TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization Supplementary Material

Sijie Zhu, Mubarak Shah, Chen Chen
Center for Research in Computer Vision, University of Central Florida
sizhu@knights.ucf.edu, shah@crcv.ucf.edu, chen.chen@crcv.ucf.edu

Overview

In this supplementary material, we provide the following items for better understanding the paper:

1. Head-to-head comparison with L2LTR.
2. Performance on CVACT.
3. Limited FoV results on CVUSA.
4. Unknown orientation results on VIGOR.
5. Example of polar transform on VIGOR.
6. Example of Non-uniform Crop in CVUSA.
7. Qualitative results.
8. Implementation details.

1. Head-to-head Comparison with L2LTR

In Table 1, we provide a detailed head-to-head comparison between the proposed TransGeo and L2LTR [10], which was published after the submission deadline. TransGeo has clear superiority over L2LTR in terms of both performance and computational efficiency. Our method is pure transformer-based, L2LTR adopts vanilla ViT [1] on the top of ResNet [2], resulting in a hybrid CNN+transformer approach. L2LTR [10] does not provide GFLOPs and GPU memory consumption, but the authors claim that L2LTR requires significantly more GPU memory and pre-training data than CNN-base methods, *i.e*. SAFA (10.82G of GPU memory). We try their code and verify that L2LTR has much large GPU memory comsumption and GFLOPs than our method. Since L2LTR does not conduct experiments on VIGOR, we compare the performance (R@1) on CVUSA. Although the performance of L2LTR can be improved to 94.05 with polar transform, the overall performance is still lower than TransGeo. Note that the polar transform does not work well when the two views are not spatially aligned (as discussed in the ablation study of main paper), *e.g*. VIGOR [11], while TransGeo generalizes well on such scenarios with clear advantages.

	L2LTR [10]	**TransGeo** (Ours)
Architecture	CNN+Transformer	Transformer
GFLOPs	44.06	11.32
GPU Memory	32.16G	9.85G
Pretrain	ImageNet-21k	ImageNet-1K
Best Accuracy	94.05	94.08

Table 1. Head-to-head comparison between TransGeo and L2LTR.

2. Performance on CVACT

As shown in Table 2, the proposed TransGeo achieves state-of-the-art result on CVACT. Although CVACT and CVUSA are both aligned scenarios, we observe that removing patches cause more performance drop on CVACT than CVUSA. One possible explanation is that the satellite images of CVACT (zoom-level=20) have different resolution from CVUSA (zoom-level=18), resulting in a smaller covering range for each image.

Method	R@1	R@5	R@10	R@1%
CVM-Net [3]	20.15	45.00	56.87	87.57
Liu [5]	46.96	68.28	75.48	92.01
SAFA [7]	78.28	91.60	93.79	98.15
L2LTR [10]	83.14	93.84	95.51	98.40
†SAFA [7]	81.03	92.80	94.84	98.17
†Shi [8]	82.49	92.44	93.99	97.32
†Toker [9]	83.28	93.57	95.42	98.22
†L2LTR [10]	84.89	**94.59**	**95.96**	**98.37**
Ours	**84.95**	94.14	95.78	**98.37**

Table 2. Comparison with previous works in terms of R@k (%) on CVACT-val. "†" indicates methods using polar transform.

3. Unknown Orientation Results on VIGOR

In Table 3, we show the performance of TransGeo and VIGOR [11] with unknown orientation, by randomly shift the panorama horizontally. TransGeo outperforms VIGOR with a large margin, indicating that TransGeo's superiority does not rely on the orientation alignment between two

	Same-Area				Cross-Area			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
VIGOR [11]	19.10	42.13	-	95.12	1.41	4.52	-	44.60
TransGeo	**47.69**	**79.77**	**86.36**	**99.29**	**5.54**	**14.22**	**19.63**	**66.93**

Table 3. Performance of TransGeo and previous work [11] on VIGOR dataset with unknown orientation.

	$FoV = 180°$				$FoV = 90°$			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
DSM [8]	48.53	68.47	75.63	93.02	16.19	31.44	39.85	71.13
TransGeo	**58.22**	**81.33**	**87.66**	**98.13**	**30.12**	**54.18**	**63.96**	**89.18**

Table 4. Performance of TransGeo and previous methods on CVUSA with limited FoV (Field of View) and unknown orientation.

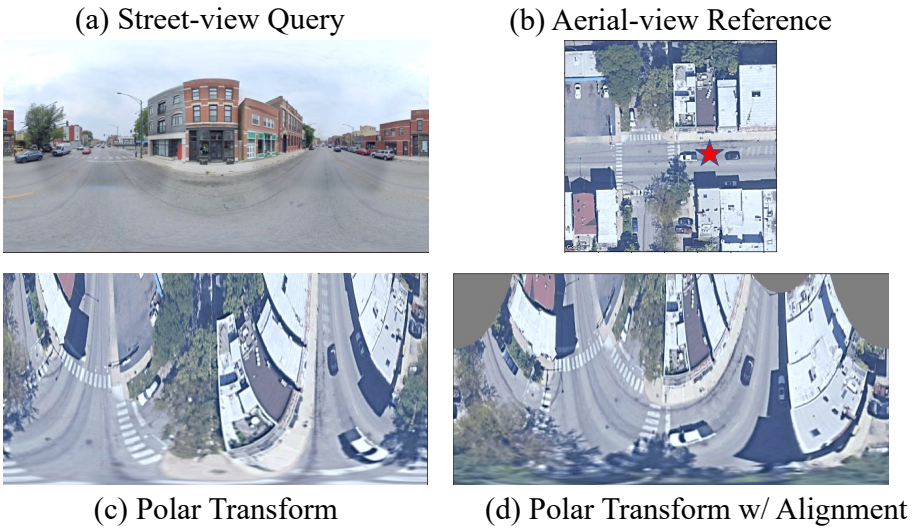

Figure 1. Example of polar transform on VIGOR. Red star denotes the location of street query in the aerial image.

views.

4. Limited FoV results on CVUSA

In Table. 4, we show the performance of TransGeo and DSM [8] on CVUSA with limited FoV (Field of View), by randomly cropping the panorama with random shift. The orientation is also unknown. TransGeo significantly outperforms DSM on $FoV = 180°$ and $FoV = 90°$, indicating that TransGeo's superiority does not rely on the wide FoV of panorama. The performance gap is more significant when the FoV is smaller.

5. Polar Transform Example on VIGOR

In Fig. 1, we show an example of polar transform on VIGOR to demonstrate why it fails in unaligned scenarios. (a) and (b) are the original street-view and aerial-view images, and the red star in (b) indicates the location of the street-view query. (c) is generated with the vanilla polar transform using the center of aerial image. VIGOR assumes that the street-view query does not lie at the center of aerial image, and we use the red star (as shown in (b)) to denote the actual location. (d) is generated by using the red star location *as the center* (*i.e.* adjustment to the spatial alignment) for polar transform, denoted as 'Polar Transform w/ Alignment'. The spatial offset of query can cause distortion in (c), and even the aligned (d) does not have a good geometric correspondence with the street-view query, due to the strong occlusion. Polar transform assumes that objects far away from the query location has a large vertical coordinates in the street-view image. However, this does not well model the geometric relationship between the two views when there are tall buildings close to the street-view query location. Besides, the roof of the building and other occluded objects occupy a large space in the transformed images (c) and (d), but they are not visible in the street-

view, thus do not help the cross-view matching.

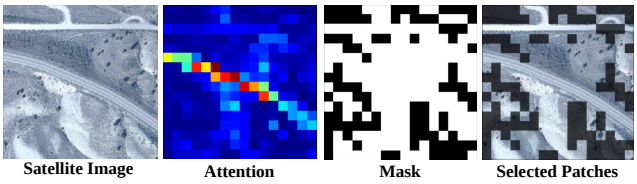

Figure 2. Example of attention map and non-uniform crop on CVUSA.

6. Example of Non-uniform Crop in CVUSA

In the main paper, we only show the example of non-uniform crop on city scenarios (VIGOR). We show the attention map and cropping selection for rural scenarios (CVUSA) in Fig. 2. The attention map in rural area looks more scattering/uniform than cities, but they still focus more on discriminative objects, *e.g.* road.

7. Qualitative Results

In Figs. 3 and 4, we include qualitative results of TransGeo on the CVUSA and VIGOR datasets. We select four queries for each dataset with the ground-truth image ranked at 1, $[2, 5]$, $[6, 100]$ and > 100, representing both success and failure cases for analysis. The ground-truth in retrieved results is marked with red box. For the first row of Figs. 3 and 4, the ground-truth is retrieved as the first one, which is very similar to the second one. This indicates the strong discriminative ability of TransGeo. The other failure cases in CVUSA are due to extreme lighting condition (too dark), lack of recognizable objects (only road and grass) with hard negative reference (the first retrieved one has very similar color to the street-view query), and different capture seasons (query was taken in winter with snow) of two views. For VIGOR, the retrieval is more challenging because of semi-positive samples [11], which cover the query image at edge area. The second and third rows both retrieve semi-positive samples as the first one. This is not considered as correct top-1 prediction, but their GPS location is actually very close to the ground-truth, resulting in good performance in meter-level evaluation. For the last row, the model fails because only trees and roads are visible in the query. They do not provide enough information to distinguish the ground-truth from other aerial images with trees.

8. Implementation Details

We use $\rho = 2.5$ for ASAM [4]. The weight decay of AdamW is set to 0.03, with default epsilon and other parameters in PyTorch [6]. The sampling strategy is the same as [11], but we re-implement it with PyTorch. Details are included in the code.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] 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 Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 1

[4] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. *arXiv preprint arXiv:2102.11600*, 2021. 3

[5] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 1

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 3

[7] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pages 10090–10100, 2019. 1

[8] 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*, pages 4064–4072, 2020. 1, 2

[9] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. 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*, pages 6488–6497, 2021. 1

[10] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[11] 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*, pages 3640–3649, 2021. 1, 2, 3

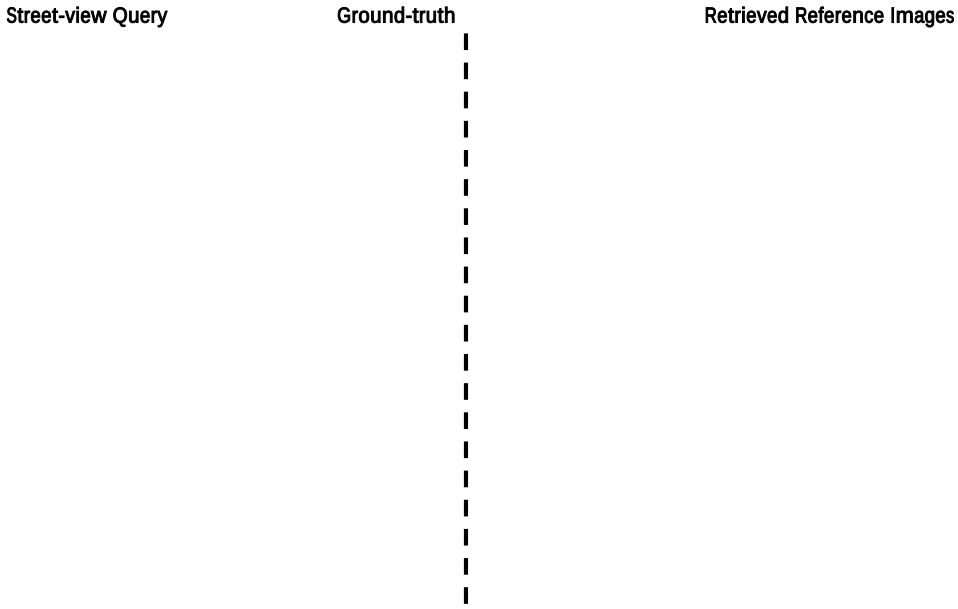

Figure 3. Qualitative results on CVUSA. Red box indicates ground-truth in retrieved results. The ground-truth is ranked at $1, 2, 6, 148$ for four queries respectively.

Figure 4. Qualitative results on VIGOR. Red box indicates ground-truth in retrieved results. The ground-truth is ranked at $1, 2, 9, 165$ for four queries respectively.