Cross-view geo-localization via Salient Feature Partition Network

Sijin He¹, Yuehuan Wang¹

¹National Key Laboratory of Science and Technology on Multispectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China

Email: yuehwang@hust.edu.cn

Abstract: Cross-view geo-localization aims to find images containing the same geographic target from images obtained from different platforms. The extreme viewpoint variations bring challenges to this task. Existing methods usually focus on mining the fine-grained features of geographic targets in images, ignoring the potential contextual information around them. In this paper, we consider that the background regions can be used as auxiliary information, which can make the image representation for geo-localization more discriminative. Specifically, we designed a classification network that divides regional features based on saliency, called Salient Feature Partition Network (SFPN), which utilizes background information in an end-to-end manner. Without using additional part estimators, SFPN divides the features into foreground and background based on saliency. It simplifies part matching and realizes the region division learning. The method proposed in this paper has achieved competitive results on the university-1652 dataset.

1. Introduction
In recent years, the chance of cross-view geo-localization is immense. For example, given a ground-view image without geographic information as a query, we need to find images acquired at the same place in the gallery of satellite-view images, which are pre-annotated with GPS tags. We can localize the building in the query image by finding the correct-match satellite-view image with the geo-tag. The cross-view based on image geo-localization can be used to improve the positioning system and realize positioning under offline conditions without GPS information.

Most works on this task have followed the traditional way of supervised deep learning methods. They aim to explore the common features between ground-view and satellite-view images. In this way, the cross-view geo-localization task can be defined as a binary classification problem. Therefore, Triplet loss[1] and Siamese architecture [2] have been widely used to deal with this task [3][4][5][6]. Furthermore, many networks which are designed with attention mechanism [5][6] and orientation information [4][7] have been proposed to solve the problem of feature alignment between images from different viewpoints. However, the performances of these methods were not such satisfactory. The main reason for this result is obvious: the visual difference between the ground-view image and the satellite-view image is so large that it is difficult to correctly match the given query image and the target image. These images with the limited tow so different viewpoints may also compromise the network model to mine the common feature between different views during the training process.

In recent years, with the rapid development of map tools such as Google Earth and the functions
providing by Google Maps API, multi-view and multi-source images with rich geographic information can be collected online. In addition, with the development of science and technology, drones are increasingly used in real life, playing a key role in areas such as agriculture, aerial photography, navigation and other fields. Geo-localization tasks such as navigation and target localization that use drone-view images have also received increasing attention. With the release of a multi-view dataset named University-1652 [8] based on drone-view, it can help image-based geo-localization tasks to get higher accuracy results by using drone-view images with rich spatial information.

With different platforms, viewpoints and an increasing number of multi-source data, the cross-view geo-localization task is no longer a binary classification problem. When extracting features, the network can use more scene data to make it more robust, which is undoubtedly practical and valuable.

Compared with some existing work such as CVMNet [3], Orientation [4] and other main benchmarks, experiments have proved that the baseline method proposed in University-1652 is more robust on sub-tasks and other smaller public datasets. However, this baseline method only uses the pooling function to obtain the global features of the image when extracting features, which does not fully mine and use the contextual information in the image that is beneficial for cross-view geo-localization.

To solve the above problems, we propose a location classification network based on salient feature partition (SFPN). The difference from existing methods is that our network part features into foreground and background based on their saliency in the image and uses both to match between images from different viewpoints. Compared with directly using the global features obtained by pooling functions, the features learned by our network are more discriminative. These features not only contain the target building information, but also use the contextual information.

2. Related Work
A lot of early research equate the geo-localization with an image retrieval problem. They aim to learn the representation which is invariant in different viewpoints to bridge the gap between these multi-view images. With the great success of the deep learning, more and more researchers have been turning their attention to extract the visual features by using convolutional neural networks (CNNs). Workman et al. [9] firstly attempt to extract features for the cross-view geo-localization task by using a pre-trained CNN. Their research shows that features from the high-level layer of CNN contain semantic information which is about the geographic location. Furthermore, Workman et al. [10] fine-tune the pre-trained network to reduce the feature distance between the ground-view images and the aerial images, which get better results.

Some works try to solve the problem caused by spatial misalignment in the matching images from ground-view and aerial view. Vo et al. design an orientation-aware network which is trained with an orientation regression loss [11]. Zhai et al. improve the semantic alignment by adopting the semantic segmentation map. Liu et al. propose a Siamese Network to explicitly involve the spatial cues, that is, orientation maps, into the training [12]. Shi et al. design a spatial-aware layer which could further improve the localization performance.

The other works concentrate on building a feature learning space for the multi-source images with metric learning. Lin et al. [13] use a modified Siamese Network, and adopt contrastive loss to optimize network parameters. Hu et al. [3] propose a CVM-Net which employs a weighted soft margin ranking loss. Recently, Liu et al. [4] put forward a loss called Stochastic Attraction and Repulsion Embedding (SARE) loss to minimize the KL divergence between the learned and the actual distributions.

With University-1652 releasing, there are several training dates from different viewpoints for each location. Therefore, we treat matching the cross-view images as a classification task, so that we can utilize classification CNNs to train our model without designing complicated loss. Zheng et al. [8], the author of university-1652, apply the instance loss [14][15] to optimize the network, and their baseline method performs well.

3. Proposed Method
As mentioned above, the images provided by University-1652 are from various viewpoints, and there are enough image samples of each target location to train the model. Based on the above background, we suppose that images acquired from the same location belong to a category, and we can treat the cross-view geo-localization as a classification task with an unknown number of categories during training our model. The network framework called SFPN is shown in Figure 1.

With the geo-localization dataset, we express the input image as \( x \), and \( y \) represents the label, which is the corresponding location i.e. the category. Since the images in University-1652 are collected from three different platforms, we apply three subscripts \( s, d, g \) to represent respectively. Specifically, \( s \) denotes the satellite-view, \( d \) denotes the drone-view, and \( g \) denotes the ground-view. For cross-view geo-localization, we aim to learn a mapping function, which could project images from different platforms to a shared feature space. In the space, the images of the same location are close, while the images from different locations are apart from each other.

3.1. Feature Extraction

As shown in Figure 1, there are three branches in SFPN, namely the satellite-view branch, the drone-view branch and the ground-view branch and the input images of each branch are from different platforms. There are many different network architectures can be used as backbones for features extraction, such as VGG, DenseNet and ResNet. In this paper, we choose ResNet-50 as our backbone because of its good performance in classification task. ResNet-50 contains five blocks named conv1, conv2, conv3, conv4, conv5, an average pooling layer, and a fully connected layer. We remove the original final average pooling layer and the fully connected layer, and use the rest layers to obtain intermediate feature maps for next processing. There are 54 drone-view images of each location, while the number of satellite-view image is one, and the too small number of samples will affect the training of the satellite-view branch. To reduce the impact of the above problems, we share weights between the satellite-view branch and the drone-view branch, considering the similarity between these two aerial viewpoints. Three branches follow the same feature extraction manner. Specifically, given an input image of size 256×256, we can acquire feature maps with the shape of 16 × 16 × 2048 in each branch. Take satellite-view branch as an example, we use \( F \) to represent this mapping function, so the process of feature extraction can be formulated as:

\[
f_s = F(x_s),
\]

where \( f_s \) stands for the extracted feature map of the input satellite-view image \( x_s \). In the same way, we can get \( f_d \) of drone-view image \( x_d \) and \( f_g \) of ground-view image \( x_g \).

![Figure 1. The Overall Structure of Our Proposed SFPN.](image-url)
The classifier module contains five type layers, which are the fully connected layer (FC), the batch normalization layer (BN), the rectified linear unit layer (ReLU), the dropout layer (Dropout), and the classification layer (Cls). In the bottom left, we show the saliency feature partition strategy.

3.2. Feature Partition Strategy

As mentioned above, many existing methods use attention mechanism to design networks so that their network focus on discriminative areas, ignoring other potential regions in the image. For example, when there is no obvious difference between two geographic targets, such as two office buildings, it is difficult to identify the true-match target. However, the task is much easier if we consider the information of the background region on the image, e.g., surrounding roads. There seems to be a simple way to solve the above problem: use a pooling function. However, the global features via pooling functions will ignore the contextual information.

To explicitly utilize contextual information, we apply a saliency-based partition strategy to divide feature maps. On the high-level feature map, the feature value of each point represents its semantic information. The strength of semantic information represents the level of saliency. The salient region of the image contains a lot of information and is the region that the network needs to focus on for the task, which can be regarded as the foreground. Meanwhile, regions with low saliency can be regarded as backgrounds containing relatively little information, and they also have a certain effect on the task. According to the saliency value, we can roughly divide the feature map into two parts: foreground and background. We observe that geographic target, i.e., foreground is generally distributed in the center of the image, and the contextual information, i.e., background, is radiated around. Based on this assumption of semantic information distribution, we select the points on the four edges of the feature map (corresponding to the edge of the original image) as the background seeds to calculate the contrast between other points and the background seeds, so as to obtain its corresponding saliency. We firstly calculate the average feature value of each background seed, then accumulate them, and finally take their average value as the feature value of the background, denoted as $v_{bg}$. Then the saliency of other points on the feature map can be calculated with the following formula:

$$S(i) = \sqrt{|v_i - v_{bg}|^2}$$

Where $v_i$ and $S(i)$ stands for the average feature value and the saliency value of the $i_{th}$ point respectively.

We normalize $S(i)$ so that $S(i) \in (0, 1)$. In this way, we get a rough saliency map. The larger the value of $S(i)$, the higher the saliency of this point and the more likely it is to belong to the foreground. We set a threshold $\gamma$ ($\gamma \in (0, 1)$), the points whose saliency value is higher than $\gamma$ belong to the foreground, and the other points belong to the background. We divide images into two parts according the saliency. In practice, we separate the global feature maps $f$ to two feature parts $f_{bg}$ and $f_{fg}$. The superscript bg represents the background, and fg represents the foreground. Then we apply the average pooling layer to transforms each part with different shapes into a 2048-dim part feature $g$.

3.3. Optimization Objective.

Now we have obtained two-part features from each source. Then we adopt a mapping function to map features of all sources into one shared feature space. The mapping function called classifier module consists of following layers: a fully connected layer (FC), a batch normalization layer (BN), a rectified linear unit function (ReLU), a dropout layer (Dropout), and a classification layer (Cls), which is a fully-connected layer. Passing through the classifier module, the dimension of the input part feature becomes the number of categories. Each value in the output column vector whose dimension equals the number of categories is the logit score of the ground-truth label $y$. Then we use the softmax function to obtain the normalized probability score. Finally, we accumulate the cross entropy losses on the image of different parts and different platforms to optimize the whole network.

4. Experiments
4.1. Implementation Details

University-1652 is a recently released dataset for multi-view multi-source geo-localization tasks. This dataset contains image data of 1652 buildings in 72 universities, and the images of each building are from three different viewpoints: satellite-view, drone-view and ground-view. There are 701 building classes with 50,218 images in the training set. In the test set, there are 37,855 query drone-view images and 951 gallery satellite-view images, meanwhile, there are 701 satellite-view images as query and 51,355 drone-view images in gallery. Besides, there are no overlapping building classes in the training and test set.

We use the Recall@K (R@K) and the average precision (AP) to evaluate the performance, which are widely used to evaluate retrieval performance. R@K represents whether the proportion of correctly matched images in the K-th images of the ranking list. And the AP reflects the precision and recall rate.

We use ResNet-50 with shared weights to extract the features of the images. The cross-entropy loss function is adopted as the objective function and we use the adam optimizer to optimize the loss function. The initial learning rate is set to 0.0001, and the learning rate is reduced with convergence. The experiments are performed on NVIDIA RTX 2080Ti GPU and the code uses the Pytorch framework.

4.2. The Results

As shown in Table 1, we compare the proposed method with the baseline method proposed by Zheng on University-1652. We can clearly see that our proposed method has a large improvement in various indicators compared to the baseline method of Zheng, which can prove that the features learned by our network are more discriminative than Zheng model.

| Task          | Method  | R@1 | R@5 | R@10 | AP  |
|---------------|---------|-----|-----|------|-----|
| Satellite to drone | Zheng   | 71.18 | 80.52 | 82.32 | 58.74 |
|               | Ours    | 80.26 | 89.34 | 93.57 | 71.58 |
| Drone to satellite | Zheng | 58.49 | 78.67 | 85.23 | 63.13 |
|               | Ours    | 70.83 | 85.07 | 91.62 | 77.36 |

5. Conclusion

In this paper, we focus on solving the problem in cross-view geo-localization task, and propose a location classification network based on salient feature partition (SFPN), to take advantage of the potential information, which is usually ignored. We design a feature partition strategy based on saliency to make our model learn more comprehensive features. Our model not only focuses on the salient regions in the image, but also effectively utilizes the beneficial information contained in the background region, which make the image representation more discriminable. We have verified the effectiveness of the proposed method on a large-scale benchmark dataset, i.e., University-1652.

6. References

[1] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In International Workshop on Similarity-Based Pattern Recognition. Springer, 84–92.
[2] 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.
[3] 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.
[4] 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.
[5] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. 2019. Optimal Feature Transport for
[6] 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.

[7] Nam N Vo and James Hays. 2016. Localizing and orienting street views using overhead imagery. In European Conference on Computer Vision. Springer, 494–509.

[8] 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.

[9] S. Workman and N. Jacobs, “On the location dependence of convolutional neural network features,” in IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[10] 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.

[11] 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.

[12] 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.

[13] 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. [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.

[14] Z. Zheng, L. Zheng, and Y. Yang, “Unlabeled samples generated by gan improve the person re-identification baseline in vitro,” in IEEE International Conference on Computer Vision, 2017.

[15] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, “Dual-path convolutional image-text embeddings with instance loss,” ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 16, no. 2, pp. 1–23, 2020.