Received: 14 September 2021 | Revised: 23 November 2021 | Accepted: 20 December 2021 | CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12077

ORIGINAL RESEARCH

Multi-scale attention encoder for street-to-aerial image geo-localization

Songlian Li[1] | Zhigang Tu[1] | Yujin Chen[2] | Tan Yu[3]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China

[2]Department of Informatics, Technical University of Munich, Garching, Germany

[3]Baidu USA, Sunnyvale, California, USA

Correspondence

Zhigang Tu, State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China.
Email: zhigangtu@whu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Number: 62106177; Wuhan University

Abstract

The goal of street-to-aerial cross-view image geo-localization is to determine the location of the query street-view image by retrieving the aerial-view image from the same place. The drastic viewpoint and appearance gap between the aerial-view and the street-view images brings a huge challenge against this task. In this paper, we propose a novel multiscale attention encoder to capture the multiscale contextual information of the aerial/street-view images. To bridge the domain gap between these two view images, we first use an inverse polar transform to make the street-view images approximately aligned with the aerial-view images. Then, the explored multiscale attention encoder is applied to convert the image into feature representation with the guidance of the learnt multiscale information. Finally, we propose a novel global mining strategy to enable the network to pay more attention to hard negative exemplars. Experiments on standard benchmark datasets show that our approach obtains 81.39% top-1 recall rate on the CVUSA dataset and 71.52% on the CVACT dataset, achieving the state-of-the-art performance and outperforming most of the existing methods significantly.

KEYWORDS

global mining strategy, image geo-localization, multiscale attention encoder, street-to-aerial cross-view

1 | INTRODUCTION

Image-based geo-localization has attracted increasingly attention and obtained constantly progress due to its great potential in the fields of autonomous driving [1, 2], robot navigation [3–5], as well as AR/VR [6]. Traditional methods aim to determine the location by matching the query street-view image with the geo-tagged street-view images in a reference database [7–13]. Relja et al. [7] successfully extended the VLAD algorithm [14] into an end-to-end trained network to exploit a NetVLAD model, which extracts rotation and scale invariant features to ensure that the model is not affected when there is a huge rotation and scale change between the query street-view image and the target street-view image. Hausler et al. [13] further improved NetVLAD by merging global descriptors and local descriptors into it to form a Patch-VLAD model, which roughly sorts the candidate street-view images in the database according to the query street-view image with the help of the global descriptors, and then uses the local descriptors to retrieve the final matching street-view image. Cao et al. [12] incorporated global and local features into a single model to form a unified feature extraction framework, which extracts more robust descriptors for image matching from street-view to street-view. However, the above methods all have a fatal problem: the database that contains the street-view images is difficult to densely cover a large area.

Thanks to the remote sensing satellite, a great number of satellite images with geo-tags have been collected. Consequently, matching street-view image to aerial-view image has gradually become the mainstream method, see Figure 1. However, the cross-view image-based geo-localization is very challenging because of the drastic domain gap, orientation uncertainty, and different scales with different viewpoints, leading to it is impossible to use the traditional methods like SIFT [15] and HOG [16] to solve this task well.

Recently, deep learning has achieved great success in the field of computer vision [17–20], so most of the current

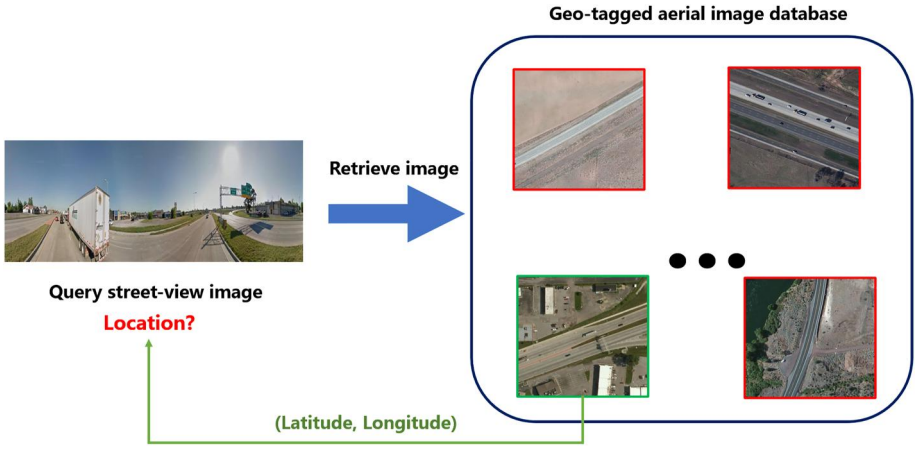

FIGURE 1 An example of street-to-aerial image geo-localization. The location of the query street-view image is determined by retrieving the geo-tagged aerial image from the same place in the reference database

works [21–26] proposed to use the convolutional neural network (CNN) to obtain robust deep feature representation between the two-view images for matching. A key measure is to find better image feature embeddings which can adaptively pull the matching image pairs closer while push the unmatching pairs far away. Encouraged by the recent success of using CNNs to learn high-level features, Workman et al. [27] introduced deep features into cross-view geo-localization, and released a massive dataset called CVUSA which contains tens of thousands pairs of cross-view images. Hu et al. [22] proposed a model named CVM-Nets, which integrated the NetVLAD into the Siamese CNN, to get robust representation of images. Inspired by hard exemplar mining [22], Cai et al. [21] explored a hard exemplar mining strategy to automatically allocate weights to triplets for cross-view image matching. To align the space layout between the street-view image and the aerial-view image, Regmi and Shah [28] introduced a GAN to convert the street-view image into the aerial-view image. Although the effectiveness of GAN is investigated in Ref. [28], it can only produce synthetic street-view images for a database which it trained on, and cannot produce the generic street-view images which are deviated from the training images to enhance the generalisation capacity. In contrast, by mining the geometric relationship between the two view images, we exploit an inverse polar transform to reduce the domain gap between them. Specifically, our approach can approximately convert any street-view image into an aerial-view image. The contributions of this paper are summarised as follows:

- We explore an inverse polar transform method to conduct a rough geometric alignment between the two view images, which is useful to reduce the domain difference and make the network easier to learn discriminative features for the two views.
- We propose a multi-scale attention encoder, and integrate it into a basic Siamese network to reduce the impact of the distortion caused by the inverse polar transform.
- We exploit a global mining strategy to discover the global hard negative samples and enable the network to focus on hard exemplars to improve the performance.

2 | RELATED WORK

There are mainly two types of related work on geo-localization: (1) Approaches based on image matching techniques. (2) Approaches that divide the Earth's surface into a large number of grids, and then calculate the probability of the image falling into the grid to achieve locating.

2.1 | Image matching techniques

Due to the significant viewpoint difference between street-view and aerial-view images, traditional hand-crafted feature methods cannot project the two view images into a unified feature space, which leads to a bottleneck in image matching performance [29–33]. Encouraged by the success of CNNs in computer vision, Workman and Jacobs [34] first transferred the pre-trained AlexNet on the Imagenet [35] and Places [36] datasets to learn deep features for cross-view geo-localization, and many experiments have been conducted to verify that deep features are far superior to hand-crafted features. Lin et al. [37] introduced the Siamese CNNs into cross-view geo-localization, and captured the deep features of street-view images and aerial-view images separately through the two-branch networks. They used an improved contrast loss to train two branches of CNN to locate cross-view images. A large number of experiments proved the descriptors extracted by the neural network are better than those hand-crafted descriptors. Vo and Hays [24] released a massive dataset called VH and tested a series of existing methods on it. To learn scale-invariant and rotation-invariant features, Hu et al. [22] integrated NetVLAD into a two branch CNN for ground-to-aerial geo-localization. Cai et al. [21] introduced an attention module to re-weight the

spatial and channel features to capture more robust image representations.

Considering that the drastic viewpoint differences between the street-view and aerial-view images is a key factor that limits the matching performance, Shi et al. [26] used feature transport to convert street-view features into aerial-view features, which is helpful to eliminate the difference between the two-view images in the feature domain. Regmi and Shah transferred the ground images into air-view images by a generative adversarial model, and then concatenated the representations of the street-view and synthesised aerial-view images to form the global representation for matching. Shi et al. [25] exploited a spatial-aware feature aggregation module to conduct feature ensemble. Since humans often take the orientation information into consideration when determining their position in daily life, Liu et al. [23] explicitly combined the orientation and RGB image, and then fed the fusion information of the two into the network. To further estimate the orientation information, Shi et al. [38] aligned the cross-view orientation information by using the DSM module during localization. Wang et al. [39] used a square ring partition method and fused these partition features as a global representation to improve the performance. Previous experiments have proved that preliminary alignment of the domain and orientation information of the street-view and aerial-view images can reduce the difficulty of network learning.

Except the huge viewpoint difference between these two views of images, to accurately locate the image spatial position, there is usually only one aerial-view image in the database used to exactly match the street-view image in the cross-view geo-localization task. It means that all other aerial-view images in the dataset will be treated as negative samples. As a result, there is a huge imbalance in the number of positive and negative samples during training. Zhu et al. [40] designed a novel binomial loss which applied anchor points to gather positive samples and push negative samples away from each other. Schroff et al. [41] tried to use the online hard negative mining to make further breakthroughs. Although this method has a certain effect on improving the accuracy, it is difficult to improve the network accuracy significantly because the online hard negative mining cannot consider the global hard negative samples.

Since global features are usually affected by noisy data during the matching process, more and more scholars are beginning to pay attention to the part-based representation learning. Li et al. [42] divided the image information into three parts: head-shoulder, upper body, and lower body, and then utilised a Spatial Transformer Network (STN) to integrate these three local information. A strong part-based Convolutional Baseline (PCB) [43] was presented to extract high-level features by developing a uniform partition strategy.

2.2 | Classification methods based on deep learning

Classification methods based on deep learning have achieved great success in different fields [44–47]. Most of the previous image-based geo-localization approaches require a large number of street-view images to cover the area of interest, which limits the application of these methods. To address geo-localization at planet-scale without any restrictions, Weyand et al. [48] introduced a PlaNet, which divided the Earth's surface into a large number of grids, and used these grids as category labels to locate the image. Muller-Budack et al. [49] introduced hierarchical knowledge given by the predictions at each scale based on the former method, and experimental results demonstrated that incorporating hierarchical knowledge in the convolutional neural network is effective. Recently, many works [46, 50–52] tried to introduce attention mechanisms and multi-scale information into the network to improve the performance of CNN on large-scale classification datasets. Woo et al. [50] combined the channel attention module with the spatial attention module to form a lightweight module, which achieved state-of-the-art performance on the datasets of ImageNet-1K, MS COCO detection, and VOC 2007 detection. Wang et al. [46] integrated residual connections in each attention module to train very deep networks, where the performance of classification was boosted on both CIFAR-10 and CIFAR-100. He et al. [51] proposed the spatial pyramid pooling layer to fuse the representations of different scales to obtain a robust global representation, which improves the performance of CNN-based image classification methods. Lin et al. [52] developed a top-down architecture with horizontal connections for constructing high-level semantic feature maps of all scales.

3 | PROPOSED METHODOLOGY

In this section, we provide an overview of our approach. Our method consists of three main components, that is, Inverse Polar Transform (IPT) (Section 3.2), Multi-scale Attention Encoder (MSAE) (Section 3.3), and Global Mining Strategy (GMS) (Section 3.4).

3.1 | Overview

Given a street-view query image, the matching aerial-view image is obtained by comparing the pair-wise Euclidean distance between their representation vectors. Since there are large appearance differences between the two view images, we try to reduce the domain gap between the viewpoints at first. Then, we design a Siamese network, which contains two independent CNN branches to extract the features of the street-view images and the aerial-view images respectively, to obtain discriminative global feature representations. Secondly, there is also a hug domain difference between these two kinds of view images, we make the two branch extractors of the Siamese network without sharing weights. Furthermore, to obtain more robust representation vectors, we introduce a multi-scale attention encoder to encode the extracted deep features. Due to our attention mechanism, the network can ignore irrelevant information.

Finally, to improve the network training process, we propose a global mining strategy to find the global hard negative exemplars. This strategy is useful to further boost the retrieval performance. An overview of our method is shown in Figure 2.

3.2 | Inverse polar transform

As we observed, there are two very important geometric correlations between cross-view images: (1) Objects on the same horizontal line of street-view images have the same depth, which means the horizontal line of street-view images corresponds to the concentric circle of aerial-view images. (2) On the vertical line of street-view images, the depth of the object increases followed with the increasing of the y coordinate, which corresponds to the radial ray on aerial-view images. It is difficult for the network to directly learn the mapping from the street-view images to the aerial-view images, therefore we make use of inverse polar transform to roughly eliminate the domain gap between these two view images. Therefore, the inverse-polar transform between the original street-view image points (x_i^s, y_i^s) and the target transformed street-view image points (x_i^t, y_i^t) is defined as:

$$y_i^s = \sqrt{2}\,\frac{H_s}{H_a}\sqrt{\left(x_i^t - \frac{W_a}{2}\right)^2 + \left(y_i^t - \frac{H_a}{2}\right)^2} \tag{1}$$

$$x_i^s = \frac{W_s}{2\pi}\,arctan2\left(\frac{y_i^t - \frac{H_a}{2}}{x_i^t - \frac{W_a}{2}}\right) \tag{2}$$

Here, the size of the original aerial images is $H_a \times W_a$, and the size of the original street-view images is $H_s \times W_s$. We use the geometric clue that the street-view image is located at the centre of the aerial-view image, accordingly a one-to-one mapping relationship between the original street-view image points $(x_i^s,\ y_i^s)$ and the synthetic air-view image points $(x_i^t,\ y_i^t)$ is established.

By explicitly mining the geometric relationship between the two-view images, the domain differences between the cross-view images can be significantly reduced, as illustrated in Figure 3. Although the neural network can theoretically learn any geometric transformation relationship, in this work, we first roughly align the domain information of the cross-view images, which can not only convert a complex cross-view matching task into a simple matching task, but also can boost the performance of the network.

3.3 | Multi-scale attention encoder

The inverse polar transform method can reduce the geometrical difference between the street-view image and the aerial-view image. However, since the transform still cannot regard the scene depth information of the street-view image, the transformed images have obvious object deformation, thus the geometrical difference problem cannot be clearly eliminated by the function transformation. Noticeably, we propose a multi-scale attention encoder, which is helpful for the network to focus on the region where the deformation of the object is small and to ignore the region where the deformation of the object is large. Moreover, since image matching requires an one-dimensional description vector, we use a fully connected layer to project the captured deep features into a global description vector as well. The proposed MSAE uses a lightweight fully convolutional network to convert multi-scale features into the attention masks. The lightweight fully convolutional network captures the context information by using convolution kernel, which is denoted as k_p. To mine multi-scale information, a set of 3 × 3, 5 × 5, and 7 × 7 convolution kernels are used to capture the feature space context information at different scales. The expression is defined as:

$$s = \bigcup_p (k_p(d) + c) \qquad p \in \{3, 5, 7\} \tag{3}$$

Here, $\bigcup_p(\cdot)$ represents the channel connection operation, d represents the input feature map, and $k_p(d)$ represents the output of the pth group of convolution kernels, and c represents the bias constant.

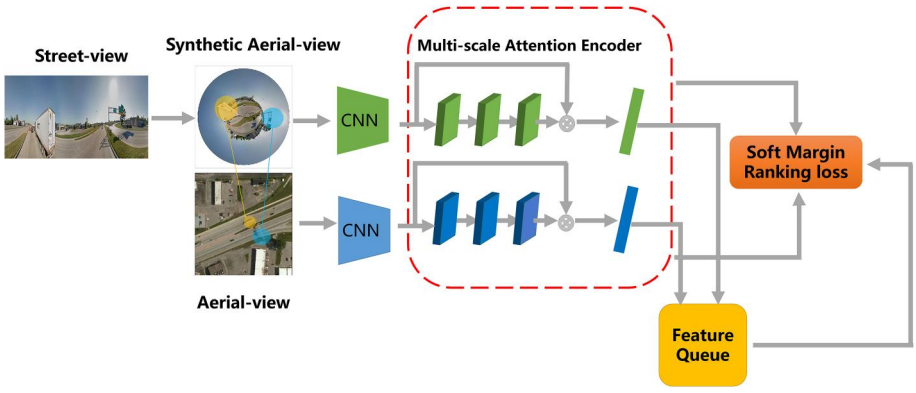

FIGURE 2 The overview of proposed pipeline

After obtaining the multi-scale context information s, a set of 1×1 convolution kernels, which are used to weight and sum the information of each spatial position channel by channel, to obtain an attention mask with a channel number of 1. Finally, we use up-sampling to restore the size of the attention mask to consistent with the size of the feature map, and utilise the restored attention mask to re-weight the feature. The detailed architecture of our MSAE module is shown in Figure 4.

3.4 | Global mining strategy

Due to the limitation of the batch size, the effect of the negative samples within a batch to the loss is gradually tending to 0, and the overall convergence speed of the model gradually slows down. Reference [21] proposed an in-batch re-weighting triplet loss, which gives more weight to hard negative exemplars and less weight to easy exemplars. Reference [22] mined difficult negative exemplars in a batch for training, and found that the difficult exemplar mining strategy can significantly boost the performance. However, the above methods have high requirement for hardware equipment, and are not applicable when the batch size is small. This is because the hard negative exemplars of a batch are difficult to represent the global hard negative exemplars within small batch size. To allow the network to consider the global hard negative exemplars as much as possible during the training process, we use the feature queue to save the global descriptor vector that is obtained from the aerial-view image via the forward propagation of the network. Specifically, to save the computation resources, the queue is set to a fixed length, and only saves the most difficult negative samples of the current batch for loss calculation in subsequent batches. After the feature queue is full, whenever there is a new hard negative exemplar comes to the queue, the sample feature at the head of the queue will be dequeued, so as to ensure the feature encoding in the queue is consistent with the network parameter update. Our network aims to learn feature embeddings to reduce the distance of matched image pairs and push the unmatched image pairs away. Denoting $f(\cdot)$ as a skeleton network with the multi-scale attention encoder, the loss function $\mathcal{L}$ of the entire network is defined as follows:

$$\mathcal{L} = \ln(1 + e^{\alpha(d_{pos} - d_{neg})}) \tag{4}$$

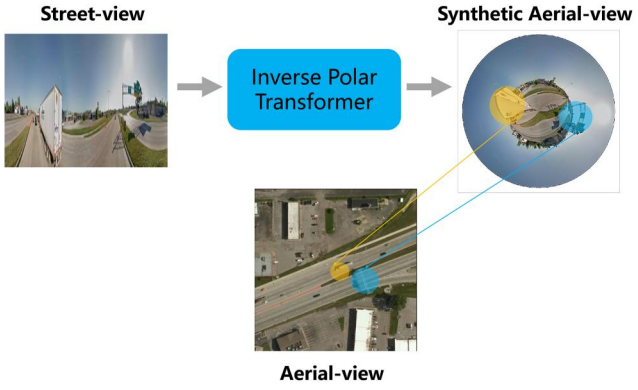

FIGURE 3 Illustration of inverse polar transform, and geometric correspondence between the synthetic and aerial images

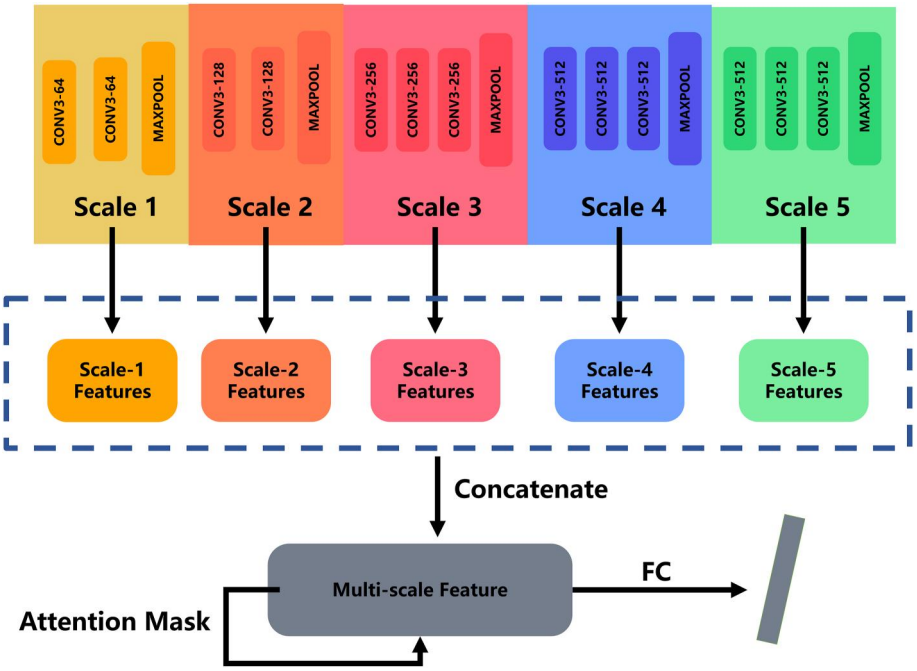

FIGURE 4 Architecture details of the proposed MSAE module. Intermediate feature maps of the backbone are concatenated along with channel to form multi-scale features. Multi-scale features are then input into a lightweight fully convolutional network to get the attention mask, where is the attention mask is used to re-weight the spatial information. MSAE, Multi-scale Attention Encoder

$$d_{pos} = \|f(I_s) - f(I_a)\|_2^2 \quad (5)$$

$$d_{neg} = \|f(I_s) - f(I'_a)\|_2^2 \quad (6)$$

where I_s is the street-view image, I_a is the matched aerial-view image, I'_a is the unmatched aerial-view image from the current batch and the queue of GMS. We set $\alpha = 10$ in this work according to Ref. [22].

4 | EXPERIMENTS AND ANALYSIS

4.1 | Datasets

We do experiments on two benchmark datasets, that is CVUSA [53] and CVACT [23]. CVUSA includes 35, 532 pairs of street-view and aerial-view images for training, and 8884 pairs of images for testing. To be consistent with the former methods, CVACT also provides the same division setting for the training set and the testing set. In these two datasets, to provide as accurate information as possible, all the street-view images are panoramic images and all the aerial-view images are high-resolution images. It is worth noting that the street and aerial view images in these two datasets are captured at different times, which also brings great challenges to this task. Figure 5 shows some samples from the two datasets.

CVUSA In this dataset, the aerial-view images are with geographic coordinates. In addition, CVUSA also provides semantic segmentation labels for the street-view images. Since our proposed method in this work does not rely on any other additional information, this semantic segmentation labels are not used.

FIGURE 5 The illustration of two benchmark datasets. The top two rows are the samples from CVUSA, and the bottom two rows are the samples from CVACT. The left column is the aerial images, and the right column is the corresponding ground-level images

CVACT is targeted for fine-grain and city-scale cross-view localization. Geo-tagged street-view panoramas and satellite images in this dataset are mainly from the Canberra city.

4.2 | Evaluation metrics

For the cross-view geo-localization problem, we use the topk recall as the evaluation metrics. For each query street-view image, if the aerial-view image matched with it is among the first k retrieval results, then it is believed that this retrieval image is correct. Top 1% is a weakly constrained evaluation metric. The existing works [21, 26, 28] have increased the top 1% to more than 95%, so the top 1% is no longer a good indicator. The top-1 accuracy is the ultimate problem that needs to be solved in cross-view geo-localization, that is, given a query image of the ground view, the only matching aerial-view image with geographic coordinates is found in the database. Therefore, top-1 has more practical application significance than top 1%.

4.3 | Comparative results

In this section, we compare the performance of our algorithm with the related prior methods [22, 23, 26, 28, 39, 54] on two benchmark datasets [23, 53]. To ensure the fairness of the comparison, we copy the results from their original reports.

As shown in Table 1, the proposed method significantly outperforms the existing state-of-the-art ground-to-aerial cross-view algorithms under the same network backbone VGG16. In our approach, the network was trained with the soft-margin ranking loss and our designed GMS loss. CVM-Net [22] is a baseline network that only uses the basic soft-margin ranking loss for training. The LPN [39] method calculates the category probability of each street view image to obtain the geographic location, that is, images from different viewpoints of the same location are classified into one category. Regmi [28] exploited the conditional GANs to produce the aerial-view image from the ground-level query. Besides, this approach also used a soft-margin ranking loss for cross-view image matching.

CVFT [26] aimed to transfer features from one domain to another, which conducts more meaningful feature similarity comparison. CVFT and our proposed method both want to reduce the difficulty of network training from the perspective of reducing the domain difference, but the CVFT reduces the domain difference from the feature space while ours from the image space. Since the domain information in the feature space is obtained by down-sampling, some detailed information will be lost, thus aligning the domain information in the image space can retain the detailed information of the original image which is beneficial for boosting the network performance. Orientation [23] tried to fuse the orientation information to improve the network performance. Implicitly using the orientation information requires the network to learn the mapping from street view to aerial view, in contrast, our inverse polar

TABLE 1 Comparison with the state-of-the-arts on the CVUSA and CVACT_val datasets

	CVUSA				CVACT_val			
Method	**r@1**	**r@5**	**r@10**	**r@1%**	**r@1**	**r@5**	**r@10**	**r@1%**
CVM-Net [22]	22.47	49.98	63.18	93.62	20.15	45	56.87	87.57
Orientation [23]	40.79	66.82	76.36	96.12	46.96	68.28	75.48	92.01
Regmi [28]	48.75	-	81.27	95.98	-	-	-	-
CVFT [26]	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [39]	78.48	92.46	95.29	99.12	-	-	-	-
Ours	**81.39**	**94.50**	**96.80**	**99.53**	**71.52**	**87.87**	**91.43**	**97.08**

Note: R@K indicates top-K retrieval rate. The bold value indicates that the method achieves the best performance with this evaluation metric.

transform can explicitly convert the street-view image to aerial-view image, which is useful to alleviate the difficulty of network learning.

4.4 | Ablation studies

We conduct the following ablation studies to test the impact of different components of our method, the corresponding experimental results are shown in Table 2. We show the retrieval recall-k accuracy for different versions of our full network.

Performance of Inverse Polar Transform (IPT). We use the original street-view images and the aerial-view images as input to train the baseline network. On the other side, the baseline with IPT takes the synthetic aerial-view images, which are obtained by applying the inverse polar transform to street-view images, as the input. As shown in Table 2, applying the inverse polar transform to street-view images can improve the top-1 recall by 2.35% (81.39% vs. 79.04%), which means roughly aligning the domain information of the two view images before extracting the description vector can effectively reduce the difficulty of network learning. On the other side, IPT causes a slight drop in the top-1%. This is because the IPT can lead the street-view image to be distorted, while it is difficult for the network to distinguish the images with large distortion.

Performance of Multi-scale Attention encoder (MSAE). To evaluate the effectiveness of our MSAE module, we remove the MSAE module from the final encoder and constructs a plain fully connected network. Both the two encoders are trained with soft margin ranking loss and our GMS loss. As shown in Table 2, the performance is boosted by over 5% (81.39% vs. 75.95%) when using our multi-scale attention module. To show the effect of the proposed multi-scale attention mechanism more intuitively, we visualise some experimental heatmaps in Figure 6. As can be seen, despite there are some confusing objects (such as the tree and car), our method can still retrieve the correct reference image and focus attention on stationary objects for example buildings and roads. The results demonstrate that our MSAE module is effective to alleviate the affect of object distortion caused by inverse polar transform, and allow the network to focus on stationary objects which are useful for distinguishing cross-view images. Although the scenes in the two-view images are shot at different time periods, the stationary objects will not change for a long time period. This allows our network to effectively combat changes in the image over time.

TABLE 2 Ablation study of different components of our proposed model

	CVUSA			
Method	**r@1**	**r@5**	**r@10**	**r@1%**
i. w/o IPT	79.04	93.75	96.47	**99.59**
ii. w/o MSAE	75.95	92.55	95.67	99.35
iii. w/o GMS	78.69	93.64	96.11	99.52
Ours	**81.39**	**94.50**	**96.80**	99.53

Note: The bold value indicates that the method achieves the best performance with this evaluation metric.

Abbreviations: GMS, Global Mining Strategy; IPT, Inverse Polar Transform; MSAE, Multi-scale Attention Encoder.

Performance of global mining strategy (GMS). Hard negative mining aims to put hard exemplars into a special set and then train the model in a targeted manner. As shown in Table 2, results reveal that the global mining strategy is able to mine the global hard negative exemplars effectively, and improves the performance by more than 2% (81.39% vs. 78.69%). This is because when the network accuracy reaches to a high-level, the ordinary exemplars can be easily distinguished by the network. At this time, it is impossible to train the ordinary exemplars to improve the accuracy, and hard negative exemplars become obstacles which limit to further improve the network accuracy. Consequently, using our global mining strategy (GMS) to discover hard negative exemplars can effectively allow the network to break through the bottleneck. As show in Figure 7, the accuracy of the network without GMS can hardly increase after the training for about 80 epochs, while the network with our GMS can continue to learn the information of difficult samples, so that the accuracy can increase steadily. At the beginning of training, due to the hard exemplars contain more nondiscriminatory information, it will bring great obstacles to network training and resulting in low accuracy.

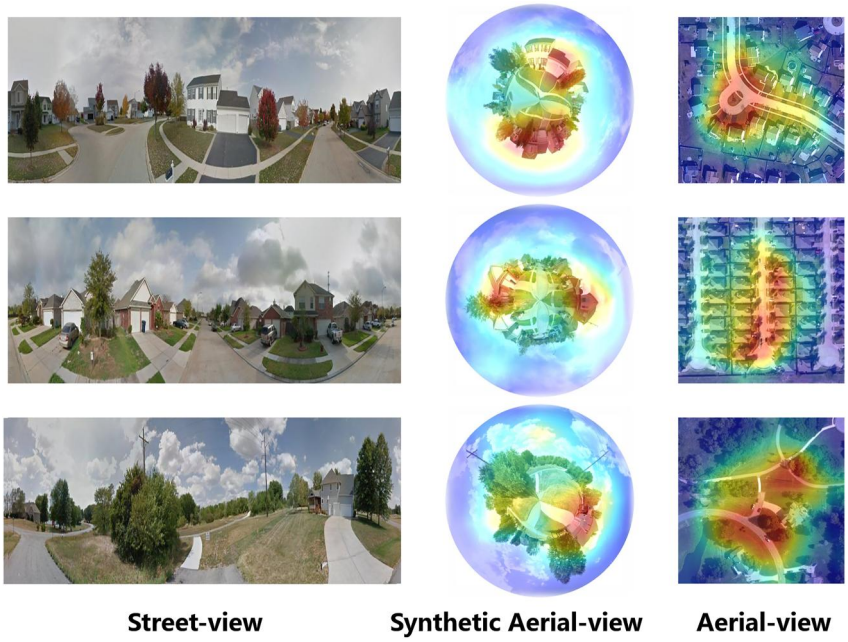

FIGURE 6 Heatmaps generated by the proposed MSAE module. The first column is the original street-view image, the second column is the attention heatmap of the synthetic aerial-view image, and the third column is the attention heatmap of the original aerial-view image. The closer the colour in the heatmaps is to red, the more the model pays attention to the region. MSAE, Multi-scale Attention Encoder

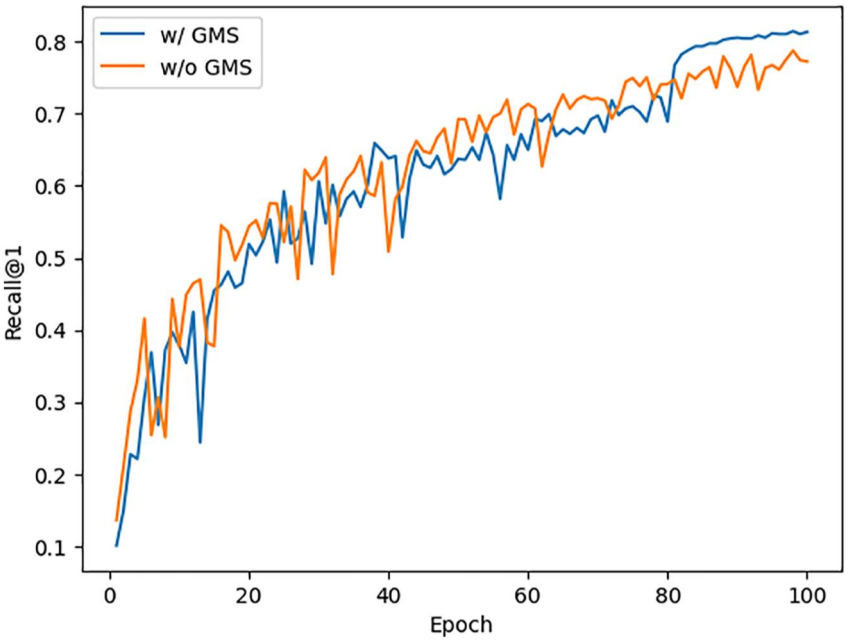

FIGURE 7 Top-1 recall accuracy curve of our baseline w/ and w/o GMS on CVUSA. GMS, Global Mining Strategy

5 | IMPLEMENTATION DETAILS

t-SNE visualization. Our network aims to learn good feature embeddings which can bring matching image pairs closer while push unmatching pairs far away. We use t-SNE [54] to visualise the image global description vector on a two-dimensional plane. As shown in Figure 8, the closer the images in the two-dimensional plane, the more similar their global description vectors are. It is also obvious from the three enlarged windows that our model distinguishes images from different areas very well. For example, the cluster of pictures in the upper left corner can be seen from the scene information that they are all taken from a road covered with more vegetation, while the cluster of pictures in the lower right corner is obviously taken from a desert area with less vegetation.

Sample localization results. To demonstrate the model's ability to retrieve and match cross-view images more intuitively, we also show some localization examples in Figure 9. From left to right are respectively the street-view query image, and the top 1-5 retrieved aerial-view images. It is obvious that the spatial patterns of the top5 images are very similar, with only the subtle difference between the distribution of buildings and the vegetation cover, which demonstrates that our model pays more attention to the unchanging spatial pattern.

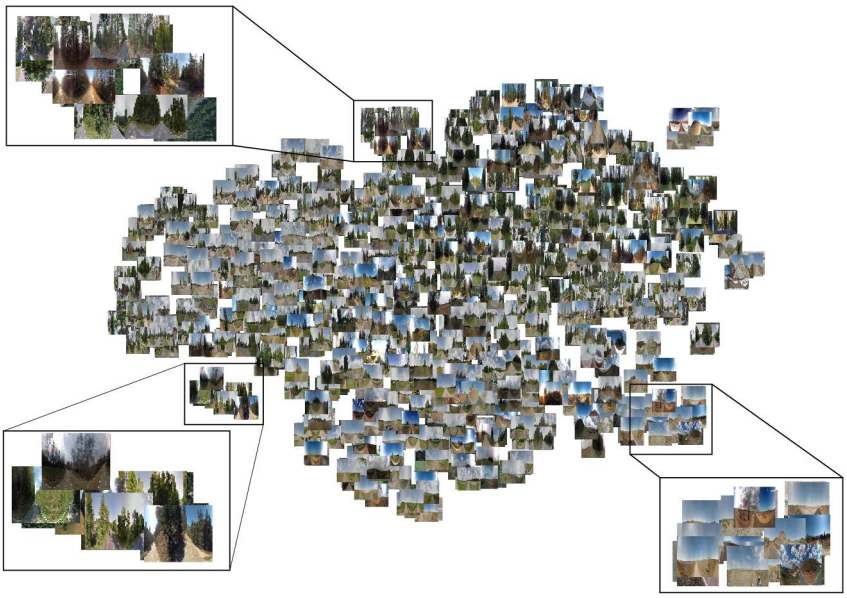

FIGURE 8 t-SNE visualization of the cross-view features learnt by our network. The appearance of the images of each cluster is very similar, which means that our network has learnt the correct discriminative features

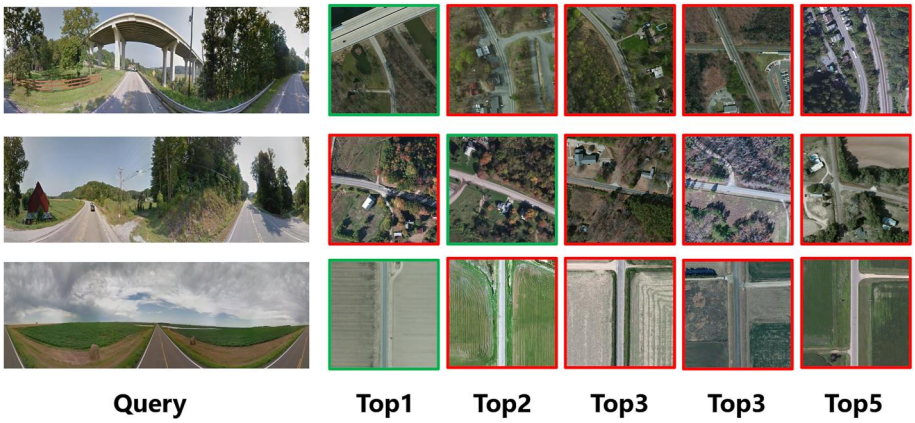

FIGURE 9 Localization results of our method on the CVUSA dataset

Training hyperparameters. We implement our network in PyTorch using Adam optimiser. The momentum parameters $\beta 1$ and $\beta 2$ are respectively set to 0.9 and 0.999, and the initial learning rate is set to 1e-4. We use StepLR scheduler to adjust the learning rate, reduce the learning rate to one-tenth every 80 epochs and train the model for 100 epochs. The resolution of both the synthetic aerial-view image and the original aerial-view image is 288 × 288. We augment the synthetic aerial-view images with random rotation. For the weighted soft-margin loss, we use the exhaustive mini-batch strategy to create the triplets within a batch. For the batch size B (we choose B = 64), we average the triplet loss for all 2B(B-1) combinations of the positive and negative pairs.

6 | CONCLUSION

In this paper, we proposed an efficient method to solve the problem of ground-to-aerial image geo-localization. We exploit three effective strategies that is the inverse polar transform approach, the multi-scale attention mechanism, and the global hard mining scheme, to reduce the domain difference and obtain more robust descriptor vectors. Specifically, the inverse polar transform approach explicitly uses the geometric relationship between the street-view image and aerial-view image, and can reduce the difficulty of the network learning by roughly aligning the domain information between the two. The multi-scale attention mechanism is able to suppress the deformed areas in the image and focuses more on the spatial layout of the image, which can further enable the network to understand the high-level semantic information. Finally, with the usage of our global mining scheme, the network can pay more attention to hard negative exemplars to break through the performance bottleneck. The state-of-the-art experimental results demonstrate the effectiveness of our proposed method.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant 62 106 177. It was also supported by the Central University Basic Research Fund of

China (No.2042020KF0016). The numerical calculation was supported by the supercomputing system in the Supercomputing Center of Wuhan University.

CONFLICT OF INTEREST

None.

DATA AVAILABILITY STATEMENT

Data available on request due to privacy/ethical restrictions.

ORCID

Zhigang Tu https://orcid.org/0000-0001-5003-2260

REFERENCES

1. McManus, C., et al.: Shady dealings: Robust, long-term visual localisation using illumination invariance. In: 2014 IEEE international conference on robotics and automation (ICRA), pp. 901–906. IEEE (2014)
2. Jun-hui, Z., Da-peng, C., Qing, L.: Research status and development trend of vehicle following control system. Computer Science. 47(8), 10 (2020)
3. Sünderhauf, N., et al.: Place recognition with convnet landmarks: viewpoint-robust, condition-robust, training-free. Robotics: Science and Systems. XI, 1–10. (2015)
4. Wei-liang, Z., et al.: Comprehensive review of autonomous taxi dispatching systems. Computer Science. 47(5), 9 (2020)
5. Kaur, A., Sattar, T., Tokhi, M.: Development of a robot for in-service radiography inspection of subsea flexible risers. Journal of Artificial Intelligence and Technology. 1(3) (2021)
6. Middelberg, S., et al.: Scalable 6-dof localization on mobile devices. In: European conference on computer vision, pp. 268–283. Springer (2014)
7. Arandjelovic, R., et al.: NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5297–5307. (2016)
8. Kim, H.J., Dunn, E., Frahm, J.M.: Learned contextual feature reweighting for image geo-localization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3251–3260. IEEE, (2017)
9. Radenović, F., Tolias, G., Chum, O.: CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In: European conference on computer vision, pp. 3–20. Springer (2016)
10. Zamir, A.R., Shah, M.: Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs. IEEE Trans. Pattern Anal. Mach. Intell. 36(8), 1546–1558. (2014)
11. Saurer, O., et al.: Image based geo-localization in the alps. Int. J. Comput. Vis. 116(3), 213–225 (2016)
12. Cao, B., Araujo, A., Sim, J.: Unifying deep local and global features for image search. In: European Conference on Computer Vision, pp. 726–743. Springer (2020)
13. Hausler, S., et al.: Patch-NetVLAD: Multi-Scale Fusion of Locally-Global Descriptors for Place Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14141–14152. (2021)
14. Arandjelovic, R., Zisserman, A.: All about VLAD. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1578–1585. (2013)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60(2), 91–110. (2004)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), Vol. 1, pp. 886–893. IEEE (2005)
17. Tu, Z., et al.: Multi-stream CNN: learning representations based on human-related regions for action recognition. Pattern Recogn. 79, 32–43. (2018)
18. Chang, Y., et al.: Clustering Driven Deep Autoencoder for Video Anomaly Detection. In: European Conference on Computer Vision, pp. 329–345. (2020)
19. Tu, Z., et al.: A survey of variational and CNN-based optical flow techniques. Signal Process. Image Commun. 72, 9–24. (2019)
20. Chen, Y., et al.: Model-based 3D Hand Reconstruction via Self-Supervised Learning. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10451–10460. (2021)
21. Cai, S., et al.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8391–8400. (2019)
22. Hu, S., et al.: 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, pp. 7258–7267. (2018)
23. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5624–5633. (2019)
24. Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: European conference on computer vision, pp. 494–509. Springer (2016)
25. Shi, Y., et al.: Spatial-aware feature aggregation for image based cross-view geo-localization. Adv. Neural Inf. Process. Syst. 32, 10090–10100. (2019)
26. Shi, Y., et al.: Optimal feature transport for cross-view image geo-localization, Aaai. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, pp. 11990, 11997–11997. (2020)
27. Workman, S., Souvenir, R., Jacobs, N.: Wide-area image geolocalization with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3961–3969. (2015)
28. Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 470–479. (2019)
29. Noda, M., et al.: Vehicle ego-localization by matching in-vehicle camera images to an aerial image. In: Asian Conference on Computer Vision, pp. 163–173. Springer (2010)
30. Bansal, M., et al.: Geo-localization of street views with aerial image databases. In: Proceedings of the 19th ACM international conference on Multimedia, pp. 1125–1128. (2011)
31. Senlet, T., Elgammal, A.: A framework for global vehicle localization using stereo images and satellite and road maps. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2034–2041. IEEE (2011)
32. Castaldo, F., et al.: Semantic Cross-View Matching. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops (2015)
33. Lin, T.Y., Belongie, S., Hays, J.: Cross-view image geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 891–898. (2013)
34. Workman, S., Jacobs, N.: On the location dependence of convolutional neural network features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 70–78. (2015)
35. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115(3), 211–252. (2015)
36. Zhou, B., et al.: Learning deep features for scene recognition using places database (2014)
37. Lin, T.Y., et al.: Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5007–5015. (2015)
38. Shi, Y., et al.: Where am i looking at? Joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4064–4072. (2020)
39. Wang, T., et al.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology (2021)
40. Zhu, S., Yang, T., Chen, C.: Revisiting street-to-aerial view image geo-localization and orientation estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 756–765. (2021)

41. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 815–823. (2015)
42. Li, D., et al.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 384–393. (2017)
43. Sun, Y., et al.: Beyond Part Models: Person Retrieval with Refined Part Pooling (and A Strong Convolutional Baseline). In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
44. Xu, Y., Qiu, T.T.: Human activity recognition and embedded application based on convolutional neural network. Journal of Artificial Intelligence and Technology. 1(1), 51–60. (2021)
45. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105. (2012)
46. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164. (2017)
47. Ullah, H., et al.: Comparative study for machine learning classifier recommendation to predict political affiliation based on online reviews. CAAI Transactions on Intelligence Technology. 6(3), 251–264. (2021)
48. Weyand, T., Kostrikov, I., Philbin, J.: Planet-photo geolocation with convolutional neural networks. In: European Conference on Computer Vision, pp. 37–55. Springer (2016)
49. Muller-Budack, E., Pustu-Iren, K., Ewerth, R.: Geolocation estimation of photos using a hierarchical model and scene classification. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 563–579. (2018)
50. Woo, S., et al.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp. 3–19. (2018)
51. He, K., et al.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1904–1916. (2015)
52. Lin, T.Y., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125. (2017)
53. Zhai, M., et al.: Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 867–875. (2017)
54. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. 9(11), 2579–2605. (2008)

How to cite this article: Li, S., et al.: Multi-scale attention encoder for street-to-aerial image geo-localization. CAAI Trans. Intell. Technol. 1–11 (2022). https://doi.org/10.1049/cit2.12077