Supplementary Material

Xiaohan Zhang[1], Waqas Sultani[2], and Safwan Wshah[1]

[1]Department of Computer Science, University of Vermont, USA
[2]Intelligent Machine Lab, Information Technology University, Pakistan
{Xiaohan.Zhang, Safwan.Wshah}@uvm.edu waqas.sultani@itu.edu.pk

In this supplementary material, we are providing additional information for the following items:

- The availability of panoramas and limited Field-Of-View (FOV) images.
- Dataset coverage map
- More implementation details
- More details about the baseline methods discussed in the main paper.
- Comparison of the number of trainable parameters between the proposed model and baseline methods
- The availability of our proposed dataset and code for the public.
- More samples from our proposed dataset.
- More qualitative results predicted by our proposed model.

1. Panorama vs Limited FOV images

As we discussed in our main paper, limited FOV images are more popular and common than panoramas. To highlight the difference, we presented the coverage areas of limited FOV images and panoramas from Mapillary [1] in Fig 1. Mapillary [1] is one of the largest crowdsourcing platforms for sharing geotagged photos. As of 2018, Mapillary [1] hosted 422 million images across the world. As observed from Fig 1, the coverage area of limited FOV images (Fig. 1b) on Mapillary is substantially greater than the coverage area of panoramas (Fig. 1a), especially in some developing areas such as Middle East , Africa and south America. We refer this to the complexity of capturing panoramic images which they need special and expensive cameras. To this end, using sequences of limited FOV images as the query is much more practical than using panoramas as the query in cross-view geo-localization.

2. More implementation details

Our model was trained in an end-to-end manner using Adam [3] with weight decay of 10^{-6} for 50 epochs on a single Nvidia V100 GPU. The learning rate is set initially to 10^{-5} and decayed linearly to 5×10^{-7} after 30 epochs. We set the γ in Equation 5 of main paper to 10. We set the ground sequence length $T = 7$ which is suitable for our dataset. We used the exhaustive mini-batch strategy [6] to construct the triplet pair with batch size set to 24.

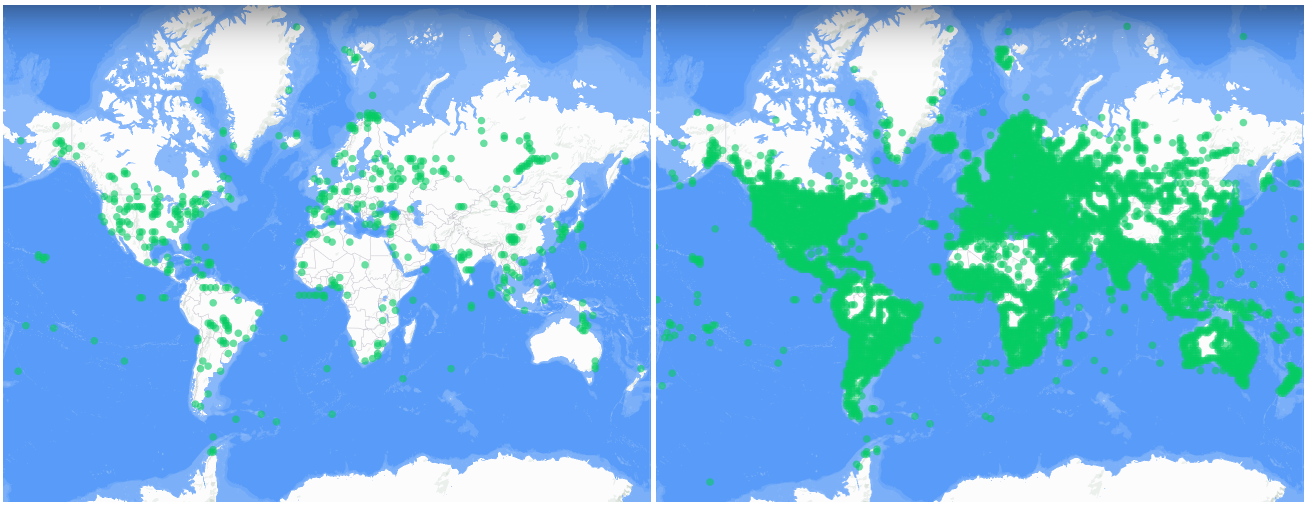
3. Baseline Methods

We employed two baseline methods for comparison, SAFA [4] and VIGOR [8]. For SAFA [4], we adopted their original code. [1] SAFA trained only on the center images of each sequence. For fair comparison, SAFA has been initialized with weights pretrained on CVUSA [7] dataset then trained on our dataset. We used same hyperparameters reported in SAFA's original paper [4] and fine-tuned the model for 10 epochs. For VIGOR [8], we used their code[2] for training. Similar to SAFA, we trained their model from all images in the sequences by setting the center ground-level image to a 'positive' sample and the others are 'semi-positive' samples as defined in their original paper. We set the hyperparameters as reported in original VIGOR paper [8] and followed their exact procedures for training.

4. Dataset Availability and Anonymity

Our proposed dataset is composed of two parts, ground-level image sequences and satellite imagery as explained in the main paper. Our ground-level images are public images collected by Vermont Agency of Transportation[3]. The private information of all ground-level images has

[1]https://github.com/shiyujiao/cross_view_localization_SAFA
[2]https://github.com/Jeff-Zilence/VIGOR
[3]https://vtrans.vermont.gov/



(a) (b)

Figure 1: Comparison of coverage area (green lines) of user uploaded street view images between panoramic (a) and limited FOV images (b) on Mapillary [1].

been anonymized. These images will be shared publicly. Our satellite images came from Google Maps. Following Google Maps Platform Terms of Service[4], we will make our dataset available for research purposes only. We will follow existing datasets, such as VIGOR [8], to distribute the collected dataset upon request.

5. Dataset Coverage Map

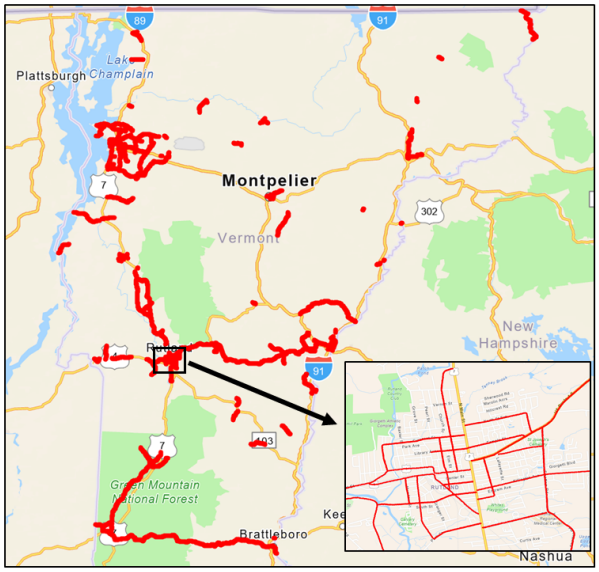


Figure 2: The coverage map of the proposed dataset. The coverage area is indicated by red lines.

[4]https://cloud.google.com/maps-platform/terms

Method	Parameters	R@1	R@10	R@1%
VIGOR [8]	395M	0.54%	4.48%	18.55%
SAFA† [4]	319M	0.68%	5.06%	21.81%
Ours w/ VGG16 [5]	2.9G	1.80%	10.36%	34.38%
Ours w/ Res50 [2]	775M	2.07%	13.16%	40.10%
Res34 [2]	240M	1.71%	11.67%	38.16%
Res18 [2]	161M	1.58%	10.14%	33.83%

Table 1: Caption

To better visualize the diversity of the proposed dataset, we visualize the coverage area in Fig. 2. As indicated by the coverage map, our dataset includes both suburban and urban areas in Vermont, US which cover most scenarios on the roads.

6. Comparison of parameters

In this section, we present the comparison of trainable parameters between the proposed model with different backbones and baseline methods in Table 1. Our model with VGG16 [5] is larger than the baselines. This is because the output dimension of VGG16 is 4096. As a result, we need wider TFAMs to handle this large latent vector. When we switch to ResNet [2] as backbone, the number of parameters is significantly less than VGG [5] as backbone. This is because the dimension of output of ResNet50 is 2048. For ResNet34 and ResNet18, the dimension of output is only 512 which cause these two models are even smaller than baselines. However, despite of the backbones, the proposed model is constantly outperforms the baseline methods. For a fair comparison with baseline methods, we

choose VGG16 as the backbone in the main script.

7. More Dataset examples

In this section, we provided 6 randomly sampled satellite and ground sequence pairs from our proposed dataset as shown in Fig. 3. As shown in Fig. 3, our dataset covers diverse locations, urban, suburban, and rural areas which we discuss in detail in our main script.

8. More Qualitative Results

In this section, we provided more retrieval examples. Fig. 4 shows correct top-1 examples predicted by our model and Fig. 5 shows top-5 retrieval examples. Each figure shows pairs of satellite and ground images ordered from top to bottom. For each pair, the bottom row is the query ground-level sequence and the upper row is the predicted top-5 satellite images ranked in descending order from left to right. The satellite images with blue boarder are the ground truth.

References

[1] Mapillary. `https://www.mapillary.com/app`.

[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 (CVPR)*, June 2016.

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[4] 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*, 32:10090–10100, 2019.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[6] Nam N. Vo and James Hays. Localizing and orienting street views using overhead imagery. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 494–509, Cham, 2016. Springer International Publishing.

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

[8] 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, June 2021.

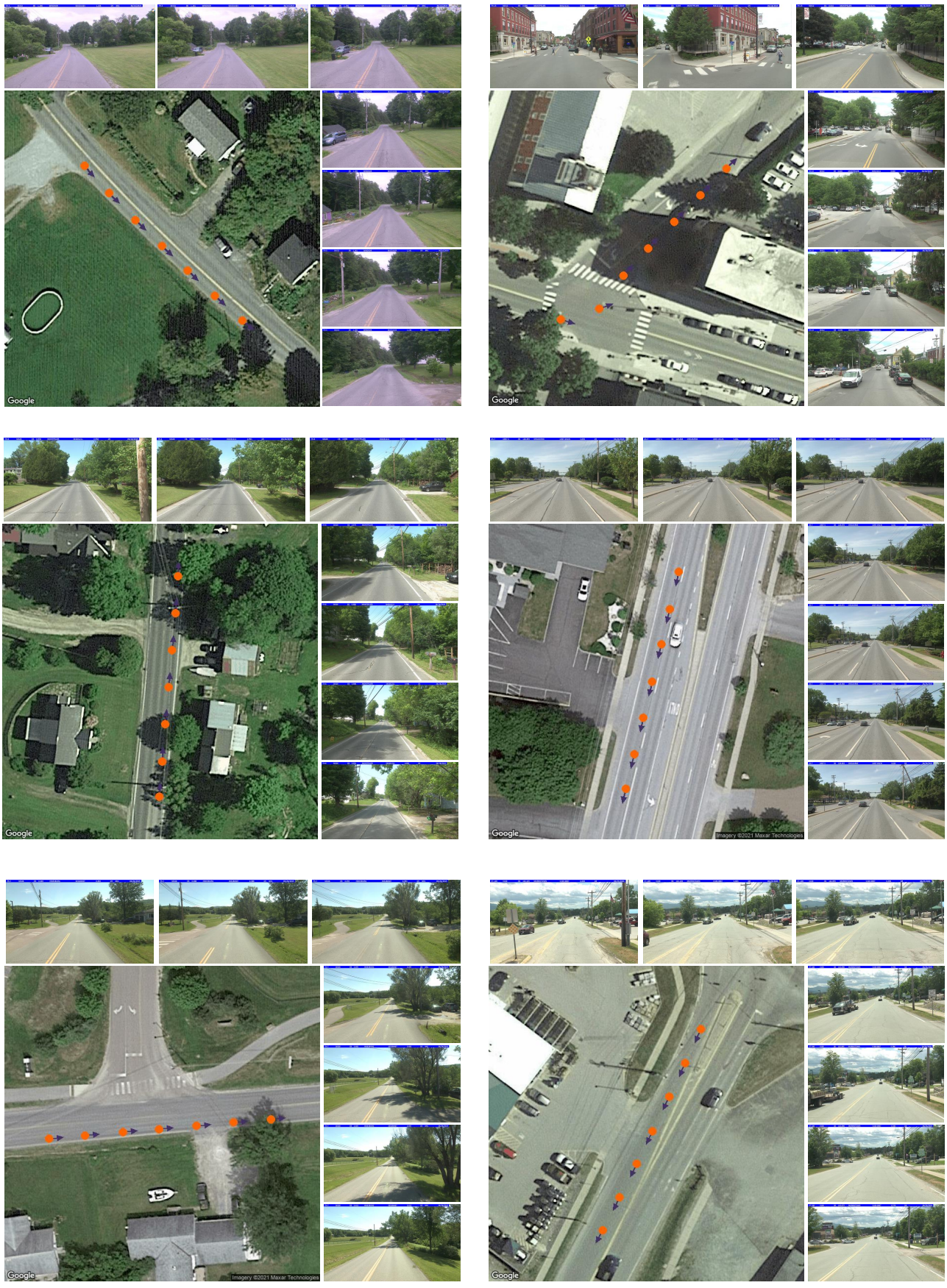


Figure 3: Six randomly sampled satellite and ground sequence pairs from our dataset.

Figure 4: Samples been correctly predicted as top-1 by our model.

Figure 5: Samples been correctly predicted as top-5 by our model.