Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization

Royston Rodrigues
Biometrics Research Laboratories
NEC Corporation, Japan
r-rodrigues@nec.com

Masahiro Tani
Biometrics Research Laboratories
NEC Corporation, Japan
masahirotani@nec.com

Abstract

When we humans recognize places from images, we not only infer about the objects that are available but even think about landmarks that might be surrounding it. Current place recognition approaches lack the ability to go beyond objects that are available in the image and hence miss out on understanding the scene completely. In this paper, we take a step towards holistic scene understanding. We address the problem of image geo-localization by retrieving corresponding aerial views from a large database of geo-tagged aerial imagery. One of the main challenges in tackling this problem is the limited Field of View (FoV) nature of query images which needs to be matched to aerial views which contain 360°FoV details. State-of-the-art method DSM-Net [19] tackles this challenge by matching aerial images locally within fixed FoV sectors. We show that local matching limits complete scene understanding and is inadequate when partial buildings are visible in query images or when local sectors of aerial images are covered by dense trees. Our approach considers both local and global properties of aerial images and hence is robust to such conditions. Experiments on standard benchmarks demonstrates that the proposed approach improves top-1% image recall rate on the CVACT [9] data-set from 57.08% to 77.19% and from 61.20% to 75.21% on the CVUSA [28] data-set for 70°FoV. We also achieve state-of-the art results for 90°FoV on both CVACT [9] and CVUSA [28] data-sets demonstrating the effectiveness of our proposed method.

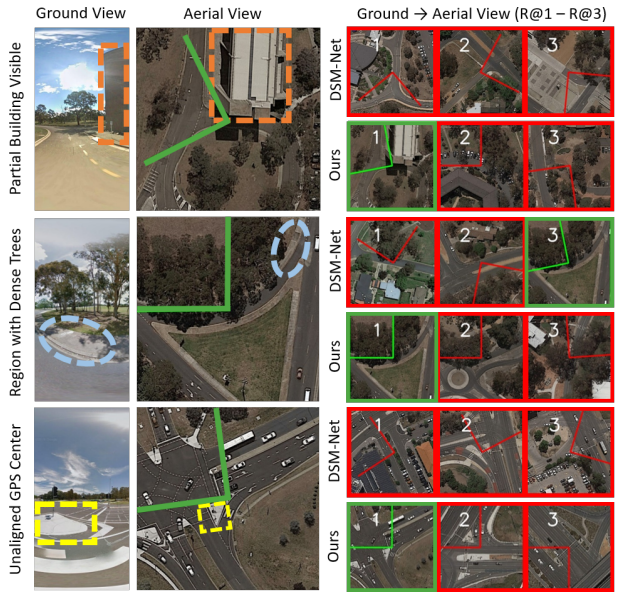

Figure 1. Given a field of view constrained query image *(Left)*, we retrieve corresponding aerial views *(Right)*. Correct and incorrect regions are indicated by *green* and *red* respectively. State-of-the-art method DSM-net [19] matches the query images locally in aerial views. We show that local matching is inadequate when buildings are visible partially *(Orange)*, dense trees cover a region *(Blue)* and when the GPS center is set incorrectly *(Yellow)*. Our work considers both global and local properties of aerial views and hence is robust to the above mentioned conditions.

1. Introduction

Consider the ground view images in Figure 1 *(Left)*. How can we determine their true location? One promising way to recognize their location is to use cross-view image geo-localization. In this method the query image who's location needs to be determined is matched against a database of geo-tagged aerial imagery. This is challenging due to large differences in visual appearance caused by extreme change in viewpoint. In this work, we consider aerial views for geo-localization of low FoV images. Images captured from standard smart phone devices and hand-held cameras have small FoVs, these devices cannot capture images with FoV larger than 180°. Images captured in portrait mode too have low FoV information. Targeting small FoV has applications in enabling geo-localization them. This paper looks at effective aerial representations to enable low field of view

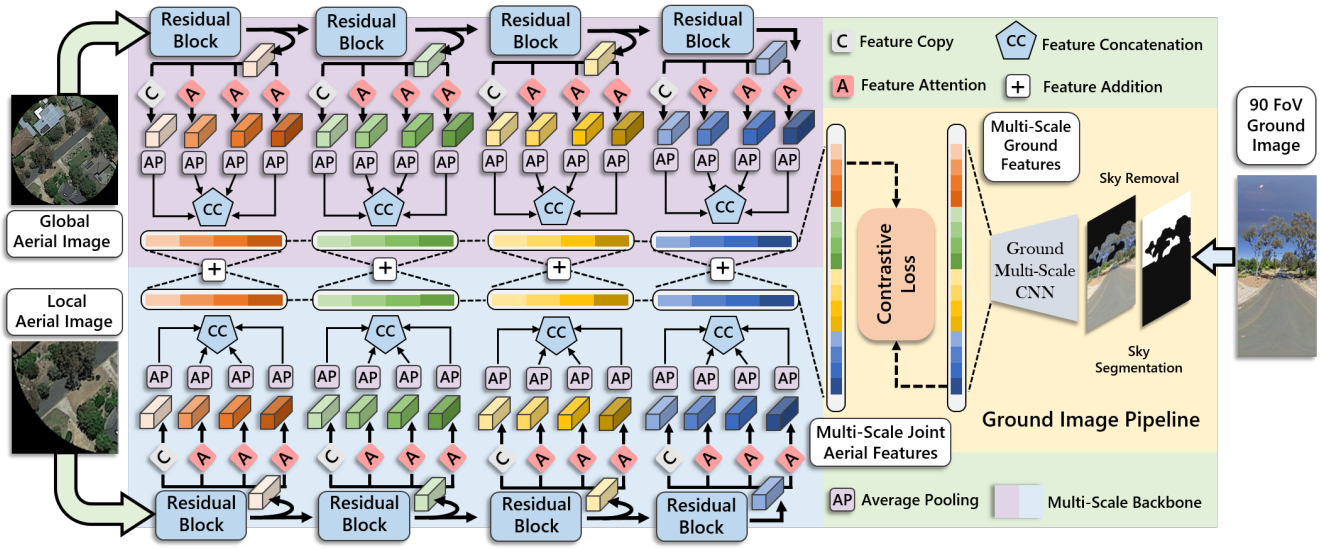

Figure 2. **Our Siamese Framework:** We propose to leverage both global *(Purple)* and local *(Blue)* information present in aerial images to enable Field of View constrained image geo-localization. Our aerial view representation consists of both global and local properties. Original and attention [25] based features are extracted from intermediate layers of a ResNet-18 [5] module. For ground view features, an off-the-shelf image segmentation module [30] is used to remove sky region from ground view images *(Yellow)*. This is important since sky region is not available in aerial images and can be ignored during matching. We use contrastive loss for learning these representations.

(i.e. 90°and 70°) cross-view image geo-localization.

Cross-view image geo-localization problem is posed as an image retrieval task. Recently, deep-learning [16, 9, 20, 18, 19, 32, 24, 22, 21, 14, 11] has enabled high cross-view image geo-localization rates for panoramic images. One of the main challenges to use such frameworks is the limited FoV nature of query images as opposed to the 360°FoV nature of aerial view images. Existing methods such as DSM-Net [19] solve for this challenge by matching locally in aerial views within fixed FoV sectors. The quest for perfectly aligned aerial views to the corresponding ground view image involves a dense search along fixed FoV sectors. Such a search demands significant compute and can be time consuming. We show that relying just on local matching is inadequate for recognizing places where complete information is not present (See Figure 1). These involve places which are covered by dense trees or landmark regions that are partially visible. For successful geo-localization of such cases one needs to look at scenes in a holistic manner. Here information about neighbouring objects beyond fixed FoV sectors must also be considered.

Our work uses 360°FoV information of aerial views to infer global information and fuses it with local aerial information, the joint representation is then used for cross-view image matching. To the best of our knowledge, this is the first work that learns joint global-local representations effectively to enable field of view constrained image-localization. We also propose a straight-forward data-augmentation method for aerial images that eliminates the compulsion to search perfectly aligned aerial views to geo-localize images. Our data-augmentation method leads to recall rates similar to dense search on CVACT dataset for 90°FoV by performing fewer search iterations. This saves cross-view image retrieval time for FoV constrained image geo-localization.

The contribution of this work are as follows:

- We propose to take advantage of 360°FoV information present in aerial view images to learn joint global and local aerial view representations for cross-view image geo-localization. This leads to holistic scene understanding and enables cross-view image matching module to infer beyond objects that are available.

- We also present a new data augmentation method to facilitate faster inference during search and retrieval. Our data augmentation enables reduction in search time for geo-localizing 8884 images from 120 mins to 8 mins for 90°FoV, while maintaining similar top-1 recall rates.

- Experiments indicate that we improve top-1% image recall rate on the CVACT [9] data-set from 57.08% to 77.19% and from 61.20% to 75.21% on the CVUSA [28] data-set for 70°FoV. We also achieve state-of-the art results for 90°FoV on both CVACT [9] and CVUSA [28] data-sets demonstrating the effectiveness of our proposed method.

2. Literature Review

In this section we review some prior and concurrent works on cross-view image geo-localization.

Cross-view image geo-localization has been researched extensively before the existence of deep-learning. Methods proposed by [4, 7, 17, 2] worked towards extracting meaningful handcrafted features for this task. With the increasing effectiveness of deep learning, end-to-end feature learning is now considered as an effective approach. [26] made use of a pre-trained CNN to discover location specific features, it was discovered that deeper layer of pre-trained CNNs include location specific semantic information. [27] explored the effect of fine tuning CNNs that were pre-trained on large data-sets by imposing an objective function to place feature vectors of images that came from the same location together. [8] made use of contrastive loss function to train Siamese networks for cross view image geo-localization task. [6] used NetVlad [1] for place recognition task. [9] made use of orientation priors by incorporating color coded information. This color coded signal was fed as an input to Siamese networks. Embedding orientation information made cross-view image retrieval frameworks more robust as it simplified the cross view image matching. [20] proposed an optimal feature transport which lead to effective cross-view image feature alignment. Geometric differences were compensated in feature space by proposing a learn-able feature alignment. [18] proposed to reduce the geometric differences between view points by proposing a polar transformation. It is interesting to note that this transform was not leanable and just involved pixel reordering. [18] also introduced a new spatial aware attention framework which lead to higher recall rates. A new data-set was proposed by [32] to handle miss-alignment across the two views. Authors in [24] proposed an image feature extraction module for capturing local properties effectively. [16] proposed a data-augmentation method to facilitate cross-view image matching across temporally varying scenes. They discovered that cross-view image pairs were not captured at the same time, leading to change in scenes. [16] also proposed an attention mechanism for extracting cross-scale features. [13] reduced the view point differences by using conditional GANs [12] to bridge the domain gap between the two views. Approach proposed by [13] could generate aerial views from ground view images. Another approach that used GANs for cross-view image generation was proposed by [22], here polar transformed images [18] were also considered in the image generation framework. GANs with geo-metric priors were trained by [10]. Authors in [29] propose drone based image geo-localization. This was considered as another valuable data source for geo-localization task.

Our approach is different from the above mentioned works as we look at the task of field of view constrained image geo-localization. Particularly we target low FoV images (*i.e.* 90°FoV and 70°FoV). The work most relevant to us in the literature is [19]. Authors in [19] proposed a local approach to field of view constrained image geo-localization. We introduce global aerial view features into cross view image matching pipeline. This is useful to match objects that are not visible in locally constrained fixed field of view sectors and provides holistic scene understanding. This makes image geo-localization systems more robust and significantly increases image recall rates.

3. Approach

In this section we introduce our Siamese framework (Figure 2). We later present our proposed data-augmentation method and state its effectiveness for field of view constrained image geo-localization.

3.1. Siamese framework for geo-localization

We use a Siamese network pipeline trained with a contrastive loss objective to facilitate cross-view image retrieval task. (Figure 2) The key idea of our framework is to incorporate global aerial representation into the feature extraction pipeline to aid field of view constrained image geo-localization.

Input Representation: In Figure 3 we illustrate the input images for field of view constrained ground view images and their corresponding global and local aerial views. In particular, for 90°FoV we use an input resolution of 312 $\times$ 156 pixels. We maintain similar height for the case of 70°FoV, here we use an input resolution of 312 $\times$ 121 pixels. Image resolution for global aerial view is 200 $\times$ 200 pixels. Local aerial view resolution is set to 100 $\times$ 100 pixels. In order to mask out regions that do not correspond to the local region, a 70°mask is used.

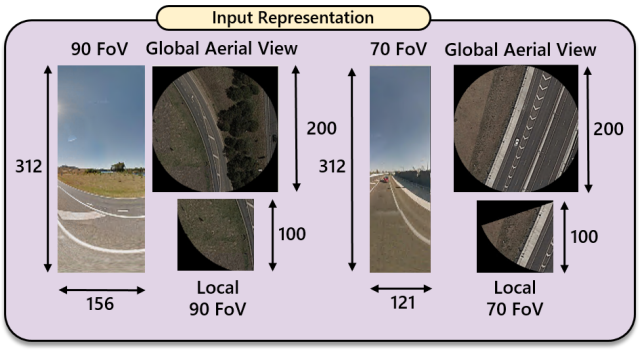

Figure 3. Input representation: Ground and aerial cross-view image pairs considered in our geo-localization framework.

Objective Function We make use of a metric learning objective to learn joint global-local aerial image representations. Contrastive loss is used to learn image representations such that positive matching pairs are brought closer to each other and non matching negative pairs are at-least

separated by a margin m. We employ the below objective function:

$$L = (1-Y)*\frac{1}{2}*(D_w)^2+(Y)*\frac{1}{2}*max(0, m - D_w)^2 \quad (1)$$

$$D_w = ||F_g(I_g) - F_a(I_a)||^2 \quad (2)$$

$$F_a(I_a) = F_{a_l}(I_{a_l}) + F_{a_g}(I_{a_g}) \quad (3)$$

here, m is a margin parameter. F_g is the feature extractor for field of view constrained ground view image I_g. F_a is joint global-local aerial view feature extractor which operates on aerial image I_a. It extracts features from local aerial image I_{a_l} using local feature extractor F_{a_l} and global aerial image I_{a_g} using global feature extractor F_{a_g}. Joint representation of aerial image feature is the addition of $F_{a_l}(I_{a_l})$ and $F_{a_g}(I_{a_g})$. Y is a binary indicator representing the input pair as positive or negative. D_w is the squared error distance computed between ground feature representation $F_g(I_g)$ and aerial representation $F_a(I_a)$.

Multi-scale CNN backbone: In-order to capture feature representation from aerial and ground views at multiple scales we utilize a Multi-scale CNN backbone (Figure 2). We use ResNet-18 [5] as our primary backbone network and use attention mechanism from [25] to extract features at multiple scales. Intermediate layers of the model are considered as features at multiple scales [15]. The attention mechanism applies both channel and spatial attention. We use one original feature and three such attentive features at each scale. The learnable parameters of ResNet-18 are initialised from Image-net pre-taining, the parameters for the learnable attention modules are randomly initialised.

3.2. Orientation robust data augmentation

Due to the miss-alignment between aerial views and images to be geo-localized (See Figure 5). Its common practice to extract features multiple times from the same aerial view along different orientations (See Figure 4, Orange). We notice the computation over-head [19] involved in extracting features from many orientations to perform cross-view image geo-localization. This computational over head can be substantially reduced by designing an orientation robust image geo-localization frame work. We intentionally introduce controlled miss-aligned global and local aerial views as positive samples during the training phase (See Figure 4, Green). This enforces the matching network to match non-aligned samples with partial objects that might not be available for matching. In Figure 4 we explain the steps to prepare positive and negative samples for our proposed cross-view image geo-localization framework. We use a zero mean Gaussian random variable with variance parameter σ to control the strength of miss-alignment. The parameter σ is FoV dependent and must be tuned separately for each FoV. We show in Table 4,5 the effect the parameter σ for varying FoV conditions.

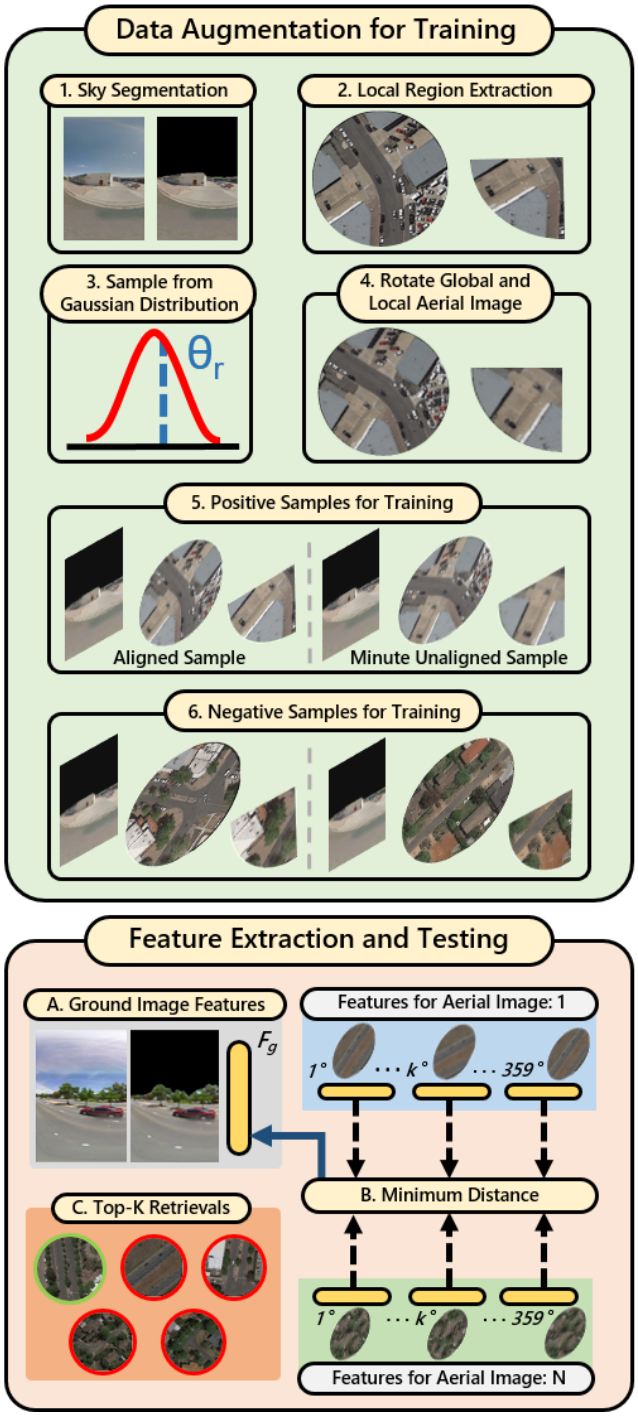

Figure 4. Illustration of our proposed data-augmentation *(Green)*. We introduce controlled miss-aligned pairs as positive samples during training. Data-augmentation is not performed on negative samples. We also show aerial image feature extraction and retrieval process during testing *(Orange)*.

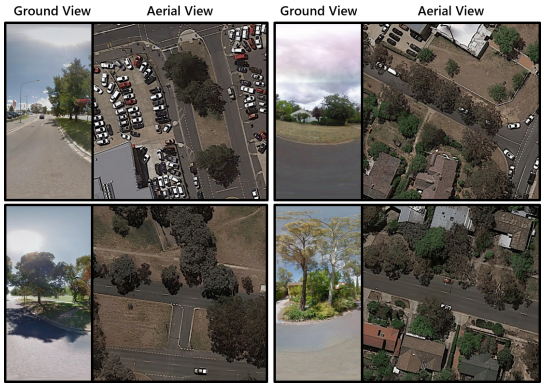

Figure 5. Ground images and their corresponding aerial views from the CVACT [9] data-set. Examples shown correspond to the case of 90°FoV. Note that its challenging to match the two since the camera orientation is unknown and the images are not aligned.

4. Experiments

We show the effectiveness of our proposed approach with qualitative and quantitative experiments. Qualitatively, (See Figure 6) we show good and poor examples of our method in comparison to DSM [19]. Our method infers about neighbouring landmarks and is able to consider aerial image features outside local regions successfully.

Data-sets: We evaluate our method using two standard benchmarks *i.e.* CVUSA [28] and CVACT [9]. Each data-set contains 35,532 training cross view image pairs and a held out set of 8,884 validation image pairs. Here, the ground views consist of full FoV panoramic images. In order to evaluate for field of view constrained image geo-localization, we create limited FoV images for training and testing from panoramic view images following the protocol defined by [19].

Implementation Details: Our Siamese framework is trained end-to-end with contrastive loss function as defined in Eq.1. We use adam as optimizer with 10^{-4} as initial learning rate for the first 30 epochs. After which, we reduce the learning rate by a factor of 5 every two epochs. This is continued for another 10 epochs. The margin m in our contranstive loss framework is set to 1. We make use of large batch size for training B positive pairs and $B(B-1)$ negative pairs. B in our experiments is set to 128. Large batch size is realised by accumulating gradients from smaller batches and updating parameters of the network after effective batch size is attained. We make use of a single RTX 8000 GPU for our experiments. Our model when trained end to end takes 12 days to complete 40 epochs with a learning rate schedule as discussed.

Evaluation metric: We follow top-k as our evaluation metric similar to [16, 18, 9, 6, 20, 19] to measure location estimation performance of our proposed method and compare it with [6, 20, 19]. Here, for an input ground view image we retrieve K corresponding aerial views with the least L_2 distance between the learned ground view and joint global-local aerial feature representation. The ground view is considered as correctly localized if the corresponding aerial image is among the top-k retrievals. The fraction of correctly localized ground views is considered as top-K.

4.1. Comparison with state-of-the-art methods

We compare our proposed method with [6, 20, 19] for the task of field of view constrained image geo-localization on the CVACT [9] and CVUSA [28] data-sets. Visual results for DSM [19] (See Figure 6) are obtained by utilizing codes open sourced by the authors. We report image recall rates at Top-1, Top-5, Top-10 and Top-1% and enumerate it in the Tables 1 and 2.

Quantitative results for CVACT: As shown in Table 1, we compare the proposed method with [6, 20, 19] on the CVACT [9] dataset for the evaluation conditions of 90°FoV and 70°FoV. The proposed method achieves 26.05% Top-1, 49.23% Top-5, 59.26% Top-10 and 85.60% Top-1% retrieval rates for 90°FoV and 14.17% Top-1, 32.96% Top-5, 43.24% Top-10 and 77.19% Top-1% rate for 70°FoV conditions. Our method surpasses all existing work on the above benchmark attaining the new state-of-the-art.

Quantitative evaluation on CVUSA: We report results on the CVUSA [28] dataset in Table 2. Our method is compared with [6, 20, 19] for 90°FoV and 70°FoV. We achieve 22.54% Top-1, 44.36% Top-5, 54.17% Top-10 and 83.59% Top-1% retrieval rates for 90°FoV and 15.20% Top-1, 32.86% Top-5, 42.06% Top-10 and 75.21% Top-1% rate for 70°FoV conditions. The results show the effectiveness of our method and achieves the new state-of-the-art on the CVUSA [28] benchmark.

Table 1. Comparison of recall rates for localizing ground view images for 90°and 70°FoV on the CVACT [9] data-set.

	CVACT [9] 90°FoV				CVACT [9] 70°FoV			
Method	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
CVM [6]	1.47	5.70	9.64	38.05	1.24	4.98	8.42	34.74
CVFT [20]	1.85	6.28	10.54	39.25	1.49	4.13	8.19	34.59
DSM [19]	18.11	33.34	40.94	68.65	8.29	20.72	27.13	57.08
Ours	**26.05**	**49.23**	**59.26**	**85.60**	**14.17**	**32.96**	**43.24**	**77.19**

Table 2. Comparison of recall rates for localizing ground view images for 90°and 70°FoV on the CVUSA [28] data-set.

	CVUSA [28] 90°FoV				CVUSA [28] 70°FoV			
Method	r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
CVM [6]	2.76	10.11	16.74	55.49	2.62	9.30	15.06	21.77
CVFT [20]	4.80	14.84	23.18	61.23	3.79	12.44	19.33	55.56
DSM [19]	16.19	31.44	39.85	71.13	8.78	19.90	27.30	61.20
Ours	**22.54**	**44.36**	**54.17**	**83.59**	**15.20**	**32.86**	**42.06**	**75.21**

Figure 6. Examples of our results compared to DSM on the CVACT data-set for 70°and 90°FoV. Successful matching results are shown in Yellow, Blue and Red. Since our aerial representation includes 360°FoV information our method understands symmetry in scenes *(Yellow)*. We enable cross-view image matching when landscapes are covered by trees *(Blue)*. Successful retrievals are obtained when just a part of landscape is visible *(Red)*. Common failures occur when similar aerial regions exist in the search data-base *(Grey)*.

Table 3. Ablation Study: Effect of considering local and global aerial view representations along with sky removal from ground view for 90°and 70°FoV conditions on the CVACT [9] data-set.

Ground View	Aerial View Representations		CVACT [9] 90°FoV				CVACT [9] 70°FoV			
Sky Remove	Local	Global	Top-1	Top-5	Top-10	Top-1%	Top-1	Top-5	Top-10	Top-1%
-	✓	-	16.31	35.70	45.46	77.60	11.59	26.53	35.13	67.04
-	-	✓	18.01	36.83	46.66	78.48	12.85	28.55	37.35	71.20
-	✓	✓	19.61	39.47	48.12	78.91	13.40	31.58	41.17	74.04
✓	✓	-	22.64	42.85	52.66	80.53	14.06	32.17	42.55	74.65
✓	-	✓	23.15	45.90	56.43	83.73	14.10	32.41	42.63	76.48
✓	✓	✓	**26.05**	**49.23**	**59.26**	**85.60**	**14.17**	**32.96**	**43.24**	**77.19**

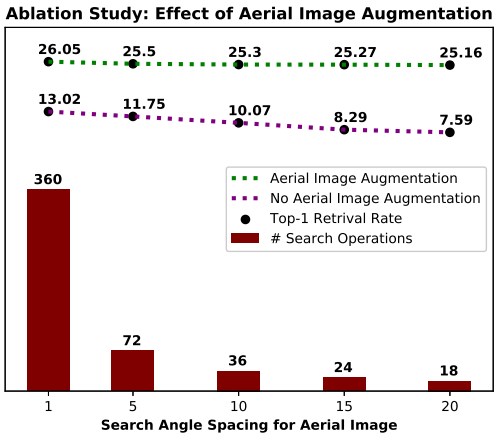

Figure 7. Ablation Study: Effect of proposed data-augmentation on Top-1 retrieval rate for 90°FoV on the CVACT [9] data-set. Smaller angle spacing leads to higher search operations within the aerial view. We compared Top-1 retrieval rates for with and without data-augmentation. Our method maintains consistent Top-1 retrieval rates even for large angle spacing enabling faster search.

4.2. Ablation Studies

In this section, we provide ablation studies performed on the CVACT [9] data-set for field of view constrained image geo-localization. In particular, we study the of effect of considering local and global aerial view features. We also validate important design choices such as the choice variance strength (σ) for data-augmentation and sky removal from ground view images.

Effect of sky removal: Top portion of all ground view images include sky region, however this region does not exist in aerial views. We remove the sky region by blacking out the pixels that correspond to sky. This is done by setting pixel values belonging to sky region to zero. From Table 3 its evident that removing sky region improves recall rates.

Effect of global and local aerial features: It is interesting to observe (See Table 3) that independent use of global aerial features outperforms the use of only local aerial features. Combining both local and global features further improves recall rates, demonstrating that the two support each other for cross-view image matching. A further improvement in retrieval rate is observed when sky region is removed. This increase is consistent across both 90°and 70°FoV.

Table 4. Ablation Study: Effect of Variance level (σ) used in data-augmentation for 70°and 90°FoV on CVACT data-set.

Variance	CVACT [9] 70°FoV			CVACT [9] 90°FoV		
Level (σ)	Top-1	Top-10	Top-1%	Top-1	Top-10	Top-1%
$\sigma = 0$	8.03	28.75	63.24	13.02	38.11	72.02
$\sigma = 15$	**14.17**	**43.24**	**77.19**	23.33	54.87	83.68
$\sigma = 20$	13.91	41.15	73.73	**26.05**	**59.26**	**85.60**
$\sigma = 40$	3.42	17.06	48.45	10.51	22.40	61.34

Effect of variance strength (σ): We show in Table 4 that variance strength (σ) i.e the amount of misalignment introduced in positive samples during training, has to be tuned for each FoV separately. Best retrival rates for 90°FoV is obtained when $\sigma = 20$ whereas for 70°FoV peak performance is obtained for $\sigma = 15$. This is expected since a large miss alignment can lead to little/no overlap between the ground and local aerial view affecting retrieval rates.

Effect of search angle spacing vs data-augmentation: It is evident from Figure 7 that the proposed data augmentation leads to stable top-1 retrieval rates on the CVACT [9] data-set for 90°FoV compared to the case with no data-augmentation, here top-1 retrieval rates decrease as search angle spacing increases. Our method is less sensitive to change in angle spacing.

Effect of search angle spacing vs retrieval time: A smaller search angle indicates dense search inside aerial views. For 1°spacing 360 search operations are performed inside a single aerial view as compared to 20°spacing case where only 18 search operations need to be performed. Larger search angle spacing indicates coarser search. We also report retrieval time for 8884 images for different angle spacing's. It takes 120 minutes to complete cross view image matching for 8884 images on the CVACT data-set for 90 FoV for the dense searching case of 1°angle spacing. This time gets significantly reduced to 8 minutes when a coarse search of 20°angle spacing is used. Time reported was evaluated using a single RTX 8000 GPU.

4.3. Large Field of View Image Geo-localization

Datasets: We consider FoVs greater than 90 degrees under the category of large FoV for image geo-localization. We evaluate the performance of large FoV image geo-localization using the above mentioned CVACT and CVUSA datasets. We also make use of additional Vo's [23] dataset for evaluating image geo-localization under this setup. Vo et al. [23] introduced a cross-view dataset which consists of about one million image pairs from 11 cities in the US. We follow the split specified in [23] using 8 cities for training and Denver city as test set for evaluation. Vo's a challenging dataset in comparison to CVUSA and CVACT, as it contains around 70k images for testing.

Large FoV Images: Since CVACT and CVUSA provide access to panoramic images 180 degree FoV crops are obtained using strategy provided by [19]. Vo's [23] dataset on the other hand provide pre-cropped limited FoV images. They donot provide the exact FoV details, visually the FoV appear greater than 90 but less than 180 degrees.

Results for 180°FoV: In Table 5 we compare our method with existing work on the CVACT [9] and CVUSA [28] dataset. Our results for 180°FoV, are just marginally high compared to the state of the art. One plausible reason could be use of similar back-bone (i.e. ResNet-18) for both large and low FoVs. Matching larger FoVs, might demand a backbone network with higher capacity. As seen in Table 5, Data-Augmentation strength (σ) needs to be carefully tuned to obtain desired results. This indicates that our proposed method is most effective for low FoV conditions such as 70°and 90°FoV. For higher FoV local aerial view might contain sufficient information, adding extra global information improves recall rates slightly.

Table 5. Comparison of recall rates for localizing ground images for 180°FoV on CVUSA and CVACT data-sets.

		CVACT [9] 180°FoV				CVUSA [28] 180°FoV			
Method		r@1	r@5	r@10	r@1%	r@1	r@5	r@10	r@1%
CVM [6]		3.94	13.69	21.23	59.22	7.38	22.51	32.63	75.38
CVFT [20]		7.13	18.47	26.83	63.87	8.10	24.25	34.47	75.15
DSM [19]		49.12	67.83	74.18	89.93	48.53	68.47	75.63	93.02
O	σ : 22	40.79	62.37	74.61	90.74	39.11	64.52	71.50	92.03
U	σ : 24	44.52	66.09	76.26	92.88	45.63	65.76	75.32	93.18
R	σ : 28	**49.93**	**68.48**	**77.16**	**93.01**	**48.91**	**69.87**	**78.50**	**95.68**
S	σ : 30	41.35	65.78	75.72	92.21	40.03	64.96	76.50	94.68

Evaluation on Vo's Data-set: We compare our method with existing works using Vo's dataset. The test set in Vo's data-set is large-scale with around 70k samples from Denver city area. It consists of additional challenges such as image examples from near by places and multiple crops that are extracted from same region. Due to the large-scale nature of this data-set we follow Top-1 % rate as our evaluation criteria similar to [27, 23, 6, 3]. From the results we observe that our method achieves results comparable to state-of-the-art. This validates that our proposed approach is most suited for low FoVs and achieves comparable performance for the condition of large FoV image geo-localization.

Table 6. Comparison of Top-1% recall rate using Vo's Dataset.

Method	Top-1%
WideArea-Net [27]	15.4%
Triplet-eDBL [23]	62.4%
CVM [6]	67.9%
Reweight [3]	78.3%
Binomial [31]	88.3%
Ours	**89.1%**

5. Conclusion

In this work we proposed a method to address the challenges of field of view constrained cross-view image geo-localization. Leveraging 360°FoV aerial representation is advantageous as it infers scenes holistically and improves image geo-localization recall rates for low FoV conditions. Our solution involves two proposals. We first propose global satellite representations to leverage full FoV aerial image information, this enables cross-view matching framework to infer beyond objects that are limited to a fixed FoV and also consider surrounding landmarks.We also introduced a new data-augmentation approach where we included controlled miss aligned cross-view pairs as positive examples for training. This made the matching network robust to miss alignments and improved retrieval speed during inference. Experiments on standard benchmarks confirm that we boost recall rates for 70°and 90°FoV demonstrating the effectiveness of our proposed solution.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[2] Mayank Bansal, Harpreet S. Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *Proceedings of the 19th ACM International Conference on Multimedia*, MM '11, 2011.

[3] 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/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[4] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. Semantic cross-view matching. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, December 2015.

[5] 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.

[6] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] T. Lin, S. Belongie, and J. Hays. Cross-view image geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[8] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Niluthpol Chowdhury Mithun, Karan Sikka, Han-Pang Chiu, Supun Samarasekera, and Rakesh Kumar. Rgb2lidar: Towards solving large-scale cross-modal visual localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 2020.

[12] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[13] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[14] Krishna Regmi and Mubarak Shah. Video geo-localization employing geo-temporal feature learning and gps trajectory smoothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[15] Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, and Subhasis Chaudhuri. Multi-timescale trajectory prediction for abnormal human activity detection. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[16] Royston Rodrigues and Masahiro Tani. Are these from the same place? seeing the unseen in cross-view image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[17] T. Senlet and A. Elgammal. A framework for global vehicle localization using stereo images and satellite and road maps. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, 2011.

[18] 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 (NIPS)*. 2019.

[19] 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 (CVPR)*, June 2020.

[20] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[21] Tsenjung Tai and Masato Toda. Adapting intra-class variations for sar image classification. In *IEEE International Conference on Image Processing (ICIP)*, 2021.

[22] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixe. 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 (CVPR)*, June 2021.

[23] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision (ECCV)*, 2016.

[24] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zhenga, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2021.

[25] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[26] Scott Workman and Nathan Jacobs. On the location dependence of convolutional neural network features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2015.

[27] 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)*, 2015.

[28] 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 (CVPR)*, 2017.

[29] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 2020.

[30] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] 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)*, 2021.

[32] 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)*, 2021.