# A Cross-View Geo-localization Method Guided By Relation-Aware Global Attention

**Jing Sun**
Dalian Minzu University
**Rui Yan**
Dalian Minzu University
**Bing Zhang**
Dalian Minzu University
**Bing Zhu**
Harbin Institute of Technology
**Fuming Sun** ( sunfuming@dlnu.edu.cn )
Dalian Minzu University

**Research Article**

**Keywords:** Cross-view geo-localization, Attention mechanism, Dual-branch structure, Dilated convolution

**Posted Date:** February 27th, 2023

**DOI:** https://doi.org/10.21203/rs.3.rs-2607140/v1



**Additional Declarations:** No competing interests reported.

**Version of Record:** A version of this preprint was published at Multimedia Systems on May 9th, 2023. See the published version at https://doi.org/10.1007/s00530-023-01101-1.

# A Cross-View Geo-localization Method Guided By Relation-Aware Global Attention

Jing Sun[1], Rui Yan[1], Bing Zhang[1], Bing Zhu[2] and Fuming Sun[1*]

[1]School of Information and Communication Engineering, Dalian Minzu University, Liaohe West Road, Dalian, 116600, Liaoning, China.

[2]Department of Information Engineering, Harbin Institute of Technology, Xidazhi Street, Harbin, 150006, Heilongjiang, China.

*Corresponding author(s). E-mail(s): sunfuming@dlnu.edu.cn;

**Abstract**

Cross-view geo-localization mainly exploits query images to match images from the same geographical location from different platforms. Most existing methods fail to adequately consider the effect of image structural information on cross-view geo-localization, resulting in the extracted features can not fully characterize the image, which affects the localization accuracy. Based on this, this paper proposes a cross-view geo-localization method guided by relation-aware global attention, which can capture the rich global structural information by perfectly integrating attention mechanism and feature extraction network, thus improving the representation ability of features. Meanwhile, considering the important role of semantic and context information in geo-localization, a joint training structure with parallel global branch and local branch is designed to fully mine multi-scale context features for image matching, which can further improve the accuracy of cross-view geo-localization. The quantitative and qualitative experimental results on University-1652, CVUSA, and CVACT datasets show that the algorithm in this paper outperforms other advanced methods in recall accuracy (Recall) and image retrieval average precision (AP).

**Keywords:** Cross-view geo-localization, Attention mechanism, Dual-branch structure, Dilated convolution

# 1 Introduction

Cross-view geo-localization can be regarded as a content-based image retrieval task [1, 2], which refers to matching the query image from one platform with the images from other platforms to find the images with the same geographic location. Previous research mainly focused on matching ground views with satellite and aerial images. Recently, drone-view images have been introduced in cross-view geo-localization with the gradual maturity of UAV technology [3], and geo-localization based on drone-view and satellite images has become the current research hotspot.

As convolutional neural networks (CNN) are widely used in visual tasks such as image classification [4, 5], object detection [6, 7], semantic segmentation [8, 9], and action recognition [10, 11], some researchers have applied CNN to cross-view geo-localization [12] and made significant progress. However, most cross-view geo-localization methods mainly consider the high-level semantic information of the target image, ignoring the impact of spatial structure information on improving the accuracy of geo-localization. Zheng et al. [13] regarded geo-localization as a classification task and measured the similarity of the image semantic features. However, this method ignores the context information of the area around the target, resulting in the extracted features are not comprehensive enough. Wang et al. [14] used the square-ring partition strategy to make the network focus on the surrounding area of the target, thus improving the accuracy of geo-localization by exploiting context information. However, this method directly divides the feature map into four scales and ignores the global structure information of the image, which leads to the false detection of similar images as the correct retrieval results in the retrieval process. Obviously, it is helpful to elevate the performance of geo-localization tasks by sufficiently exploring the structural information of the geographic target images.

To alleviate the impacts of existing algorithms that fail to fully consider the image structure information on the matching accuracy of cross-view geo-localization, this paper proposes a cross-view geo-localization method guided by relation-aware global attention. Specifically, this method adopts the deep residual network [15] as the backbone network, and exploits the relation-aware global attention module (RGA) [16] to capture more robust global structure information of the image for image feature matching. Meanwhile, a dual-branch network is designed to capture deep features with rich semantic information and local features with multi-scale context information, respectively. Among them, the local branch employs the dilated convolution [17] to increase the receptive field of the feature map while adopting the square-ring partition strategy [14] to divide the feature map at four scales. Moreover, our method converts the feature map of each branch into a column vector and obtains its prediction category through the classifier. Finally, the cross-entropy loss function [18, 19] is exploited to learn the image prediction category for improving the training accuracy of the network.

The contributions of this paper mainly include the following aspects:

(1) A cross-view geo-localization method guided by relation-aware global attention is proposed. This method exploits the relation-aware global attention module to learn the relationship between image feature nodes, which can sufficiently mine the global structural information of the image, thus extracting more robust features for image feature matching.

(2) A dual-branch structure is designed, where the deep residual network is exploited in the global branch to extract deep features for obtaining the feature maps containing abundant semantic information, while the dilated convolution is employed in the local branch to capture local features with richer multi-scale context information, which further enhances the precision of geo-localization.

(3) The method achieves superior positioning accuracy than other advanced models on the three datasets of Univers- ity-1652, CVUSA, and CVACT, which proves the effectiveness of the proposed method in geo-localization.

The remainder of this paper is organized as follows. Section 2 introduces the related work of cross-view geo-localization. Section 3 details the method and network structure. Section 4 describes the experimental results and analyzes the results, and conducts the ablation experiment and followed by the summary of the full text and prospects for future research directions in section 5.

## 2 Related work

The research content of early geo-localization is mainly based on ground view and aerial view images. Workman et al. [20] adopted two publicly available pre-trained models to extract image features, and proved that deep features can discriminate images from different geographic locations. However, this method only focuses on image feature extraction at a single scale, and fails to effectively utilize the multi-scale information, resulting in insufficient matching features extracted by the network. On this basis, Workman et al. [21] constructed a CVUSA (Cross-View USA) dataset to perform a multi-scale fusion of aerial image features and improved the cross-view localization results. Lin et al. [22] employed publicly available data to build 78,000 street view and 45° aerial image pairs and then adopted the deep siamese network to extract features for conducting cross-view localization. Vo [23] et al. trained the network by exploiting the distance based logistic layer (DBL) and rotation invariance to evaluate different deep learning methods and improve the localization accuracy. Considering that the image semantic information is more robust to viewpoint changes, Tian et al. [24] used object detection technology to extract buildings in the image for building matching, and obtained the final geo-localization results. Altwaijry et al. [25] focused on the matching task of aerial image pairs, and they exploited data-driven methods to learn discriminant representations from image pairs, thus solving the problem of ultra-wide baseline image matching. Furthermore, Zhai et al. [26] first extracted aerial image features, then mapped them to the ground view by employing adaptive transformation, and finally minimized the difference between the semantic feature predicted from the ground view and those directly extracted from the

ground images through an end-to-end learning method. Hu et al. [27] combined the siamese networks with NetVLAD [28] to encode local features for obtaining global descriptors and accelerated network convergence by introducing weighted soft margin ranking loss, thus improving network performance. Shi et al. [29] believed that existing methods ignore the differences in appearance and geometry between ground view and aerial view images, so they utilized the polar coordinate transform to approximately align aerial images with ground view images. In order to further solve the problem of orientation alignment in the cross-view, Shi et al. [30] designed a dynamic similarity matching network (DSM), which makes the image matching results more accurate. Liu et al. [31] believed that geometric cues (such as orientation) can be used for localization, so they designed a siamese network to integrate the orientation information of each pixel from the image into the network model, which can enable the network to learn both appearance and geometric information, and improve the recall accuracy and precision of the network. To solve the problem of scene changes over time, Rodrigues et al. [32] proposed a semantic-driven data augmentation technique aimed at simulating the phenomenon of the scene change in cross-view image matching, and then employed the multi-scale attention module to match the image, and improved the network performance. Regmi et al. [33] first applied generative adversarial networks (GANs) [34] to cross-view localization, and synthesized aerial images from ground views by using GANs for image matching, but this method is not an end-to-end method. Toker et al. [35] employed polar coordinate transformation on satellite views to synthesize the ground views followed by image retrieval, and achieved advanced geo-localization performance by integrating the two steps in an end-to-end architecture. The above methods mainly focus on the matching task between ground view and aerial view images, they only consider two views for geo-localization and do not pay attention to the drone-view image, so the feature learning of multi-view matching task is ignored.

Recent research on cross-view geo-location believes that adding the viewpoint can improve the accuracy of geo-localization, so the drone images are introduced to solve geo-localization problem. Zheng et al. [13] constructed the University-1652 dataset, including satellite view images, ground view images, and drone view images, and they considered all view images at the same location as a category to complete the geo-localization task in a classified manner, while optimizing the model by applying the instance loss [36]. Nevertheless, this method only concentrates on the semantic information and does not consider the impact of the detailed information on cross-view geo-localization. To solve this problem, Wang et al. [14] proposed a local pattern network (LPN), which takes the contextual information of the image as an auxiliary clue and divides the feature image to make the network notice the environment around the target building, thus effectively solving the problems of ignoring the image details in the method [13] and achieving better matching results. Ding et al. [37] adopted the location classification (LCM) to achieve image matching, which solves the problem of sample imbalance between satellite images and drone

images and improves the image matching accuracy. The attention mechanism has been widely applied in the field of computer vision [16, 38–40], which aims to enable the network to pay more attention to discriminative features while filtering out some irrelevant information, thereby improving the training effect of the model. Zhang et al. [16] integrated relation-aware global attention into the person re-identification network, which enhances the feature representation ability by capturing the global structural information of the image, and improves the performance of person re-identification. In order to avoid the impact of target offset and view scaling on image matching, Zhuang et al. [38] proposed a multi-scale block attention (MSBA) structure to enhance the salient features of different regions. Lin et al. [39] introduced the unit subtraction attention module (USAM), which makes the model focus on the salient areas in the image by detecting key points in the feature map, and improves the performance of the model with fewer parameters. Dai et al. [40] believed that some operations based on CNN would lead to the loss of fine-grained image information, so the Transformer structure [41] is introduced in the cross-view localization and designed the feature segmentation and region alignment method (FSRA), which segments the feature map into different regions on the basis of the heat distribution for classifying and supervising each region, thus effectively realizing the cross-view localization.

The above methods provide a new research idea for solving the problem of inaccurate geo-localization. Inspired by this, this method fully combines the attention mechanism with the feature extraction network to mine structural information from a global perspective. Meanwhile, the dual-branch structure is designed for joint training, and the dilated convolution is fused in the local branch to increase the receptive field of the feature map, which can capture richer multi-scale context information and further improve the accuracy of cross-view localization.

# 3 Method

## 3.1 Overview architecture

The overview framework is shown in Figure 1. The entire network structure is divided into the global branch and the local branch, which share the network weights. First, this model employs ResNet50 as the backbone network while removing the average pooling layer and classification layer, and then extracts features of the input images. At the same time, the relation-aware global attention module is introduced after extracting the shallow features, which can sufficiently capture the global structure information of the image. Then, a dual-branch structure is exploited to process the output features of the previous stage respectively, which can effectively focus on global and local information. Among them, the global branch is adopted to extract the high-level semantic information of the whole image, while the local branch is employed to focus on the context features of the network, thereby retaining more image detail

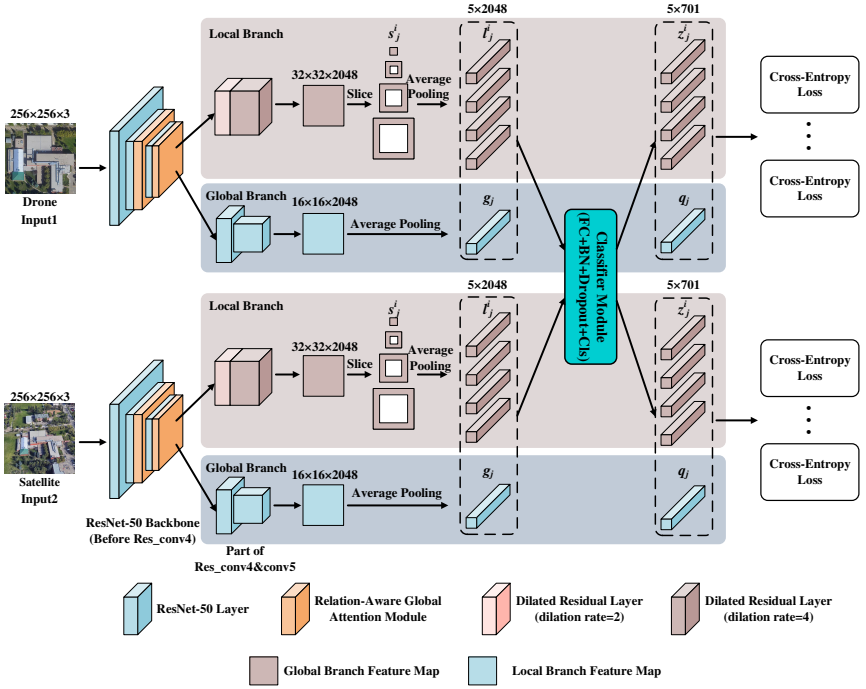


**Fig. 1** The framework of the proposed method.

information. Meanwhile, so as to combine the valuable environmental information of the region around the building, the feature map is divided into four distinct regions by using a square-ring partition strategy in the local branch. Finally, the image high-level features are converted into column vector descriptors through global average pooling. The classifier module is utilized in the training process to get the predicted category probability of each column vector descriptor, and the cross-entropy loss function is employed to minimize the difference between the predicted class and the true one. The Euclidean distance is adopted to calculate the similarity between the query image and the database image during the test, and finally, the retrieved images are sorted according to the similarity.

## 3.2 Relation-aware global attention

In the cross-view geo-localization task, the RGA module can make the network notice the differences in image features to help identify buildings with a similar appearance. This paper combines the RGA module with the deep residual network to construct a feature extraction network guided by relation-aware global attention, which calculates the attention weights by learning the relationship between feature nodes, thus making the network sufficiently mine the feature of discriminant region. The relation-aware global attention is shown in Figure 2. The feature vector in the feature map is represented as the feature nodes

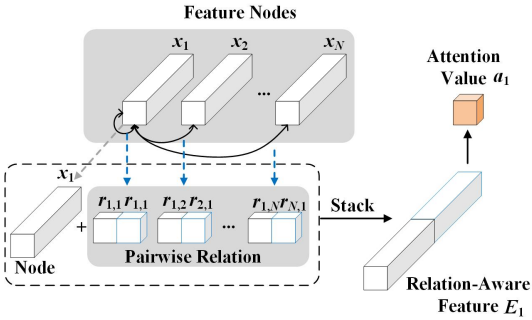


**Fig. 2** The structure of relation-aware global attention.

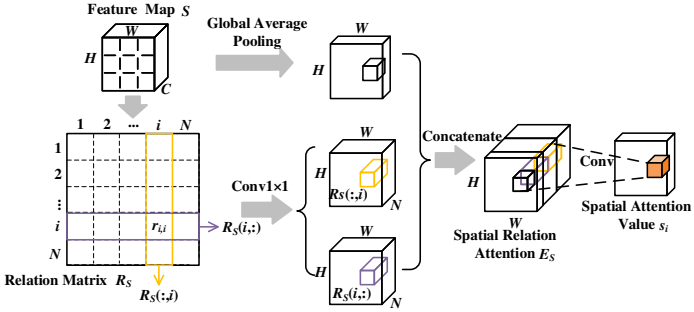


**Fig. 3** The structure of spatial relation-aware global attention.

$x_i$, where $i = 1, 2, \cdots N$, and $N$ is the number of feature nodes. For a feature node $x_i$, calculate the correlation relationships $r_{i,j}$ and $r_{j,i}$ between the current node and all other nodes, where $j = 1, 2, \cdots N$, thus the relationship vector of the feature node $x_i$ is $r_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,N}, r_{1,i}, r_{2,i}, \cdots, r_{N,i}]$. Then, the feature node $x_i$ and the relationship vector $r_i$ are concatenated to acquire the relation-aware feature $E_i$, and the attention weight $a_i$ of the current feature node is inferred.

### 3.2.1 Spatial relation-aware global attention

The spatial relation-aware global attention (RGA-S) learns the correlations among all feature nodes in the spatial dimension of the feature map to enable the network to capture the features of the salient target. The RGA-S is shown in Figure 3.

Specifically, for the feature map $S \in \mathbb{R}^{C \times H \times W}$ obtained from the neural network, the $C$-dimensional feature vector of each spatial position is taken as a feature node to form a graph $G_s$ with a total of $N = W \times H$ nodes and the feature node is represented as $x_i$, where $i = 1, 2, \cdots N$. The correlation $r_{i,j}$ between the feature nodes $x_i$ and $x_j$ can be obtained through the dot product

operation, which can be defined as Equation (1):

$$r_{i,j} = f_s\left(x_i, x_j\right) = (\mathrm{Re}LU(BN(Conv(x_i))))^T(\mathrm{Re}LU(BN(Conv(x_j)))), \quad (1)$$

where $f_s(\cdot)$ represents the dot product operation, $\mathrm{Re}LU(\cdot)$is the modified linear unit activation function, $BN(\cdot)$denotes the batch normalization layer, $Conv(\cdot)$ represents the 1×1 convolution operation, and the dimensionality reduction ratio is controlled by a predefined positive integer. Similarly, the correlation $r_{j,i}$ between feature nodes $x_j$ and $x_i$ can be obtained, and $(r_{i,j}, r_{j,i})$ is used to represent the pairwise relationship between feature nodes $x_i$ and $x_j$. Finally, the correlation between all nodes can be represented by a relation matrix $R_S \in \mathbb{R}^{N\times N}$, where $r_{i,j} = R_S(i,j)$.

Stack the relationships between the $i^{th}$ feature node and all nodes in a fixed order to obtain the spatial relationship vector $r_i = [R_S(i,:), R_S(:,i)] \in \mathbb{R}^{2N}$, where $R_S(i,:)$ represents the correlation between the $i^{th}$ feature node and all nodes, and $R_S(:,i)$ represents the correlation between all nodes and the $i^{th}$ node. In order to enable the network to sufficiently exploit the global structural information, the spatial relationship vector $r_i$ is concatenated with the feature node itself $x_i$ to get the spatial relation-aware feature $E_s$, which can be defined as Equation (2):

$$E_s = C(x_i, r_i) = (pool_c(\mathrm{Re}LU(BN(Conv(x_i)))), (\mathrm{Re}LU(BN(Conv(r_i))))), \quad (2)$$

where $C(\cdot)$ represents concatenation operation, $pool_c(\cdot)$ represents the global average pooling on the channel dimension, reducing the channel dimension to 1. The spatial attention weight $a_i$ can be calculated through $E_s$, which is defined as Equation (3):

$$a_i = sigmoid(BN(Conv_2(\mathrm{Re}LU(BN(Conv_1(E_s)))))), \quad (3)$$

where $sigmoid(\cdot)$ represents *sigmoid* activation function, $Conv_2(\cdot)$ converts the number of channels to 1, and $Conv_1(\cdot)$ reduces dimensions at a fixed ratio.

### 3.2.2 Channel relation-aware global attention

The channel relation-aware global attention (RGA-C) learns the correlations between all feature nodes in the channel dimension of the feature map to assign different weights for each channel. The RGA-C is shown in Figure 4.

Specifically, for the acquired feature map $S \in \mathbb{R}^{C\times H\times W}$, the feature map on each channel is considered as a feature node to form a graph $G_C$ with a total of $C$ nodes, and each feature node is denoted as $x_i$, where $i = 1, 2, \cdots C$.

For the input feature graph $S$, it is first compressed into $S' \in \mathbb{R}^{(HW)\times C\times 1}$ in space, and then the correlation $r_{i,j}$ between the feature node $x_i$ and $x_j$ can be obtained similar to the RGA-S, which is defined as Equation (4):

$$r_{i,j} = f_c\left(x_i, x_j\right) = (ReLU(BN(Conv(x_i))))^T(ReLU(BN(Conv(x_j)))), \quad (4)$$

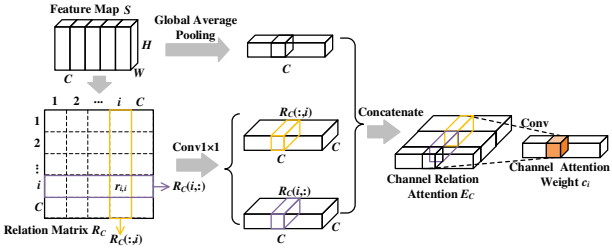


**Fig. 4** The structure of channel relation-aware global attention.

where $f_c(\cdot)$ represents the dot product operation. Similarly, the correlation $r_{j,i}$ between the feature nodes $x_j$ and $x_i$ can be obtained, and the pairwise relationship between all nodes is expressed by matrix $R_C \in \mathbb{R}^{C\times C}$. The relationship between the $i^{th}$ feature node and all nodes is stacked to obtain the channel relationship vector $r_i = [R_C(i,:), R_C(:,i)] \in \mathbb{R}^{2N}$, which is similar to Equation (2)(3), and the final channel attention weight $c_i$ can be obtained.

## 3.3 Local branch

Since the rich multi-scale context information and detailed spatial structure information can assist the network to match the image of the same geographical place to optimize the precision of cross-view geo-localization, the dilated convolution [17] with multiple dilation factor is adopted in the local branch to increase the receptive field of the feature map without losing image details, thereby the model can capture more robust multi-scale information. Meanwhile, the feature map is divided into four scales by using the square-ring partition strategy to obtain rich spatial context information.

The dilated convolution expands the receptive field of the convolution kernel by inserting $r-1$ values with weight 0, where $r$ is the dilation factor. The structure of the standard and dilated convolutions is shown in Figure 5, where (a) represents the standard convolution and (b) represents the dilated convolution with the dilation factor 2. Using the convolution kernel of 3×3 under the same conditions, the receptive fields of the standard and dilated convolutions are 3×3 and 5×5, respectively. Compared with standard convolution, the dilated convolution can capture richer multi-scale context information for image matching.

Specifically, this module employs the dilated convolution with dilation factors 2 and 4 to increase the receptive field of the feature map, and the stride of both the convolutional layer and the downsampling layer in the last residual block of ResNet50 is adjusted to 1. When the resolution of the input image is 256×256, the resolution of the feature image output by the backbone network is 8×8, while that of the output feature image using the dilated residual network is 32×32.

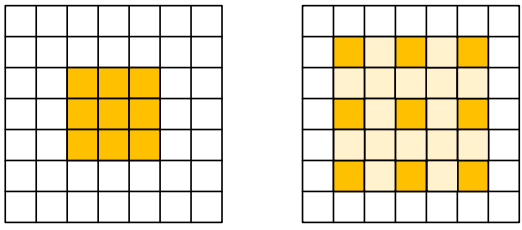


**Fig. 5** The standard convolution and the dilated convolution. (a) represents the standard convolution; (b) represents the dilated convolution with dilation factor 2.

To help the network better discriminate the images in different geographical locations, the environment around the target building is used as auxiliary information. Meanwhile, the feature image is divided into four parts by adopting the square-ring partition strategy in the local branch, to obtain the feature representation of distinct regions. Then, the obtained image features are converted into 2048-dimensional feature vectors through the average pooling operation, represented by Equation (5):

$$l_j^i = Avgpool\left(s_j^i\right), \tag{5}$$

where $Avgpool(\cdot)$ represents the average pooling operation, $s_j^i (i \in [1,4];\ j \in [1,2])$ denotes the feature maps of the local branch divided in the different view platforms, and $l_j^i$ represents the 2048-dimensional feature vector of the four local branches after pooling.

## 3.4 Global branch

Since the semantic information focused by the deep network is also an important part of the cross-view geo-localization task, a global branch structure is designed which is parallel to the local branch. The deep residual network is exploited in the global branch to extract and refine the large-scale features for obtaining the feature map $f_j$ containing rich semantic information. Then, the average pooling method is applied to obtain the 2048-dimensional feature vector and enables the network to recognize the categories of image features, which is expressed by Equation (6):

$$g_j = Avgpool\left(f_j\right), \tag{6}$$

where $g_j$ denotes the feature vector of the global branch after pooling.

## 3.5 Classification of learning and loss function

The classifier module is introduced after the feature extraction stage to predict the category of each feature vector, where the classifier consists of a fully

connected layer (FC), a batch normalization layer (BN), a dropout layer, and a classification layer (Cls). This module takes the local feature vector $l_j^i$ and the global feature vector $g_j$ as input to predict the category to which each feature vector belongs, and finally obtains the local and global prediction probability distribution vector $z_j^i$ and $q_j$ respectively.

The method adopts cross-entropy loss as a loss function to measure the distribution difference between the image predicted probability and the real one, which can learn more robust image features and enhance the network training accuracy. The cross-entropy loss can be expressed by Equation (7)):

$$Loss = \sum_{i,j} -\log(\hat{p}(y|x_j^i)) - \sum_{j} \log(\hat{q}(y|x_j)), \tag{7}$$

where $x_j^i (i \in [1,4];\ j \in [1,2])$ denotes the corresponding original image after segmentation, $x_j (j \in [1,2])$ represents the input image, $j = 1$ represents the UAVs platform, and $j = 2$ represents the satellite platform; $y$ denotes the true category of the input image, $\hat{p}(y|x_j^i)$ and $\hat{q}(y|x_j)$ respectively represent the normalized probability score of $x_j^i$ and $x_j$ belonging to the true category, which is defined by Equation (8) and Equation (9),

$$\hat{p}(y|x_j^i) = \frac{\exp(z_j^i(y))}{\sum_{c=1}^{C} \exp(z_j^i(c))}, \tag{8}$$

$$\hat{q}(y|x_j) = \frac{\exp(q_j(y))}{\sum_{c=1}^{C} \exp(q_j(c))}, \tag{9}$$

where $C$ represents the number of all geo-tagged categories in the database.

# 4 Experiments

## 4.1 Datasets

In this paper, three datasets of University-1652 [13], CVUSA [21], and CVACT [31] are exploited to train and test the proposed method.

(1) University-1652 is a multi-view and multi-source dataset, including drone view, satellite view, and ground view images of 1652 buildings in 72 universities, and the images in the training dataset and the test dataset are not duplicates. Our method uses this dataset to study the two tasks of drone-view target localization and drone navigation. There are 701 image categories of drone view query images in the drone-view target localization task, and each category corresponds to a real matching satellite image. In the drone navigation task, there are a total of 701 image categories in the satellite view query dataset, and each category corresponds to 54 real matching drone images.

(2) CVUSA dataset includes satellite and panoramic ground images, in which there are 35532 training image pairs and 8884 test image pairs.

(3) CVACT is a larger benchmark dataset, which also includes 35532 training image pairs, except that 8884 image pairs are employed for validation and additional 92802 image pairs are used as a test set.

## 4.2 Experimental detail

Our method is implemented on the Linux server with the Ubuntu20.04 operating system, and all performance comparisons are based on the results under this configuration. The server configuration is GTX 3090 GPU with 24G memory capacity. The proposed model is implemented based on the Pytorch framework. Before training, the size of all input images is adjusted to 256×256, and horizontal flipping and random rotation are used for data augmentation. The SGD optimizer with 0.9 momentum and 0.0005 weight decay is adopted to update the model and we set the initial learning rate to 0.001. To accelerate the network convergence, the model training epoch is 140 for University-1652 dataset, and 100 for CVUSA and CVACT datasets. During testing, the feature vectors of each branch are spliced to obtain the final feature representation, so as to complete the image matching.

## 4.3 Performance comparison

### 4.3.1 Quantitative comparison

In this paper, recall accuracy at top K (Recall@K) and image retrieval average precision (AP) are adopted as the performance metrics of image retrieval. Recall@K refers to the ratio of the true-matched images in the top k retrieved results to all the real matching images in the database. In this paper, the case of k=1 is mainly considered. AP refers to the ratio of the real matching images retrieved to the total number of retrieved results. The larger the value of Recall@K and AP, the higher the precision of image retrieval.

We compare our method with other CNN-based algorithms on three datasets, University-1652, CVUSA, and CVACT. Among them, the comparison results on the University-1652 dataset are shown in Table 1. The comparison methods include Instance Loss [13], LCM [37], LPN [14], Instance Loss+USAM [39], LPN+USAM [39].

As can be seen from Table 1, our method achieves the best results in both tasks on the University-1652 dataset. In the task of drone-view target localization, i.e., Drone→Satellite, the performance of the proposed method on R@1 and AP reached 81.06% and 83.74%, respectively, our method achieves the improvement of 3.99% and 3.65% on each of the two metrics compared to the suboptimal method LPN+USAM [39]. In the drone navigation task, i.e., Satellite→Drone, the performances of 89.58% and 79.63% are achieved on the R@1 and AP, respectively. Compared with the suboptimal method LPN [14], it has improved by 3.13% and 4.84% in the two metrics, which proves that our method has significant advantages in image retrieval performance.

The comparison with other approaches on the CVUSA and CVACT_val datasets are detailed in Table 2. Since the ground images in these two datasets

**Table 1** Quantitative test results on University-1652 dataset. Where method [22, 27, 42] is the result obtained by replacing the loss function on the basis of method [13]. The optimal and suboptimal results of the evaluation indicators are indicated in red and blue font, respectively.

| Method | Drone→Satellite | | Satellite→Drone | |
|---|---|---|---|---|
| | R@1↑ | AP↑ | R@1↑ | AP↑ |
| Instance Loss [13] | 58.49 | 63.13 | 71.18 | 58.74 |
| Contrasive Loss [22] | 52.39 | 57.44 | 63.91 | 52.24 |
| Triplet Loss(M=0.3) [42] | 55.18 | 59.97 | 63.62 | 53.85 |
| Triplet Loss(M=0.5) [42] | 53.58 | 58.60 | 64.48 | 53.15 |
| Soft Margin Triplet Loss [27] | 53.21 | 58.03 | 65.62 | 54.47 |
| LCM [37] | 66.65 | 70.82 | 79.89 | 65.38 |
| LPN [14] | 75.93 | 79.14 | 86.45 | 74.79 |
| Instance Loss+USAM [39] | 65.63 | 69.68 | 78.32 | 64.87 |
| LPN+USAM [39] | 77.07 | 80.09 | 85.16 | 74.06 |
| Ours | 81.06 | 83.74 | 89.58 | 79.63 |

**Table 2** Quantitative test results on CVUSA and CVACT_val datasets. The optimal and suboptimal results of the evaluation indicators are indicated in red and blue font, respectively. † represents that the method uses additional orientation information.

| Method | CVUSA | | CVACT_val | |
|---|---|---|---|---|
| | R@1↑ | R@Top1%↑ | R@1↑ | R@Top1%↑ |
| CVM-Net [27] | 18.80 | 91.54 | 20.15 | 87.57 |
| Orientation† [31] | 27.15 | 93.91 | 46.96 | 92.04 |
| Instance Loss [13] | 43.91 | 91.78 | 31.20 | 85.27 |
| Regmi [33] | 48.75 | 95.98 | - | - |
| Siam-FCANet [43] | - | 98.30 | - | - |
| CVFT [12] | 61.43 | 99.02 | 61.05 | 95.93 |
| LPN [14] | 85.79 | 99.41 | 80.37 | 96.40 |
| Instance Loss+USAM [39] | 52.50 | 96.52 | 40.53 | 89.12 |
| LPN+USAM [39] | 85.97 | 99.43 | 80.50 | 96.37 |
| Ours | 88.00 | 99.47 | 80.98 | 96.53 |

are panoramic images, a sequential partition strategy [14] is adopted to divide the images. The comparison methods include the CVM-Net [27], Orientation† [31], Instance Loss [13], Regmi et al. [33], Siam-FCANet [43], CVFT [12], LPN [14], Instance Loss+USAM [39], LPN+USAM [39], where the results of LPN [14], LPN + USAM [39]algorithms are generated by adopting the publicly released codes for training, while the other methods directly use the results provided by the authors.

It can be observed that on the CVUSA dataset, the proposed method achieves 88.00% and 99.47% on the evaluation indicators R@1 and R@Top1%, respectively. Compared with the other nine advanced models, this method has achieved evident promotion in retrieval performance, especially in the R@1 indicator, the performance is improved by 2.03%. On the CVACT_val dataset, the proposed method reached 80.98% and 96.53% on R@1 and R@Top1%, both of which achieved optimal results, thus demonstrating the effectiveness of our method.

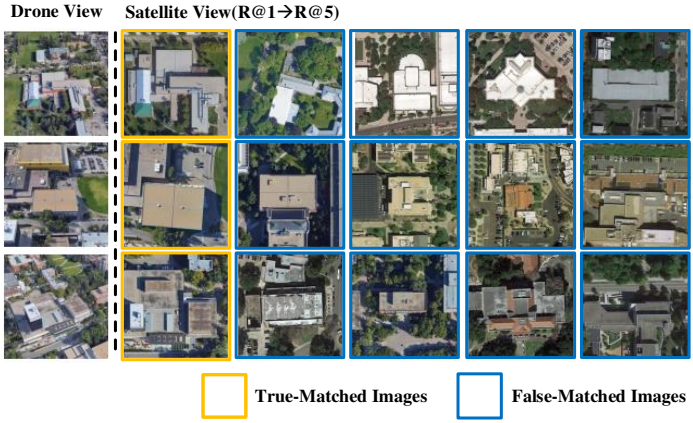


**Fig. 6** University-1652 (Drone Localization).

Except for integrating the relation-aware global attention in the feature extraction network to capture the rich global structural information, our method also designs a joint training structure with parallel global branch and local branch to fully mine multi-scale context features, which is the key that the proposed method outperforms other cross-view geo-localization models.

### 4.3.2 Qualitative results

Figures 6 and 7 are the retrieval results of our method on the University-1652 dataset, which respectively visualizes the results from the tasks of drone-view target localization and drone navigation; and Figure 8 is the retrieval results on the CVUSA dataset. In these qualitative results, each row represents the retrieval result of a position, the first image is the query image, and the top images in the matching results are shown on the right side of the dotted line, where the yellow box represents the true retrieval and the blue box denotes the false retrieval.

For the drone-view target localization task, there is only one truly matched image in the first five images that showed the matching results in Figure 6, this is because each drone view image has only one matching satellite image, which proves that our method can correctly retrieve the matched image under the interference of similar images. For the drone navigation task, the top five images of the matching results in Figure 7 are all correctly matched images, because each satellite image has 54 drone view images matched with it. Since each ground image in the CVUSA dataset corresponds to one real satellite image, the first image in the retrieval result of each query image in Figure 8 is the correctly matched image. Through the analysis of qualitative results, it can be found that our method can retrieve the correct results on both datasets, which further demonstrates the effectiveness of our method.

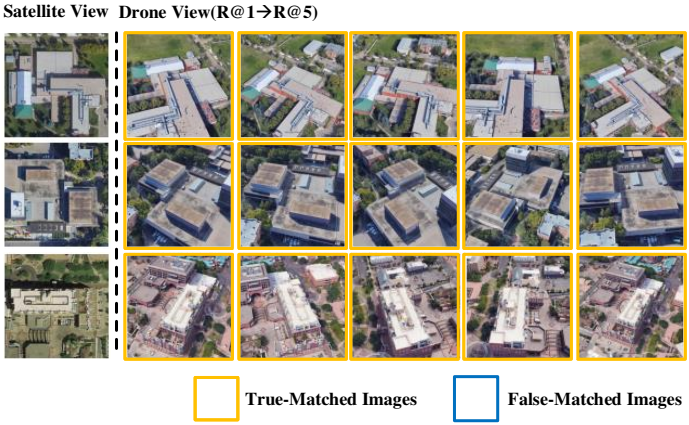


**Fig. 7** University-1652 (Drone Navigation).

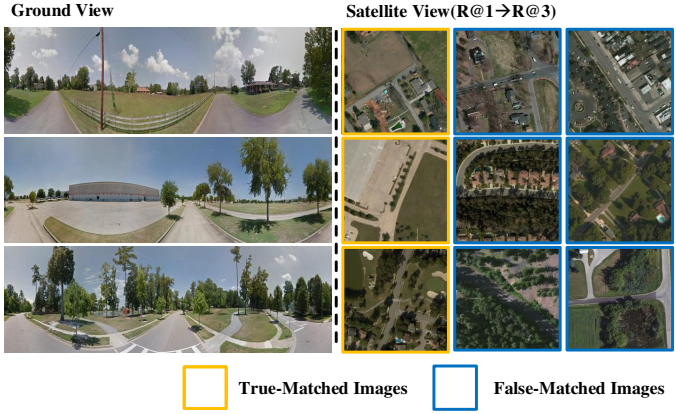


**Fig. 8** CVUSA.

## 4.4 Ablation experiment

To verify the effectiveness of each component, we conduct several ablation experiments on the University-1652 dataset.

### 4.4.1 Effectiveness of the relation-aware global attention

To verify the effectiveness of the relation-aware global attention module, two ablation experiments are conducted in this subsection. The first experiment is to remove the relation-aware global attention module and only use the network with a dual-branch structure for image feature extraction. The second experiment is to add the SE attention module in SENet [44] to the network based on the first experiment for obtaining the attention of the image in the channel dimension.

**Table 3** Comparison results of ablation experiments on the relation-aware global attention module. (a) indicates without attention module; (b) indicates the addition of the SE attention module; (c) indicates the addition of the RGA module. The best results are represented in red font.

| Method | Drone→Satellite | | Satellite→Drone | |
|---|---|---|---|---|
| | R@1↑ | AP↑ | R@1↑ | AP↑ |
| (a) Without Attention Module | 76.95 | 80.02 | 84.88 | 75.84 |
| (b) +SE Attention Module | 79.16 | 82.04 | 89.44 | 78.65 |
| (c) +RGA(ours) | 81.06 | 83.74 | 89.58 | 79.63 |

**Table 4** Comparison results of ablation experiments on the dilated convolution. (a) indicates the dilation factors in the local branch residual blocks are 1 and 1 respectively; (b) indicates the dilation factors are 1 and 2 respectively; (c) indicates the dilation factors are 2 and 2 respectively; (d) indicates the dilation factors are 2 and 4 respectively. The best results are represented in red font.

| Method | Drone→Satellite | | Satellite→Drone | |
|---|---|---|---|---|
| | R@1↑ | AP↑ | R@1↑ | AP↑ |
| (a) (1,1) | 79.02 | 82.80 | 87.73 | 77.94 |
| (b) (1,2) | 80.04 | 82.80 | 88.02 | 78.18 |
| (c) (2,2) | 80.77 | 83.07 | 89.02 | 79.20 |
| (d) (2,4) (ours) | 81.06 | 83.74 | 89.58 | 79.63 |

According to the results of Table 3, it can be observed that compared with not adding any attention mechanism and adding an SE attention module, using the relation-aware global attention module can make the network pay attention to the discriminative features of the image while capturing more robust global structure information, which improves the retrieval ability of the network and achieves the better performance.

### 4.4.2 Effectiveness of the dilated convolution

In this part, we conducted three ablation experiments to verify the effectiveness of the dilation convolution, that is, we adjusted the dilated convolution in the local branch residual blocks and adopted different dilation factors to extract image features. It can be seen from the results in Table 4 that increasing the receptive field of the feature map by using the dilated convolution can effectively capture more detailed information of the image and mine the potential features. When the dilation factors are 2 and 4, respectively, the performance of the model is optimal.

### 4.4.3 Effectiveness of the dual-branch structure

The dual-branch structure is an important component of this method. Therefore, two ablation experiments are conducted to verify its effectiveness, that is, using different branches to extract features for subsequent matching. According to the results in Table 5, it can be found that using the dual-branch for joint training can adequately exploit the semantic and multi-scale context information of the image, thus obtaining the optimal retrieval performance.

**Table 5** Comparison results of ablation experiments on the dual-branch structure. (a) indicates the local branch structure for training; (b) indicates the global branch structure for training; (c) indicates the dual-branch structure for joint training. The best results are represented in red font.

| Method | Drone→Satellite | | Satellite→Drone | |
|---|---|---|---|---|
| | R@1↑ | AP↑ | R@1↑ | AP↑ |
| (a) Local Branch | 75.77 | 78.87 | 86.59 | 75.49 |
| (b) Global Branch | 58.96 | 63.54 | 74.32 | 57.56 |
| (c) Dual-branch (ours) | 81.06 | 83.74 | 89.58 | 79.63 |

**Table 6** The effect of input images with different resolutions on the performance. (a) indicates that the input image size is 224×224; (b) indicates that the input image size is 256×256; (c) indicates that the input image size is 320×320; (d) indicates that the input image size is 384×384. The best results are represented in red font.

| Image Size | Drone→Satellite | | Satellite→Drone | |
|---|---|---|---|---|
| | R@1↑ | AP↑ | R@1↑ | AP↑ |
| (a) 224×224 | 77.69 | 80.64 | 88.02 | 77.15 |
| (b) 256×256 | 81.06 | 83.74 | 89.58 | 79.63 |
| (c) 320×320 | 83.85 | 86.10 | 91.73 | 82.72 |
| (d) 384×384 | 83.61 | 85.92 | 90.58 | 82.10 |

### 4.4.4 Effect of the input image size on the results

In real-world applications, training models with high-resolution images can achieve better accuracy but require more computational resources and time. Due to limited resources, it is necessary to exploit low-resolution input images in the actual operation, which will reduce the accuracy of image matching. Therefore, this paper designs a set of ablation experiments to observe the impact of input images with different resolutions on the model performance. It can be found from the results in Table 6 that while increasing the size of input image from 224 to 320, the R@1 and AP values of the network are both improved; and when the image size increases to 384, the performance of the network decreases slightly.

## 5 Conclusion

In this paper, we proposed the cross-view geo-localization method guided by relation-aware global attention, which exploits the relation-aware global attention to capture the global structural information and extract more robust image features for geo-localization. Meanwhile, the dual-branch strategy is designed for joint training, and the dilated convolution is adopted in the local branch to increase the receptive field of the feature map while dividing the feature map into four scales, which obtains the feature representation containing semantic and context information to calculate the image category probability, and higher accuracy is obtained in geo-localization. The experimental results show that on the three datasets of University-1652, CVUSA, and CVACT, our method has achieved significant improvements in both Recall@K and AP. In

addition, the algorithm can also avoid the interference of similar buildings and retrieve the correct image. In future research, we will consider the complexity of images in real scenes, further study the cross-view geo-localization methods that can adapt to complex scenes, and explore approaches to effectively elevate the precision of geo-localization.

# 6 Acknowledgments

This work was partly supported by the National Natural Science Foundation of China under Grant 61976042 and 61972068, the Innovative Talents Program for Liaoning Universities under Grant LR2019020, the Liaoning Revitalization Talents Program under Grant XLYC2007023, and the Applied Basic Research Project of Liaoning Province under Grant 2022JH2/101300279.

# References

[1] Wang, Z., Qin, J., Xiang, X., Tan, Y.: A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing. Multim. Syst. **27**(3), 403–415 (2021)

[2] Saritha, R.R., Paul, V., Kumar, P.G.: Content based image retrieval using deep learning process. Cluster Computing **22**(2), 4187–4200 (2019)

[3] Outay, F., Mengash, H.A., Adnan, M.: Applications of unmanned aerial vehicle (uav) in road safety, traffic and highway infrastructure management: Recent advances and challenges. Transportation Research Part A: Policy and Practice **141**, 116–129 (2020)

[4] Zhao, X., Huang, P., Shu, X.: Wavelet-attention CNN for image classification. Multim. Syst. **28**(3), 915–924 (2022)

[5] Wang, P., Fan, E., Wang, P.: Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters **141**, 61–67 (2021)

[6] Wang, H., Song, Y., Huo, L., Chen, L., He, Q.: Multiscale object detection based on channel and data enhancement at construction sites. Multim. Syst. **29**(1), 49–58 (2023)

[7] Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)

[8] Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Proceedings of the European Conference on Computer Vision, pp. 173–190 (2020)

[9] Hao, S., Zhou, Y., Guo, Y.: A brief survey on semantic segmentation with deep learning. Neurocomputing **406**, 302–321 (2020)

[10] Jaouedi, N., Boujnah, N., Bouhlel, M.S.: A new hybrid deep learning model for human action recognition. Journal of King Saud University-Computer and Information Sciences **32**(4), 447–453 (2020)

[11] Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal pyramid network for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 588–597 (2020)

[12] Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11990–11997 (2020)

[13] Zheng, Z., Wei, Y., Yang, Y.: University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1395–1403 (2020)

[14] Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., Yang, Y.: Each part matters: Local patterns facilitate cross-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(2), 867–879 (2021)

[15] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[16] Zhang, Z., Lan, C., Zeng, W., Jin, X., Chen, Z.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3186–3195 (2020)

[17] Yu, F., Koltun, V., Funkhouser, T.: Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 472–480 (2017)

[18] Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM transactions on multimedia computing, communications, and applications **14**(1), 13–11320 (2018)

[19] Li, X., Yu, L., Chang, D., Ma, Z., Cao, J.: Dual cross-entropy loss for small-sample fine-grained vehicle classification. IEEE Transactions on Vehicular Technology **68**(5), 4204–4212 (2019)

[20] 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)

[21] 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)

[22] Lin, T.-Y., Cui, Y., Belongie, S., Hays, J.: Learning deep representations for ground-to-aerial geolocalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5007–5015 (2015)

[23] Vo, N.N., Hays, J.: Localizing and orienting street views using overhead imagery. In: Proceedings of the European Conference on Computer Vision, pp. 494–509 (2016). Springer

[24] Tian, Y., Chen, C., Shah, M.: Cross-view image matching for geolocalization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3608–3616 (2017)

[25] Altwaijry, H., Trulls, E., Hays, J., Fua, P., Belongie, S.: Learning to match aerial images with deep attentive architectures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3539–3547 (2016)

[26] Zhai, M., Bessinger, Z., Workman, S., Jacobs, N.: Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 867–875 (2017)

[27] Hu, S., Feng, M., Nguyen, R.M., Lee, G.H.: 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)

[28] Arandjelovic, R., Gronát, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5297–5307 (2016)

[29] Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for cross-view image based geo-localization. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp. 10090–10100 (2019)

[30] Shi, Y., Yu, X., Campbell, D., Li, H.: Where am i looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.

4064–4072 (2020)

[31] Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5624–5633 (2019)

[32] Rodrigues, R., Tani, M.: Are these from the same place? seeing the unseen in cross-view image geo-localization. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, pp. 3753–3761 (2021)

[33] Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: Proceedings of the IEEE International Conference on Computer Visio, pp. 470–479 (2019)

[34] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

[35] Toker, A., Zhou, Q., Maximov, M., Leal-Taixé, L.: Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6488–6497 (2021)

[36] Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications **16**(2), 1–23 (2020)

[37] Ding, L., Zhou, J., Meng, L., Long, Z.: A practical cross-view image matching method between uav and satellite for uav-based geo-localization. Remote Sensing **13**(1), 47 (2020)

[38] Zhuang, J., Dai, M., Chen, X., Zheng, E.: A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization. Remote Sensing **13**(19), 3979 (2021)

[39] Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., Sebe, N.: Joint representation learning and keypoint detection for cross-view geo-localization. IEEE Transactions on Image Processing **31**, 3780–3792 (2022)

[40] Dai, M., Hu, J., Zhuang, J., Zheng, E.: A transformer-based feature segmentation and region alignment method for uav-view geo-localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4376–4389 (2022)

[41] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez,

A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, vol. 30, pp. 1–11 (2017)

[42] Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11**(3), 1109–1135 (2010)

[43] Cai, S., Guo, Y., Khan, S., Hu, J., Wen, G.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 8391–8400 (2019)

[44] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)