VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval Supplementary Material

Sijie Zhu, Taojiannan Yang, Chen Chen
University of North Carolina at Charlotte
{szhu3, tyang30, chen.chen}@uncc.edu

Abstract

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

1. *Additional details about our dataset.*
2. *Effect of mining.*
3. *Meter-level evaluation without orientation alignment.*
4. *Results on Manhattan without sample balancing.*
5. *Simulation results on CVUSA.*

1. Additional Details of VIGOR Dataset

Data Cleaning. We collect all the panoramas in the Areas of Interest (AOI) in four target cities. To make sure that the raw GPS is accurate, **we only consider the panoramas collected by Google with car-based industrial-level GPS**. This results in about $70,000$ panoramas for each city. However, some images are still not usable, *e.g.* subway or indoor images with poor light conditions. We first filter out most of the subway images according to the elevation data. Then we filter out images with poor light conditions by the average HSI (hue, saturation, intensity) values.

Distribution of Positive Samples. In our original dataset, the panoramas can be very dense at the street intersection areas. One aerial reference image may cover lots of panoramas, but only part of them are positive samples (lie in the central area). We show the distribution of panoramas for each reference covered by each aerial reference image in Fig. 1 (including positive and semi-positive samples). The distribution of positive samples are included in Fig. 2. Reference images covering zero panoramas are preserved as distraction samples. We perform sample balancing by randomly sample 2 positive panoramas for each reference image in our experiment to ensure that the panoramas have distinguishable distance to each other.

2. Effect of Mining

We show the retrieval accuracy w.r.t. each epoch to illustrate the effect of the mining strategy. Fig. 3 shows the accuracy-epoch curves of same-area and cross-area settings. The results of the cross-area evaluation tends to over-fit after 20 epochs, which is possibly due to the distribution discrepancy between different cities, we thus stop the training on 20 epochs. For the same-area evaluation, the accuracy barely changes after 45 epochs.

3. Without Orientation Alignment

To further show the applicability of the proposed methods without orientation information, we randomly shift the panoramas and conduct meter-level evaluations based on different search scopes in Sec. 7 of our paper. Our experiment shows that the SAFA [1] block does not work without the orientation alignment, we thus use the simple Siamese-VGG from [3] in our pipeline. As shown in Fig. 4, for the same-area setting, the localization results still have great potential for practical applications. For the cross-area, few information can be used for cross-city generalization, thus the full-scope localization ("All") is even worse than the "Original" noisy GPS signal. However, with a smaller search scope, our method is useful for complementing the noisy phone-grade noisy GPS.

4. Without Sample Balancing

Since the original dataset without balancing contains too many panoramas for experiment, we only conduct this experiment on New York (Manhattan) with same-area protocol. In Table 1, we show the retrieval accuracy of different methods for comparison and the results are consistent with our paper.

5. Simulation on CVUSA

To show the effect of multiple-reference retrieval, we simulate the seamless coverage and overlap sampling on CVUSA [2] dataset. Each aerial image (750×750) is

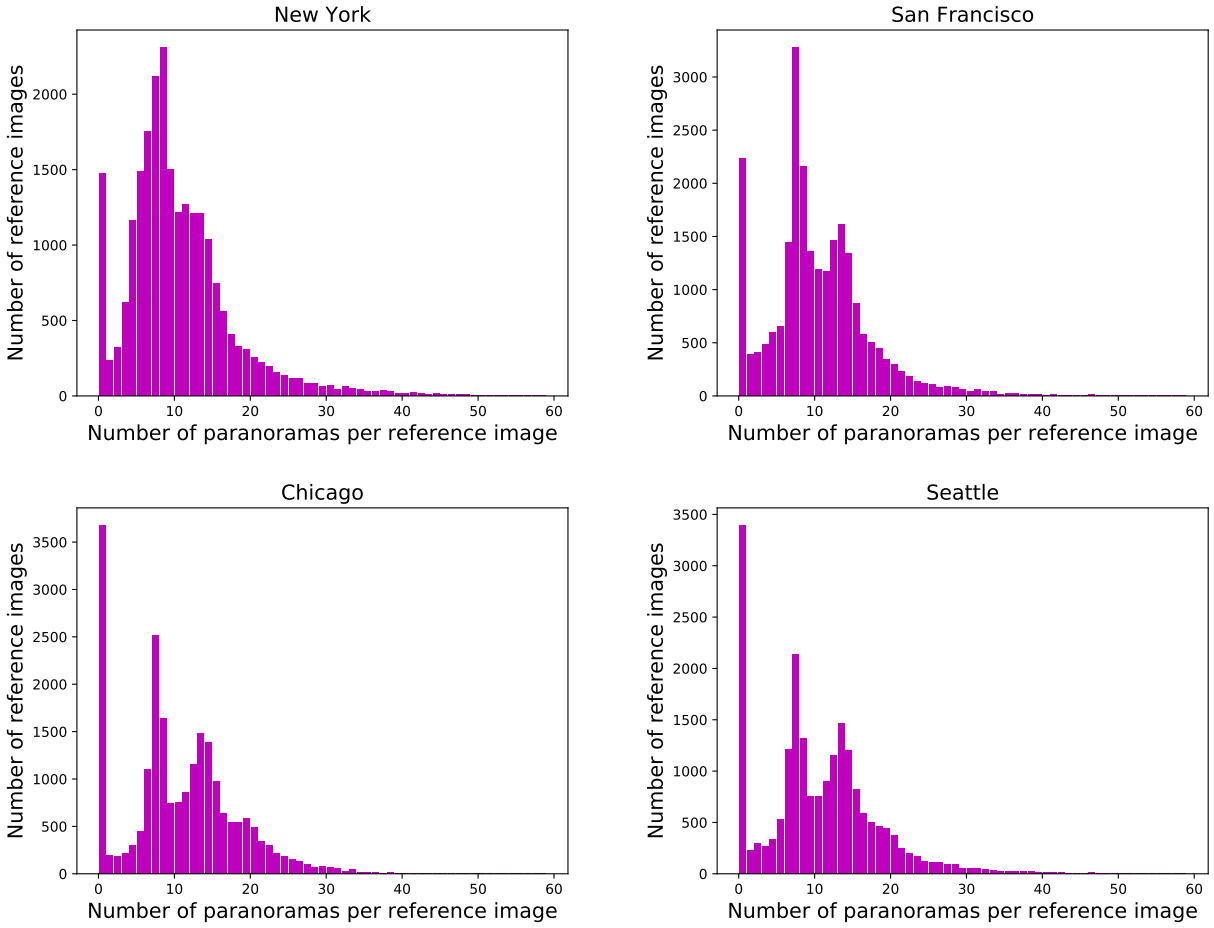

Figure 1. Number of panoramas covered by each reference image.

	top-1	top-5	top-1%	hit rate
Siamese-VGG	35.4	69.0	97.5	45.0
SAFA [1]	48.0	75.3	97.9	55.0
SAFA + Mining [3, 1]	56.0	78.6	96.4	66.6
Ours	**63.3**	**83.6**	**98.1**	**71.6**

Table 1. Retrieval accuracy (%) of different methods without balancing sampling in New York with same-area protocol.

re-sampled as five 300×300 images with 50% overlap at [central, left, right, top, bottom]. Random offsets in $[\pm 150, \pm 150]$ are applied to the central crop which is considered as the positive sample and ground-truth. This protocol is denoted as "SAFA-Our Protocol" in Table 2. The results with only the aligned central crop 300×300 is denoted as "SAFA-Center Crop" and "SAFA-Random Crop" means the only one crop with random offset. As shown in Table 2, the crop size 300×300 does not have much influence on the performance, but the accuracy drops dramatically when the aerial image is not always perfectly aligned at the location of panoramas. The accuracy further drops if multiple reference images exist as distractions for retrieval. Our dataset enables realistic research on this problem.

	top-1	top-5	top-1%	hit rate
SAFA [1]	89.0	-	-	-
SAFA-Center Crop	81.4	-	-	-
SAFA-Random Crop	32.5	57.2	81.7	38.5
SAFA-Our Protocol	23.2	35.5	70.1	24.1

Table 2. Retrieval accuracy (%) of different settings on CVUSA.

References

[1] 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, 2, 4

[2] 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*, pages 867–875, 2017. 1

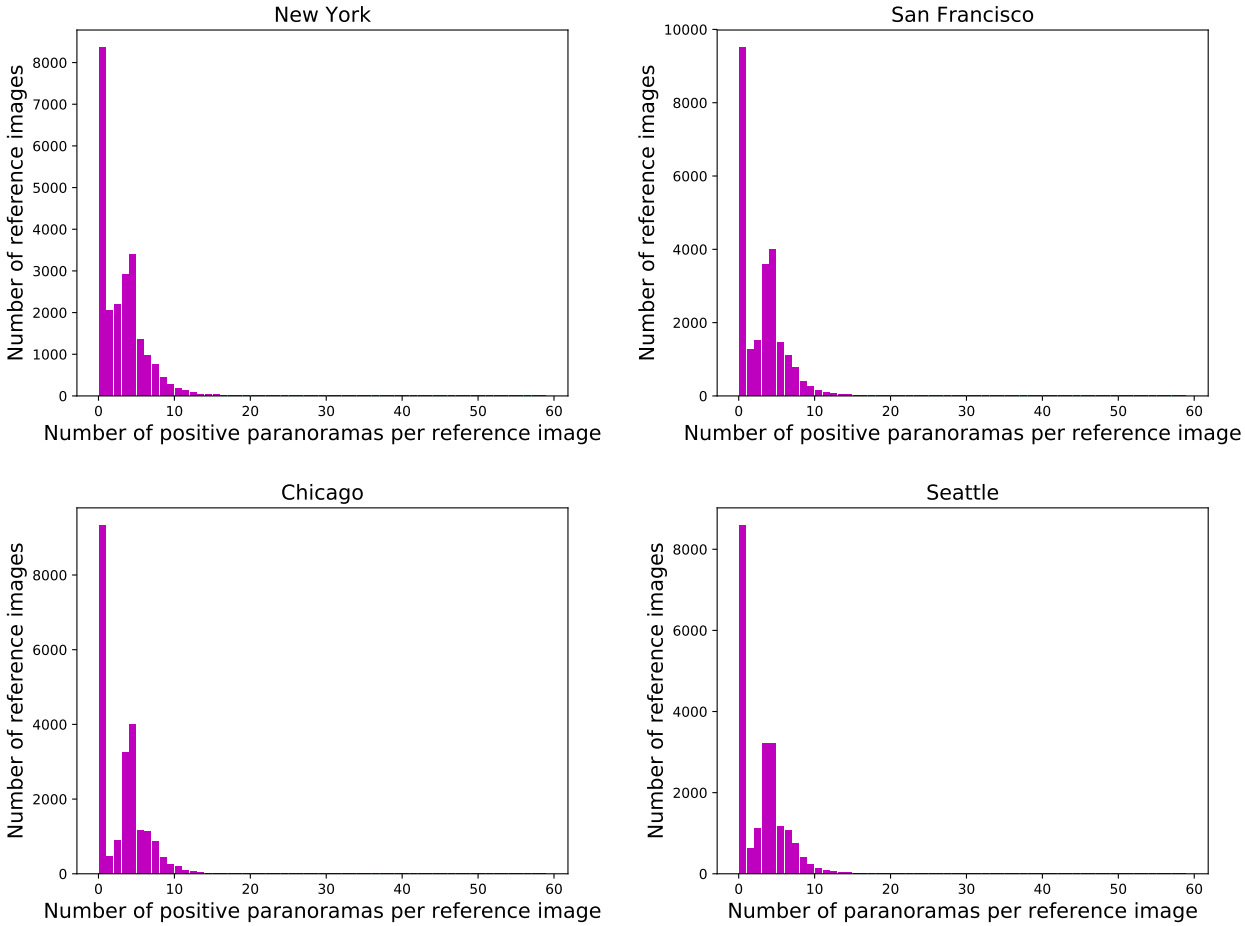

Figure 2. Number of positive panoramas for each reference image.

[3] 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)*, pages 756–765, January 2021. 1, 2, 4

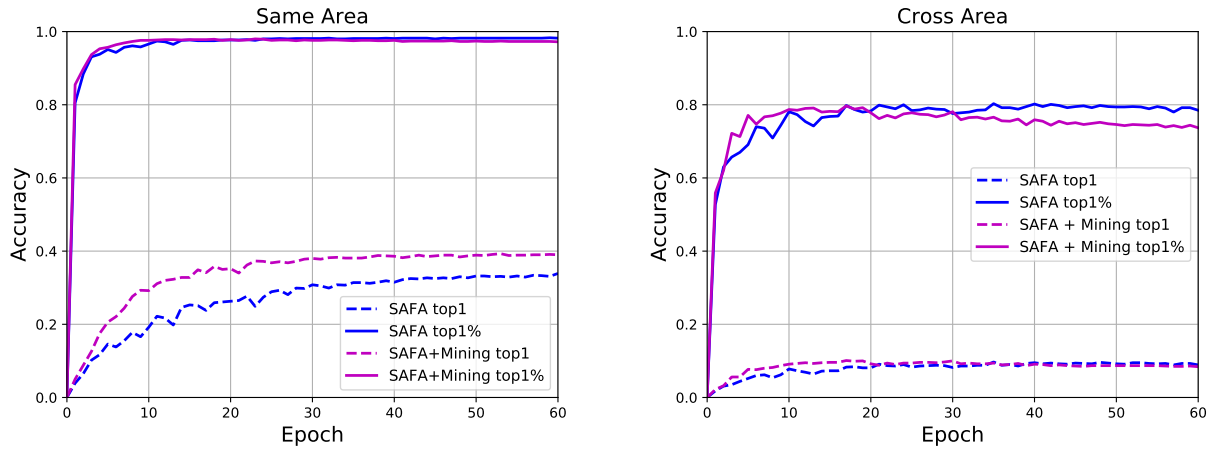

Figure 3. Accuracy-epoch curves for SAFA [1] and SAFA [1] + Mining [3].

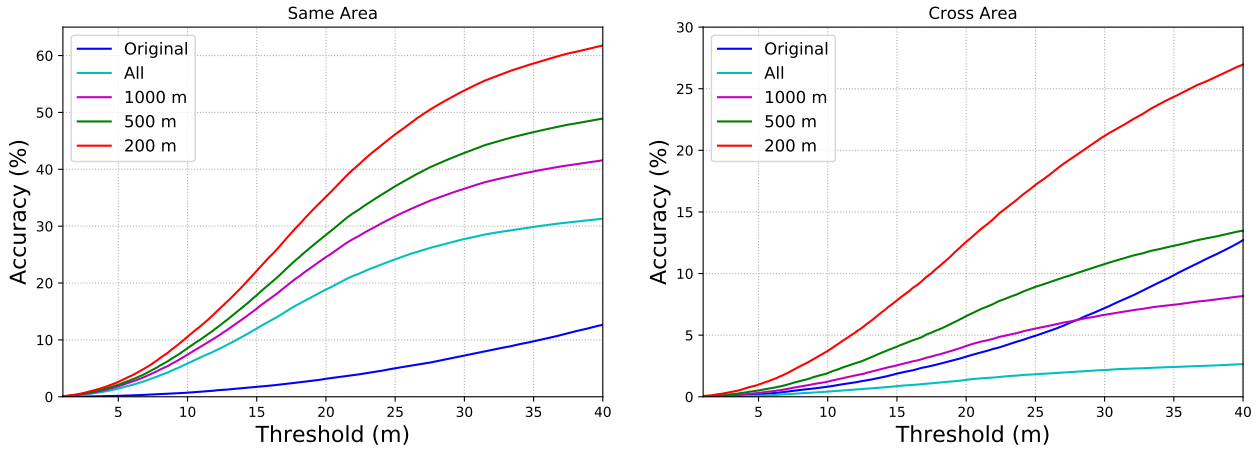

Figure 4. Meter-level evaluation of our method w/o orientation alignment given the noisy GPS signal.