Unleashing Unlabeled Data: A Paradigm for Cross-View Geo-Localization (Supplementary Material)

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

Overview

In this supplementary material, we provide the following items for a better understanding of our main paper.

1. Visual comparison of different projections.
2. Pseudo-Labels at different thresholds.
3. Results in semi-supervised settings.
4. Activation maps.
5. Implementation details.

1. Visual comparison of different projections

During the initial cold-start stage, our objective is to generate fake images in order to establish cross-view image pairs, which are essential for learning cross-view consistency.

Why do we need a novel projection? As depicted in Fig. 1, existing supervised methods [5, 7] typically project satellite-view images to ground-view perspectives. However, the transformed images (*e.g.*, (B-C) in Fig. 1) may suffer from severe distortions due to the limited overlaps between the two views and the presence of sky regions in ground images. Therefore, supervised refinements by ground-truth correspondences become necessary for them. Although the advanced cross-view synthesis method [3] project successfully satellite-view images to ground-view (*e.g.*,(D) in Fig. 1), it needs fully supervised training and does not work in crowed cities(*i.e.*, VIGOR).

Different from existing supervised settings, we propose a correspondence-free projection, which projects ground images to satellite view without any ground-truth labels. The transformed images, such as (F) in Fig. 1, resemble highly satellite images.

Why doesn't the projection work in the case of VIGOR? The projection is very important to start robust cross-view learning. Due to the complex scenes and severe occlusions in the cities of VIGOR, it is hard to synthesize cross-view images without extra 3D information, even in a fully supervised method [3].

Our projection is based on homography projection like other methods [4, 5], assuming that each point on the ground image is located on the ground in the satellite image, disregarding the height of buildings and other 3D volumetric effects. So our projection is not suitable for crowded cities with many tall buildings. For example, there are many high buildings in the left three images of Fig. 2, which leads to distinct views in the ground and satellite images. When we project ground images into a satellite view, it is hard to synthesize the blocked area and the height of buildings without any 3D information.

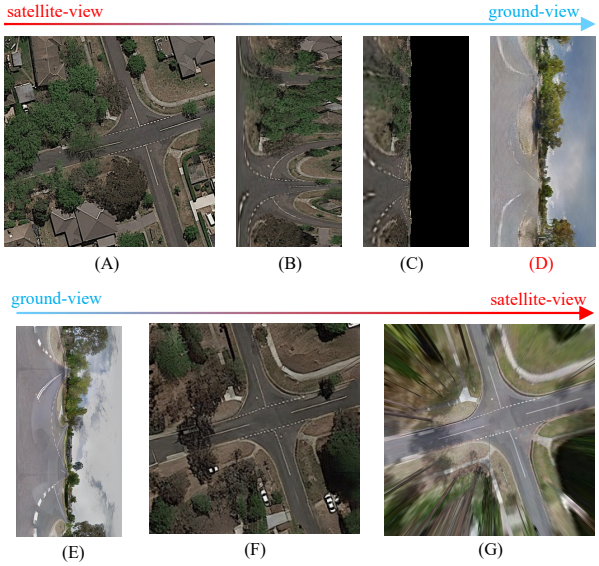

Figure 1. **Different Projection.** (A) Satellite and (E) Ground images are ground truth. Others are (B) Polar Transform [5], (C) Projective Transform [7], (D) Sat2Density [3], (F) Our CFT, and (G) Spherical [6, 9] Transform. (B-D) needs ground-truth correspondences between ground and satellite images. Our CFT is correspondence-free and transformed images are the most similar to ground truth.

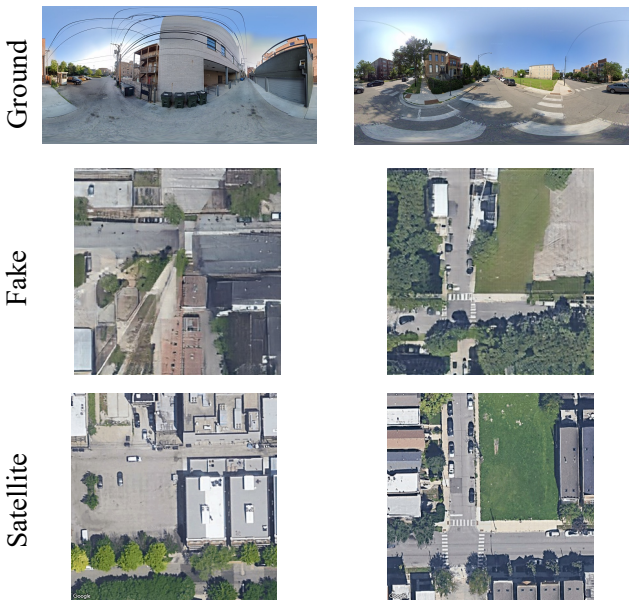

Figure 2. **Our projection in VIGOR.** The performance of our projection is poor in the left three images due to significant occlusion caused by numerous buildings. However, in the right three images, our generated fake image closely resembles the satellite image, thanks to the expansive field of view. Therefore, our projection is well-suited for open scenes.

2. Pseudo-Labels at different thresholds

In the left two images of Fig. 6 of our main paper, we analyze the function of our *threshold-filter* by comparing the two filter strategies of CVACT. Similarly, in the left two images of Fig. 3, we show the same function in CVUSA. Our *threshold-filter* ensures the higher the threshold, the better the quality of pseudo-labels.

In the right two images of Fig. 3, we show the effect of different thresholds in our *threshold-filter*. In (C) CVUSA, it's clear that a low threshold brings a low correct ratio of pseudo-labels, hindering the entire training process. Conversely, the higher threshold brings a higher correct ratio of pseudo-labels, enhancing the effectiveness of the entire training process. In (D) CVACT, the correct ratio of pseudo-labels is high even with a low threshold, so our *threshold-filter* has little effect on the training in CVACT. Totally, the *threshold-filter* is very important when the initial correct ratio of pseudo-labels is low.

3. Results in semi-supervised settings

In Tab. 1, we compare our semi-supervised method with the advanced supervised method[2]. Firstly, when we have a few labeled images (*e.g.*, 1% in CVACT), our method can take full advantage of unlabeled images and start a robust retrieval system with good performance (R@1 is 68.29). But Sample4Geo has a poor performance, with only 2.42 R@1. Secondly, we observe that the difficult samples are very important for this task. For example, there are many similar scenes in CVACT, so the performance has poor improvement when we give the model more labeled images from 5% to 20%. Differently, the scenes are more complex in cities of VIGOR, so the performance has significant improvement when we give the model more labeled images from 5% to 30%. Lastly, our method achieves comparable performance in semi-supervised settings compared to fully-supervised settings. For example, R@1 is 78.10 with 5% labeled images and R@1 is 84.44 with 100% labeled images in CVACT. Similarly, R@1 is 60.42 with 30% labeled images which is close to 68.40 with 100% labeled images in VIGOR (Chicago).

In Tab. 2, we provide the results with 30% labeled images in the same-area and cross-area settings of VIGOR in 4 cities. They still have good performances.

4. Activation maps

In Fig. 4, we compare the active maps of our method and Sample4Geo[2]. Sample4Geo tends to pay attention to road markings or trees, but our method tends to pay attention to scene-level information. Similar to most unsupervised methods[8], our model learns a more general representation with a bigger receptive field than supervised Sample4Geo[2] because it needs to focus on more areas to enhance its discriminative features with unlabeled data instead of ground-truth labels.

5. Implementation details

We report the result of the last epoch, not the best result in all experiments. The learning rate and weight decay of AdamW are set to 0.0001 and 0.03. Although we train the encoders with 100 epochs (including 40 epochs for the cold-start stage and 60 epochs for the semi-supervised stage), it is easy to get good results with 40 epochs (including 10 epochs for the cold-start stage and 30 epochs for the semi-supervised stage) like Sample4Geo [2]. Due to the entire training being noise, we use label smoothing to smooth the noise like Sample4Geo [2]. We don't use the hard negative sample sampling in [2], but this sampling is also useful for our setting.

References

[1] 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)*, pages 14960–14969, 2021. 3

CVACT							VIGOR (Chicago)						
GT	w/o AMM			Ours			GT	w/o AMM			Ours		
Ratio	R@1	R@5	R@10	R@1	R@5	R@10	Ratio	R@1	R@5	R@10	R@1	R@5	R@10
1%	2.42	6.88	10.45	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%	69.51	86.62	90.44	78.88	91.31	93.53	20%	-	-	-	55.90	75.44	81.20
20%	-	-	-	79.60	91.98	93.96	30%	44.49	65.96	73.31	60.42	80.12	84.88
100%	-	-	-	84.44	94.85	98.53	100%	-	-	-	68.40	88.49	92.44

Table 1. **Results in semi-supervised settings.** In the same area setting [10], there is little difference for training with all 4 cities and one city, so we report the result in Chicago for simplicity and time-saving. "GT Ratio" denotes the ratio of ground-truth labels used for training. The total number of image pairs is 35532 for CVACT and 12740 for VIGOR in Chicago. "w/o AMM" denotes our method without the semi-supervised stage, *i.e.*, training only on the fixed labeled images. As shown, our semi-supervised method improves performance by a large margin.

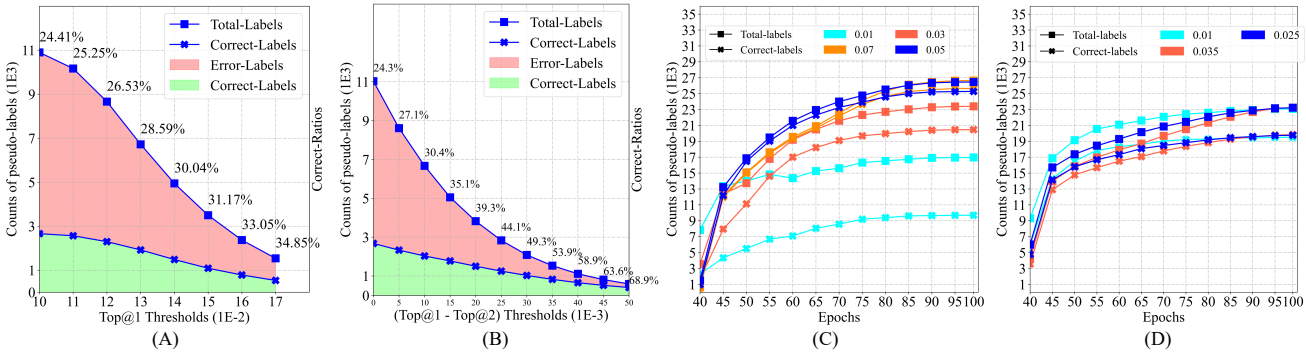

Figure 3. **Pseudo-labels.** The two left figures denote pseudo-labels produced by the highest or the difference value between the highest and second-highest retrieved similarity scores after the cold-start stage in CVUSA. The right two figures are the trend chart of pseudo-labels' counts at different thresholds in (C) CVUSA and (D) CVACT. We use blue to represent the strategies used in our method.

Type	GT ratio	R@1	R@5	R@10	R@1%
same-area	0.3	50.12	73.16	79.90	98.79
cross-area	0.3	32.76	55.09	63.63	94.04

Table 2. **Results with 30% labeled images in VIGOR.**

Type	GT ratio	R@1	R@5	R@10	R@1%
Sample4Geo [2]	0%	0.023	0.08	0.20	1.36
Classification [1]	0%	0.0	0.10	0.20	1.60
Ours	0%	82.96	92.96	94.43	97.37

Table 3. **Results of different methods on CVACT Val.** We apply the supervised CVGL and classification-based instance retrieval, *i.e.*, Sample4Geo [2] and [1], to the unlabeled data. As shown, existing methods are not suitable for solving the huge spatial and imaging gaps between ground and satellite images.

[2] 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)*, pages 16847–16856, 2023. 2, 3, 5

[3] 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. 1

[4] 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)*, pages 17010–17020, 2022. 1

[5] 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 (NIPS)*, 32, 2019. 1

[6] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(12):10009–10022, 2022. 1

[7] 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)*, 45 (3):2682–2697, 2022. 1

[8] Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhi-

nav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7486–7496, 2023. 2, 5

[9] 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

[10] 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)*, pages 3640–3649, 2021. 3

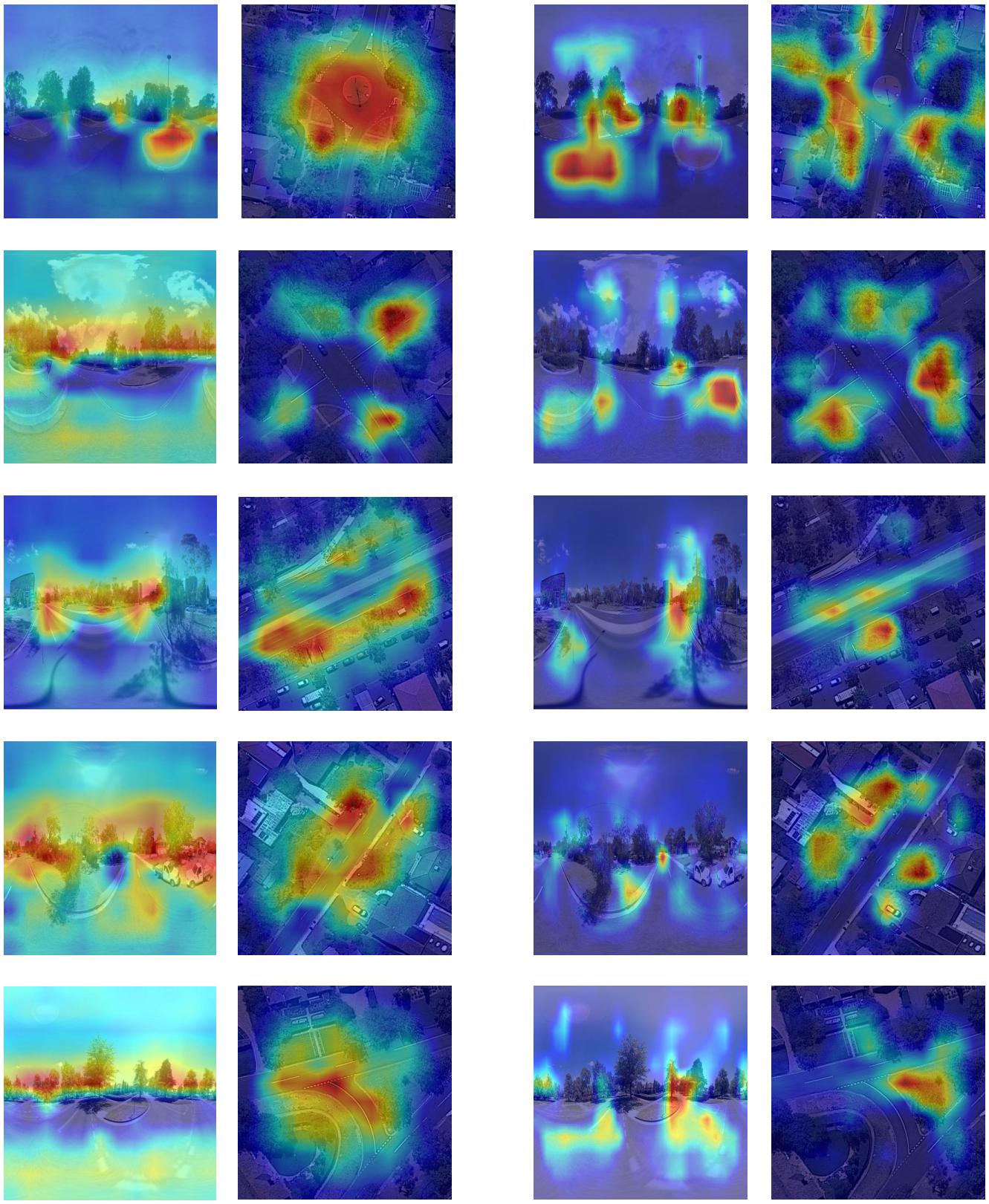

Figure 4. **Active Maps.** The supervised Sample4Geo [2] is more semantically rich (*e.g.*, trees), but our unsupervised method is more general, with a bigger receptive field and comparable active features at the scene level [8].