fail when very limited information is provided in the query image. The first case is that the query image is of a very small FoV. As can be seen from Table 1 of the main paper, the retrieval performance under very limited FoV (*e.g.* 70°) is significantly lower than other settings. The second case is that the query image contains merely single and common layouts. Some samples exhibit greater sensitivity to the robustness test, making it more challenging for the model to uphold its robustness on such samples. The model will struggle to distinguish the corresponding aerial image from similar counterparts given common layouts.

We show the examples on the right part of Fig. S2 when the rank of ground truth annotations are 14,000, and 8,000. In general, when the query image only contains grass or trees, the similarity score of the ground truth aerial image may be drowned in other similar candidates. Analysing the sensitivity of different samples can also help improve model robustness and can be an interesting direction for future research.

S8 Discussion

In this work, we propose ConGeo to boost the base models' robustness across ground view variations. Without specialized training, ConGeo outperforms state-of-the-art methods on unknown orientation and limited FoV benchmarks and demonstrates adaptability to three different base models and generalizability to various unseen ground view variations. However, as mentioned in the *Limitations* section of the main paper, several aspects and challenges remain to be addressed for future research.

First, compared to models trained specifically for the North-aligned setting, ConGeo faces an almost unavoidable slight performance drop in this setting. Indeed, when performing orientation-specific training, North-alignment is a key assumption and models often take shortcuts by using the spatial correspondence in the data stemming from this alignment. However, ConGeo aims to relax it and achieve robustness in more challenging settings, where such shortcuts will reduce the model's generality. Moreover, it is also worth noting that the North-aligned setting is a special case of FoV=360°, where the orientation angle is set to 0°. Therefore, the performance drop in the North-aligned setting is almost unavoidable when improving the unknown orientation setting, which also indicates a trade-off between orientation-specific or FoV-specific spatial correspondence and the robust semantic information across ground view variations. Despite this slight decrement, ConGeo serves as a practical solution to real-world scenarios: When the orientation is known, ConGeo performs comparably to the state-of-the-art methods, while the model significantly improves the performance when the orientation is unknown.

Second, in this work, we increase the models' robustness by proposing a new learning pipeline; however, other aspects of the model can face robustness issues and could be discussed, such as network architecture and sample-level information. As mentioned in Section 5.3 of the main paper and Section S5.1 in this supplementary material, ConGeo can consistently improve the models' robustness regardless of the backbones. Future research can further study the differences between different backbones and different components in the base models that lead the model to prefer learning shortcuts in the data over robust features. As discussed in the previous work in image classification [5], some backbones are more likely to rely on spurious correlations or cues inessential to the object. Moreover, as we showed and discussed in Section S7.3 in this supplementary material, some samples with less salient semantic information might be more sensitive to ground view variations. This is further supported by prior research on content-based adversarial attacks [1], which indicates that certain elements within an image are more susceptible to variations. Thus, further research might improve robustness across ground view variations from the data perspective.

Finally, in this work, we mainly focus on ground view orientation and reduced FoV variations. However, we also validate the effectiveness of the model under several unseen variations (Table 8 in the main paper), such as blur and zooming. Expanding the robust geo-localization across a broader range of ground view variations (*e.g.*, nighttime or extreme weather conditions) can further widen the applicability of this research to real-world use cases.

References

1. Chen, Z., Li, B., Wu, S., Jiang, K., Ding, S., Zhang, W.: Content-based unrestricted adversarial attack. Advances in Neural Information Processing Systems **36** (2024)
2. 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)
3. Gildenblat, J., contributors: Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam (2021)
4. 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)
5. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
6. 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)
7. Kwon, J., Kim, J., Park, H., Choi, I.K.: ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In: International Conference on Machine Learning. pp. 5905–5914 (2021)
8. 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)
9. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE International Conference on Computer Vision. pp. 10012–10022 (2021)
10. 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)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
13. 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)
14. 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)
15. 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)
16. Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to Earth: Satellite-to-street view synthesis for geo-localization. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6488–6497 (2021)
17. 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)

18. 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)
19. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: International Conference on Machine Learning. pp. 12310–12320 (2021)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)

Algorithm 1 ConGeo: PyTorch-like Core Pseudocode

```
# ==========================Operations==============================
# E_q: query encoder: backbone + proj mlp
# E_r: reference encoder: backbone + proj mlp
# T_q: transformations: random shifts + FoV cropping
# T_r: transformations: data augmentations for the aerial view
# ==========================Parameters==============================
# theta: shift angle
# alpha: training FoV angle
# w_1, w_2: loss weights
# tau: learnable temperature in the loss function
# ==========================Objectives==============================
# ctr(Q, R): vanilla contrastive objective
# ctr(Q*, Q): ground view contrastive objective
# ctr(R*, R): aerial view contrastive objective
# ctr(Q*, R): cross-view contrastive objective

for I_Q, I_R in zip(dataloader_q, dataloader_r):
    # I_Q: a batch of query images {I_q}
    # I_R: a batch of reference images {I_r}
    I_Q* = []
    I_R* = []
    for I_q in I_Q:
        I_q* = T_q(I_q | theta, alpha)
        I_Q*.append(I_q*)
        I_r* = T_r(I_r)
        I_R*.append(I_r*)
    Q, Q* = E_q(I_Q), E_q(I_Q*) # queries: [N, C]
    R, R* = E_r(I_R), E_r(I_R*) # references: [N, C]

    loss = ctr(Q, R)+w_1*ctr(Q*, Q)+w_2*ctr(R*, R)+w_3*ctr(Q*, R)
    # the total loss
    loss.backward()

    update(E_q) # optimizer update: E_q
    update(E_r) # optimizer update: E_r

# contrastive objective
def ctr(Q, R):
    # Q: a batch of query features {q}
    # R: a batch of reference features {r}
    logits_1 = tau * mm(Q, R.t()) # [N, N] pairs
    logits_2 = logits_1.t()
    labels = range(N) # positives are in diagonal
    loss_1 = CrossEntropyLoss(logits_1, labels)
    loss_2 = CrossEntropyLoss(logits_2, labels)
    loss = (loss_1+loss_2)/2
    return loss
```

Notes: `mm` is matrix multiplication. `R.t()` is `R`'s transpose.

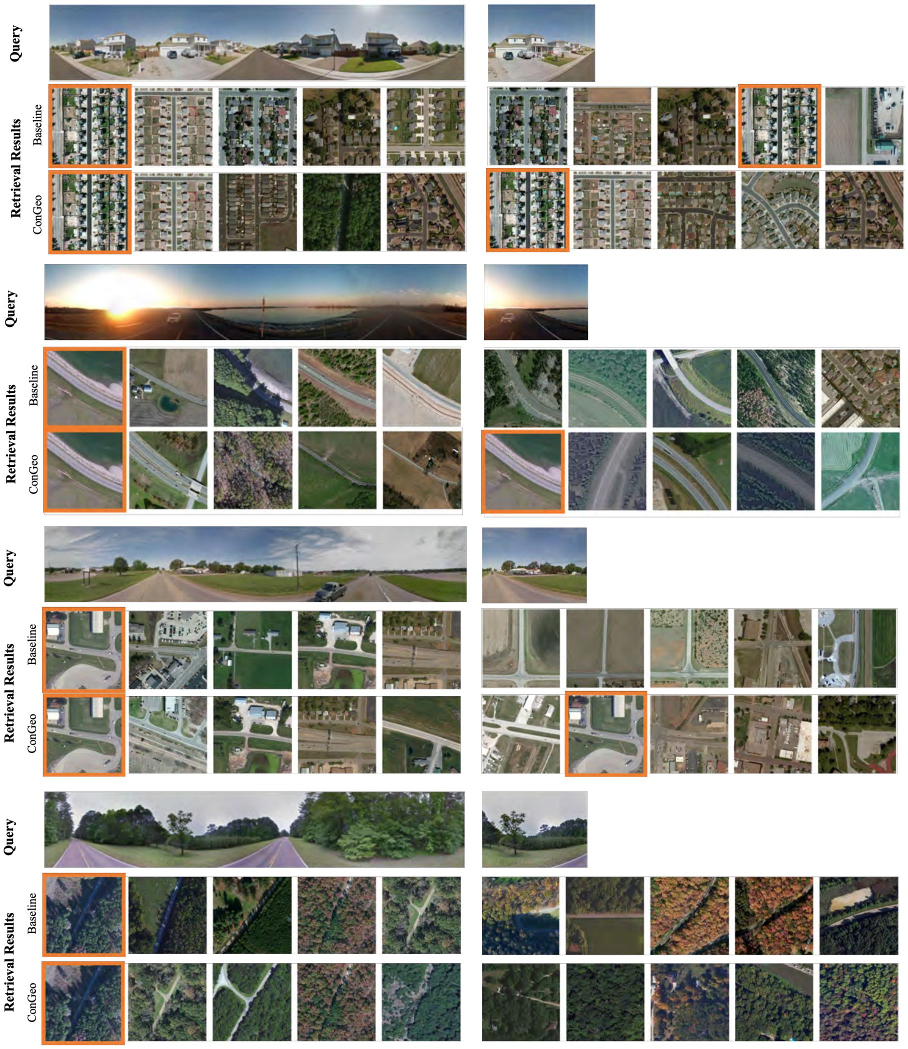
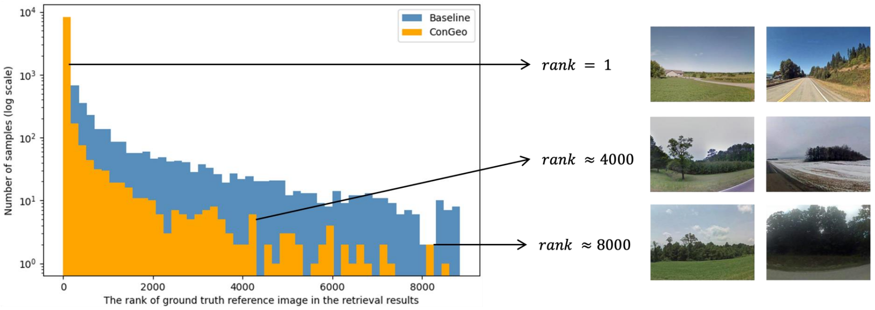


Fig. S1: Retrieval results of the baseline method and ConGeo in the North-aligned setting and FoV=90°. Retrieval results are ranked by the similarity score. Images marked in yellow denote the correct retrieval result.



(a) The distribution

(b) Examples of query images

Fig. S2: Distribution of the rank of ground truth reference images in the retrieval results when FoV=90° (left) and some examples of query images according to the retrieval results of ConGeo (right).

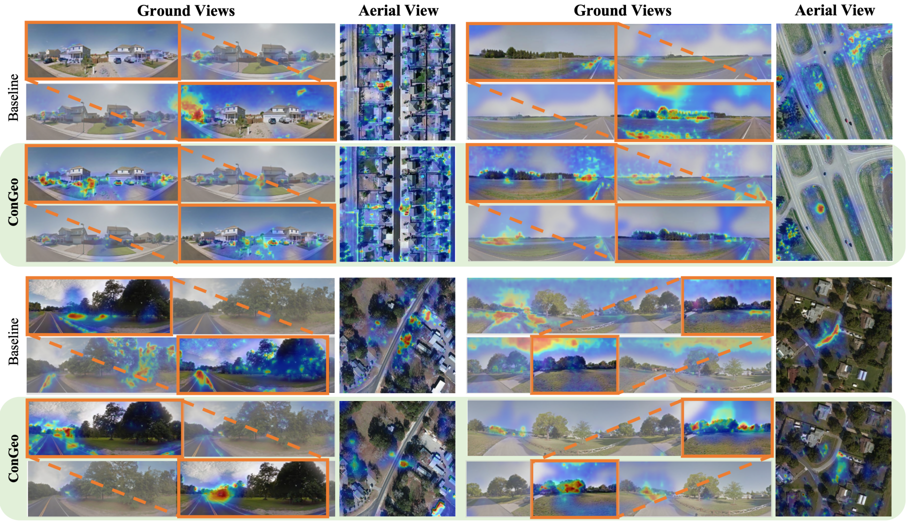


Fig. S3: Additional examples of the Grad-CAM activation maps of the base model and ConGeo on the North-aligned setting (the first row for each sample) and the unknown orientation setting (the second row for each sample). The orange box indicates the same area in different ground view variants.

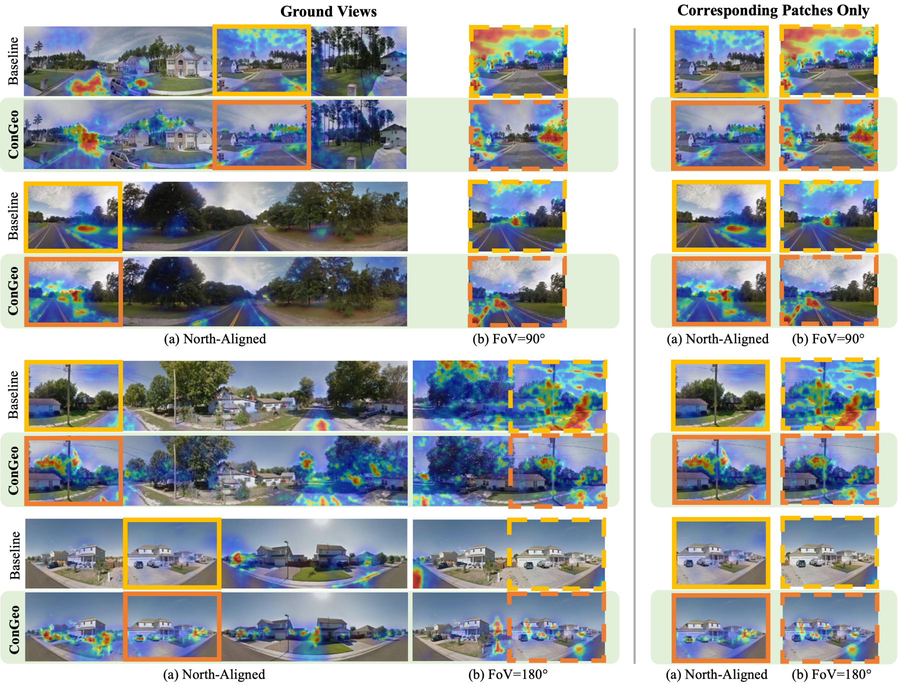


Fig. S4: Additional examples of the Grad-CAM activation maps of the base model (top row) and ConGeo (bottom row) on the North-aligned setting (a) and limited FoV setting (b) on the left. Corresponding patches from the two settings are shown on the right.