

Multi-view Drone-based Geo-localization via Style and Spatial Alignment

Siyi Hu

siyi.hu@monash.edu

Faculty of Information and Technology, Monash
University
Melbourne, VIC

Xiaojun Chang

xiaojun.chang@monash.edu

Faculty of Information and Technology, Monash
University
Melbourne, VIC

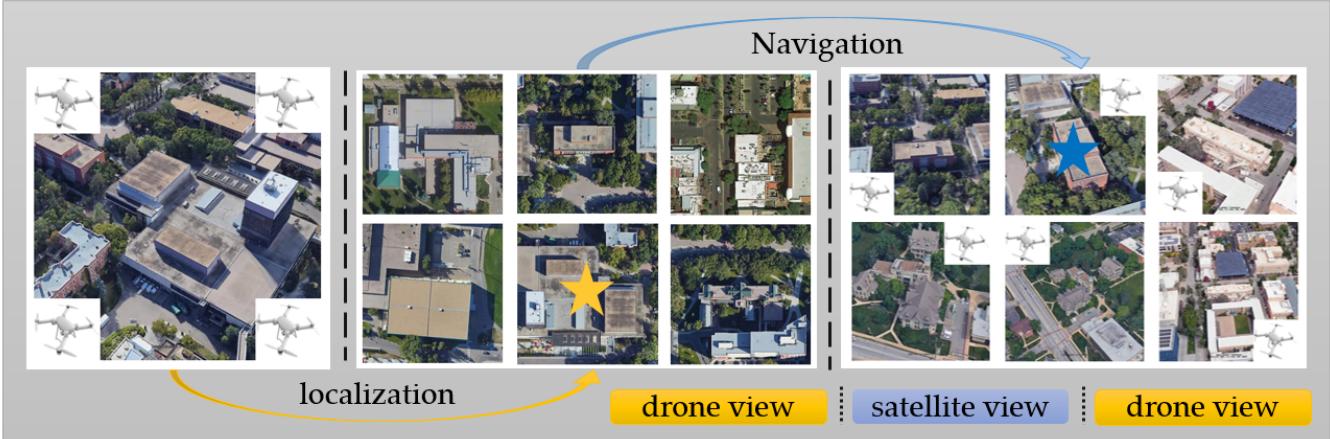


Figure 1: Drone-based geo-localization and navigation. Stars indicate the correct match.

ABSTRACT

In this paper, we focus on the task of multi-view multi-source geo-localization, which serves as an important auxiliary method of GPS positioning by matching drone-view image and satellite-view image with pre-annotated GPS tag. To solve this problem, most existing methods adopt metric loss with an weighted classification block to force the generation of common feature space shared by different view points and view sources. However, these methods fail to pay sufficient attention to spatial information (especially viewpoint variances). To address this drawback, we propose an elegant orientation-based method to align the patterns and introduce a new branch to extract aligned partial feature. Moreover, we provide a style alignment strategy to reduce the variance in image style and enhance the feature unification. To demonstrate the performance of the proposed approach, we conduct extensive experiments on the large-scale benchmark dataset. The experimental results confirm the superiority of the proposed approach compared to state-of-the-art alternatives.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Seattle '20, October 12–16, 2020, Seattle, USA

© 2018 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

CCS CONCEPTS

- Computing methodologies → Artificial intelligence; • Computer vision tasks → Visual content-based indexing and retrieval.

KEYWORDS

deep learning, geo-localization, image retrieval, drone navigation

ACM Reference Format:

Siyi Hu and Xiaojun Chang. 2018. Multi-view Drone-based Geo-localization via Style and Spatial Alignment. In *Seattle '20: ACM MULTIMEDIA, October 12–16, 2020, Seattle, USA*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/1122445.1122456>

1 INTRODUCTION

Cross-view geo-localization has attracted increasing attention in the past few years [20] [34] [33] [31] [36] [11] [22] [23] [28] [33] [18] [2] [14] [8] [15] [17] [32] [41] [3]. This task aims at localizing the target using only images with pre-annotated GPS tags. Given a query image, we can match the paired satellite-view images and use the GPS tag to determine the location of the user (ground-view image). The cross-view image based geo-localization task shows us the offline localization without GPS information is possible when we are able to match images from different views.

Existing work on this task has followed the traditional approach to supervised deep learning methods [20] [34] [33] [31] [36]. The main purpose of these work is to mine the shared features between ground-view and satellite-view images. In this way, the geo-localization task can be defined as a binary classification problem.

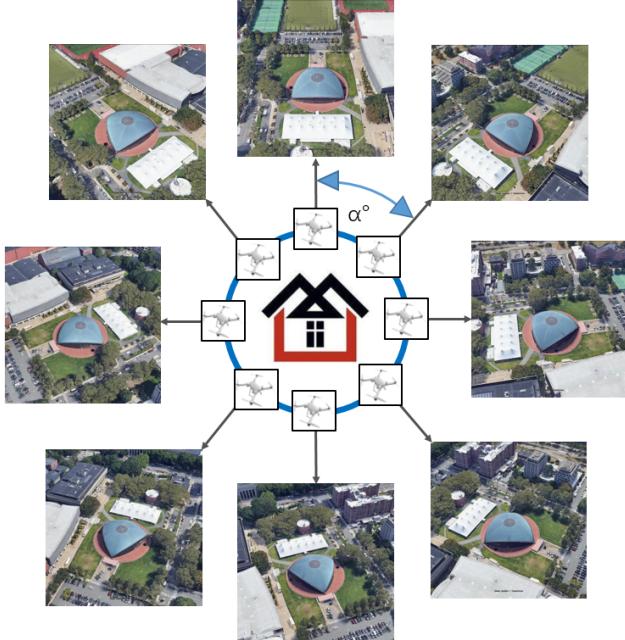


Figure 2: Drone-based geo-localization is a multi-view task. When flying around the building, the drone-view camera can capture rich information about the target, including scale and viewpoint variance.

Therefore, Triplet loss [10] and Siamese architecture [4] has been widely used to handle this task [11] [22] [23] [28]. Based on this, a large number of attention mechanisms have been proposed to improve the feature alignment from different views [23] [28]. These approaches did not perform well on this task. The main reason for this is obvious: it is difficult even for a human to find the correct match between a given query ground-view image and a target satellite-view image. Some researchers has used orientation information to further improve the model performance[22], however, there is still a gap between these models and real-world use.

In recent years, with the fast development of the map tools like Google Earth and functions provided by Google Maps API [5] [6], multi-view multi-source images with rich geo-information have become available for online collection. Moreover, drone based tasks are becoming more and more important and have come to play the key roles in areas such as agriculture, aerial photography, navigation, event detection and accurate delivery. Drone-view-based geo-localization tasks such as navigation and target localization are also gaining more attention.

With the release of drone-based multi-view image dataset named University-1652[37], the geo-localization task with rich spatial information towards higher accuracy on image-based geo-localization task has become possible.

With different platforms, viewpoints and increase amount of multi-source multi-view data, drone-view based geo-localization tasks is no longer a binary classification problem. The feature extraction can be made more robust and cover more scenarios, which

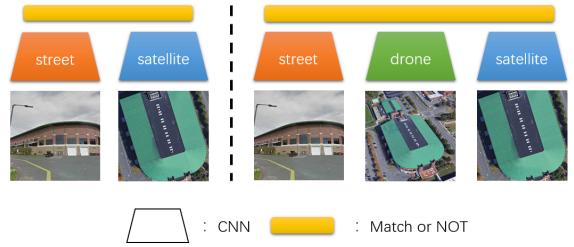


Figure 3: Illustration of the existing geo-localization pipeline (left) and ours (right). Matching between a ground-view image and satellite-view image is difficult even for human. However, drone-view and satellite-view image share more common features with only viewpoint-caused occlusion and style variance.

is more practical and valuable. At the same time, ideas and methods from other image retrieval task like person re-identification can be adopted or learned from.

The implementation of the vanilla method on University-1652[37] shows the robustness on sub-tasks and other small released dataset [24] [25] [26] compared to existing work such as CVMNet [11], Orientation[22] and other main benchmarks. However, the vanilla method of multi-source multi-view task suffers from low ability to extract spatial information caused by viewpoints variance. Besides, variance in image style including illumination and fuzziness prejudice the feature unification in both the training and testing stages.

To resolve these problems, we adopt three strategies that significantly improve the drone-based geo-localization performance:

- To handle the variance in the image style, we provide a style alignment strategy to transform the raw image, which helps to enhance the feature unification.
- To help CNN capture spatial information about the target with its surroundings, we adopt an orientation based method to align the part feature with a novel crop method.
- To enhance the feature extraction, we provide a series of partition strategies to extract partial features. Moreover, we analyze the factor of improvement using different partition strategies .

By applying three strategies, we achieve significant performance improvement compared to the vanilla method, which is a large step towards real-world use.

2 RELATED WORK

2.1 Geo-localization Datasets

To handle the image based geo-localization problem , several datasets have been built including [20], CVUSA [36] and CVACT [22]. [20] was the first well-known cross-view image dataset containing 78k image pairs from two views (i.e. 45° bird view and ground view). Later, CVUSA was released to study the problem of matching the panoramic ground-view image and satellite-view image. CVUSA made the first attempt to conduct user localization when Global



Figure 4: Result of style alignment on satellite-view images from University-1652. The left side presents raw images, while the right side presents images following style alignment. As we can see, the style of raw images with orange dotted frames (left, third row) are warmer than the second row. The red color scale channel is then uniformed to be reduced on amplitude and the style becomes cooler (right, third row). The results show that our method is robust on different styles.

Positioning System (GPS) is not available and serves as an auxiliary localization method. The main difference between CVUSA and [20] is that the former focus on localizing the user. CVACT is another dataset that differs slightly from CVUSA, as CVACT provides user orientation for the ground-view image, which can serve as additional information for better localization performance.

However, these datasets are all paired, which limits the geo-localization problem to a binary classification problem. Worse yet, the viewpoints is fixed in these datasets. In addition, panoramic ground-view image cannot be easily obtained. These disadvantages make it difficult to progress in paired image geo-localization tasks as well as to put them into practical use.

To relieve this limitation, the first drone-based multi-view multi-source image dataset, which was released only recently, is named University-1652 [37]. This dataset has three main characteristics, as follows:

Multi-source University-1652 contains data from three different platforms, namely, satellites, drones and phone cameras. To the best of our knowledge, University-1652 is the first geo-localization dataset to contain drone-view images.

Multi-view University-1652 contains data from different viewpoints. The ground-view images are collected from different facets of the target buildings. In addition, the synthetic drone-view images capture the target building from various distances and orientations.

More images per class Unlike the existing datasets that provide image pairs, University-1652 contains 71.64 images per location on average. During training, the increase amount of multi-source multi-view data can help the model to understand the target structure as well as learn the viewpoint-invariant features.

Compared to CVUSA and CVACT, University-1652 focuses on the relation between images from different sources and different views. Moreover, the task aims at localizing the target in the image (which) rather than localizing the user (where).

Another contribution of the University-1652 dataset is that it can handle the challenges in real-world drone image collection contexts, considering both the high cost and the privacy and safety issues by using automatically collected synthetic drone-view images. Nevertheless, there is a domain gap between real drone-view image and synthetic drone-view images; the author proves that real drone

view images can also work well on models trained by synthetic images [37].

2.2 Geo-localization Method

Most previous works treat the geo-localization as an image retrieval problem [20] [34] [33] [31] [36] [11] [22] [23] [28]. The key to geo-localization is learning the view-point invariant representation, which aims to bridge the gap between images of different views. Following the development of the deeply learned model, convolutional neural networks (CNNs) have been widely applied to extract the visual features [29] [9] [27] [12] [35].

One line of works focuses on metric learning and builds a shared space for images collected from different platforms. Workman et al. show that the classification CNN pre-trained on the Place dataset [39] can be very discriminative by itself without explicitly fine-tuning [34]. The contrastive loss, pulling the distance between positive pairs, could further improve the geo-localization results [20]. Recently, Liu et al. propose Stochastic Attraction and Repulsion Embedding (SARE) loss, minimizing the KL divergence between the learned and the actual distributions [23].

Another line of works focus on the spatial misalignment problem in the ground-to-aerial matching. Vo et al. evaluate different network structures and propose an orientation regression loss to train an orientation-aware network [33]. Zhai et al. utilize the semantic segmentation map to improve the semantic alignment [36], while Hu et al. insert the NetVLAD layer [1] to extract the discriminative features [11]. Furthermore, Liu et al. propose a Siamese Network to explicitly involve the spatial cues, i.e., orientation maps, into the training [22]. Similarly, Shi et al. propose a spatial-aware layer to further improve the localization performance[28]

Following the release of University-1652, since each location has a number of training data points from different views, the model can be trained using a classification CNN with regular cross-entropy loss. The author of university-1652 provides a novel baseline using instance loss to extract the common features from different views and sources. In this way, the viewpoint-invariant feature can be learned in a robust method.

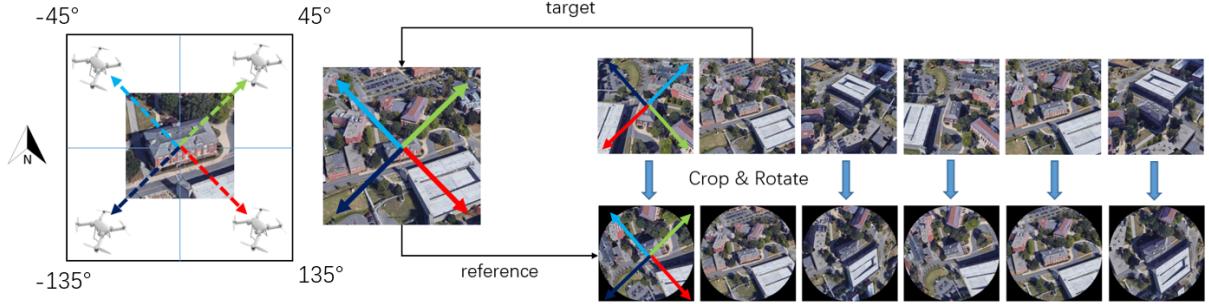


Figure 5: Illustration of ‘crop and rotate’ image transform method. Picking a raw drone-view image from the first row as a reference, we first crop all images into circles and rotate the other images according to the orientation between itself and the reference image. To ensure compatibility with satellite-view images, we usually adopt the orientation of the satellite as the reference in practice. Colored arrows indicate the orientation reference of the chosen image. After the ‘crop and rotate’ method, the orientation of other images are aligned. Moreover, the second-row images are somehow more ‘similar’ to each other than those in the first row, as the surroundings of the target building stay in the same part of the image (for example, the white building with rectangular roof is always in the lower right corner in all six images in the second row).

3 METHODS

3.1 Style Alignment

Image-based geo-localization tasks aims at find a robust way of representing the features shared by images of a same place. The learned feature should be invariant despite differences in the view-points or sources. Thus an obvious method is to minimize the style invariance both inter-domain and intra-domain.

University-1652 provides us 1652 buildings with various viewpoints and view sources. As the drone-view images are synthesized from the virtual drone engine of Google Earth , the style of drone-view images do not vary significantly. However, the style does vary substantially in satellite-view images. This style variance comes from 1) the satellite-view images being captured in different seasons. (e.g., in winter the image style will be cooler and darker, while in summer and autumn the map style can be warmer and brighter); 2) the satellite-view images being synthesized from multi-color layers with preprocessing.

Existing methods provide us with several kinds of style transfer or style uniform solutions [7] [16] [19] [21] [40] [13]. However, these approaches are all deep learning-based and therefore require a large amount of training data. In our case, there are only 1652 satellite-view images of different university buildings without annotated tags showing which style or class each belongs to. The unsupervised method also does not fit well in this case, as the content of satellite-view images varies widely and the features from the pre-trained model cannot be used directly used without fine-tuning.

Hence, in order to force the satellite-view images to be in a uniform style, we provide a simple color scale-based method that uniforms the image style in a statistical way. Each image has 4 dimension of color scale including: ‘S’;‘R’;‘G’;‘B’. The ‘S’ scale can also be regard as the light scale, as this scale is the main factor in deciding whether a picture is light bias or dark bias. ‘R’;‘G’;‘B’ represent the red scale, green scale and blue scale, which can decide whether an image is warmer or colder. In addition, some images may have color distortion or color cast caused by an unbalanced color

scale. Our style alignment method is simple and efficient for style uniformity on a small-scale dataset without sufficient annotation.

We first compute the mean value of four color scales and record them as the channel mean.

$$S_{i,j} = \frac{1}{3} \sum_{i,j \in p} R_{i,j} + G_{i,j} + B_{i,j}$$

$$S_{cm} = \frac{1}{p} \sum_{i,j \in p} S_{i,j}$$

$$R, G, B_{cm} = \frac{1}{p} \sum_{i,j \in p} R, G, B_{i,j}$$

Here ‘cm’ stands for channel mean, while ‘p’ stands for pixels of the whole image. We then compute the color bias of the different scales and then rescale value of all channels.

$$R, G, B_{bias} = R, G, B_{cm} - S_{cm}$$

$$scale = \sum_{i,j \in p} S_{i,j} / S_{ave}$$

We rescale the RGB channel of each pixel by rescale value and uniform the color scale by color bias. Considering that the color scale ranges from 0 to 255, we apply a clip method to prevent color distortion .

$$R, G, B_{uni} = R, G, B_{raw} \times scale \times (1 + R, G, B_{bias} / S_{ave})$$

$$R, G, B_{final} = F_{clip}(R, G, B_{uni})$$

We show the result of style alignment in figure 4. In our experiments, we will see the performance improvement obtained by using style alignment strategy. Moreover, our style alignment strategy can be applied to single test image with no prerequisites.

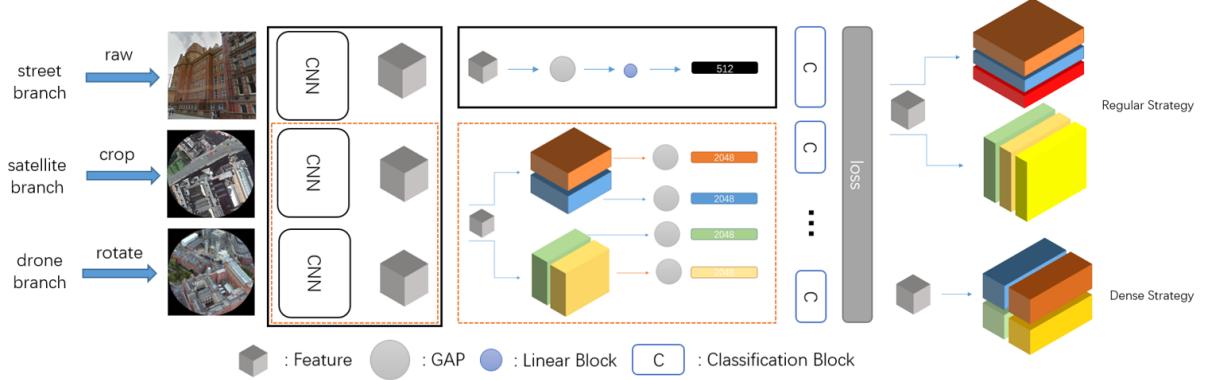


Figure 6: Illustration of our pipelines. Three different sources of images are fed into the model with different data preprocessing strategies. The gray cube represents a feature after the pooling layer of the CNN backbone. The features of satellite branch and drone branch will be fed into the partial feature model (orange dotted frame) followed by a partition strategy. Features of all three branches will be fed into the global feature model (black line frame). The weight of classification block is shared by the global feature model to generate a common space. In practice, considering that there is only one satellite view image per building, we share the weight of the CNN for the satellite branch and drone branch. We show the two partition strategies on the right side of the pipelines.

3.2 Feature Alignment with Partial Feature Extraction

University-1652 provides us with a vanilla multi-branch model that uses instance loss to guide the CNN in learning the common feature shared by different views. The vanilla model has two main disadvantages: 1) it only focuses on global features and 2) it ignores the spatial misalignment according to the variant viewpoint. It can only tell whether the given image contains target building with corresponding surroundings; however, it is not able to distinguish between the relative locations among these buildings due to viewpoint variance.

A great model should capture the information of relative locations among the target building and its surroundings. Luckily, the drone-view image from University-1652 are collected following a constant angle step of about 20°. Thus, the most efficient way to get the features aligned is to rotate the raw image. Moreover, the satellite-view images are all captured in the same orientation, which means the feature between the satellite-view and drone-view images can also be aligned by simply rotating a certain number of degree according to the image index.

In order to rotate the images smoothly, we crop the raw image in a novel way. A typical description of a target building's surroundings will resemble e.g. 'within a hundred meters, there is a hospital to the east and a school to the west.' Thus, we only consider information within a circle. We then mask the raw image with a circle such that the radius equals half of the side length.

Once the raw images are cropped to a circle shape, the rotation of each image is smooth and easy to implement. Here, we provide a group of raw images alongside a group of images after feature alignment in Figure 5. It is clear that after the 'crop and rotate' procedure, the feature of the drone-view image is aligned. It is worth mentioning that, in real-world test stage, although there is no index to show which angle we should take to rotate the captured

image, the orientation can be easily obtained without any GPS information. Instead, a compass is enough.

Once the feature alignment is complete, the aligned drone-view images boost the feature extraction performance of the CNN. As we can see in the experiment section, even with the vanilla model, the performance is significantly improved with no modifications.

There is another pipeline of aligned feature extraction based on aligned partial features called part-based feature CNN [30]. This method usually conducts uniform partitions on the conv-layer for learning partial features. However, it does not explicitly partition the images; instead, it takes a whole image as the input and outputs a convolutional feature. Thus, the architecture of partial feature extraction network is concise, with slight modifications on the backbone network.

In line with the above, we introduce an additional branch of the vanilla approach. In departure from [30], we do not abandon the global feature branch as although there is no feature alignment procedure that can be done for the street-view images or additional Google pictures, the additional information provided by these image sources is still useful for providing guidance for common feature space generation.

We mainly adopt two partition methods to get the partial feature: the regular partition and dense partition. Considering the equal importance of the target surroundings, we treat each partition equally and apply no coefficient. In section 4.3.2, we will demonstrate the performance of global feature under the guidance of partial features is greatly improved when compared to the vanilla method. We will also discuss different partition strategies to analyze which one is better for extracting the partial features.

4 EXPERIMENT

4.1 Dataset

University-1652 is a recently released dataset for multi-view multi-source geo-localization tasks. This dataset covers 1652 architectures of 72 universities around the world as target locations. Thus, there are 1,652 classes of different source images, including drone-view, satellite, ground-view and common-view images. In total, there are 5,580 street-view images and 21,099 common-view images from Google Map and Google Images, respectively. The images collected from Google Images only serve as an extra training set, not a test set. Every building has one satellite-view image on average. The images were cropped from the drone-video every 15 frames, which is around 20 degrees, resulting in 54 drone-view images. Overall, every building has a total of 58.38 reference images. If using extra Google-retrieved data, there will be 16,644 ground view images per building for training. There are 701 classes with 50,218 images in the training set and 951 classes with 51,355 drone-view images and 793 classes with 2,921 ground-view images in the gallery set, including 701 classes of drone-view images in the query set. There is no overlap between the 33 universities in the training set and 39 universities in test sets.

4.2 Implementation Details

We implement the model on the ResNet50[9] backbone with several optimizations relative to the original one in University-1652. We add an additional branch before the global average pooling (GAP) layer. As illustrated in Figure 4, we maintain the vanilla branch extracting global feature with no modifications. Moreover, we have one or more copies of 3D tensor which represents features after layer 4 of the ResNet50. For example, for the '3+3' partition in Table3, we partition the 3D tensor on both vertical and horizontal dimensions to get six equal parts along both the height and width of the 3D tensor. After dividing the 3D tensor, we use average pooling to average all the column vectors in the same stripe into a single part-level vector. We then employ a convolutional layer to reduce the vectors to a 512-dim vector; finally, we use a classifier implemented with a fully connected (FC) layer and follow the softmax function to predict the identity of the input. During testing, we employ the 2048-dim feature rather than the 512-dim feature to compute similarity, as the experiment shows that 512-dim features are lower than 2048-dim features on accuracy metrics.

4.3 Ablation Study

4.3.1 Crop, Rotate and Style Alignment. Table 1 shows the ablation study results of our data preprocessing method. We add 'crop', 'rotate' and 'style alignment' consequently on the raw data. Note that we here adopt 512-dim feature to compare with the model performance on raw data. We make no modifications to the vanilla model using only instance loss. We still take Google images and street-view images as an extra source to guide model using the same method as before. The experimental results show that all three data preprocessing methods gain significant performance improvement. The most effective method is 'rotate', which gains about 10 percent improvement on Recall@1 and 7 percent improvement on mAP.

4.3.2 Partial Feature Guidance. We evaluate the performance improvement with partial feature extraction on different input sizes. As we can see from Table 2, employing partial feature in the basic global feature model significantly improves the model performance on accuracy. All three input sizes achieved significant improvement. Input size 384 shows the highest accuracy on R@1 R@5 and R@10 under '3+3' partition strategy. Input size 512 shows the highest accuracy on mAP under this strategy. The result indicates that input size is an important factor on model performance under same partition strategy. We will discuss this in next section.

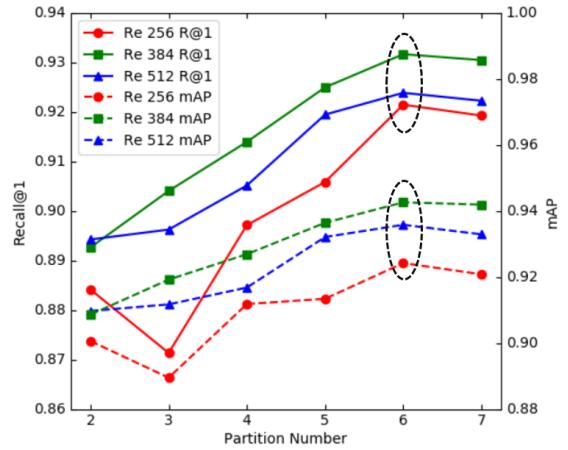


Figure 7: Recall@1 and mAP with different input sizes and partition numbers using regular partitioning. Here, we only show the drone-satellite branch. The black dashed circle indicates the highest accuracy.

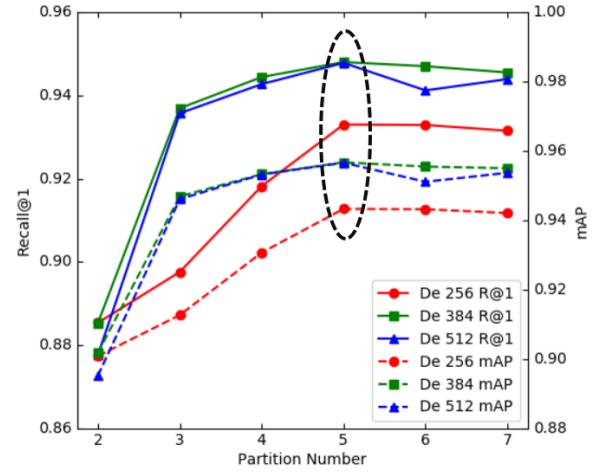


Figure 8: Recall@1 and mAP with different input sizes and partition numbers using dense partitioning. Here, we only show the drone-satellite branch. The black dashed circle indicates the highest accuracy.

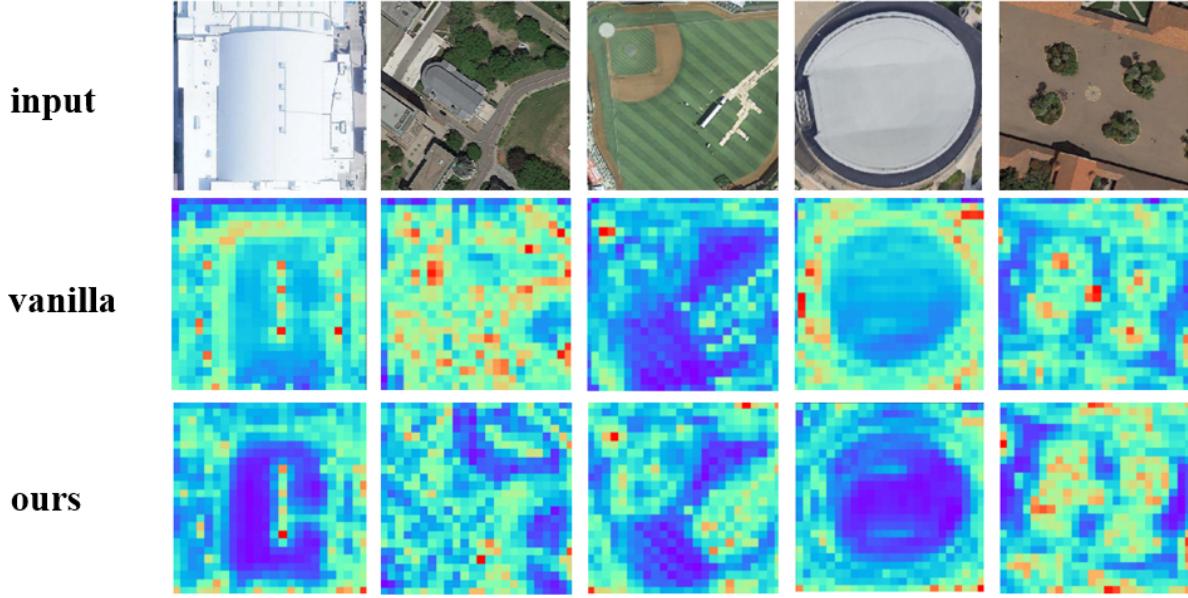


Figure 9: Embedding visualizations. Here we adopt [38] to visualize the CNN embeddings. The heatmap of our method is clearer than the vanilla method. Moreover, the key points of the target building and its surroundings are more specific and accurate, and we can easily find the segment border.

Method	R@1	R@5	R@10	mAP
vanilla	58.49	78.67	85.23	63.13
vanilla+C	60.59	82.73	87.51	66.13
vanilla+C+R	70.03	86.94	91.18	73.86
vanilla+C+R+A	71.70	88.79	93.23	76.12

Table 1: We evaluate the different data preprocessing methods. The basic method is from University-1652 using instance loss. C represents circle crop, R represents rotation and A represents style alignment. Note that we here use only the global feature after the bottleneck of linear block to test the model performance and we only show the drone-satellite branch.

4.4 Partition Strategy and More

The use of different partition strategies results in different model performance. Considering the equal importance of all surroundings, the partitioning should be symmetrical. We adopt two groups partition strategies with different numbers of parts to evaluate their relative effectiveness. The first group is the regular partition group. In this group, we equally divide the 3D tensor into n equal parts both vertically and horizontally to get $2 \times n$ parts of vectors. The second group is the dense partition group; here we divide the 3D tensor into n^2 parts, as illustrated in Figure 6. Note that the dense partition group only needs one copy of the 3D tensor from global feature extraction branch while regular partition group needs two copies. In table 4, we show the different partition strategies for these two groups with different input sizes and partition numbers.

The results for the different partition strategies in table 3 indicates two main factors to model performance, which are partition number and input size. Here we have three input size including 256, 384, and 512; these equate to 1/2, 3/4 and equal to the raw image side length respectively. In each group, we apply several different partition numbers. We first test our model performance

Method	size	R@1	R@5	R@10	mAP
basic	256	70.03	86.94	91.18	73.86
basic	384	70.04	86.70	90.71	73.82
basic	512	66.62	86.39	90.59	71.06
basic+P	256	87.14	95.08	96.61	88.96
basic+P	384	90.41	97.13	98.15	91.93
basic+P	512	89.63	96.33	97.62	91.18

Table 2: We evaluate the performance improvement resulting from feature extraction on different input sizes. The basic method contains our data preprocessing method including 'crop' and 'rotate' without 'style alignment' as we treat the style alignment method as an independent process in the testing stage. 'P' stands for partial feature, which is the only difference from the basic method. Here, we choose 256, 384 and 512 as the input side length of images considering the raw image size (512x512). The partition strategy is fixed to be '3+3' regular partition which divides the 3D tensor equally into 6 parts both vertically and horizontally. Note that we use only the feature vector with 2048-dim in the basic+P group to test the model performance, and we only show the drone-satellite performance.

using regular partitioning. The model achieves the best accuracy when using a 6+6 partitioning strategy with 384 or 512 input size (figure 7). We then apply dense partitioning to train the model. In this partition case, the model achieves the best accuracy when using a 5x5 partition strategy (figure 8). Meanwhile, it is interesting to see that the dense partition model gains significant improvement when the partition number goes up. By contrast, the improvement under regular partitioning increases relatively slowly. We can guess that the reason lies in how ‘independent’ the part feature is. Considering the overlap between different feature vectors of the regular partition, the ability to cover more features of the surroundings will be weakened, leading to lower performance when representing partial features in the testing stage compared to the dense partition strategy.

4.5 Performance Evaluation

Here, we present the visualization heatmap in Figure 9, while the best performance between the vanilla method, the basic method with all data pre-processing methods applied and the full model using different partition strategies are represented in Table 3. Compared to the vanilla method with instance loss, our approach is far beyond the baseline, with about 40 percent improvement on both Recall@1 and mAP. Figure 9 shows the heatmap for the vanilla method and our full approach, which reveals that our approach is more robust and covers more surroundings of the target building to assist feature representation.

5 CONCLUSION

In this paper, we have proposed style and spatial alignment approaches for multi-view drone-based geo-localization. Specifically, we have proposed an elegant orientation-based method to align the patterns and introduced a new branch to extract aligned partial feature. In addition, we have provided a style alignment strategy to reduce the variance in image style and enhance the feature unification. We have verified the effectiveness of the proposed approach on the large-scale benchmark dataset. Besides, we have conducted ablation studies to confirm the influence of each component. From the

strategy	drone-satellite		satellite-drone	
	R@1	mAP	R@1	mAP
V+CL	52.39	57.44	63.91	52.24
V+TL	55.18	59.97	63.62	53.85
V+IL	58.23	62.91	74.47	59.45
Ours+Re	93.17	94.27	96.86	90.96
Ours+Re+A	93.21	94.38	97.05	91.23
Ours+De	94.78	95.67	98.15	93.74
Ours+De+A	94.84	95.80	98.35	94.02

Table 3: Comparison to the existing methods on University-1652. ‘V’ represents the vanilla method, while ‘CL’,‘TL’ and ‘IL’ represent contrastive loss, triplet loss and instance loss respectively. Our full model with style alignment strategy ‘A’ can bring the drone-view-based geo-localization task into practical use.

experimental results, we observe that all the components contribute significantly to the overall improvement.

REFERENCES

- [1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5297–5307.
- [2] Mayank Bansal and Kostas Daniilidis. 2014. Geometric urban geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3978–3985.
- [3] Cross-View Image based Geo-Localization. [n.d.]. Spatial-Aware Feature Aggregation for Cross-View Image based Geo-Localization. ([n. d.]).
- [4] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a “siamese” time delay neural network. In *Advances in neural information processing systems*. 737–744.
- [5] Google Cloud. [n.d.]. *Google satellite image api*. <https://developers.google.com/maps/documentation/maps-static/intro>
- [6] Google Cloud. [n.d.]. *Google street view api*. <https://developers.google.com/maps/documentation/streetview/intro>
- [7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
- [8] Riad Hammoud, Scott Kuzdoba, Brian Berard, Victor Tom, Richard Ivey, Renu Bostwick, Jason HandUber, Lori Vinciguerra, Nathan Shnidman, and Byron Smiley. 2013. Overhead-based image and video geo-localization framework. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 320–327.
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [10] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.
- [11] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. 2018. 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*. 7258–7267.
- [12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4700–4708.
- [13] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.
- [14] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2015. Predicting good features for image geo-localization using per-bundle vlad. In *Proceedings of the IEEE International Conference on Computer Vision*. 1170–1178.
- [15] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2136–2145.
- [16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*. Springer, 694–711.
- [17] Dong-Ki Kim and Matthew R Walter. 2017. Satellite image-based localization via learned embeddings. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2073–2080.
- [18] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. 2017. Learned contextual feature reweighting for image geo-localization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3251–3260.
- [19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. In *Advances in neural information processing systems*. 386–396.
- [20] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. 2015. Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5007–5015.
- [21] Chong Liu, Xiaojun Chang, and Yi-Dong Shen. 2020. Unity Style Transfer for Person Re-Identification. *arXiv preprint arXiv:2003.02068* (2020).
- [22] Liu Liu and Hongdong Li. 2019. Lending Orientation to Neural Networks for Cross-view Geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5624–5633.
- [23] Liu Liu, Hongdong Li, and Yuchao Dai. 2019. Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2570–2579.
- [24] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.

Group	input size	strategy	drone-satellite				satellite-drone			
			R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Re-Partition	256	2+2	88.42	95.65	97.04	90.07	94.29	96.01	96.86	86.52
		3+3	87.14	95.08	96.61	88.96	94.15	96.43	97.00	85.53
		4+4	89.72	96.27	97.54	91.19	94.29	96.29	97.43	86.86
		5+5	90.59	96.57	97.64	91.35	93.58	96.58	97.29	87.65
		6+6	92.15	96.73	97.77	92.43	95.15	97.43	98.00	89.04
		7+7	91.93	96.68	97.84	93.09	94.29	97.29	97.72	88.59
		2+2	89.27	96.34	97.50	90.87	95.44	97.43	97.72	88.41
	384	3+3	90.41	97.13	98.15	91.93	96.01	97.29	98.15	88.13
		4+4	91.40	97.05	98.09	92.70	95.58	97.57	98.43	88.22
		5+5	92.50	97.51	98.34	93.65	95.86	97.72	98.00	90.42
		6+6	93.17	97.97	98.69	94.27	96.58	98.29	98.57	90.96
		7+7	93.05	98.06	98.79	94.20	96.72	98.86	99.14	90.32
		2+2	89.43	96.33	97.54	90.98	94.86	97.00	97.43	87.62
		3+3	89.63	96.33	97.62	91.18	94.01	96.72	97.57	87.29
De-Partition	512	4+4	90.52	96.86	97.88	91.69	95.44	97.29	98.15	87.94
		5+5	91.95	97.51	98.36	93.22	96.72	98.15	98.29	89.20
		6+6	92.39	97.50	98.55	93.59	95.86	97.72	98.15	89.05
		7+7	92.23	97.73	98.69	93.30	96.86	98.43	98.86	90.75
		2x2	88.54	95.21	96.63	90.09	95.72	97.15	98.29	87.61
		3x3	89.76	96.21	97.32	91.27	94.29	96.43	97.00	88.10
		4x4	91.82	97.32	98.30	93.07	96.01	97.72	98.29	90.81
	384	5x5	93.30	97.82	98.56	94.33	96.01	97.72	98.43	91.41
		6x6	93.29	97.72	98.54	94.32	96.15	97.86	98.57	91.53
		2x2	88.53	95.76	97.09	90.18	94.86	98.29	98.57	86.92
		3x3	93.69	98.07	98.64	94.69	97.29	98.57	99.14	92.53
		4x4	94.44	98.24	98.79	95.33	97.15	98.57	98.86	92.77
		5x5	94.80	98.61	99.25	95.67	97.57	99.14	99.57	93.50
		6x6	94.70	98.41	99.17	95.55	97.57	99.29	99.57	93.74
	512	2x2	87.76	95.41	96.85	89.52	93.72	97.15	98.15	85.93
		3x3	93.57	98.11	98.76	94.61	97.15	98.72	99.43	91.48
		4x4	94.27	98.34	98.89	95.21	96.72	98.29	98.86	92.47
		5x5	94.78	98.63	99.12	95.66	97.00	98.72	99.14	92.93
		6x6	94.12	98.45	99.11	95.11	98.15	99.29	99.57	93.24

Table 4: Automatic Evaluation Score. ‘Re’ represents ‘regular’ and ‘De’ represents ‘dense’.

- [25] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 1–8.
- [26] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5706–5715.
- [27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.
- [28] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. 2019. Optimal Feature Transport for Cross-View Image Geo-Localization. *arXiv preprint arXiv:1907.05021* (2019).
- [29] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [30] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.
- [31] Yicong Tian, Chen Chen, and Mubarak Shah. 2017. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3608–3616.
- [32] Anirudh Viswanathan, Bernardo R Pires, and Daniel Huber. 2014. Vision based robot localization by ground to satellite matching in GPS-denied situations. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 192–198.
- [33] Nam N Vo and James Hays. 2016. Localizing and orienting street views using overhead imagery. In *European Conference on Computer Vision*. Springer, 494–509.
- [34] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2015. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*. 3961–3969.
- [35] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1492–1500.
- [36] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. 2017. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 867–875.
- [37] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. *arXiv preprint arXiv:2002.12186* (2020).
- [38] Zhedong Zheng, Liang Zheng, and Yi Yang. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 1 (2017), 1–20.
- [39] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1452–1464.

- [40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.
- [41] Ryan R Zunker, Atreyee Sinha, and Sugata Banerji. 2019. House Hunting: Image-based Geo-Localization of Buildings Within a City. In *Proceedings of the 2019 5th International Conference on Computing and Data Engineering*. 100–104.