Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization

Guopeng Li[1], Ming Qian[2], Gui-Song Xia[1,2,†]
[1]School of Computer Science, Wuhan University [2]State Key Lab. LIESMARS, Wuhan University
{guopengli, mingqian, guisong.xia}@whu.edu.cn

Abstract

This paper investigates the effective utilization of unlabeled data for large-area cross-view geo-localization (CVGL), encompassing both unsupervised and semi-supervised settings. Common approaches to CVGL rely on ground-satellite image pairs and employ label-driven supervised training. However, the cost of collecting precise cross-view image pairs hinders the deployment of CVGL in real-life scenarios. Without the pairs, CVGL will be more challenging to handle the significant imaging and spatial gaps between ground and satellite images. To this end, we propose an unsupervised framework including a cross-view projection to guide the model for retrieving initial pseudo-labels and a fast re-ranking mechanism to refine the pseudo-labels by leveraging the fact that "the perfectly paired ground-satellite image is located in a unique and identical scene". The framework exhibits competitive performance compared with supervised works on three open-source benchmarks. Our code and models will be released on https://github.com/liguopeng0923/UCVGL.

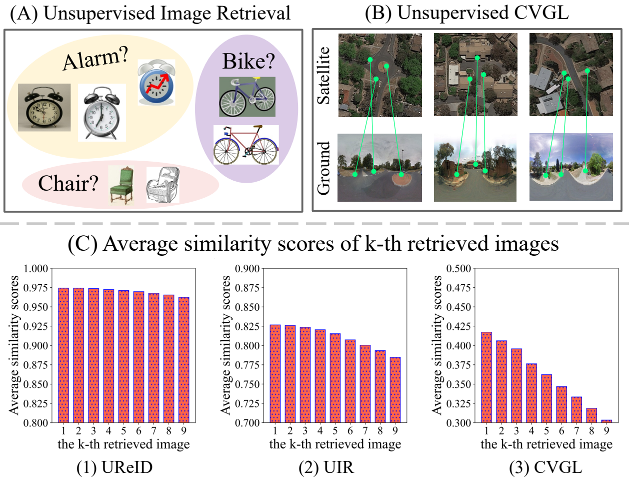


Figure 1. **Task Settings.** (A): Unsupervised Image Retrieval [16] has object-level images and applies class-level clusters by class semantics. (B): Unsupervised CVGL has scene-level images and applies cross-view alignments by spatial correspondences (*i.e.*, green lines). Compared with supervised CVGL, GPS labels and paired annotations (*i.e.*, correspondences between ground and satellite images) are not accessible in UCVGL. (C): Common image retrievals are class-level (*e.g.*, UReID [15] and UIR [16]), but CVGL [9] aims to align cross-view images in the same scene, which is fine-grained instance-level [9]. The Top-k most similar images are more discriminative without class semantics in CVGL.

1. Introduction

Large-area Cross-View Geo-Localization (CVGL) aims to determine the localization of ground images by retrieving the most similar GPS-tagged satellite images [53]. The easy availability of open-source GPS-tagged satellite images, such as Google Maps, helps users get accurate ground localization at a low cost [9, 40] and provides CVGL with great potential for various real-life applications, including person localization, automatic navigation, and augmented reality [9, 21, 27].

However, existing CVGL methods are usually supervised training with ground-truth correspondences of Ground images (Grds) and Satellite images (Sats), known as SCVGL. This label-driven nature brings some practical limitations. Firstly, obtaining accurately located ground images requires expensive devices (*e.g.*, lidar devices [11, 19]), and matching ground and satellite images brings extra human costs [51]. Secondly, numerous non-corresponding ground and satellite images without GPS tags are available online but cannot be directly used for SCVGL. Lastly, SCVGL requires re-annotation for new or changed scenes, increasing the cost of human effort and resources. Given these limitations, an important question arises: can we uncover hidden correlations from cross-view images without label information? This motivates us to study how to utilize unlabeled data to start a retrieval system for CVGL.

Regrettably, just as a ship needs a compass to find its way, without labels, the model may lose its optimized direction. This phenomenon is usually called *cold-start problem* [26] for un- and semi-supervised learning.

[†] Corresponding author

To mitigate this issue, most unsupervised image retrievals (UIRs) label an informative subset of images first, guided by some common prior knowledge (*e.g*., category [16] or instance [5, 15] information). Compared to them, CVGL has some unique characteristics illustrated in Fig. 1 and more details in Sec. 2.2. Specifically, existing retrievals [5, 8, 15, 16] have object-level cluster centers, such as some alarm, person or object ID, which includes many images from the same category or instance and enlighten them to apply cluster technologies [10, 13] to get initial label sets, as can be seen in the upper row of Fig. 1. Differently, most CVGL [9, 53] methods are trained by one-to-one cross-view image pairs to learn spatial correspondences, which means, "class" information is not essential in the training process and CVGL is a fine-grained retrieval. The reference satellite images can only be retrieved by ground images in the same scene position by spatial correspondences, just like green lines in Fig. 1(B). Besides, in Fig. 1(C), we collect the similarity scores between each image and its k-th retrieved image and average the scores of more than 5000 image pairs in different tasks. It's clear that compared to existing instance retrieval [15] and image retrieval [16], the average similarity scores of CVGL [9] are lower and drop fast from 1-st to the right. This comparison further suggests our task is non-class and fine-grained. Therefore, we cannot cold-start the unsupervised CVGL using existing image retrieval methods.

In this paper, we design a framework for unsupervised and semi-supervised CVGL. Following a common design [7, 16], in the unsupervised setting, we separate our framework into two components: the cold-start stage to produce initial pseudo-labels and the semi-supervised stage to refine labels. Moreover, the semi-supervised stage can also be utilized alone for CVGL (details in Sec. 4.4). In our framework, non-class and spatial corresponding characteristics are fully utilized, making the unsupervised CVGL possible and the semi-supervised CVGL a good performance.

In the cold-start stage, the significant imaging and view gaps make ground and satellite images distinct features[32] in the same scene. To address this issue, we attract cross-view images in the same position by spatial corresponding prior knowledge. Specifically, we project the ground panorama into a 3D coordinate system and obtain its bird's-eye-view image through a spherical transform [35, 40]. We then fill the unknown regions and reduce imaging gaps by a CycleGAN model [50], resulting in a fake satellite image (*e.g*., Fig. 2(B)) aligned with the ground panorama. Finally, we replace unknown ground-satellite pairs with known ground-fake pairs as trained positive pairs, enabling successful retrieval of more than 40% image pairs on CVACT and achieving cold-start initialization.

In the semi-supervised stage, the initial labels may be noisy or insufficient, which limits the final performance. We re-rank pseudo-labels to enhance the quality and quantity of pseudo-labels by using non-class characteristics. Specifically, the perfectly paired cross-view images are two views of the same scene, which should be mutually retrievable and dissimilar to other images, as shown in Fig. 1(C-3). We filter out many "error" labels that do not meet this requirement with a threshold, improving the correctness ratios of the remaining labels from around 30% to 80% on CVACT. However, the number of pseudo-labels is only about 20% of the dataset. Therefore, as the model's knowledge increases, we progressively introduce more challenging samples by adjusting our filtering mechanisms to get more pseudo-labels. Eventually, our model achieves competitive performance with recent supervised approaches [9, 53, 54].

2. Releated Work

2.1. Large-area Cross-View Geo-Localization

Ground images (Grds) and Satellite images (Sats) are usually captured by different sensors with huge perspective differences and different environmental conditions, making CVGL challenging.

In order to simplify CVGL, many works leverage two assumptions in popular datasets [22, 42, 51]. 1) Localization Alignment [54]: the localization of Grds is aligned to a known position of Sats. 2) Orientation Alignment [33]: the geographical orientation of Grds is aligned to Sats. By leveraging these alignments, some works [2, 22, 29, 34, 38] implicitly enhance the high-level features supervised by ground-truth correspondences. Recently, some methods [32, 37, 39, 46] improve the spatial alignments of input images by the explicit polar transform for better performance, but the transformed images suffer from severe distortions that also need to be corrected by full supervisions.

To make CVGL more practical, some methods try to get rid of one of two assumptions. Firstly, [9, 32, 43, 53, 54] have achieved good results without relying on Localization Alignment, but they still rely highly on ground-truth correspondences or GPS labels [9]. Secondly, although some works [18, 27, 33, 52] make efforts to remove the Orientation Alignment, they need to align the rotation angle of ground images and polar transformed images [32, 33] with known cross-view image pairs [53].

Different from existing works, in this paper, we no longer rely on ground-truth labels and instead propose a more challenging and practical setting called Unsupervised CVGL (UCVGL) to unleash unlabeled data.

2.2. Unsupervised Image Retrieval (UIR)

Unsupervised Image Retrieval (UIR) is a well-studied area, but Unsupervised Cross-View Geo-Localization (UCVGL) presents unique challenges and characteristics that require specialized approaches. Existing image retrieval methods

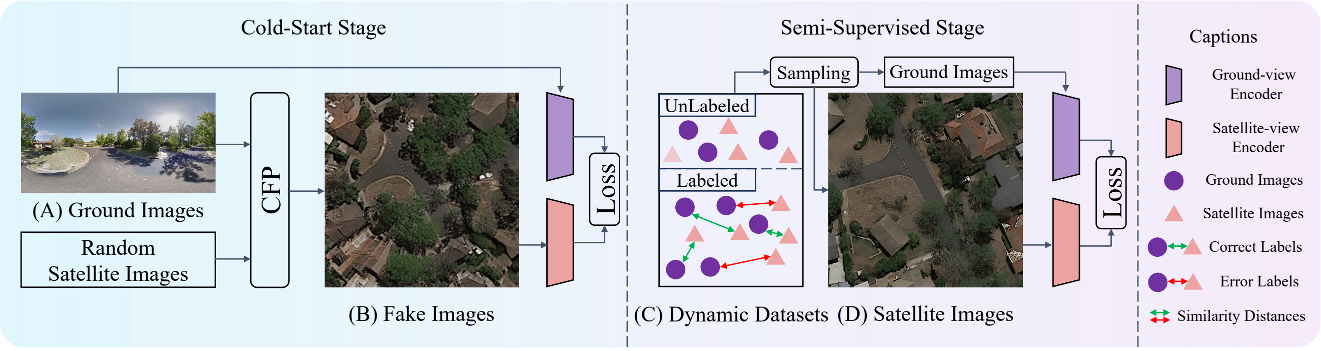


Figure 2. **Pipeline Overview.** Firstly, we train two separate encoders with ground panoramas and projected images to initialize a cross-view feature space for solving cold-start problems. Secondly, we train ground-satellite image pairs by sampling from adaptive pseudo-labels.

are not directly applicable to UCVGL due to the specific requirements of fine-grained, cross-domain, and cross-view retrieval [3, 6, 8, 16, 17, 30, 47].

Firstly, traditional instance-level methods aim to retrieve all reference images containing the object depicted in the query image [15, 16], generating pseudo labels through clustering similar features. However, this approach is not suitable for UCVGL, which necessitates fine-grained retrieval based on unique spatial correspondences [53].

Secondly, Unsupervised Cross-Domain Image Retrieval [16] aligns two different domains by reducing the distances between limited class centroids using self-supervised pre-trained models [6, 16, 17, 41, 45]. However, aligning cross-view images in UCVGL is challenging due to the absence of class semantics and large spatial gaps. Besides, self-supervised pre-trained models are unnecessary for UCVGL.

Lastly, existing cross-view retrieval methods often focus on handling minor view variances through feature matching technologies [30] or cluster algorithms [3, 8]. However, these methods are inadequate for CVGL as ground and satellite images have significant perspective differences, occlusions, and illumination variations.

In summary, the objectives and settings of existing retrievals do not align with the requirements of UCVGL. This necessitates the development of specialized approaches that consider the unique challenges posed by fine-grained, cross-domain, and cross-view retrieval in UCVGL.

3. Method

Problem Statement: Given the set of ground images $\{I_{\text{grd}}\}$ and satellite images $\{I_{\text{sat}}\}$, the objective of CVGL is to learn an embedding space where each I_{grd} is close to its *nearest* corresponding ground-truth I_{sat} [53]. In SCVGL [9], each ground-view image and its corresponding satellite image are trained as a positive pair, and other cross-view pairs are trained as negative. Differently, we propose a more practical setting, *i.e*. Unsupervised CVGL (UCVGL), which trains the models without GPS labels and the ground-truth correspondences between ground and satellite images.

Method Overview: Our framework can implement cold-start initialization for unsupervised learning and separate semi-supervised learning in CVGL, as illustrated in Fig. 2. Without any labels, we design a cold-start stage to align cross-view images and label some data as pseudo-labels. Subsequently, we refine this pseudo-label set or a given label set, and inject more samples adaptively as the models' knowledge increases.

Soft Symmetrical InfoNCE Loss: Our encoders are trained with Symmetric InfoNCE Loss [9, 14]:

$$\mathcal{L}(q,R)_{\text{InfoNCE}} = -\log\left(\frac{\exp(q\cdot r_+/\tau)}{\sum_{i=0}^{R}\exp(q\cdot r_i/\tau)}\right) \quad (1)$$

Here, q denotes an encoded query image, and R is a set of encoded reference images. The InfoNCE loss is low when the query q and the positive match r_+ are similar and high when they are dissimilar. The temperature parameter τ is learnable. Given the prevalence of noisy labels throughout the training process, such as fake images in (B) and error labels in (C) of Fig. 2, we additionally smooth the loss using label smoothing [36].

3.1. Cold-Start for Unsupervised Learning

Supervised methods [9, 54] enable robust retrievals by human-annotated labels, but unsupervised approaches struggle to understand what retrieval feature spaces users want without labels. This phenomenon, known as the *cold-start problem* [26, 49], poses one of the most challenging obstacles in unsupervised learning for UCVGL. It implies the absence of direct information to align cross-view images in the same position with unlabeled ground-satellite databases. To this end, we design a Correspondence-Free Projection to project ground panoramas to satellite view by leveraging the spatial correspondences of cross-view images and imaging principle. The transformed images, such as (B) in Fig. 2, have the same position as ground images

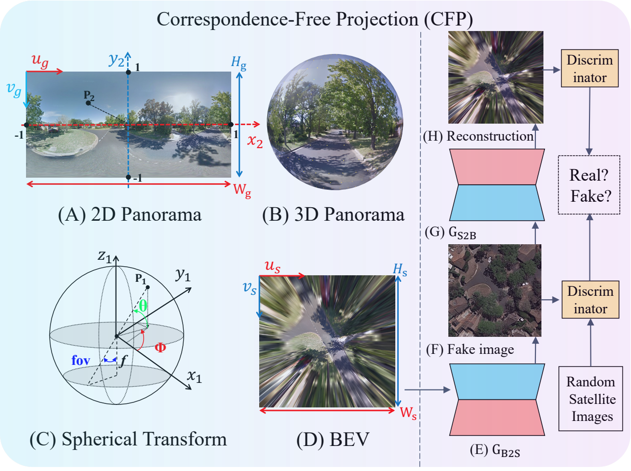


Figure 3. **Projections.** We project geometrically ground panoramas to BEV-view images on the left and transform BEV into fake images that resemble satellite images on the right.

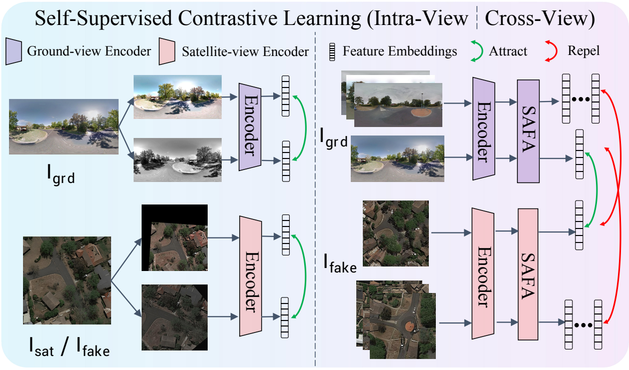
and resemble satellite images ((D) in Fig. 2). Then, we cleverly design a self-supervised objective to attract ground and projected images, leading to robust cross-view feature alignments. Lastly, our model successfully predicts more than 40% cross-view image pairs and provides 6012 pseudo labels with about 80% correct ratios (CVACT in Tab. 2).

3.1.1 Correspondence-Free Projection (CFP)

Without correspondences between $\{\mathrm{I}_{\mathrm{grd}}\}$ and $\{\mathrm{I}_{\mathrm{sat}}\}$, the implicit spatial alignments are hard to learn which forces us to leverage unique data-driven knowledge. We observe two differences in cross-view images: (a) they have very different perspectives because they are captured in the front-view and top-view camera positions. (b) they have different styles of images caused by different camera sensors and imaging processes. In this section, we bridge the two gaps by Geometric Projecting and Imaging Projecting without any ground-truth correspondences, and the projected images achieve explicit spatial alignments from ground-view images to satellite-view images.

Geometric Projection [35, 40]: As shown in (A-D) of Fig. 3, we denote the points of the 3D world by $\mathrm{P}_1 = (x_1, y_1, z_1)$ or $\mathrm{P}_1 = (\phi, \theta)$, the field of view of the bird's-eye view (BEV) by fov, the points of panoramas (the width and height are W_g and H_g) by $\mathrm{P}_2 = (u_g, v_g)$, and the points of BEV-images (the width and height are W_s and H_s) by $\mathrm{P}_3 = (u_s, v_s)$. The focal length of the BEV in the imaging process is $f = 0.5\mathrm{W}_s/tan(fov)$. We can transform the points P_2 to P_3 according to the captured principles in the 3D spherical world:

$$\begin{cases} u_g = [1-\arctan2(\mathrm{W}_s/2 - u_s, \mathrm{H}_s/2 - v_s)/\pi]\mathrm{W}_g/2 \\ \quad v_g = [0.5 - \arctan2(-f, d/\pi)]\mathrm{H}_g \end{cases} \tag{2}$$



Self-Supervised Contrastive Learning (Intra-View | Cross-View)
Ground-view Encoder
Satellite-view Encoder
Feature Embeddings
Attract
Repel
$\mathrm{I}_{\mathrm{grd}}$
Encoder
Encoder
$\mathrm{I}_{\mathrm{sat}}$ / $\mathrm{I}_{\mathrm{fake}}$
$\mathrm{I}_{\mathrm{grd}}$
Encoder
SAFA
$\mathrm{I}_{\mathrm{fake}}$
Encoder
SAFA

Figure 4. **Self-supervised contrastive learning.** We learn intra-view discriminative features by attracting two self-augmented images and cross-view alignments by attracting cleverly ground images and projected fake images.

where $d = \sqrt{(\mathrm{W}_s/2 - u_s)^2 + (\mathrm{H}_s/2 - v_s)^2}$ denotes the distance between points of BEV and the camera's projective center. After the Geometric Projection, we obtained BEV images $\{\mathrm{I}_{\mathrm{bev}}\}$ that exhibit high geometric similarity to satellite images (*e.g.*, (d) in Fig. 3).

Imaging Projection [12, 50]: While we get BEV images that resemble satellite images in geometry, these two kinds of images still have some imaging gaps such as different distortions, illumination, occlusion, weather, and so on. To this end, we apply a generative CycleGAN model [20, 25, 48, 50] to decrease further the imaging gaps between the two-view images illustrated in (D-H) of Fig. 3. Specifically, we train the model with unpaired BEV and satellite images, after the training, the model can fill the distorted unknown areas in BEV and project the style of BEV to satellite-view images. By now, we obtain projected fake images $\{\mathrm{I}_{\mathrm{fake}}\}$ that exhibit high similarity to the remote sensing images (*e.g.*, Fig. 2 (B)), suggesting the explicit projection aligns the spatial and imaging difference of cross-view images. Note that our projection is differentiable, correspondence-free, and performs one-to-one mappings from ground to fake images.

3.1.2 Self-Supervised Contrastive Learning

Supervised CVGL methods [9] learn both discriminative instance-level and aligned spatial features by ground-satellite pairs. In UCVGL, the ground-satellite pairs are not accessible, but we can replace them with the previous ground-fake pairs produced by CFP. Specifically, we apply two self-supervised contrastive learnings illustrated in Fig. 4: intra-view learning to make each scene more discriminative and cross-view learning to align the spatial features of cross-view images in the same scene position.

Intra-View: Each same-view image is captured in an approximately independent scene in CVGL [9]. For this ob-

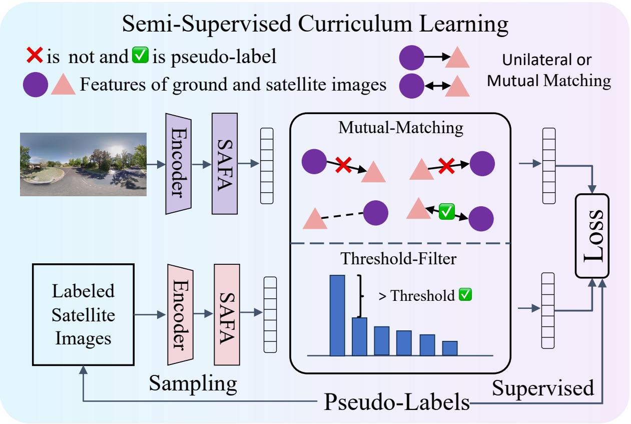


Figure 5. **Semi-supervised curriculum learning.** Ground images and satellite images are attracted and supervised with the guidance of adaptive pseudo-labels.

jective, we apply instance-wise contrastive learning methods that take only augmented images of the same image as positive pairs (*i.e.*, q and r_+ in Eq. (1)), while all other same-view samples in the dataset are regarded as negatives like [4]. After intra-view learning, two encoders have an initial feature space to distinguish different scenes, which facilitates the convergence of after steps and benefits for the quality of pseudo-labels (details in Tab. 2).

Cross-View: In CVGL [9], each cross-view image pair is aligned because they approximately represent the same scene, such as ground-satellite image pairs. Although the ground-satellite image pairs are not accessible, our ground-fake image pairs generated by CFP are actually cross-view representations for the same scene. So we think the model can learn robust spatial features to align the two views by training on ground-fake pairs. Specifically, a ground image and the corresponding fake image are a positive pair (*i.e.*, q and r_+ in Eq. (1)) and this ground image and other fake images are negative pairs. Advanced SCVGL methods ([9, 53]) learn spatial features without any spatial enhancement module [32, 33, 54] by aligning ground-truth pairs, but it is hard for our UCVGL because of the noisy pseudo-labels and unavoidable inconsistency between fake and satellite images. Therefore, an additional spatial attention module SAFA [32] is incorporated after the encoders to enhance spatial robustness.

3.2. Semi-Supervised Curriculum Learning

While the model trained by self-supervised contrastive learning successfully predicts certain ground-truth correspondences between $\{\mathrm{I}_{\mathrm{grd}}\}$ and $\{\mathrm{I}_{\mathrm{sat}}\}$, the predictions exhibit excessive noise (only about 30% predictions are correct on CVACT in Fig. 6 (C-a)). Some methods apply re-ranking mechanisms [3, 30] to refine the pseudo-labels according to their tasks, but they are slow to solve one hundred thousand level retrieval tasks (*i.e.*, CVGL). In this paper, we propose a rapid re-ranking mechanism termed Adaptive-Mutual-Matching (AMM) to refine labels by leveraging non-class and fine-grained characteristics of CVGL (*e.g.*, Fig. 1(C)). Based on these pseudo-labels (about 80% labels are correct), we start semi-supervised learning to align the features of ground and satellite images.

Mutual-Matching: Due to the ground and satellite images are actually one scene with different views, the truly aligned ground-satellite correspondences should be retrievable mutually. Formulaly, given the extracted features of i-th ground image $\mathrm{I}_{\mathrm{grd}}^{i}$ and the j-th satellite image $\mathrm{I}_{\mathrm{sat}}^{j}$, they are considered as a truly positive pair (*i.e.*, q and r_+ in Eq. (1)) if and only if $\mathrm{I}_{\mathrm{grd}}^{i}$ and $\mathrm{I}_{\mathrm{sat}}^{j}$ are mutually the most similar to each other. In this paper, we use the cos similarity to evaluate the similarity like [53].

Threshold-Filter: Although we filter most error labels by mutual matching, current pseudo-labels are also slightly noisy, especially for CVUSA (about 25% correct ratio of labels after *mutual-matching*). In CVGL, the objective is to find the most similar reference image to the query image [53], so the retrieved image pairs should have high similarity scores and be dissimilar from other images like (C-3) in Fig. 1. Therefore, we refine them using a threshold criterion. Specifically, we retain pseudo-pairs where the difference between the first-largest similarity and the second-largest similarity exceeds a threshold. This allows us to establish high-quality pseudo labels (more details in Fig. 6 (B)). However, the filtered labeled image pairs are often considered easy for models, suggesting a potential risk of overfitting if we solely rely on training with these high-quality pseudo-labels.

Curriculum-Learning: When humans learn knowledge, they often learn easy courses first and then difficult courses after they have a certain foundation [1]. Motivated by this, we train a "stupid" model on easy samples, which are obtained by filtering the pseudo-labels using a high threshold. As the model's knowledge advances, we introduce progressively more difficult samples by lowering the filtering threshold and establishing more mutual-matching pairs. This iterative process ensures that our training samples diversify over time, mitigating the risk of overfitting. Besides, we also introduce other unlabeled data to learn negative pairs. Finally, we get a surprising performance that is comparable with the nearest supervised approach [9].

4. Experiment

4.1. Datasets and Metrics

Datasets. We conducted experiments on three cross-view panorama datasets, *i.e.*, CVUSA [42, 44], CVACT [22], and VIGOR [51]. CVUSA and CVACT consist of various images captured in rural and urban areas, each with 35,532 ground-satellite pairs for training and 8,884 pairs for

Approach	GT Ratio	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 [18]	100%	22.47	49.98	63.18	93.62	20.15	45.00	56.87	87.57	5.41	14.79	25.63	54.53
Liu [22]	100%	40.79	66.82	76.36	96.12	46.96	68.28	75.48	92.01	19.21	35.97	43.30	60.69
SAFA [32]	100%	81.15	94.23	96.85	99.49	78.28	91.60	93.79	98.15	-	-	-	-
SAFA† [32]	100%	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17	-	-	-	-
DSM† [33]	100%	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	-	-	-	-
L2LTR [43]	100%	91.99	97.68	98.65	99.75	83.14	93.84	95.51	98.40	58.33	84.23	88.60	95.83
GeoDTR [46]	100%	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
TransGeo [53]	100%	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
Sample4Geo‡ [9]	100%	97.83	99.63	99.75	99.89	87.49	96.56	97.50	98.98	60.57	89.50	92.99	98.92
Ours-small	100%	93.53	98.42	99.18	99.77	84.44	94.85	96.15	98.53	57.71	86.35	90.40	98.49
Ours-small	0%	87.90	95.86	97.51	99.63	82.96	92.96	94.43	97.37	58.85	84.27	88.16	97.45
Ours-base	0%	92.56	97.67	98.55	99.61	84.58	93.95	95.29	97.59	60.53	86.35	89.77	97.52
Ours-base*	10%	94.88	98.80	99.36	99.77	87.89	95.27	96.40	98.37	65.30	89.47	92.15	98.27

Table 1. **Quantitative comparisons.** † denotes satellite images are projected to ground view by polar transform [32]. ‡ denotes results without hard negative sample sampling in [9]. $*$ denotes we fine-tune unsupervised models on 10% labeled images. The encoders of ConvNeXt-Small and ConvNeXt-Base are shortened as small and base [23].

evaluation. Besides, CVACT has an additional test dataset including 92,802 image pairs. In the two datasets, each satellite-view image has one corresponding street-view image, and all pairs are aligned with the similar spatial localization [54] and orientation [33]. We evaluate the results of our un- and semi-supervised settings in the two datasets. VIGOR has 105,214 panoramas and 90,618 aerial images from four cities. Due to the complex city scenes, the projection between ground and satellite images is hard on VIGOR, even in supervised settings [28]. So it is difficult for us to solve the cold-start problem by simple homography projection [31] in unsupervised settings. We provide a semi-supervised method for it. To the best of our knowledge, we are the first to remove GPS labels and ground-satellite pair annotations in CVGL.

Evaluation Metrics. We evaluate the retrieval performance by top-k recall accuracy (R@k) [9]. Specifically, for each query ground image, we retrieve the k nearest reference neighbors in the embedding spaces based on cosine similarity. A retrieval is considered correct if the ground-truth satellite image appears among the top k retrieved images.

4.2. Implementation Details

We train the CFP with the CycleGAN model [50] for 50 epochs. We use ConvNeXt-Base [23] with 384 × 384 input images in Ours-base of Tab. 1 and ConvNeXt-Small with 224 × 224 input images in other experiments on the CVUSA and CVACT datasets. We use ConvNeXt-Base [23] encoders with 384 × 768 like Sample4Geo [9] on VIGOR(Chicago). We use the AdamW [24] optimizers for 40 epochs of the cold-start stage, and 60 epochs of the semi-supervised stage. The threshold of our filter in Sec. 3.2 is 0.05 for CVUSA, 0.025 for CVACT, and 0.035 for VIGOR. Generally speaking, the higher threshold means the better quality of pseudo-labels (details in Fig. 6 (B)).

4.3. Results of Unsupervised Learning

As shown in Tab. 1, our unsupervised approach performs comparably to recent works on CVUSA, CVACT Val, and the large CVACT Test. This suggests that our model learns a robust retrieval and generalization ability without the supervision of human-annotated labels. When we fine-tune our unsupervised models on a few labeled images (10%), the models perform better on CVACT Val and CVACT Test. This further demonstrates the potential of our methods for unleashing unlabeled data.

4.3.1 Analysis of Cold-Start Stage

In the cold-start stage, we aim to initialize a feature space where cross-view images from the same scene are aligned and thereby enable the models to predict some pairs as pseudo labels. So we remove some components to validate their functions by analyzing the R@1 evaluated in the validation sets and corresponding quality of pseudo labels in the training stage in Tab. 2.

Ablation	CVUSA			CVACT		
(w/o)	R@1 ↑	Labels	Correct ↑	R@1 ↑	Labels	Correct ↑
(a) BEV	0.0225	14	0	0.0	6	0
(b) Fake	9.18	119	98	22.78	1349	923
(c) Intra	15.58	1476	798	42.98	5539	4281
(d) Cross	0.011	21	0	0.011	48	0
(e) Ours	17.96	1477	968	44.81	6012	4752

Table 2. **Ablation studies in the cold-start stage.** "Labels" and "Correct" denote the total and correct labels after the first stage.

Projection: In (a,b,e) of Tab. 2, we show the effect of our correspondence-free projections on CVUSA and CVACT. When we remove the geometric projection (*i.e.*, without BEV in (a)), the performance of our model drops significantly and fails to label a set, suggesting the model cannot

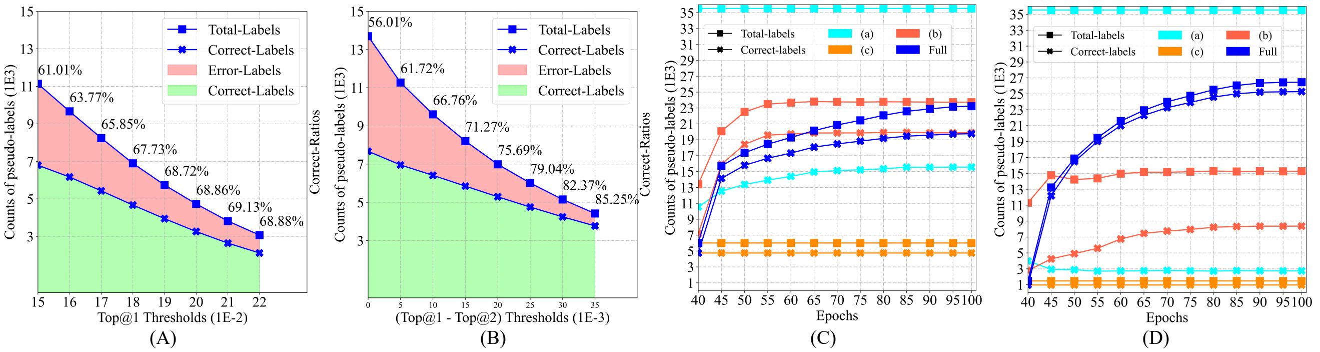


Figure 6. **Pseudo-labels.** (A) and (B) denote pseudo-labels produced by the highest or the difference value between the highest and second-highest retrieved similarity scores after the cold-start stage in CVACT. The right two figures are the trend chart of pseudo-labels' counts of Tab. 3 in (C) CVACT and (D) CVUSA. (a-c) denote our framework without *mutual-matching*, *threshold-filter*, and *curriculum-learning*.

align cross-view spatial information without prior geometry knowledge. When we remove the imaging projections (*i.e*., without Fake in (b)), R@1 of the model also drops about half, because the different imaging representations of the same scene in different views hinder their alignments. Besides, the ground panoramas are not completed [40] on CVUSA, which makes the geometric projection not work correctly and so the R@1 is lower than that on CVACT.

Self-supervised contrastive learning: In (c,e) of Tab. 2, intra-view learning has been shown to enhance the quality of pseudo labels, primarily attributed to its ability to enhance scene discriminability and mitigate same-view feature overlappings. In (d) of Tab. 2, without cross-view learning, the toward-zero performance suggests the importance of spatial correspondences again in CVGL.

4.3.2 Analysis of Semi-Supervised Stage

In the semi-supervised stage of our study, our objective is to initiate robust training processes using pseudo-labeled ground-satellite image pairs and dynamically refine the pseudo-labels. Therefore, we delve into the design choices during the assignment and refinement of pseudo labels, evaluating the different designs based on the final performance in the validation sets and the changes observed in the correct labels during training.

Ablation	CVUSA ↑		CVACT ↑	
(w/o)	R@1	R@10	R@1	R@10
(a) Mutual-Matching	14.40	35.43	61.90	82.80
(b) Threshold-Filter	36.52	56.10	82.46	94.04
(c) Curriculum-Learning	32.56	54.35	63.14	82.83
Ours	87.90	97.51	82.96	94.43

Table 3. **Ablation studies in semi-supervised learning.** "Ours" denotes our Adaptive-Mutual-Matching.

Final performance: In Tab. 3, we conduct some ablation studies to demonstrate the effectiveness of key technologies in the proposed adaptive mutual matching (AMM).

Firstly, we use the ground images and their most similar satellite images as initial pseudo-labels without our *mutual-matching*, which gets poor performance, especially for CVUSA. This is because the proportion of error labels is too large in Fig. 6 (C-a) (about 70% on CVACT), misleading the optimization direction. By digging into the unique characteristics of CVGL, we find that cross-view images are different representations of the same scene, so they should be successfully retrieved from each other. Therefore, we use the mutually matched cross-view images as pseudo-labels and improve their correct ratio from 11% to 25% on CVUSA and from 30% to 53% on CVACT.

Secondly, despite the improvement made by *mutual-matching*, the correct ratio of assigned pseudo-labels is still low. With these pseudo-labels, CVACT has good performance because it has enough correct labels (about 7,000 correct and error pairs), but CVUSA has bad performance with more than 10,000 error labels and 2,000 correct labels. After filtering the labels by our *threshold-filter* strategy, the correct ratio is improved from 25% to 65% in CVUSA and from 53% to 80% on CVACT (*e.g*., Fig. 6 (B)), and they both have a better final performance. So the *threshold-filter* is very important for starting a robust retrieval system.

Lastly, we freeze the initial pseudo-labels, and the performance drops a lot because the model only learns limited knowledge based on fixed samples. So we introduce more hard samples by lowering the threshold of our *threshold-filter* and matching more image pairs as the model's knowledge increases. After that, we get comparable performance compared to the recently supervised methods [9].

Labels change: In the left two images of Fig. 6, we compare two different filter choices in CVACT. It's clear that filtering pseudo-labels by the maximum similarity scores has low accuracy, but our methods not only have a higher correct ratio but also allow the correct ratios to rise gradually as the threshold increases. In other words, we don't need to set a well-chosen threshold because the higher the threshold, the better the quality of pseudo-labels. Actually, the design of our *threshold-filter* is motivated by a non-class

GT Ratio	CVACT			GT Ratio	VIGOR(Chicago)		
	R@1	R@5	R@10		R@1	R@5	R@10
10%	56.10	81.69	88.18	30%	36.82	65.16	74.28
1%	68.29	85.18	88.80	5%	25.82	42.81	49.81
5%	78.10	90.87	93.11	10%	44.17	63.30	69.81
10%	78.88	91.31	93.53	20%	55.90	75.44	81.20
20%	79.60	91.98	93.96	30%	60.42	80.12	84.88
100%	84.44	94.85	98.53	100%	68.40	88.49	92.44

Table 4. **Results in semi-supervised settings.** The first gray line denotes the results of supervised Sample4Geo [9] without hard negative sample sampling and the others are our results. “GT Ratio” denotes the ratio of ground-truth labels used for training.

characteristic like Fig. 1(C), each ground image should have the highest similarity with the corresponding satellite image and is dissimilar to images captured by other scenes.

In the right two images of Fig. 6, it is observed that our *mutual-matching* technology yields more precise initial pseudo-labels than direct predictions. The application of a *threshold-filter* strategy further enhances the quality of these labels, while the incorporation of *curriculum learning* ideas ensures a gradual increase in the number of pseudo-labels with a consistent ratio of correctly labeled samples. Moreover, our *threshold-filter* strategy is more important when initial labels are extremely noisy (*e.g.*, CVUSA). Although it discards many pseudo-labels, the ratio of correct labels is guaranteed, which guides a robust and positively optimized direction for our model.

4.4. Results of Semi-Supervised Learning

As shown in Tab. 4, we use a part of ground-truth labels to start semi-supervised learning on CVACT and VIGOR(Chicago). Firstly, with the same labeled images, *i.e.*, 10% labeled images on CVACT and 30% labeled images on VIGOR(Chicago), our method has better results than Sample4Geo [9]. In other words, we can improve the performance of existing supervised methods with our semi-supervised stage by unleashing unlabeled data. Secondly, we get good performance with only 355 images (1%) on CVACT and 1274 images (10%) on VIGOR(Chicago). This suggests our methods can work guided by a few labels and leverage unlabeled data to refine it for better performance. Lastly, the more ground-truth labels have a limited improvement on CVACT (*e.g.*, from 5% to 20%), suggesting our method learns some common spatial layouts to automatically label most other images by leveraging a few samples. However, we need more samples (*e.g.*, 30%) to learn complex spatial layouts in city scenes of VIGOR.

5. Discussion and Limitation

The advanced supervised approach [9] discards spatial aggregation modules like SAFA [32] and uses the share-weight encoders for training. But, without labels, discarding them brings some additional learning burdens which lead to suboptimal results in Tab. 5 (a-b). Therefore, we still apply SAFA [32] and two non-share-weight encoders to help the model better understand and align the cross-view features, leading to a better retrieval performance. Besides, the entire training process is noisy, so the performance drops a lot without label smoothing in Tab. 5 (c).

Ablations	CVACT ↑			
	R@1	R@5	R@10	R@1%
(a) w/o safa	73.48	89.13	92.18	97.18
(b) w/ share-weight	81.06	92.04	94.06	97.33
(c) w/o label-smoothing	71.88	85.60	88.82	94.83
(d) w/ random-shift	54.49	69.70	74.65	86.81
(e) w/ random-rotate	50.32	68.50	75.08	88.70
(f) Ours	82.96	92.96	94.43	97.37

Table 5. **Discussions of some common components.**

In this work, we also utilize the localization and orientation alignment of ground and satellite images to simplify our unsupervised task like most supervised works [9, 32, 53], but the futural UCVGL should also get rid of these assumptions. So we also test the performance trained with random-shift or random-rotate images in Tab. 5, and we find our model is robust for them.

While we make the unsupervised way work in CVGL, our method has some limitations. Firstly, to learn cross-view alignments without labels, we project ground panoramas to satellite-view by leveraging the prior knowledge of approximate spatial projection between cross-view images, but a fixed geometric projection is not enough for simulating various cross-view relationships in the real world. Secondly, our projection is homography like many works [31, 32, 40], which is limited to represent the real transforms perfectly between two views only by images without extra 3D information (*e.g.*, depth), hindering the applications on complex scenes of VIGOR. For example, the advanced supervised cross-view image synthesis method [28] does not work in cities of VIGOR because of the complex scene and serve occlusions.

6. Conclusion

In this paper, we propose the first framework to utilize unlabeled data in cross-view geo-localization. Compared to the recent supervised methods, our method achieves comparable performance in an unsupervised setting without any labels and in a semi-supervised setting with few labels.

7. Acknowledgement

This work is supported by National Natural Science Foundation of China grants under contracts NO.62325111 and No.U22B2011. We thank the reviewers and ACs for their efforts. We want to thank the insightful discussions of Xiangyu Li, Jinwei Han, Jiakun Xu, Wuxuan Shi, Caoyuan Ma, Yajing Luo, Han Feng, Jiayu Li and Chao Pang.

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning (ICML)*, 2009. 5

[2] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[3] Hao Chen, Benoit Lagadec, and Francois Bremond. Ice: Inter-instance contrastive encoding for unsupervised person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3, 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning (ICML)*, 2020. 5

[5] Wei Chen, Yu Liu, Weiping Wang, Erwin M Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. Deep learning for instance retrieval: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[7] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Zelu Deng, Yujie Zhong, Sheng Guo, and Weilin Huang. Insclr: Improving instance retrieval with self-supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2, 3

[9] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[10] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, 1996. 2

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 2013. 1

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems (NeurIPS)*, 2014. 4

[13] Greg Hamerly and Charles Elkan. Learning the k in k-means. *Advances in neural information processing systems (NIPS)*, 2003. 2

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020. 3

[15] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 1, 2, 3

[16] Conghui Hu and Gim Hee Lee. Feature representation learning for unsupervised cross-domain image retrieval. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3

[17] Conghui Hu, Can Zhang, and Gim Hee Lee. Unsupervised feature representation learning for domain-generalized cross-domain image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[18] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[19] Xun Huang, Hai Wu, Xin Li, Xiaoliang Fan, Chenglu Wen, and Cheng Wang. Sunshine to rainstorm: Cross-weather knowledge distillation for robust 3d object detection. *arXiv preprint arXiv:2402.18493*, 2024. 1

[20] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations (ICLR)*, 2019. 4

[21] Ang Li, Huiyi Hu, Piotr Mirowski, and Mehrdad Farajtabar. Cross-view policy learning for street navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1

[22] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019. 2, 5, 6

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[25] Caoyuan Ma, Yu-Lun Liu, Zhixiang Wang, Wu Liu, Xinchen Liu, and Zheng Wang. Humannerf-se: A simple yet effective approach to animate humannerf with diverse poses, 2023. 4

[26] Evelyn J Mannix and Howard D Bondell. Cold paws: Unsupervised class discovery and the cold-start problem. *arXiv preprint arXiv:2305.10071*, 2023. 1, 3

[27] Niluthpol Chowdhury Mithun, Kshitij S Minhas, Han-Pang Chiu, Taragay Oskiper, Mikhail Sizintsev, Supun Samarasekera, and Rakesh Kumar. Cross-view visual geo-localization for outdoor augmented reality. In *2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 2023. 1, 2

[28] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground

image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6, 8

[29] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2020. 3, 5

[31] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6, 8

[32] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 5, 6, 8

[33] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 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 (CVPR)*, 2020. 2, 5, 6

[34] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[35] Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 4

[36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016. 3

[37] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixe. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[38] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision ECCV*, pages 494–509. Springer, 2016. 2

[39] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2021. 2

[40] Xiaolong Wang, Runsen Xu, Zuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *arXiv preprint arXiv:2308.16906*, 2023. 1, 2, 4, 7, 8

[41] Yuting Wang, Jinpeng Wang, Bin Chen, Ziyun Zeng, and Shu-Tao Xia. Contrastive masked autoencoders for self-supervised video hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 3

[42] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 2, 5

[43] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 6

[44] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[45] Min Zhang, Siteng Huang, Wenbin Li, and Donglin Wang. Tree structure-aware few-shot image classification via hierarchical aggregation. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[46] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 2, 6

[47] Zhongyan Zhang, Lei Wang, Yang Wang, Luping Zhou, Jianjia Zhang, and Fang Chen. Dataset-driven unsupervised object discovery for region-based instance image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 3

[48] Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023. 4

[49] Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 2, 4, 6

[51] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5

[52] Sijie Zhu, Taojiannan Yang, and Chen Chen. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[53] Sijie Zhu, Mubarak Shah, and Chen Chen. 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 (CVPR)*, 2022. 1, 2, 3, 5, 6, 8

[54] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023. 2, 3, 5, 6