ArcGeo: Localizing Limited Field-of-View Images using Cross-view Matching

Maxim Shugaev* Maxim.Shugaev@BlueHalo.com
Ilya Semenov* Ilya.Semenov@BlueHalo.com
Kyle Ashley* Kyle.Ashley@BlueHalo.com
Michael Klaczynski* Michael.Klaczynski@BlueHalo.com
Naresh Cuntoor* Naresh.Cuntoor@BlueHalo.com
Mun Wai Lee* MunWai.Lee@BlueHalo.com
Nathan Jacobs† jacobsn@wustl.edu

Abstract

Cross-view matching techniques for image geo-localization attempt to match features in ground-level query images against a collection of satellite images to determine their positions of origin. We present ArcGeo, a novel cross-view image matching approach which introduces a batch-all angular margin loss and several train-time strategies including large-scale pretraining and FoV-based data augmentation. This allows our model to perform well even in challenging cases with limited field-of-view (FoV). Further, we evaluate multiple model architectures, data augmentation approaches and optimization strategies to train a deep cross-view matching network, specifically optimized for limited FoV cases. In low FoV experiments (FoV = 90°) our method improves top-1 image recall rate on the CVUSA dataset from 30.12% to 43.08%. We also demonstrate improved performance over the state-of-the-art techniques for panoramic cross-view retrieval, improving top-1 recall from 95.43% to 96.06% on the CVUSA dataset and from 64.52% to 79.88% on the CVACT test dataset. Lastly, we evaluate the role of large-scale pretraining for improved robustness. With appropriate pretraining on external data, our model improves top-1 recall dramatically to 66.83% for the FoV = 90° test case on CVUSA, an increase of over twice what is reported by existing approaches.

1. Introduction

Image geolocation techniques look to identify the locations of images by correlating features between a query image and set of reference images. These methods have several applications including autonomous driving, and metadata enrichment for indexing and retrieval. Recently, cross-view matching approaches have shown success matching features between aerial and ground images using datasets of paired images [1–10]. The reference aerial images are usually obtained using the Google Maps API [11] or Bing Maps API [12] which provide high resolution imagery spanning the entire planet. However, there is a significant domain gap between aerial and ground images including an extreme shift in perspective and variation in transient attributes (e.g., season, weather and lighting).

Prior research addresses these challenges using a combination of strategies including polar transformation of aerial images [2, 4] and attention mechanisms for cross-view feature correlation [6, 7]. However, most existing approaches use features from 360° ground view panoramas which include a full perspective of matchable features [6, 7, 10]. In many real-world applications, panoramic imagery is not available. For example, the FoV in typical cell phone cameras ranges from ∼130 degrees (0.5x magnification) to ∼20 degrees (3x magnification). Existing research in cross-view matching for restricted FoV is limited and shows a significant gap in performance between panoramic and limited-FoV test cases [8]. The technique providing state-of-the-art performance for 90° FoV only obtains 30.12% top-1 recall, a third of the 94.08% it reports for 360° FoV [7].

DSM [4] accounts for limited FoV and reports improved results, but due to the co-dependence between the aerial and ground embeddings it uses to estimate view direction, the model does not use batch-all triplet loss, and is instead trained with individual triplet loss. More recent approaches using batch-all loss nearly double its performance [7].

We introduce ArcGeo loss, which further improves model performance. By replacing batch-all triplet loss with ArcGeo, our model's r@1 improved from 28.59% to 44.18% (see Table 4). ArcGeo loss addresses the shortcomings of triplet loss in performing local optimization of the embedding space by replacing triplet approximation with true SoftMax, thereby enabling global embedding space op-

*BlueHalo, 15400 Calhoun Dr #190, Rockville, MD
†Washington University, 1 Brookings Dr, St. Louis, MO
Approved for public release, **NGA-U-2023-00446**

timization. ArcGeo loss is motivated by Arcface [13], an additive margin Softmax loss function that uses an angular penalty to improve class separation. ArcGeo adapts this concept to cross-view matching by using model predictions instead of centroids to approximate the global embedding space, as centroids may not synchronize well in use cases where there are few images per class, such as in cross-view matching, where each location only has one ground-aerial image pair. ArcGeo loss provides new SOTA both for FoV of 360° as well as limited FoV and unknown view direction. The improvement is especially pronounced under conditions far from the model performance saturation, e.g., large evaluation sets (see Table 6), and low FoV (Table 4).

Additionally, we observe that many popular cross-view matching datasets do not contain sufficient diversity for low FoV matching due to their limited size. Specifically, the full CVUSA dataset contains ∼1.2M image pairs, however [1–10] use only a small subset of ∼35k image pairs for training and ∼9k for validation. We show that proper pretraining on larger datasets yields significant performance improvements.

The main contributions of this work are:

- We introduce a novel ArcGeo loss function for robust cross-view matching. It outperforms triplet and Arcface loss in a variety of tests including different FoVs and known/unknown view directions.
- Using pretraining with an extended version of the CVUSA dataset [14], we demonstrate that (a) large scale cross-view matching is feasible, and that (b) it yields 6.22% improvement in top-1 recall.
- We establish a new SOTA in cross-view matching on multiple datasets for a variety of test cases.

2. Related Work

Ground-view feature matching for geolocalization: Early experiments for image geolocalization leveraged handcrafted features as a foundation for retrieval in large datasets of GPS-tagged ground imagery [15]. Later, it was shown that modeling geolocation as a classification task could provide improved performance by discretizing the world into bins corresponding to regions of the planet and training a CNN to classify images accordingly [16]. Recent variants of these approaches have incorporated hierarchical modeling [17] semantic context [18] and transformer networks [19] to improve performance. However, these techniques rely on ground-view reference databases which are geographically sparse and biased toward densely populated areas and tourist sites making these approaches limited for global usage. In contrast, satellite imagery provides global coverage, making it a more attractive foundational data source for matching.

Cross-view matching for geolocalization: One of the first cross-view matching approaches leveraged handcrafted features for matching overhead and ground images [20]. The work was furthered in [14] which had two main contributions—introduction of a new aerial-ground image pair dataset and feature matching using deep neural networks. Early methods for cross-view matching applied pretrained deep networks for feature extraction in aerial imagery [21] but were limited due to a lack of domain-specific training and the large domain gap between aerial and ground images. Later works leveraged 2-branch CNNs which were cast in an embedding learning formulation between aerial and ground imagery [1, 22, 23]. These benefit from separate branches specialized for aerial and ground feature extraction, and often include shared parameters for feature aggregation [3]. Recently, transformer networks have provided a boost in performance using multi-head attention [6, 7].

Addressing the perspective gap: A major barrier for effective cross-view matching is the domain gap between aerial and ground images. Attempts to address the perspective gap usually seek to align aerial and ground features in image space, for example using polar transformation [2], conditional synthesis strategies based on semantic content [24], or generative adversarial learning [5, 25–27]. Transformer-based approaches rely on spatial attention mechanisms to learn correspondences between features in aerial and ground images [6, 7]. The state-of-the-art panoramic cross-view matching technique uses a geometric disentanglement approach, learning spatial and semantic correspondence through counterfactual learning [10].

Orientation-aware cross-view matching: Knowing orientation (e.g., ground level image is looking north) is a key factor in the cross-view matching process. Existing approaches incorporate orientation information to the matching process through learned UV-mapping [22] or dynamic similarity matching to calculate correlation in spatially aware features [4]. Alternatively, it is possible to learn a joint embedding of global and local aerial views for the corresponding limited FoV ground image [8].

Cross-view matching datasets: The largest existing cross-view matching dataset, CVUSA [14], contains over a million aerial and ground images. However, existing deep-learning based approaches use only a small subset of the dataset containing 44k image pairs. Other datasets have been collected, but are focused on city-scale applications and contain dense imagery for a handful of geographic regions [22, 23, 28]. It is common to find misalignment in aerial and ground images in existing datasets. Ground image locations are typically gathered from consumer-grade GPS (e.g., smartphone sensors) which are subject to noise. Small errors can be corrected by comparing image similarity between ground-view images and polar projected satellite images [25], but misalignment is difficult to overcome.

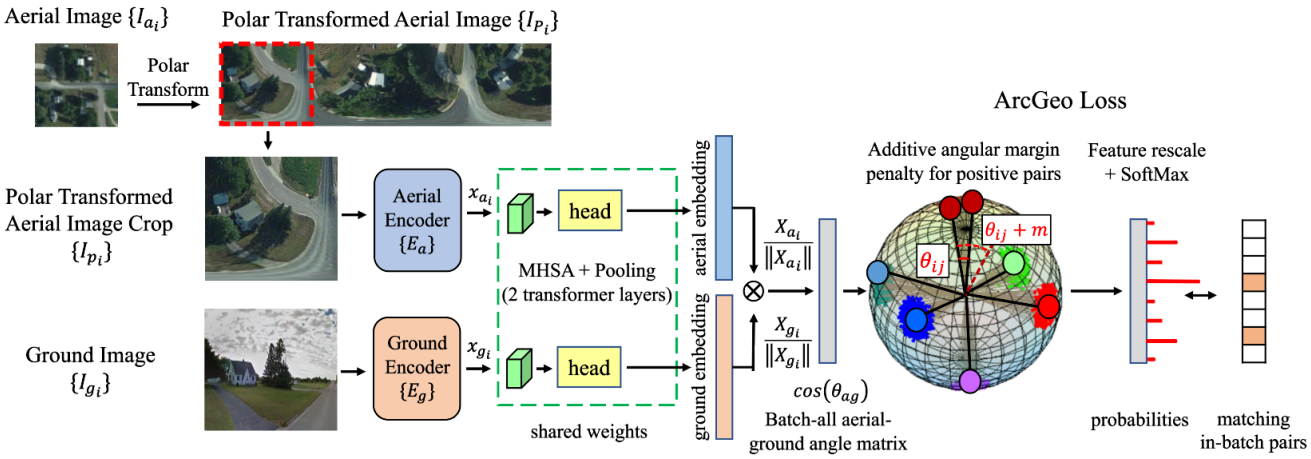

Figure 1. Schematic illustration of our aerial-ground image retrieval model. The two-branch model is shown on the left and ArcGeo loss on the right.

Optimization for cross-view retrieval networks: Negative mining is a popular approach for learning a precise local structure in embedding learning and is easily applied to triplet loss. Existing work shows that in-batch negative mining can improve top-1 recall of cross-view matching networks [29, 30] and can be further improved using global negative mining strategies [28, 31].

Cross-view matching with limited FoV: Comparatively fewer investigations have been made which account for differences in FoV. One technique for handling low FoV imagery fuses local and global aerial features and incorporates variable FoV data augmentation to learn a robust representation for low FoV matching [8] achieving a top-1 recall of 22.54% on the CVUSA dataset using cropped imagery corresponding to FoV = 90°. Still, the highest performance reported for FoV = 90° imagery [7] achieves top-1 recall of 30.12% in comparison with 94.08% for FoV = 360° [7], showing a significant gap in performance between panoramic and limited FoV imagery.

3. Methodology

We provide an overview of our approach and describe the cross-view matching problem formation in Section 3.1. Next, we describe the model architecture (Section 3.2), and our novel ArcGeo loss for learning robust cross-view embeddings (Section 3.3). Lastly, we describe our model pretraining and data augmentation strategy to obtain robust performance in low FoV settings (Section 3.4).

3.1. Problem Statement

We formulate the image geolocation task as a cross-view image retrieval problem. Given a set of ground-view query images $\{I_g\}$ and aerial-view reference images $\{I_a\}$ we define positive image pairs as aerial and ground images taken from the same location $\{I_{g_i}, I_{a_i}\}$, with all others being considered as negative pairs. Our objective is to learn a latent representation where embeddings for positive pairs are close, while embeddings for negative pairs are distant.

3.2. Model Overview

Figure 1 shows a diagram of our model architecture. We leverage a two-branch model with separate encoders for aerial and ground images (E_a and E_g respectively). Each encoder takes the corresponding input image and generates a set of aerial and ground features. Next, we apply a set of shared layers including a multi-head self-attention module (represented as two transformer layers), which models global correlation between input tokens. In this case, input tokens are deep features belonging to either the aerial or ground images. The network head consists of two linear layers with GeM pooling [32] to aggregate the features into a final embedding vector as described in Section 4.4. The model is optimized using our novel ArcGeo loss which includes an angular margin loss applied in a batch-all manner to optimize both the local and global embedding landscape. For more details of our loss function, see Section 3.3.

Our design has several benefits. Because aerial and ground embeddings are calculated independently, it is more efficient than approaches which evaluate correlations between feature maps to identify view direction [4]. The joint consideration of feature maps makes the generated embeddings unique to specific input combinations, and requires re-computation of aerial reference features for each query, thereby making it infeasable for large-scale image retrieval applications such as geolocalization. In addition, the proposed ArcGeo loss uses batch-all evaluation, which requires each aerial embedding in the batch to be compared to each ground embedding. If the aerial and ground embeddings were co-dependant, they would have to be extracted B^2

times per batch, where B is the batch size. While the feature maps extracted by the backbones in each calculation could be reused, the overhead of running the feature mixing B^2 times would still be significant, especially for the large batches required by ArcGeo loss. Therefore, recent cross-view matching studies [1, 3, 5, 6, 8, 22, 31] similar to ours consider the interdependent generation of aerial and ground embeddings to be more suitable for batch-all losses.

Our approach is agnostic to aerial image processing—it can be applied to retrieval strategies using polar transformed aerial images (*e.g.*, as used in [2,4,6]) or using aerial images directly (*e.g.*, [7, 8]). ArcGeo does not assume known orientation (*i.e.*, view direction) or FoV but can use this information if available to improve retrieval accuracy. This has allowed us to conduct extensive ablations comparing a variety of model architectures and data preparation strategies with other approaches which rely on specific inputs [8, 10].

3.3. ArcGeo Loss

One of the key differences of our approach from previous works is the use of ArcGeo loss, proposed in our study instead of batch-all triplet loss [33], which is typically used in image retrieval training. Triplet loss performs local optimization of the embedding space by considering one positive and one negative example. Batch-all formulation and negative mining seeks to improve local specificity by a search for hard triplets containing negative examples located nearby in the embedding space. Meanwhile, the global structure of the embedding space is optimized indirectly through triplet-based updates.

An alternative approach, performing direct optimization of the global embedding space structure, was suggested in the ArcFace paper [13], which revolutionized image retrieval for face recognition. In this approach the embedding space structure is approximated with learnable centroids assigned to each particular class, and Softmax based updates are applied to identify where the particular embedding should be pushed. To improve class separation, an angular margin term is added. Similar to the margin in triplet loss, this term ensures that the embeddings predicted by the model are located closer to the ground truth centroids than to the neighboring ones, improving local compactness of the produced embedding. In contrast to face recognition, however, in cross-view matching each location is often represented by a single aerial-ground pair. Therefore, during training the positions of the centroids, approximating the embedding space structure may experience a misalignment with the actual model output for given locations due to a very limited number of samples per each location. This misalignment leads to incorrect updates of model weights and may explain low performance of cross-view training with ArcFace loss (Table 4).

To address this problem, we propose ArcGeo loss, which approximates the global structure of the embedding space with examples provided in each batch using a batch-all formulation instead of relying on learnable cluster centers (see Figure 1). Then, following the standard ArcFace formulation, we add an angular margin to positive pairs, scale the angles, apply SoftMax (flattening all pairs in the batch), and compute cross-entropy loss using matching pairs as a set of positive labels:

$$L = -\frac{1}{N}\Sigma_N \log\left(\frac{e^{s\cos(\theta_{ii}+m)}}{e^{s\cos(\theta_{ii}+m)} + \Sigma_{N,j\neq i} e^{s\cos(\theta_{ij})}}\right)$$

where N is the batch size assuming N aerial and N ground images provided with the matching index corresponding to matching pair, θ_{ij} is the angle between corresponding aerial and ground embeddings in the batch, s is the scaling factor, and m is the angular margin penalty. To improve the numerical stability, we clip the cosine tensor and use the Li-ArcFace [34] formulation instead of the original ArcFace formulation [13]. We select $s = 20$ and $m = 0.5$ based on our experiments.

Even though ArcGeo loss uses a sparse approximation of the global structure of the embedding space, at sufficiently large batch size this approximation is sufficient to significantly outperform triplet loss and Arcface loss (Table 4).

3.4. Model Pretraining and Data Augmentation

Existing cross-view matching datasets are limited in geographic scope and scale. For example, the subset of CVUSA [14] that is typically used for comparative evaluation contains only 35,532 training image pairs and 8,884 validation image pairs. CVACT [22] is similarly sized but contains an additional 92,802 test images for larger evaluation experiments. We show that model pretraining and data augmentation can be used to ensure robustness in performance for large scale tests.

Current approaches focus on model architecture improvements to address gaps in performance [3, 6, 7, 30]. While these methods demonstrate improvement for panoramic cross-view matching, they show only incremental improvement for experiments with limited FoV. We show that the size of the training datasets is insufficient for training robust cross-view matching models. Our method uses a training dataset that is an order of magnitude larger than the smaller training subsets used previously. By training on larger data, we dramatically improve recall, especially for test cases with limited FoV (see Section 4.5).

Additionally, we observe that existing methods train separate models for each test-time FoV, where the training FoV is precisely aligned with the test FoV [1, 3, 4, 6, 8]. This training regime can yield models which perform well for a narrow range of test-time FoV but require FoV to be known during test time to obtain their expected performance. In

many practical settings, FoV is not precisely known (e.g. digital zoom can yield multiple FoV for a single type of sensor). Hence, we leverage additional FoV-based data augmentation where FoV is varied randomly during training. This, combined with proper pretraining yields a single model which can perform well in a variety of test-time FoV requiring no knowledge of FoV during test time.

4. Experiments

We demonstrate our proposed approach with a set of quantitative and qualitative experiments on several common cross-view matching benchmark datasets. The experimental results demonstrate the efficacy of various model configurations and ablation studies of feature extraction backbone, pretraining and training settings.

4.1. Datasets

We evaluate our method using two standard cross-view matching benchmarks, CVUSA [14] and CVACT [22]. These datasets each contain 35,532 training image pairs and 8,884 validation image pairs. We also report results for the CVACT full validation (test) set which contains an additional large test set of 92,802 image pairs. The ground images are cropped at the top and bottom to have similar appearance to CVUSA data, and the image size in these experiments is identical.

Additionally, we conduct an extensive pretraining experiment on the extended CVUSA-full dataset which contains 1.2M additional aerial-ground image pairs. In this extended version of the dataset, street-view images have a maximum FoV of 325 degrees compared to the 360-degree panoramas present in the standard CVUSA dataset. We select the subset of imagery from this dataset containing only pairs whose ground image was sourced from street-view style images which contains 413,740 image pairs. We exclude all images corresponding to CVUSA validation from CVUSA-full dataset. The aerial images are rotated and cropped and the ground images are cropped at the top and at the bottom to match the appearance of CVUSA data. We refer to this extended dataset as CVUSA-full. The dataset can be obtained via email request to the authors of the original CVUSA dataset [14]. For more detail, see Section 4.6.

4.2. Data Preparation

We define a set of data pre-processing steps we use in this work. Note that not every step will be used for training a given model.

Polar Transform: We follow the commonly adopted polar image transformation introduced by Shi *et al*. [2] which has been shown as an effective means for bridging the cross-view perspective gap.

Data Augmentation: We apply several standard data augmentation strategies for training our model including flip, rotation (by up to 20 degrees for ground and 360 for aerial images), up to 20% rescale, shift, cutout, perspective change as well as hue, saturation, value, and brightness variation.

Known/Unknown View Direction: We consider two cases for known and unknown view direction. When view direction is known, aerial images are south-aligned corresponding the seam of the panoramic ground image, resulting in a view aligned polar transform. When view direction is unknown, the ground images are rotated randomly along the azimuth direction yielding image pairs with no view direction alignment.

Image Size: Images are cropped and resized according to desired FoV (shown in Table 1 of the Supplementary Materials). For common configurations (e.g. FoV = 360° and FoV = 90°) we adopt image sizes from existing methods [8] to make our results comparable. In the case of unknown view direction and polar transform, the entire 896×224 aerial image is given as an input to the model regardless of the ground image FoV. We perform horizontal cropping for limited FoV training and evaluation, while preserving the vertical image size unless otherwise stated (see the Supplementary materials).

Negative Mining: Negative mining has been shown to be a critical part of many cross-view matching methods due to the large imbalance between positive and negative pairs. In many cases, there is only a single positive example per location making it difficult to estimate the boundary between positive and negative pairs. Our best performing model uses a 2-step negative mining approach which incrementally increases the difficulty of batches as training progresses. Our strategy provides image pairs which are easy to separate in the early stages of the training process, and gradually introduces more difficult negative samples to refine the learned embedding space.

4.3. Metrics

We consider top-k recall as our primary metric following existing work [4, 6–8, 28] in cross-view matching. For this metric we retrieve the K nearest aerial views as measured by cosine similarity of learned aerial and ground embeddings. The ground-view query image is considered correctly localized if the corresponding aerial image is within the set of top-k retrieved images. We adopt the common notation of r@K and evaluate our models for multiple values of K. For limited FoV cases we randomly crop a section of the corresponding panorama to use for evaluation. Reported metrics are an average of 10 evaluation runs to eliminate noise due to selection of random image crops.

We also consider mAR@5 (Mean Average Recall at 5), which is expressed as the following:

$$mAR@5 = \frac{1}{U}\Sigma_U \Sigma_{k=1}^{\min(n,5)} R(k)$$

where U is the number of images, $R(k)$ is recall at cutoff of k, and n is the number of predictions per image (the model predictions are sorted). This metric can be considered as weighted model performance for making a correct prediction within top-1 to top-5. Specifically, the metric assigns 1.0 if the model makes the correct prediction from the first attempt, 0.5 for the second attempt, and so on until 0.2 for top-5 prediction. If the model does not predict a correct image within top five results, the metric value for the image is zero. So, this metric can be interpreted as a weighted performance measure within top-5, putting an emphasis on prediction within fewer attempts. This has the benefit of robustly describing model performance for cases where the model correctly ranks imagery at position $k + 1$.

4.4. Feature Extraction

We conduct experiments with two popular feature extraction backbones, the CNN-based model ResNeXt-50 and a pure transformer model ViT-B, to demonstrate the applicability of our approach regardless of model architecture. We follow a typical image retrieval training format where batches of aerial and ground image pairs $\{I_{g_i}, I_{a_i}\}$ are passed to their respective encoders to obtain local features $\{x_{g_i}, x_{a_i}\}$. For ResNeXt-50 based models we add a multi-head self-attention module (MHSA) which serves to aggregate and attend to feature maps extracted from the two branches. For our transformer-based experiments we report results for two ViT-B pretraining configurations (BEiT-v2 [35] and DEiT-v3 [36]). For these we omit the shared feature aggregation and pooling layers and produce the embedding through modification and propagation of the class token, as is typically used in transformer-based image retrieval [37]. Details of feature extraction configurations are in the Appendix.

Table 1 shows results for several models trained using our ArcGeo loss with a standard training configuration. As expected, r@1 increases with increasing number of network parameters. Interestingly, we observe significant improvement in using a BEiT-v2-B backbone compared to DEiT-v3-B, which use the same ViT-B model and differ mainly in pretraining strategy. We argue that model architecture selection is less important than selection of proper loss function and pretraining strategy which we further demonstrate in ablations (Section 4.6). For computational reasons, most of our experiments use a ResNeXt-50 backbone.

4.5. Comparison with State-of-the-art Approaches

We compare our proposed approach with several existing methods including approaches [2, 4, 7, 8, 28]. Because existing work in limited FoV cross-view matching is limited, we also report performance for the panoramic test case. Though the focus of our work is not on improving performance for panoramic images, our improvements appear to apply universally, increasing performance over baseline approaches even for FoV = 360° input images.

CVUSA, FoV = 90°, known view-direction

Method	# Parameters	r@1 (%)	r@5 (%)	r@1% (%)
Ours (ResNeXt-50) †	62.77M	83.80	94.34	96.48
Ours (DEiT-v3-B) †	172.17M	87.40	95.83	99.55
Ours (BEiT-v2-B) †	172.17M	90.10	96.91	99.66
Ours (ResNeXt-50) †◊*	**62.77M**	**93.47**	**98.12**	**99.81**

Table 1. Quantitative results showing multiple model architectures for FoV = 90° test case on CVUSA with known view-direction. The † indicates models which use polar transformation. The ◊ indicates the use of the ASAM optimizer and global negative mining. The * indicates pretraining on the larger CVUSA-full dataset.

Localization for Panoramic Imagery: We evaluate three model configurations including versions which use a ResNeXt-50 CNN backbone as well as a BEiT-v2-B (ViT-B) backbone, demonstrating the flexibility of our approach to adapt to a variety of backbone feature extraction models. As shown in Table 3, when using a transformer network for our backbone, our approach provides higher top-k recall compared to baseline methods.

Meanwhile, performance using CNN-based ResNeXt-50 backbone achieves competitive performance using standard pretraining techniques. We suspect performance of our ResNeXt-50 based model to be limited due to the relatively limited scope of transfer learning available from standard ImageNet pretraining. Hence, we also report performance for an identical model which was pretrained on a larger set of cross-view image pairs from the CVUSA-full dataset. This model significantly outperforms baseline approaches and indicates that cross-view matching models may benefit from large-scale pretraining to help address the domain gap between aerial and ground images.

Note that our BEiT-v2-B does not leverage any specialized cross-view pretraining, ASAM optimizer, or global negative mining. These techniques improve performance for ResNeXt-50 but were omitted due to the increase in computational cost.

Localization for Limited FoV/Unknown View Direction: The main goal of the proposed approach is to improve performance for limited FoV test cases where view direction is unknown, as illustrated in Figure 2. We evaluate our approach alongside existing works which also report performance for limited FoV [1,3,4,6–8]. In some cases, the baseline techniques were not necessarily focused on low FoV matching so performance varies across methods.

Our model incorporates ArcGeo loss to create a robust embedding landscape well-suited for matching in low-FoV scenarios. We demonstrate this by evaluating a single instance of our model across multiple test FoVs, (Table 4). In contrast, existing approaches align train and test FoV (noted as "Matching" in Table 2, requiring FoV to be known during

CVUSA, unknown view-direction

Method	Pretraining	Training FoV	Test FoV = 180° r@1 (%)	r@5 (%)	r@10 (%)	Test FoV = 90° r@1 (%)	r@5 (%)	r@10 (%)	Test FoV = 70° r@1 (%)	r@5 (%)	r@10 (%)
CVM-Net [1]	ImageNet-1k	matching, fixed	7.38	22.51	32.63	2.76	10.11	16.74	2.62	9.30	15.06
CVFT [3]	ImageNet-1k	matching, fixed	8.10	24.25	34.47	4.80	14.84	23.18	3.79	12.44	19.33
DSM [4] †	ImageNet-1k	matching, fixed	48.53	68.47	75.63	16.19	31.44	39.85	8.78	19.90	27.30
GAL [[8]]	ImageNet-1k	matching, w/ aug	48.91	69.87	78.50	22.54	44.36	54.17	15.20	32.86	42.06
L2LTR [6] †	ImageNet-1k	matching, fixed	56.69	80.86	87.75	26.92	50.49	60.41	13.95	33.07	43.86
TransGeo [7]	ImageNet-1k	N/A	58.22	81.33	87.66	30.12	54.18	63.96	-	-	-
Ours (ResNeXt-50) †	ImageNet-1k, ‡	90°, fixed	57.56	83.24	89.76	37.22	64.83	74.98	30.86	57.51	68.54
Ours (ResNeXt-50) †◊	**ImageNet-1k, ‡**	**90°, fixed**	**63.68**	**85.39**	**90.61**	**44.18**	**70.33**	**78.84**	**37.71**	**64.41**	**73.64**
Ours (ResNeXt-50) †◊	**CVUSA-full**	**90°, fixed**	**83.63**	**96.20**	**97.85**	**66.83**	**88.12**	**92.52**	**59.89**	**83.63**	**89.44**

Table 2. Quantitative results for limited FoV test case on CVUSA. The † symbol indicates models which leverage polar transformation. The ‡ symbol indicates models which were pretrained using self-supervised learning on ImageNet-1k / IG-1B [38] for ResNeXt-50 and ImageNet-21k for BEiT-v2. The ◊ symbol indicates our models which were trained with global negative mining.

CVUSA, FoV = 360°, known view-direction

Method	mAR@5	r@1 (%)	r@5 (%)	r@1% (%)
SAFA [5] †	-	89.84	96.93	99.64
DSM [4] †	-	91.93	97.50	99.67
CDE [8] †	-	92.56	97.55	99.57
L2LTR [6] †	-	94.05	98.27	99.67
TransGeo [7] ◊	-	94.08	98.36	99.77
SEH [10] †	-	95.11	98.45	99.78
GeoDTR [14] †	-	95.43	98.86	99.86
Ours (ResNeXt-50) †	96.07	94.32	98.51	99.80
Ours (BEiT-v2-B) †	**97.26**	**96.06**	**98.89**	**99.88**
Ours (ResNeXt-50) †◊*	**98.33**	**97.47**	**99.48**	99.67

Table 3. Comparison of our approach with baseline methods on CVUSA dataset. The † symbol indicates models which use polar transformation. The ◊ indicates global negative mining. The * indicates pretraining on the larger CVUSA-full dataset.

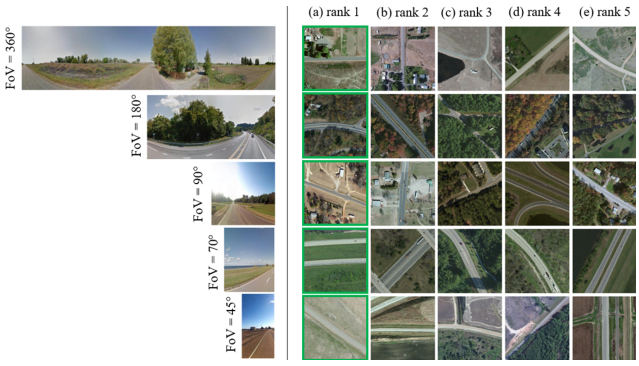

Figure 2. Visualization of query images under different test-time FoV (left) and top-5 retrieved aerial images predicted from our model. Ground truth aerial pairs are marked in green.

test time to achieve the reported performance. In practical applications, FoV may not be known and requires complex approaches to estimate [39].

We improve top-1 recall from 30.12% to 44.18% using our ResNeXt-50 based model for FoV = 90° case using semi-weakly supervised ImageNet pretraining. Notably, we achieve a massive boost in top-1 recall to 66.83% by pretraining on additional data in the CVUSA-full dataset.

4.6. Ablation Studies

Effects of ArcGeo Loss: We conducted a set of experiments to measure the effects of the proposed ArcGeo loss compared to standard triplet loss used by existing approaches. We trained models for known and unknown view direction and evaluated performance on several limited FoV test cases (Table 5). We observe a significant improvement in performance using ArcGeo loss, resulting in an 8.87% improvement in r@1 for FoV = 90° when view direction is known and a 15.59% improvement in r@1 for FoV = 90° when view direction is known. This demonstrates the effectiveness of our ArcGeo loss to be applied broadly for improved cross-view matching with limited FoV, apart from other performance increases obtained from negative mining and data augmentation.

Effects of large-scale pretraining: Existing techniques for cross-view matching typically use the CVUSA dataset, specifically, the subset of ∼44k image pairs which were proposed for use in early cross-view matching methods [1, 14]. Since then, significant advances have been made with modern techniques achieving impressive results with $>$ 94% r@1 for FoV = 360° test case. However, performance on limited FoV quickly declines for these models, yielding only ∼30% r@1 or less for FoV = 90°. We argue that one possible reason for such a drastic decrease in performance on low FoV imagery is a lack of diverse image features in the small CVUSA training dataset. To maintain high accuracy for limited FoV test cases, models must learn robust representations, matching as many features as possible, not just the most discriminative set which would be suitable for FoV = 360° matching. Without sufficient diversity in the training data, the robustness of the learned embedding is limited, potentially manifesting as rapidly decreasing performance with lowered FoV.

We designed a set of experiments to measure the effects of pretraining with a larger dataset, specifically for performance improvement in low FoV. We pretrain our cross-view matching network using CVUSA-full which contains an ad-

Method	View Direction	Loss Function	Test FoV = 180°			Test FoV = 90°			Test FoV = 70°		
			r@1 (%)	r@5 (%)	r@10 (%)	r@1 (%)	r@5 (%)	r@10 (%)	r@1 (%)	r@5 (%)	r@10 (%)
Ours (ResNeXt-50)	Unknown	Triplet	44.31	72.73	81.85	28.59	55.92	67.32	23.82	49.91	61.71
Ours (ResNeXt-50)	Unknown	Arcface	8.52	15.32	19.69	8.02	13.96	17.87	7.87	13.60	17.38
Ours (ResNeXt-50)	**Unknown**	**ArcGeo**	**63.68**	**85.39**	**90.61**	**44.18**	**70.33**	**78.84**	**37.71**	**64.41**	**73.64**
Ours (ResNeXt-50)	Known	Triplet	84.99	95.56	97.43	78.19	92.10	95.05	70.86	87.85	91.98
Ours (ResNeXt-50)	**Known**	**ArcGeo**	**93.72**	**98.39**	**99.03**	**87.06**	**95.34**	**97.09**	**81.08**	**92.17**	**94.78**

Table 4. Quantitative results comparing ArcGeo loss with conventional triplet loss for limited FoV test case on CVUSA with known view-direction. Each model was trained using polar transformation with FoV=90°, and global negative mining.

Method	View Direction	Pretraining	Test FoV = 180°			Test FoV = 90°			Test FoV = 70°		
			r@1 (%)	r@5 (%)	r@10 (%)	r@1 (%)	r@5 (%)	r@10 (%)	r@1 (%)	r@5 (%)	r@10 (%)
Ours (ResNeXt-50)	Unknown	ImageNet-1k ‡	63.68	85.39	90.61	44.18	70.33	78.84	37.71	64.41	73.64
Ours (ResNeXt-50)	**Unknown**	**CVUSA-full**	**83.63**	**96.20**	**97.85**	**66.83**	**88.12**	**92.52**	**59.89**	**83.63**	**89.44**
Ours (ResNeXt-50)	Known	ImageNet-1k ‡	93.72	98.39	99.03	87.06	95.34	97.09	81.08	92.17	94.78
Ours (ResNeXt-50)	**Known**	**CVUSA-full**	**97.44**	**99.54**	**99.66**	**93.47**	**98.12**	**98.85**	**89.42**	**96.48**	**97.76**

Table 5. Quantitative results for limited FoV test case on CVUSA using various pretraining strategies. Each of model was trained using polar transformation with FoV=90°, and global negative mining. The ‡ symbol indicates models which were pretrained using self-supervised learning on IG-1B [46] for ResNeXt-50.

ditional 1.2M aerial-ground image pairs. We select the subset of 413,740 pairs belonging to street view imagery similar to the CVUSA dataset. We perform basic image preprocessing including vertical cropping to align the imagery to be similar to CVUSA. Training is performed identically to models shown in Table 5.

Pretraining on CVUSA-full was found to be beneficial for all models, including known and unknown view-direction cases. Table 6 shows how this pretraining strategy increases r@1, with a larger effect on low FoV test case. For example, performance for FoV = 90° test case was increased from 87.06% r@1 to 93.49 r@1 (known view-direction). Similarly, for unknown view-direction performance FoV = 90° was increased from 44.18% to 66.83%.

Performance on CVACT: To demonstrate that our approach is applicable for multiple data sources we performed additional experiments on the CVACT dataset. We use BEiT-v2 training procedure identical to our CVUSA experiments with test FoV = 360° and known view direction. Table 6 shows a comparison of our results with several baseline approaches.

Similar to improvements observed on CVUSA, our approach provides a massive improvement of the performance on both CVACT_val and CVACT_test sets in comparison to previously reported values. The difference is especially apparent in the larger evaluation set yielding nearly 15% improvement over baseline models.

Method	CVACT_val Known view-direction Test FoV = 360°			CVACT_test Known view-direction Test FoV = 360°		
	r@1 (%)	r@5 (%)	r@10 (%)	r@1 (%)	r@5 (%)	r@10 (%)
CVM-Net [1]	20.15	45.00	56.87	-	-	-
Liu *et al.* [22]	46.96	68.28	75.48	-	-	-
CVFT [3]	61.05	81.33	86.52	26.12	45.33	53.80
SAFA [2]	81.03	92.8	94.84	-	-	-
DSM [4] †	82.49	92.44	93.99	35.63	60.07	69.10
L2TR [6] †	84.89	94.59	95.96	60.72	85.85	89.88
TransGeo [7]	84.95	94.14	95.78	-	-	-
SHE [9] †	84.75	93.97	95.46	-	-	-
GeoDTR [10] †	86.21	95.44	96.72	64.52	88.59	91.96
Ours (BEiT-v2) †	**90.90**	**95.84**	**96.77**	**79.88**	**90.97**	**92.94**

Table 6. The performance on CVACT dataset using test FoV = 360° with known view direction. The † symbol indicates models which use polar transformation

5. Conclusion

To address the challenges of cross-view matching under limited FoV, we propose a novel ArcGeo loss and large-scale pretraining to improve model robustness. We demonstrate that a relatively simple model is capable of state-of-the-art performance and can out-perform more complicated models when ArcGeo loss is applied. Our pretraining experiments with CVUSA-full indicate that existing approaches may be limited by the relatively small dataset size of CVUSA and CVACT. We also identify critical training techniques for achieving high accuracy in limited FoV test cases (e.g. ASAM optimizer and global negative mining). The result is a single model which is capable of accurate cross-view image retrieval across a wide range of test FoV.

6. Acknowledgement

This research is supported by the National Geospatial Intelligence Agency (NGA) via Contract No. HM047621C0004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NGA or the U.S. Government. Approved for public release, **NGA-U-2023-00446**.

References

[1] R. M. Nguyen S. Hu, M. Feng and G. Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 4, 6, 7, 8

[2] Y. Shi, L. Liu, X. Yu, and H. Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Neurips*, 2019. 1, 2, 4, 5, 6, 8

[3] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li. Optimal feature transport for cross-view image geo-localization. *arXiv:1907.05021*, 2019. 1, 2, 4, 6, 7, 8

[4] Y. Shi, X. Yu, D. Campbell, and H. Li. Where am i looking at? joint location and orientation estimation by cross-view matching. *arXiv:2005.03860*, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[5] A. Toker, Q. Zhou, M. Maximov, and L. Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. *arXiv preprint arXiv:2103.06818*, 2021. 1, 2, 4, 7

[6] H. Yang, X. Lu, and Y. Zhu. Cross-view geo-localization with layer-to-layer transformer. *Neurips*, 2021. 1, 2, 4, 5, 6, 7, 8

[7] S. Zhu, M. Shah, and C. Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. *arXiv:2204.00097*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[8] R. Rodrigues and T. M. Global assists local: Effective aerial representations for field of view constrained image geo-localization. *WACV*, 2022. 1, 2, 3, 4, 5, 6, 7

[9] Y. Guo, M. Choi, K. Li, and F. Boussaid. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE Transactions on Image Processing*, 2022. 1, 2, 8

[10] X. Zhang, X. Li, W. Sultani, Y. Zhou, and S. Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. *arxiv pre-print, arXiv:2212.04074*, 2022. 1, 2, 4, 7, 8

[11] Google maps api. https://developers.google.com/maps/documentation/maps-static. 1

[12] Bing maps api. https://learn.microsoft.com/en-us/bingmaps/rest-services. 1

[13] J. Deng, J. Guo, J. J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. Recognition, arcface: Additive angular margin loss for deep face. *arXiv:1801.07698*, 2018. 2, 4

[14] R. Souvenir S. Workman and N. Jacobs. Wide-area image geolocalization with aerial reference imagery. *IEEE International Conference on Computer Vision*, 2018. 2, 4, 5, 7

[15] J. Hays and A. Efros. Im2gps: estimating geographic information from a single image. *CVPR*, 2018. 2

[16] T. Weyand, I. Kostrikov, and J. Philbin. Planet - photo geolocation with convolutional neural networks. *arXiv:1602.05314*, 2016. 2

[17] E. Muller-Budack, K. Pustu-Iren, and R. Ewerth. Geolocation estimation of photos using a hierarchical model and scene classification. *ECCV*, 2018. 2

[18] J. Theiner, E. Muller-Budack, and R. Ewerth. Interpretable semantic photo geolocation. *WACV*, 2022. 2

[19] S. Pramanick, E. Nowara, J. Gleason, C. Castillo, and R. Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. *ECCV*, 2022. 2

[20] T.-Y. Lin and S. Belongie. Cross-view image geolocalization. *CVPR*, 2013. 2

[21] S. Workman and N. Jacobs. On the location dependence of convolutional neural network features. *CVPR*, 2015. 2

[22] L. Liu and H. Li. Lending orientation to neural networks for cross-view geo-localization. *CVPR*, 2019. 2, 4, 5, 8

[23] N. Vo and J. Hays. Localizing and orienting street views using overhead imagery. *ECCV*, 2016. 2

[24] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs. Predicting ground-level scene layout from aerial imagery. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 2

[25] Y. Shi, D. Campbell, X. Yu, and L. Hongdong. Geometry-guided street-view panorama synthesis from satellite imagery. *arXiv:2103.01623*, 2020. 2

[26] K. Regmi and M. Shah. Bridging the domain gap for ground-to-aerial image matching. *ICCV*, 2019. 2

[27] K. Regmi and A. Borji. Cross-view image synthesis using conditional gans. *CVPR*, 2018. 2

[28] S. Zhu, T. Yang, and C. Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. *CVPR*, 2021. 2, 3, 5, 6

[29] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting. *ICCV*, 2019. 3

[30] B. Sun, C. Chen, Y. Zhu, and J. Jiang. Geocapsnet: Aerial to ground view image geo-localization using capsule networks. *arXiv preprint arXiv:1904.06281*, 2019. 3, 4

[31] S. Zhu, T. Yang, and C. Chen. Revisiting street-to-aerial view image geo-localization. *WACV*, 2020. 3, 4

[32] F. Radenović, G. Tolias, and O. Chum. Fine-tuning cnn image retrieval with no human annotation. *arXiv:1711.02512*, 2017. 3

[33] E. Hoffer and N. Ailon. Deep metric learning using triplet network. *arXiv:1412.6622*, 2014. 4

[34] X. Li, F. Wang, Q. Hu, and C. Leng. Airface: Lightweight and efficient model for face recognition. *arXiv:1907.12256*, 2019. 4

[35] Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv:2208.06366*, 2022. 6

[36] H. Touvron, M. Cord, and H. Jégou. Deit iii: Revenge of the vit. *arXiv:2204.07118*, 2022. 6

[37] A. Dosovitskiy, L. Beyer, and A. Kolesnikov. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*, 2020. 6

[38] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv:1905.00546*, 2019. 7

[39] O. Bogdan, V. Eckstein, F. Rameau, and J.-C. Bazin. Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. *CVMP*, 2018. 7