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

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

Abstract

Cross-view image geo-localization aims to determine the locations of street-view query images by matching with GPS-tagged reference images from aerial view. Recent works have achieved surprisingly high retrieval accuracy on city-scale datasets. However, these results rely on the assumption that there exists a reference image exactly centered at the location of any query image, which is not applicable for practical scenarios. In this paper, we redefine this problem with a more realistic assumption that the query image can be arbitrary in the area of interest and the reference images are captured before the queries emerge. This assumption breaks the one-to-one retrieval setting of existing datasets as the queries and reference images are not perfectly aligned pairs, and there may be multiple reference images covering one query location. To bridge the gap between this realistic setting and existing datasets, we propose a new large-scale benchmark –VIGOR– for cross-View Image Geo-localization beyond One-to-one Retrieval. We benchmark existing state-of-the-art methods and propose a novel end-to-end framework to localize the query in a coarse-to-fine manner. Apart from the image-level retrieval accuracy, we also evaluate the localization accuracy in terms of the actual distance (meters) using the raw GPS data. Extensive experiments are conducted under different application scenarios to validate the effectiveness of the proposed method. The results indicate that cross-view geo-localization in this realistic setting is still challenging, fostering new research in this direction. Our dataset and code will be released at https://github.com/Jeff-Zilence/VIGOR.

1. Introduction

The objective of image-based geo-localization is to determine the location of a query image by finding the most similar image in a GPS-tagged reference database. Such technologies have proven useful for accurate localization with noisy GPS signals [4, 26] and navigation in crowded cities [12, 9]. Recently, there has been a surge of interest in cross-view geo-localization [24, 22, 7, 17, 29, 21], which uses GPS-tagged aerial-view images as reference for street-view queries. However, the performance may suffer from a large appearance gap between query and reference images.

Recent works [7, 17, 29] have shown that the performance of cross-view image matching can be significantly improved by feature aggregation and sample mining strategies. When the orientation of street-view (or ground-view) image is available (provided by phone-based compass), state-of-the-art methods can achieve a top-1 retrieval accuracy over 80% [17], which shows the possibility of accurate geo-localization in real-world settings. However, existing datasets [24, 27, 11] simply assume that *each query ground-view image has one corresponding reference aerial-view image whose center is exactly aligned at the location of the query image*. We argue this is not practical for real-world applications, because the query image can occur at arbitrary locations in the area of interest and the reference images should be captured before the queries emerge. In this case, perfectly aligned one-to-one correspondence is not guaranteed.

In light of the novelty of this problem, we propose a new benchmark (VIGOR) to evaluate cross-view geo-localization in a more realistic setting. Briefly, given an area of interest (AOI), the reference aerial images are densely sampled to achieve a seamless coverage of the AOI and the street-view queries are captured at arbitrary locations. In total, $90,618$ aerial images and $238,696$ street panoramas are collected from 4 major cities in the United States (see details in Sec. 3). The new dataset gives rise to two fundamental differences between this work and prior research.

Beyond One-to-one: Previous research mainly focuses on the one-to-one correspondence because existing datasets consider perfectly aligned image pairs as default. However, VIGOR enables us to explore the effect of reference samples that are not centered at the locations of queries but still cover the query area. As a result, there could be multiple reference images partially covering the same query location, breaking the one-to-one correspondence. In our geo-localization method, we design a novel hybrid loss to take advantage of multiple reference images during training.

Beyond Retrieval: Image retrieval can only provide image-level localization. Since the center alignment is not guaranteed in our dataset, after the retrieval, we further employ a within-image calibration to predict the offset of the query location inside the retrieved image. Therefore, the proposed joint-retrieval-and-calibration framework provides a coarse-to-fine localization. The whole pipeline is end-to-end, and the inference is fast as the offset prediction shares the feature descriptors with the retrieval task. Moreover, our dataset is also accompanied with raw GPS data. Thus a more direct performance assessment, *i.e*. localization accuracy in terms of real-world distance (*e.g*. meters), can be achieved on our dataset.

Our main contributions can be summarized as follows:

- We introduce a new dataset for the problem of cross-view image geo-localization. This dataset, for the first time, allows one to study this problem under a more realistic and practical setting and offers a testbed for bridging the gap between current research and practical applications.
- We propose a novel joint-retrieval-and-calibration framework for accurate geo-localization in a coarse-to-fine manner, which has not been explored in the past.
- We develop a new hybrid loss to learn from multiple reference images during training, which is demonstrated to be effective in various experimental settings.
- We also validate the potential of the proposed cross-view geo-localization framework in a real-world application scenario (assistive navigation) by simulating noisy GPS.

2. Related Work

Cross-view Datasets. A number of datasets have been proposed for cross-view geo-localization [10, 25, 24, 27, 22, 11]. Lin *et al*. [10] consider both satellite images and land cover attributes for cross-view geo-localization. $6,756$ ground-view images and 182,988 aerial images are collected from Charleston, South Carolina. Although the aerial images are densely sampled, they force a one-to-one correspondence between two views and evaluation in terms of distance is not available. The original CVUSA [25] is a massive dataset containing more than 1 million ground-level and aerial images from multiple cities in the United States. Zhai *et al*. [27] further make use of the camera's extrinsic parameters to generate aligned pairs by warping the panoramas, resulting in 35,532 image pairs for training and 8,884 image pairs for testing. This version of CVUSA is the most widely used dataset in recent research [7, 17, 29, 14, 23] and we refer to it as CVUSA if not specified. Vo [24] consists of about one million image pairs from 11 cities in the United States. The authors randomly collect street-view panoramas and generate several crops from each panorama along with spatially aligned aerial images from Google Maps. Similar to CVUSA, CVACT [11] also consists of aligned panoramas and aerial images with orientation information. It has $35,532$ image pairs for training and $92,802$ pairs for testing. In a nutshell, *all these datasets consider one-to-one retrieval and none of them provide raw GPS data for localization evaluation in terms of meters*.

Cross-view Geo-localization. Early works [10, 25, 24, 22] of cross-view geo-localization suffer from low retrieval accuracy mainly because of the significant appearance gap between two views and poor metric learning techniques. With tailored feature extractors and a modified loss function, Hu *et al*. [7] show the possibility of achieving accurate localization with end-to-end deep neural networks. Several recent methods [14, 17] aim to reduce the domain gap by leveraging GANs [6] and polar transformations [18]. Regmi *et al*. [14] propose to generate the synthetic aerial-view image from the ground-view query with a conditional GAN and adopt feature fusion to achieve better performance. SAFA [17] further takes advantage of the geometric prior knowledge by applying a polar transformation on the query image and replacing the global pooling with feature aggregation blocks. The top-1 accuracy of [17] on CVUSA [27] is almost 90% if the orientation information is given. Other approaches [5, 29] exploring metric learning techniques (*e.g*. hard samples mining strategy) also show promising results on popular datasets, and they are not restricted by the geometric assumptions. *However, none of these methods consider a sub-image level localization beyond the image-level retrieval or multiple reference images for training*.

3. VIGOR Dataset

Problem Statement. Given an area of interest (AOI), our objective is to localize an arbitrary street-view query in this area by matching it with aerial reference images. To guarantee that any possible query is covered by at least one reference image, the reference aerial images must provide a seamless coverage of the AOI. As shown in Fig. 1 (a), coarsely sampled reference images (black square boxes) are not able to provide full coverage of the AOI, and an arbitrary query location (the red star) may lie in the area between reference samples. Even if the query location (the yellow star) lies at the edge of a reference aerial image, this reference image only shares partial (at most half) scene with the one whose center is at the query location, which may not provide enough information to be distinguished from other negative reference images. These queries can be covered by adding additional overlapping samples (the green box). As shown in Fig. 1 (b), if query locations (red stars) lie at the **central area** (the black dotted box) of the $L \times L$ aerial image, the query and reference images are defined as positive samples *for each other*. Other queries (blue stars) outside the central area are defined as semi-positive samples. To guarantee that any arbitrary query has one positive reference image, we propose to densely sample the aerial images with 50% overlap along both latitude and longitude directions as

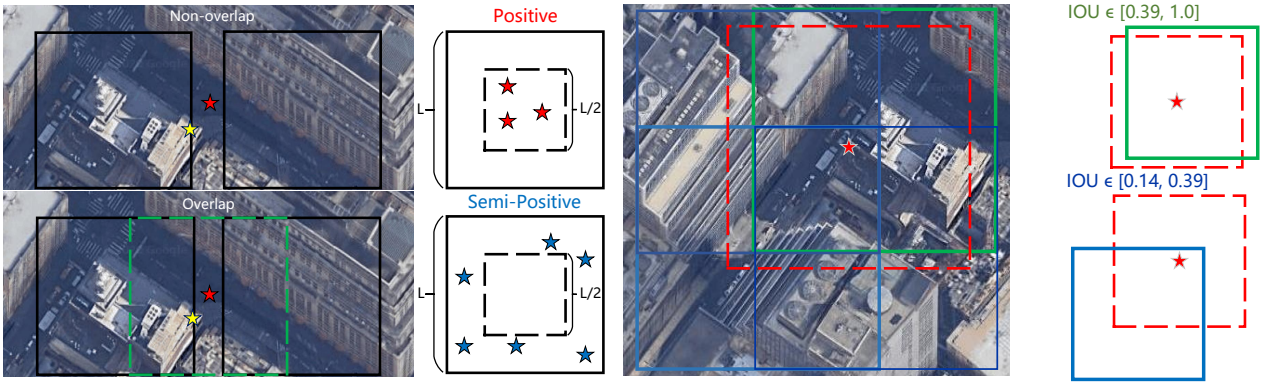

Figure 1. The sampling strategy of the proposed dataset. The stars denote the query locations.

demonstrated in Fig. 1 (c). By doing so, any arbitrary query location (the red star) in the AOI is covered by four reference images (size $L \times L$). The green box denotes the positive reference and the other three semi-positive references are denoted as blue boxes. The positive reference is considered as ground-truth, because it has the nearest GPS to the query and contains the most shared objects with the query image. The red box denotes the perfectly aligned aerial image. Based on the definitions of positive and semi-positive as illustrated in Fig. 1 (b), we can easily see that all positive reference images have an IOU (Intersection Over Union) greater than 0.39 with the perfectly aligned reference (see Fig. 1 (d)). The IOU of a typical positive sample (offset relative to the center equals to $(\pm\frac{L}{8}, \pm\frac{L}{8})$) is 0.62. The IOU between the semi-positive samples and the aligned reference falls in $[\frac{1}{7} \approx 0.14, \frac{9}{23} \approx 0.39]$.

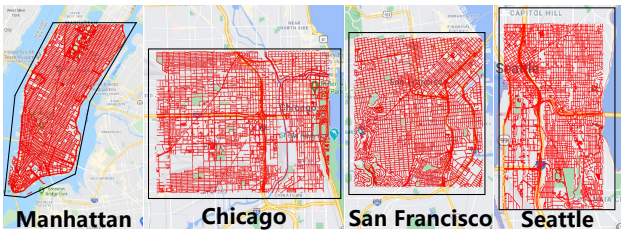

Figure 2. Aerial image coverage (black polygon) in four cities and the distributions of panoramas (red dots).

Data Collection. As shown in Fig. 2, we collect $90,618$ aerial images covering the central areas of four cities, *i.e.* New York City (Manhattan), San Francisco, Chicago, and Seattle, as the AOI using the Google Maps Static API [2]. Then $238,696$ street-view panorama images are collected with the Google Street-View Static API [1] at zoom level 2 on most of the streets. All the GPS locations of panorama images are unique in our dataset, and the typical interval between samples is about 30 m. We perform data balancing on the original panoramas to make sure that each aerial image has no more than 2 positive panoramas (see Fig. 3,

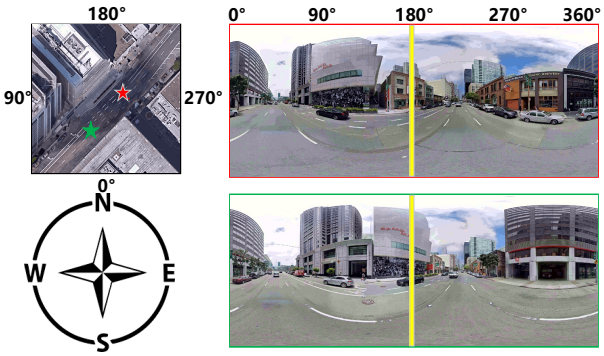

Figure 3. An example of positive samples (stars) and the orientation correspondence between aerial and ground views. The yellow bar indicates North.

the distributions are included in the supplementary material). This procedure results in $105,214$ panoramas for the geo-localization experiments. Also, around 4% of the aerial images cover no panoramas. We keep them as distraction samples to make the dataset more realistic and challenging. The zoom level for satellite images is 20 and the ground resolution is around 0.114 m. The raw image sizes for aerial-view and ground-view are 640×640 and 2048×1024, respectively. Industrial-grade GPS tags for both aerial-view and ground-view images are provided for meter-level evaluation. The panoramas are then shifted according to the orientation information so that North lies in the middle. Fig. 3 shows an example of orientation correspondence between a pair of aerial and street-view images.

Head-to-head Comparison. Table 1 shows a comparison between our dataset and previous benchmarks. The most widely used dataset, CVUSA [27], consists of images mainly collected at suburban areas. Our dataset, on the other hand, is collected for urban environments. In practice, the GPS signal is more likely to be noisy in urban areas than suburban (*e.g.* the phone-based GPS error can be up to 50 meters in Manhattan [4]). Therefore, our dataset

	Vo [24]	CVACT [11]	CVUSA [27]	VIGOR (proposed)
Satellite images	$\sim 450,000$	$128,334$	$44,416$	$90,618$
Panoramas in total	$\sim 450,000$	$128,334$	$44,416$	$238,696$
Panoramas after balancing	-	-	-	$105,214$
Street-view GPS locations	Aligned	Aligned	Aligned	Arbitrary
Full panorama	✗	✓	✓	✓
Multiple cities	✓	✗	✓	✓
Orientation information	✓	✓	✓	✓
Evaluation in terms of meters	✗	✗	✗	✓
Seamless coverage on area of interest	✗	✗	✗	✓
Number of references covering each query	1	1	1	4

Table 1. Comparison between the proposed VIGOR dataset and existing datasets for cross-view geo-localization.

has more potential application scenarios, *e.g.* vision-based mobile assistive navigation. Besides, urban areas are more crowded with tall buildings. The mutual semantics between ground and aerial views are significantly reduced by occlusions and shadows, making our dataset more challenging than CVUSA. Furthermore, previous datasets simply adopt one-to-one retrieval for evaluation, which is not the case of real-world scenarios, because it is impossible to predict the location of an arbitrary query and capture an aligned reference image there beforehand. Our dataset considers arbitrary query locations, and even the ground-truth reference image does not have the same GPS location as the query; thus it is more realistic but challenging for retrieval. Our dataset also provides the raw GPS data for meter-level evaluation which is the ultimate goal of localization applications. We believe that our dataset is a great complement to the existing cross-view image datasets, and can be served as a testbed for bridging the gap between current research and practical applications.

		Same-Area		Cross-Area	
		Number	City	Number	City
Train	Aerial	90,618	All	44,055	New York
	Street	52,609	All	51,520	Seattle
Test	Aerial	90,618	All	46,563	San Francisco
	Street	52,605	All	53,694	Chicago

Table 2. The evaluation splits of VIGOR in two settings.

The Evaluation Protocol. We design two evaluation settings for the experiments, *i.e.* same-area and cross-area evaluation, according to different application scenarios.

Same-area: If one plans to build an aerial-view reference database for arbitrary street queries in an AOI, the goal of model training is to handle arbitrary new queries. Therefore, the best solution would be collecting GPS-tagged queries in the same area for training rather than training in other areas with cross-area transfer. In this case, the aerial images in four cities are all included as the reference data for both training and testing. Then all the street panoramas are randomly split into two disjoint sets (see Table 2).

Cross-area: For cities where no GPS-tagged queries are available for training, the cross-area transfer is necessary. For this setting, all the images from New York and Seattle are used for training, and images from San Francisco and Chicago are held out for evaluation.

4. Coarse-to-fine Cross-view Localization

In this section, we propose a joint-retrieval-and-calibration framework for geo-localization in a coarse-to-fine manner. Sec. 4.1 introduces a strong baseline built with state-of-the-art techniques based on *only the positive samples*. Sec. 4.2 proposes an IOU-based semi-positive assignment loss to leverage the supervision information of semi-positive samples. With the retrieved best matching reference image, Sec. 4.3 aims to estimate the offset of the query GPS location relative to the center of the retrieved aerial-view image as a meter-level calibration.

4.1. Baseline Framework

To achieve satisfactory results on the proposed dataset, it is important to adopt state-of-the-art techniques to build a strong baseline. Therefore, we employ the feature aggregation module of SAFA (spatial-aware feature aggregation) [17] with the global negative mining strategy from [29].

Feature Aggregation. SAFA [17] is a combination of polar transformation, Siamese backbone and feature aggregation blocks. However, the polar transformation assumes that the ground-view GPS is at the center of the corresponding aerial-view reference image, which does not apply in our case. Therefore, we only adopt the feature aggregation in our framework (see Fig. 4). The main idea of the feature aggregation block is to re-weight the embeddings in accordance with their positions. The spatial-aware block provides a significant performance gain when the orientation information of query images is available.

Mining Strategy. Metric learning literature [15, 20, 13] has revealed the importance of mining hard samples during training, as the model would suffer from poor convergence when most samples barely contribute to the total loss. For cross-view geo-localization, [29] further shows the importance of mining global hard samples instead of mining

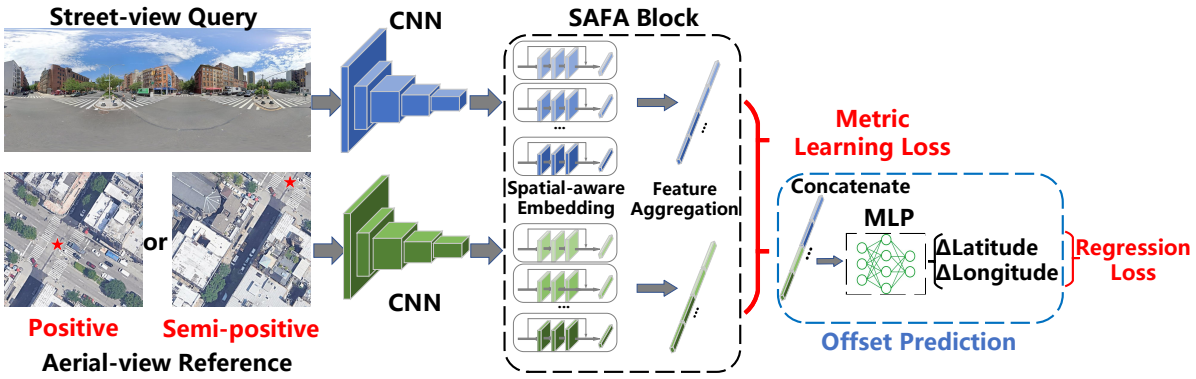

Figure 4. An overview of the proposed end-to-end framework. The Siamese network provides embedding features for retrieval as a coarse image-level localization. The offset prediction further generates refined localization in terms of meters.

within a mini-batch. The key idea is to build a first-in-first-out mining pool to cache the embedding of the hardest sample and refresh the pool along with back propagation efficiently. In a mini-batch, the first half images are randomly selected and the global hard samples with respect to each of them are mined from the mining pool to form the other half of the batch. We adopt this efficient global mining strategy [29] in the baseline to further improve its performance.

4.2. IOU-based Semi-positive Assignment

If we only consider the positive samples, the retrieval problem can be tackled with standard metric learning. For the baseline, we adopt the widely used triplet-like loss proposed in [7]:

$$\mathcal{L}_{triplet} = log\left(1 + e^{\alpha(d_{pos} - d_{neg})}\right), \quad (1)$$

where d_{pos} and d_{neg} denote the squared l_2 distance of the positive and negative pairs. In a mini-batch with N ground-view and aerial-view image pairs, we use the exhaustive strategy [15] to build $2N(N-1)$ triplets, thereby making full use of all the input images. Following [7], we adopt l_2 normalization on the output embedding features.

In addition to positive samples, it can be beneficial to take advantage of the supervision information of semi-positive samples. However, simply assigning semi-positive samples as positive would hurt the retrieval performance. For a street-view query, the semi-positive aerial images only contain a small part of the scene at the query location, thus the similarities in the feature embedding space between semi-positive samples and the query should not be as high as those of positive samples. An intuitive idea is to assign the similarity according to the IOU between the reference image and the aligned one (see Fig. 1 (d)). Therefore, the IOU-based semi-positive assignment loss is expressed as:

$$\mathcal{L}_{IOU} = \left(\frac{S_{semi}}{S_{pos}} - \frac{IOU_{semi}}{IOU_{pos}}\right)^2, \quad (2)$$

where S_{pos} and S_{semi} denote the cosine similarity of the positive and semi-positive pairs in the embedding space. IOU_{pos} and IOU_{semi} denote the IOU of positive and semi-positive pairs. This loss forces the ratio of the similarities in the embedding space to be close to the ratio of IOUs. Other assignment strategies for the semi-positive samples are also investigated and evaluated in the ablation study.

4.3. Offset Prediction

With the top-1 retrieved reference aerial image, we employ an auxiliary task to further refine the localization inside the aerial-view image in a unified framework (see Fig. 4). With image retrieval, the minimal interval between retrieved reference images in our dataset is half of the width of aerial images ($L/2$). To achieve more fine-grained localization, we apply an MLP (Multilayer Perceptron) to predict the offset of the query location relative to the center of the retrieved reference image. As shown in Fig. 4, the auxiliary MLP consists of two fully connected layers and takes the concatenated embedding features as input. Here we use regression to generate the prediction, while we also provide a comparison with classification in the experiments. The offset regression loss is formulated as:

$$\mathcal{L}_{offset} = (lat - lat^*)^2 + (lon - lon^*)^2, \quad (3)$$

where lat and lon denote the predicted offset of the query GPS location relative to the reference GPS in latitude and longitude directions, and lat^* and lon^* denote the ground-truth offset. They are all converted into meters and normalized with L during training. The final hybrid loss function is given by:

$$\mathcal{L}_{hybrid} = \mathcal{L}_{triplet} + \mathcal{L}_{IOU} + \mathcal{L}_{offset}. \quad (4)$$

5. Experiments

Implementation Details. All the experiments are deployed based on Tensorflow [3]. Ground-view panoramas and aerial-view images are resized to 640×320 and 320×320 respectively before being fed into the network. VGG-16 [19] is adopted as the backbone feature extractor and 8 SAFA blocks are used by following [17]. The mining strat-

	Same-Area				Cross-Area			
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
Siamese-VGG ($\mathcal{L}_{triplet}$)	18.1	42.5	97.5	21.2	2.7	8.2	61.7	3.1
SAFA ($\mathcal{L}_{triplet}$)	33.9	58.4	98.2	36.9	8.2	19.6	77.6	8.9
SAFA+Mining (baseline, $\mathcal{L}_{triplet}$)	38.0	62.9	97.6	41.8	9.2	21.1	77.8	9.9
Ours ($\mathcal{L}_{hybrid}$)	**41.1**	**65.8**	**98.4**	**44.7**	**11.0**	**23.6**	**80.2**	**11.6**

Table 3. Retrieval accuracy (percentage) of different methods.

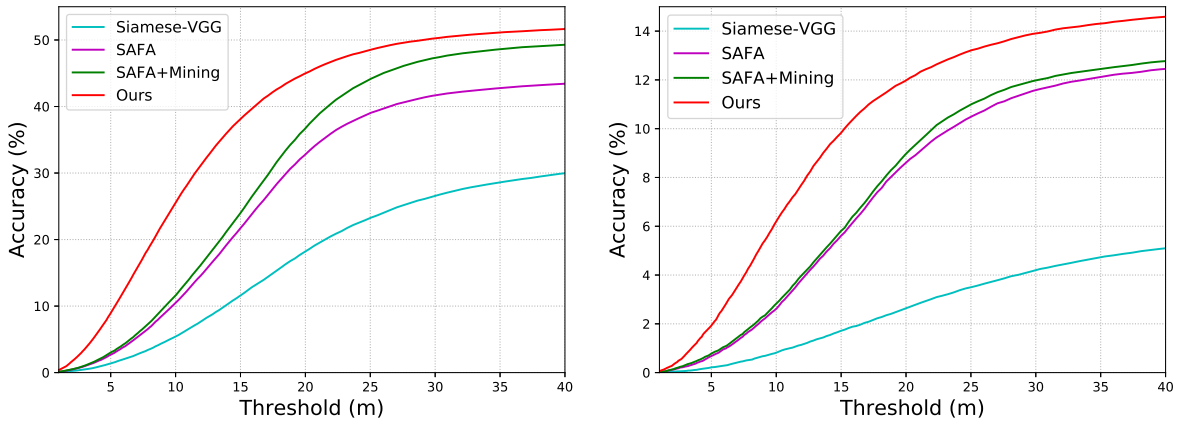

Figure 5. Same-area (left) and cross-area (right) meter-level localization accuracy of different methods.

Semi-positive Assignment	Same-Area				Cross-Area			
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
No semi-positive (*i.e.* baseline, $\mathcal{L}_{triplet}$)	38.0	62.9	97.6	41.8	9.2	21.1	77.8	9.9
Positive ($\mathcal{L}_{triplet}$)	20.3	45.7	97.9	25.4	2.7	7.6	58.2	3.1
IOU ($\mathcal{L}_{triplet}$+$\mathcal{L}_{IOU}$)	**41.1**	**65.9**	98.3	**44.8**	**10.7**	**23.5**	**79.3**	**11.4**
Positive+IOU ($\mathcal{L}_{triplet}$+$\mathcal{L}_{IOU}$)	31.1	58.3	**98.6**	36.7	5.3	13.6	69.4	6.0

Table 4. Retrieval accuracy (percentage) of the proposed method with different semi-positive assignment strategies.

egy parameters are set the same as in [29]. Following [7], we set α in the $\mathcal{L}_{triplet}$ loss to 10. The Adam optimizer [8] is used with a learning rate of 10^{-5}. Our method is first trained with $\mathcal{L}_{triplet}$. Then it switches to the hybrid loss (Eq. 4) after 30 epochs for the same-area setting and 10 epochs for the cross-area setting. The baseline (Sec. 4.1) for comparison is only trained with $\mathcal{L}_{triplet}$.

Evaluation Metrics. We first evaluate the retrieval performance with the top-k recall accuracy following previous works [7, 17]. For each test query, its closest k reference neighbors in the embedding space are retrieved as prediction. One retrieval is considered correct if the ground-truth image is included in the top-k retrieved images. If the retrieved top-1 reference image covers the query image (including the ground-truth), it is considered as a hit and the hit rate is also provided for retrieval evaluation. Moreover, we compute the real-world distance between the top-1 predicted location and the ground-truth query GPS as meter-level evaluation.

Main Results. On the proposed VIGOR dataset, we compare the proposed method with previous approaches under both same-area and cross-area settings. "Siamese-VGG" [7] is a simple Siamese-VGG network with global average pooling, and is trained with $\mathcal{L}_{triplet}$. "SAFA" and "SAFA + mining" denote the SAFA [17] architecture w/o and w/ the mining strategy in [29] using $\mathcal{L}_{triplet}$. As shown in Table 3 and Fig. 5, the proposed method constantly outperforms previous approaches in terms of both retrieval and meter-level evaluation. Compared with "SAFA+Mining" (the baseline), the relative improvements of our method for the 10-meter-level accuracy (see Fig. 5) are 124% (11.4% $\rightarrow$ 25.5%) in the same-area setting, and 121% (2.8% $\rightarrow$ 6.2%) in the cross-area setting. The substantial improvements reveal the superiority of the proposed hybrid loss.

6. Ablation Study

Semi-positive Assignment. We compare four semi-positive assignment strategies. "No semi-positive" denotes the baseline which ignores the semi-positive samples. "Positive" means assigning semi-positive samples as positive and using $\mathcal{L}_{triplet}$ (Eq. 1). "IOU" denotes our IOU-based assignment (Eq. 2). "Positive+IOU" means including the semi-positive samples as positive in $\mathcal{L}_{triplet}$ along with the IOU-based assignment loss. The results in Table 4 show that only IOU-based assignment ("IOU") boosts the per-

Offset Prediction	Same-Area				Cross-Area			
	Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
None (retrieval-only)	41.1	65.9	98.3	44.8	10.7	23.5	79.3	11.4
Regression	41.1	65.8	**98.4**	44.7	**11.0**	**23.6**	**80.2**	**11.6**
Classification	**41.5**	**66.3**	**98.4**	**45.2**	10.7	23.2	79.3	11.4

Table 5. Retrieval accuracy (percentage) of the proposed method with different offset prediction schemes.

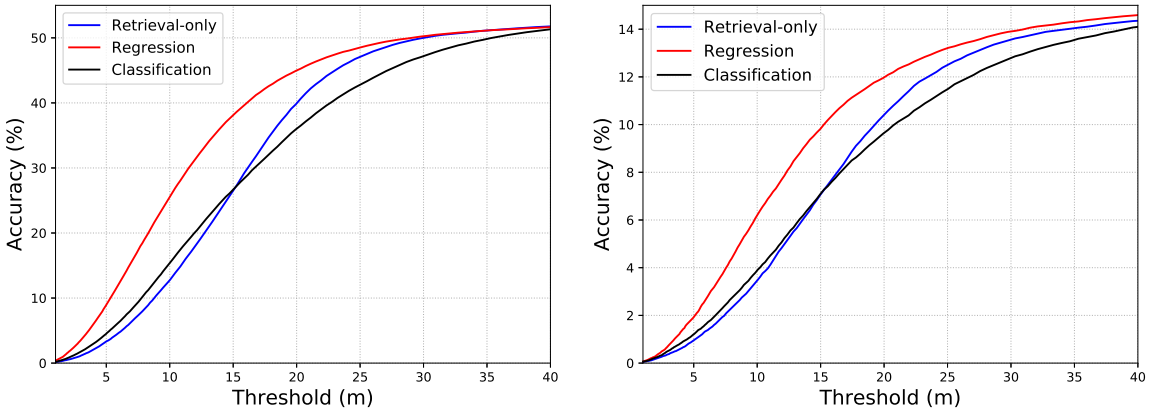

Figure 6. Same-area (left) and cross-area (right) meter-level localization accuracy of different offset prediction methods.

formance compared with the baseline. Based on the results of "Positive" and "Positive+IOU", assigning semi-positive samples as positive hinders the retrieval performance whether the IOU-based loss is used or not.

To further illustrate the difference between positive and semi-positive matching, we conduct **visual explanation**. Specifically, we use Grad-CAM [16, 28] to show which regions contribute the most to the cosine similarity of the embedding features of two views. As presented in Fig. 7, for a query image, we select the ground-truth reference aerial image (*i.e.* positive) and a semi-positive image (the query GPS location lies at its edge area), and generate the activation maps of both views. For the positive matching case, the surrounding objects (buildings, roads and trees) are all available to provide high contribution to the similarity between two views. However, in the case of semi-positive matching, two views only share half of the scene and the building on the west of the query (around 90° in panorama) does not contribute to the similarity, because it is not in this semi-positive image. This example shows how the intersection semantics of two views affect the image matching, which agrees with the design of our IOU-based assignment.

Offset Prediction. We investigate both regression and classification for offset prediction in our method. For classification, we split the central area of an aerial image (offset in $[-L/4, L/4]$) into a 10×10 grid, leading to 100 classes for classification. As shown in Table 5, both regression and classification have negligible improvement on the retrieval performance. However, as evident from Fig. 6, regression-based calibration significantly boosts the meter-level accuracy in both settings. For example, the regression

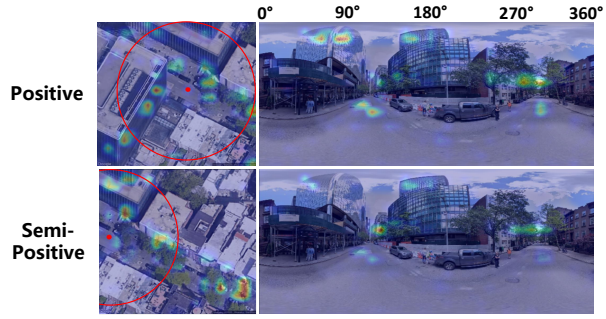

Figure 7. Visualization results of the query image matched with positive and semi-positive reference aerial images. Red circle denotes the query region.

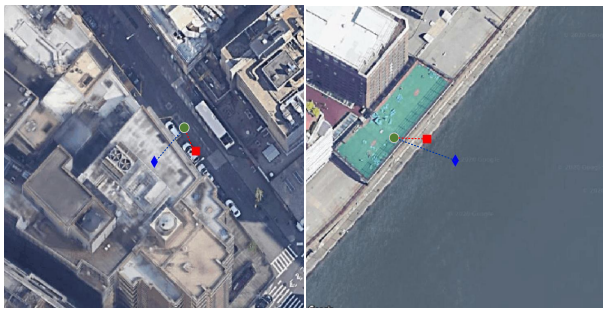

Figure 8. Case study on meter-level refinement within the retrieved aerial image. Red square, green circle and blue diamond denote the final prediction with regression, ground-truth, and center (*i.e.* the prediction with only retrieval), respectively.

method almost doubles the 10-meter-level localization accuracy. However, classification does not work well for calibration possibly due to the ambiguous supervision of grid-based classification. We provide a case study in Fig. 8 to show examples of predicted offset on aerial images.

	Positive per Aerial Image	Same-Area				Cross-Area			
		Top-1	Top-5	Top-1%	Hit Rate	Top-1	Top-5	Top-1%	Hit Rate
Baseline	2	38.0	62.9	97.6	41.8	9.2	21.1	77.8	9.9
	3	46.0	70.8	98.5	50.8	10.6	23.5	79.5	11.5
Ours	2	41.1	65.9	98.3	44.8	10.7	23.5	79.3	11.4
	3	48.5	72.9	98.9	52.6	11.5	24.8	80.8	12.2

Table 6. Retrieval accuracy (percentage) of the proposed method with different sample balancing settings.

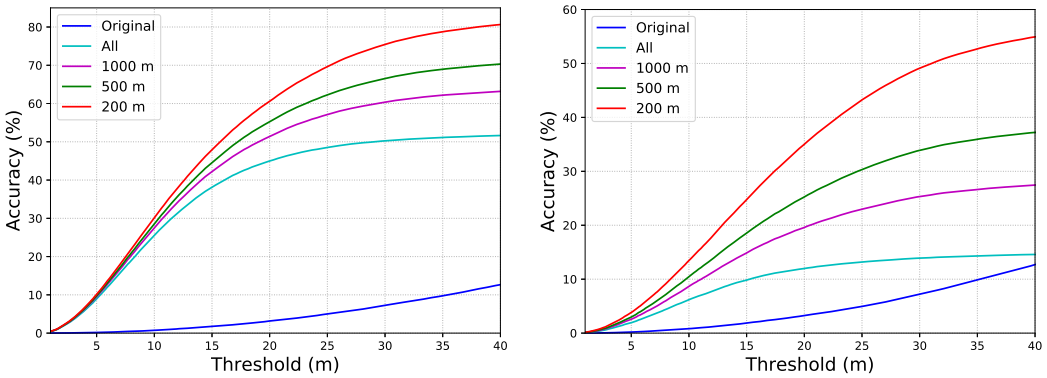

Figure 9. Same-area (left) and cross-area (right) meter-level localization accuracy of different search scopes given noisy GPS signal.

Sample Balancing. To investigate the effect of sample balancing in the pre-processing procedure, we compare "balancing-2" with "balancing-3" in Table 7. The results in Table 6 show that more densely sampled panoramas bring slightly better performance on both settings, while the improvement is consistent across different balancing settings.

	Balancing-2	Balancing-3
Positives per aerial image	2	3
Number of panoramas	105,214	149,869
Number of aerial images	90,618	90,618

Table 7. The proposed dataset with different balancing numbers.

7. Application: Assistive Navigation

The GPS data provided by commercial devices such as phones could be noisy in urban environments (*e.g.* the phone-based GPS error can reach up to 50 meters in Manhattan [4]). Image geo-localization can assist mobile navigation. To further validate the potential of cross-view geo-localization given noisy GPS signals [4, 26], we simulate noisy GPS signals by adding random offsets in $[-100m, 100m]$ to the ground-truth GPS data (latitude and longitude) in our dataset. **In the inference stage**, the query image can be matched with only a small sub-set of reference images around the noisy GPS locations instead of the entire reference database (denoted by "All"). For a noise level of $100m$, a search scope of $200m$ is sufficient to cover all possible references. To better illustrate the navigational assistance provided by our image geo-localization, we compare the results of multiple scopes with simply using the noisy GPS signals ("Original"). As shown in Table 8 and Fig. 9, smaller search scopes generate better results for both retrieval and meter-level evaluation because there are less negative reference samples. The same-area evaluation even yields an accuracy higher than 70% for $30m$-level localization. Moreover, as compared to the original noisy GPS, our cross-view geo-localization method significantly improves the localization accuracy, demonstrating its practicality in real-world applications.

Search Scope	Same-Area		Cross-Area	
	Top-1	Top-5	Top-1	Top-5
All	41.1	65.8	11.0	23.6
1000 m	49.2	76.7	19.9	41.5
500 m	54.1	82.6	26.4	53.3
200 m	60.9	90.6	37.7	72.0

Table 8. Retrieval accuracy (percentage) of the proposed method with noisy GPS signals.

8. Conclusion

We propose a new benchmark for cross-view image geo-localization beyond one-to-one retrieval, which is a more realistic setting for real-world applications. Our end-to-end framework first coarsely localizes the query with retrieval, and then refines the localization by predicting the offset with regression. An IOU-based hybrid loss is designed to leverage the supervision of semi-positive samples. Extensive results show great potential of the proposed method in realistic settings. Our proposed dataset offers a new testbed for cross-view geo-localization and inspires novel research in this field.

Acknowledgement. This work is partially supported by the National Science Foundation under Grant No. 1910844.

References

[1] *https://developers. google.com/maps/documentation/ streetview/intro.* 3

[2] *https://developers.google.com/maps/documentation/maps-static/intro*. 3

[3] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016. 5

[4] Eli Brosh, Matan Friedmann, Ilan Kadar, Lev Yitzhak Lavy, Elad Levi, Shmuel Rippa, Yair Lempert, Bruno Fernandez-Ruiz, Roei Herzig, and Trevor Darrell. Accurate visual localization for automotive applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 3, 8

[5] 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 International Conference on Computer Vision*, pages 8391–8400, 2019. 2

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[7] 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, 2, 5, 6

[8] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[9] Ang Li, Huiyi Hu, Piotr Mirowski, and Mehrdad Farajtabar. Cross-view policy learning for street navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8100–8109, 2019. 1

[10] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2

[11] 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, 2, 4

[12] Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, pages 2419–2430, 2018. 1

[13] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. *arXiv preprint arXiv:2003.08505*, 2020. 4

[14] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 470–479, 2019. 2

[15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 4, 5

[16] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 7

[17] 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, 5, 6

[18] 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. 2

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[20] Yumin Suh, Bohyung Han, Wonsik Kim, and Kyoung Mu Lee. Stochastic class-based hard example mining for deep metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7251–7259, 2019. 4

[21] Bin Sun, Chen Chen, Yingying Zhu, and Jianmin Jiang. Geocapsnet: Ground to aerial view image geo-localization using capsule network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 742–747. IEEE, 2019. 1

[22] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017. 1, 2

[23] Sebastiano Verde, Thiago Resek, Simone Milani, and Anderson Rocha. Ground-to-aerial viewpoint localization via landmark graphs matching. *IEEE Signal Processing Letters*, 27:1490–1494, 2020. 2

[24] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pages 494–509. Springer, 2016. 1, 2, 4

[25] 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*, pages 3961–3969, 2015. 2

[26] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, pages 255–268. Springer, 2010. 1, 8

[27] 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, 2, 3, 4

[28] Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual explanation for deep metric learning. *arXiv preprint arXiv:1909.12977*, 2019. 7

[29] 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, 5, 6