Electronic Research Archive

http://www.aimspress.com/journal/era

ERA, 31(7): 4119–4138.
DOI: 10.3934/era.2023210
Received: 22 March 2023
Revised: 27 April 2023
Accepted: 21 May 2023
Published: 29 May 2023

Research article

AENet: attention efficient network for cross-view image geo-localization

Jingqian Xu[1,2], **Ma Zhu**[1,2,*], **Baojun Qi**[1,2], **Jiangshan Li**[1,2] **and Chunfang Yang**[1,2]

[1] Henan Provincial Key Laboratory of Cyberspace Situational Awareness, Zhengzhou 450001, China
[2] Zhengzhou Science and Technology Institute, Zhengzhou 450001, China

* **Correspondence:** Email: qingling800@163.com.

Abstract: To address the problem that task-irrelevant objects such as cars, pedestrians and sky, will interfere with the extracted feature descriptors in cross-view image geo-localization, this paper proposes a novel method for cross-view image geo-localization, named as AENet. The method includes two main parts: an attention efficient network fusing channel and spatial attention mechanisms and a triplet loss function based on a multiple hard samples weighting strategy. In the first part, the EfficientNetV2 network is used to extract features from the images and preliminarily filter irrelevant features from the channel dimension, then the Triplet Attention layer is applied to further filter irrelevant features from the spatial dimension. In the second part, a multiple hard samples weighting strategy is proposed to enhance the learning of hard samples. Experimental results show that our proposed method significantly outperforms the state-of-the-art method on two existing benchmark datasets.

Keywords: cross-view image geo-localization; attention mechanism; filter; hard sample

1. Introduction

Image-based geo-localization refers to finding out the geographic coordinates of a given query image [1] and has broad application prospects in many fields such as autonomous driving [2], augmented reality [3] and mobile robotics [4].

The traditional image-based geo-localization method is to match the query image of the ground view with the geo-tagged ground view image from the reference database. This method is also dubbed as ground-to-ground image geo-localization [5–8]. However, since most of the available reference images are captured in densely populated areas, e.g., famous tourist attractions, business zones, etc., few or no reference images are captured in sparsely inhabited and remote areas. Therefore, the ground-to-ground image geo-localization method often fails in sparsely populated and remote areas. In recent years, with the rapid development of the space industry, high-resolution satellite images with GPS (Global Positioning System) tags have been easily obtained. Cross-view image geo-localization means

matching the query ground image with the reference satellite images to determine its geographic location. With the advantages of easy accessiblity and wide coverage of the reference satellite images, cross-view image geo-localization can be extended to large areas or even globally. Therefore, cross-view image geo-localization has attracted wide attention from researchers and has become a primary research direction in current image geo-localization.

The early cross-view image geo-localization methods mainly match the hand-crafted features of the ground images and satellite images, then use the position tags of the satellite image matched best as the estimated position of the ground image [9–12]. For example, in 2010, Noda et al. [9] extracted the SIFT [13] and SURF [14] feature descriptors from the images captured by the in-vehicle camera and GPS satellite images, and matched them to localize the vehicle. In 2011, Lin et al. [12] extracted four feature descriptors such as HoG [15], self-similarity [16], gist [17], and color histograms from ground and satellite images for localization. However, due to the significant geometric differences between satellite images and ground images of the same geographic location, traditional hand-crafted features lack viewpoint invariance. They cannot bridge the spatial layout differences between the ground image and the satellite image, i.e., the relative position of the same objects in different views may be different. This makes it so that the methods based on traditional hand-crafted features geo-localize the ground image with very low accuracy.

Inspired by the success of deep learning in many computer vision tasks, Workman and Jacobs [18] first applied deep learning methods to cross-view image geo-localization in 2015. Since then, cross-view image geo-localization based on deep learning has become the mainstream method in this direction, and researchers have proposed a series of deep models for cross-view image geo-localization with excellent performance. According to whether the image viewpoint is transformed before feature extraction, these models can be roughly classified into two categories: end-to-end-based methods and viewpoint transformation-based methods.

The end-to-end-based cross-view image geo-localization method directly feeds the ground and satellite images to the deep network to extract discriminative image features for cross-view localization. For example, in 2015, Workman and Jacobs [18] directly used the pre-trained AlexNet [19] to extract deep features of ground and satellite images for cross-view image matching. After that, researchers proposed several end-to-end deep networks such as CVM-Net [20], GeoCapsNet [21], Siam-FCANet [22], and CVFT [23], which used VGG [24] or ResNet [25] as the backbone to extract the deep features of images. Such methods mainly rely on the image appearance content to learn discriminative image features, ignoring the impact of spatial layout differences between ground and satellite images. This might make the ground image be geo-localized to the wrong position, where the satellite image contains many semantic objects similar to them in the ground image.

The viewpoint transformation-based cross-view image geo-localization method first transforms ground or satellite images to another viewpoint, then inputs the transformed image and another untransformed image into a deep network for matching. For example, in 2019, Regmi and Borji [26] fed the ground image into cGANs [27] to synthesize its corresponding satellite image and used this synthesized satellite image as auxiliary information to minimize the difference between the query ground image and the satellite image. Shi et al. [28] applied a polar transform to satellite images to generate pseudo-ground panoramic images, thereby bridging the spatial layout discrepancies. Later, the polar transform algorithm was adopted by many cross-view image geo-localization methods [29–31], while [32] fused the GAN network synthesis method [26] and the polar transform algorithm [28] on

satellite images to synthesize the corresponding ground panoramic image closer to the real ones. Such methods achieve spatial layout alignment between ground image and satellite image features through viewpoint transformation, reduce the geometric differences caused by the drastic changes of two viewpoints, and have significantly increased retrieval accuracy compared to other methods.

In summary, cross-view image geo-localization methods based on viewpoint transformation have become the main development direction in current cross-view image geo-localization research. However, such existing methods do not consider the interference of irrelevant contents in ground or satellite images on features. For example, the ground images may contain transient moving objects and backgrounds such as cars, pedestrians and sky, and the satellite images may contain redundant content beyond the coverage of ground images due to their wide range of coverage. These task-irrelevant contents will interfere with the extracted features and reduce their discriminative power, thus seriously affecting the accuracy of cross-view image geo-localization in the real environment.

To address the above problems, this paper proposes a novel cross-view image geo-localization method, named AENet. Firstly, the EfficientNetV2 [33] network containing the channel attention mechanism as the backbone is used to extract useful local features. Then the task-irrelevant features are further filtered out from spatial dimensions by a Triplet Attention [34] layer. Moreover, this paper proposes a multiple hard samples weighting (MHNW) strategy, which enhances the learning of the network on multiple hard negative samples in each training batch. The contributions of this paper are as follows:

- For the cross-view image geo-localization task, we introduce the EfficientNetV2 network to the cross-view image matching task, and propose a novel cross-view image geo-localization method AENet, which could focus more on useful features by filtering irrelevant features from the channel and spatial dimension.
- A multiple hard samples weighting (MHNW) strategy is proposed to optimize the training of the network, which emphasizes the influence of multiple hard samples when calculating the loss in the current batch, thus enhancing the learning ability of the network for cross-view image pairs.
- Extensive experiments on two benchmark datasets show that the proposed AENet performs significantly better than state-of-the-art algorithms for cross-view image geo-localization.

The rest of this paper is organized as follows. In Section 2, related works are discussed. Section 3 describes the detailed structure of AENet and the basic principle of MHNW strategy. Section 4 supplies the experimental results of AENet and the existing cross-view image geo-localization methods. Section 5 summarizes the paper and discusses the direction of further research.

2. Related works

In the existing cross-view image geo-localization methods based on deep learning, the geo-localization networks can be classified as four common structures mainly composed of transformation module, feature extraction CNN module, feature processing module and loss function module (as shown in Figure 1). By realizing each module in different ways under different composition structures, researchers have proposed a series of high-performing cross-view image geo-localization methods as shown in Table 1. The specific roles of each module and the existing realization are described in detail below.

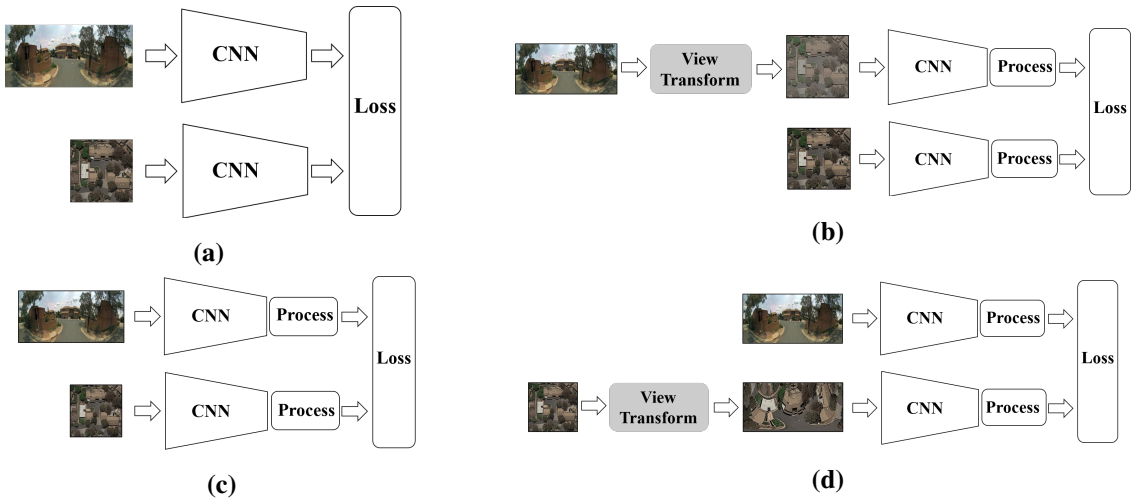


Figure 1. Common network structures in cross-view image geo-localization.

Table 1. Overview and properties of cross-view geo-localization methods.

Method	Publication	View Transform Module	Feature Extraction Module	Feature Processing Module	Loss Function
Vo and Hays [35]	ECCV2017	-	AlexNet	Orientation Regression	DBL Loss
Workman and Jacobs [18]	ICCV2015	-	AlexNet	-	Euclidean Loss
Liu and Li [36]	CVPR2019	-	VGG	-	WSMRL
CVM-Net [20]	CVPR2018	-	VGG	NetVLAD	WSMRL
CVFT [23]	AAAI2020	-	VGG	Optimal Feature Transport	WSMRL
GeoCapsNet [21]	ICME2019	-	ResNet	Capsule Network	Soft-TriHard Loss
Siam-FCANet [22]	ICCV2019	-	ResNet	FCBAM	HERTL
Rodrigues and Tan [37]	WACV2021	-	ResNet	Multi-scale Attention	Contranstive Loss
SAFA [28]	NeurIPS2019	PT	VGG	Spatial-aware Feature Aggregation	WSMRL
DSM [29]	CVPR2020	PT	VGG	Dynamic Similarity Matching	WSMRL
LPN [31]	TCSVT2021	PT	VGG	Sequential/Column Partition	WSMRL
Polar-L2LTR [30]	NeurIPS2021	PT	ResNet	Transformer	WSMRL
Regmi and Shah [26]	CV2019	GANs	GANs	Feature Fusion	WSMRL
Toker et al. [32]	CVPR2021	PT+GANs	ResNet+GANs	Spatial Attention	WSMRL

PT : polar transform, *DBL* : distance-based logistic, *WSMRL* : weighted soft-margin ranking loss, *HERTL* : hard exemplar reweighting triplet loss

Viewpoint transformation module: This module mainly transforms the ground view image to the corresponding satellite view image, or vice versa, to reduce the huge difference caused by the drastic difference between the two viewpoints. As can be seen in Table 1, most existing methods directly input the ground and satellite images into a CNN (convolutional neural network), and their network structures are shown in Figure 1(a),(c). Since 2019, researchers began to transform the input ground or satellite images to the image in another viewpoint, and there are two main methods of viewpoint transformation. One is to transform the ground view image to the satellite view image by cGANs network [26], as shown in Figure 1(b). The other is to apply a polar transformation on the satellite view image to obtain the pseudo-ground panorama image [28], as shown in Figure 1(d).

Feature extraction CNN module: The role of this module is to extract the local features of the query ground image and the reference satellite images separately through the deep network. In the existing works, the following four CNN networks are commonly used to extract the local features: the first one is the AlexNet used by methods [18] and [35], the second one is the VGG or its fine-tuned version used by methods [20, 23, 28, 29, 31, 36], the third one is the ResNet used by methods [21, 22, 30, 37], the fourth one is the GAN used by methods [26, 32].

Feature processing module: This module is used to process the local features extracted by the feature extraction module to obtain more discriminative image descriptors. As shown in Table 1, most of the methods process the extracted local features, except the methods in [18, 36]. The adopted feature processing methods can be roughly divided into two categories: feature processing based on attention mechanisms and feature processing based on spatial layout learning. The attention-based feature processing method mainly learns the salient image features through the attention mechanism [22, 28, 30, 32, 37]. The feature processing module based on spatial layout learning aims at reducing the spatial layout difference between ground image's local features and satellite image's local features by learning the orientation or spatial position relationships [21, 23, 29, 31, 35].

Loss function module: This module aims to measure the similarity between the features extracted from the query ground image and the reference satellite image. The goal is that, the closer the geographic location of the two images, the higher the similarity. In 2017, Vo and Hays [35] proposed a function called Distance-based Logistic loss (DBL loss). In 2018, Hu et al. [20] proposed a loss function called weighted soft-margin ranking loss (WSMR) based on DBL loss to speed up the training of the network. Moreover, Hu et al. [20] used a hard sample mining strategy proposed by Hermans et al. [38] to find the hard sample pairs of satellite image and ground image which do not match but look alike, then repeatedly learned the hard sample pairs to improve the generalization ability of the network. As can be seen from Table 1, the loss function proposed by Hu et al. [20] has been used by many methods since then [21, 23, 26, 28–32, 36]. Additionally, Cai et al. [22] proposed a hard exemplar reweighting triplet loss function to mine valuable hard samples for the network to learn and thus improve the performance of the network.

In summary, in existing deep learning-based cross-view image geo-localization methods, the designed network structures can be summarized as a framework similar to the Siamese network, as shown in Figure 2. The framework consists of four parts: viewpoint transformation module, feature extraction CNN module, feature processing module, and loss function module.

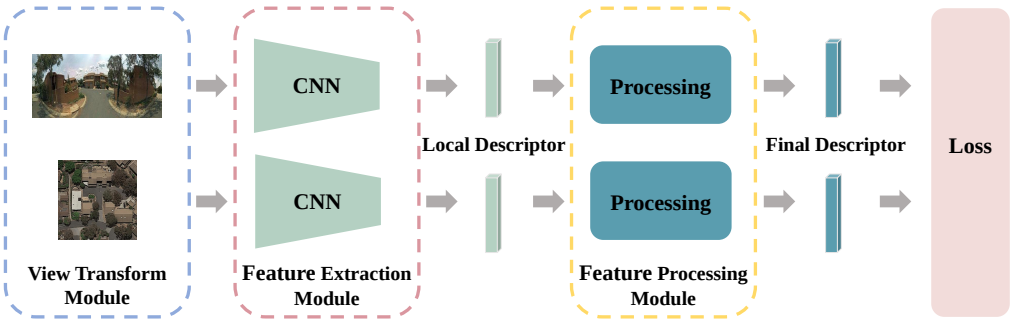


Figure 2. Overall framework of the cross-view image geo-localization method.

3. AENet framework

In image geo-localization tasks, objects such as cars, pedestrians and sky in images not only do not provide useful information, but may also interfere with the extracted image descriptor. The shallow networks can only learn features such as contour, color and texture, but can not obtain the high-level semantic features which can be used to discriminate the above objects. Although the deep networks such as ResNet and VGG commonly used in existing methods can learn richer high-level semantic features, it is still difficult to focus on the high-level semantic features important for image geo-localization. Therefore, in order to improve the localization accuracy, how to pay more attention to the objects related to geo-localization in the learning process for networks and eliminate the interference of useless features as much as possible, is an important problem to be solved in this section.

To address the above problems, this section proposes a cross-view image geo-localization method based on attention efficient networks, as shown in Figure 3. Firstly, the polar transformation is used to transform the satellite images into pseudo-ground panoramic images. Then the pseudo-ground panoramic image and the actual ground panoramic image are separately input into an attention efficient network which fuses the channel and spatial attention mechanisms, called AENet. It uses EfficientNetV2 to extract the high-level semantic features of the input images, then leverages the Triplet Attention (TA) module to determine the importance of different semantic features in order to quickly focus on the semantic features that are important for image geo-localization. Finally, in the loss function module, we first determine whether the number of hard samples is >1. If it is >1, use loss function based on a multiple hard samples weighting (MHNW) strategy to measure the similarity between the high-level semantic features extracted from the actual ground panoramic images and the pseudo-ground panoramic images. If it is not, use weighted soft-margin ranking loss (WSMR) proposed by Hu et al. [20].

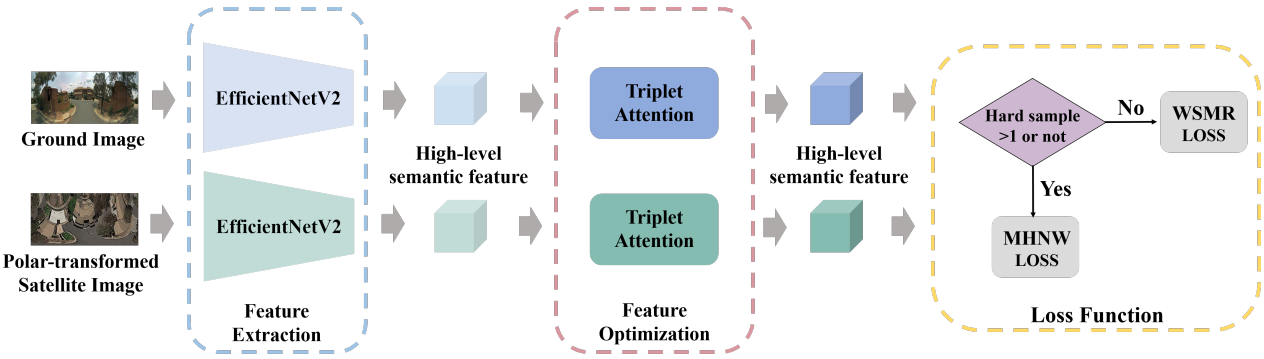


Figure 3. Overall architecture of AENet.

The core of the proposed method is the attention efficient network fusing channel and spatial attention mechanisms, which is used to extract the high-level semantic features of the actual ground panoramic image and the pseudo-ground panoramic image features. Although the pseudo-ground panoramic images of satellite images obtained by a polar transform are similar to the actual terrestrial panoramic images, there are still obvious visual differences. For example, the pseudo-ground panoramic image is difficult to clearly display, or even fails to show the facade of objects which are visible from the ground view but difficult to see from the satellite view. Therefore, we do not share

the parameters in the two branches extracting the high-level semantic features from the pseudo-ground panoramic image and the actual ground panoramic image.

The subsequent subsections describe the following three key modules of the proposed method in detail: high-level semantic feature extraction module, high-level semantic feature optimization module and loss function module.

3.1. High-level semantic feature extraction

In cross-view image matching tasks, one way to improve the accuracy is to optimize the CNN used in the feature extraction module. There are three factors that affect the performance of the CNN, including the depth of the network, the width of the network and the resolution of the input image. Increasing the values of these three factors can obtain richer and more complex features, but recklessly increasing them may magnify training difficulty and computational cost. In recent years, the EfficientNet networks [33, 39] have achieved great success in image classification. The improved version of EfficientNet, viz. EfficientNetV2, has become the most accurate model compared to other models with the same number of parameters. Compared to ResNet and VGG, EfficientNetV2 is generated by using a neural network structure search technique [40,41] to automatically learn and balance the above three factors. It mainly consists of Fused-MBConv blocks [42] and MBConv [39, 43] blocks, as shown in Figure 4.

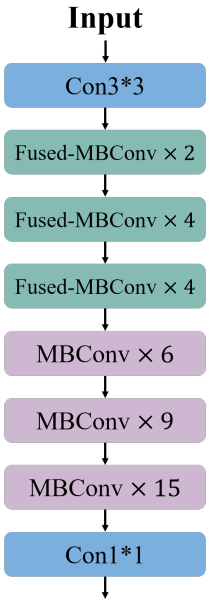


Figure 4. The structure of EfficientNetV2 network.

In this paper, the EfficientNetV2 is introduced into the field of cross-view image geo-localization. It is used as the backbone in the feature extraction module to extract rich high-level semantic features. This is mainly because its MBConv module can use the channel attention module SENet [44] to initially filter the features from the channel dimension. The structure of the MBConv is shown in Figure 5, which mainly consists of a 1×1 convolution, Depthwise convolution, SENet, 1×1 convolution, and residual connection, where the SENet is shown in the bottom part of Figure 5.

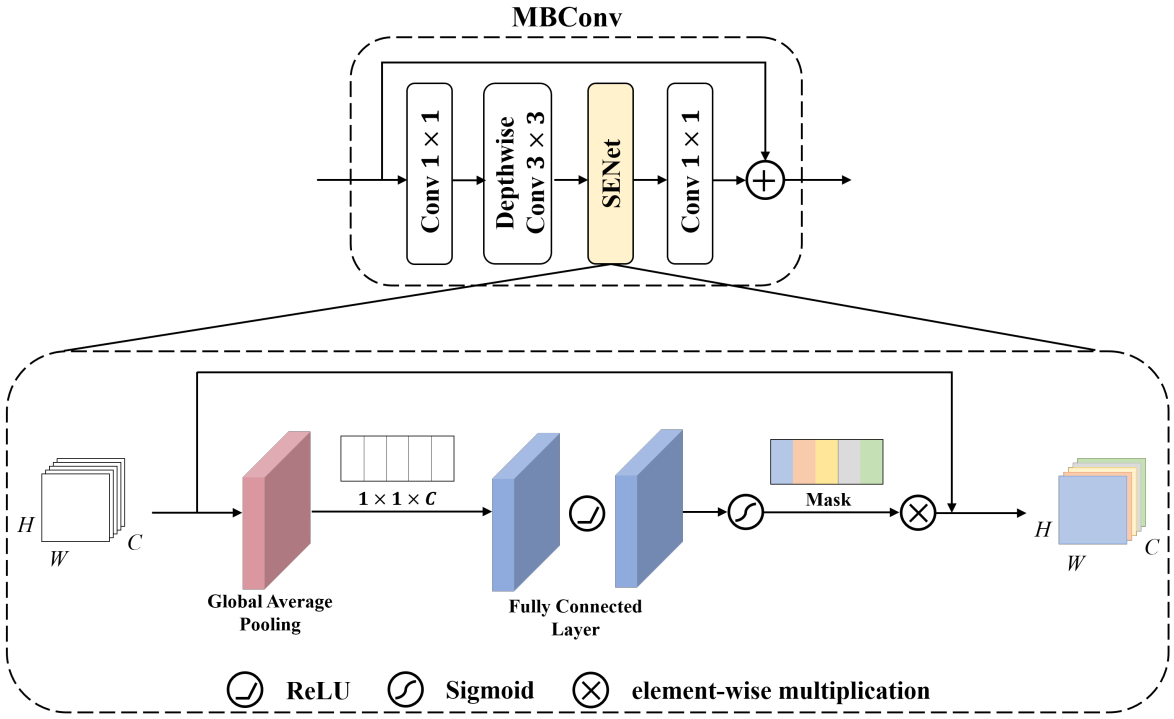


Figure 5. MBConv and SENet structure.

In the MBConv module, the Depthwise convolution can preserve the information of each channel as much as possible by adopting a separate convolution kernel for each channel. This can extract information such as the contour, shape and color of each object in the image as much as possible. We denote the output feature map of the Depthwise convolution as $f_{H\times W\times C}$. SENet performs global average pooling on $f_{H\times W\times C}$ to obtain a feature map $f_{1\times 1\times C}$ of size $1 \times 1 \times C$. This could compress the feature space dimension to capture the energy in each feature channel dimension and thus obtain the high-level semantic features. Then, $f_{1\times 1\times C}$ is fed to two fully connected layers to model the correlation between channel features and get the weight of each channel, and then obtain the weight mask by sigmoid activation. Specifically, task-irrelevant features have smaller weights, even close to 0; task-relevant features have larger weights, close to 1. Finally, $f_{H\times W\times C}$ is multiplied by the weight mask multiplication to generate the final weighted feature map $f'_{H\times W\times C}$.

3.2. High-level semantic feature optimization

Although the EfficientNetV2 can better extract the semantic object characteristics in images, the importance of different semantic objects for image geo-localization varies, and cross-view image geo-localization also needs to consider the spatial layout relationship between the semantic objects in images. For example, when judging whether the ground image and the satellite image (as shown in Figure 6) belong to the same geographical location, people usually first judge whether the two images contain geographical landmarks with the same shape and color, such as houses, highways and rivers, and ignore the objects that are not geographically representative, such as vehicles and pedestrians, then determine whether the two images belong to the same geographical location according to whether the

spatial layout relationships between the same content objects in two images are consistent (as shown in Figure 6(b),(c), the highways in both images are in the middle of the forest and the river).

(a) Ground Image

(b) Matching Satellite Image

(c) Unmatching Satellite Image

Figure 6. Cross-view image pairs.

The attention mechanism can make the network focus more on task-relevant content objects by assigning different weights to different regions of the feature maps. Therefore, in this paper, we choose to add the attention mechanism layer after the EfficientNetV2 network. The traditional attention mechanism module BAM [45] and CBAM [46] fuse channel and spatial attention mechanisms, but the channel attention and spatial attention of both are separated. However, in 2021, Triplet Attention (TA) proposed by Misra et al. [34] establishes the connection between channel attention and spatial attention through rotation operations and residual transformations to capture the dependency between the spatial dimension and the channel dimension of the input tensor, and let the network quickly focus on the task-relevant object.

In view of this, this subsection chooses to utilize the Triplet Attention module to optimize the high-level semantic features extracted by the EfficientNetV2. The Triplet Attention module is used to determine the weights of different semantic objects at different positions, so that the finally obtained descriptors can consider both the importance of different semantic objects and the spatial layout relationship. The structure of the TA module is shown in Figure 7. For the input local feature map $f'_{C\times H\times W}$ with a shape of $1280\times4\times16$ extracted by the EfficientNetV2, the TA module first permutes it to obtain two other feature maps $f'_{W\times H\times C}$ and $f'_{H\times C\times W}$. Secondly, three feature maps pass through three branches with the same structure to obtain the attention weighted feature maps $f^*_{C\times H\times W}$, $f^*_{W\times H\times C}$ and $f^*_{H\times C\times W}$, respectively. For example, in the first branch, the feature map $f'_{C\times H\times W}$ is fed to the max-pooling layer and the avg-pooling layer, respectively. Then the obtained results are concatenated as $f_{2\times H\times W}$ with a shape of $2\times H\times W$. Then, $f_{2\times H\times W}$ is fed to a structure sequence composed of a 7×7 convolution, BN layer and Sigmoid activation to obtain the weight mask. The input $f'_{H\times W\times C}$ is multiplied by the weight mask to generate the attention weighted feature map $f^*_{H\times W\times C}$. The other two branches are analogous to the first branch to obtain $f^*_{W\times H\times C}$, $f^*_{H\times C\times W}$. Finally, $f^*_{W\times H\times C}$ and $f^*_{H\times C\times W}$ are inverted to the feature maps of original size $H\times W\times C$ and the element-wise addition and average operations are performed on the three feature maps of the same size to obtain the final image feature descriptors.

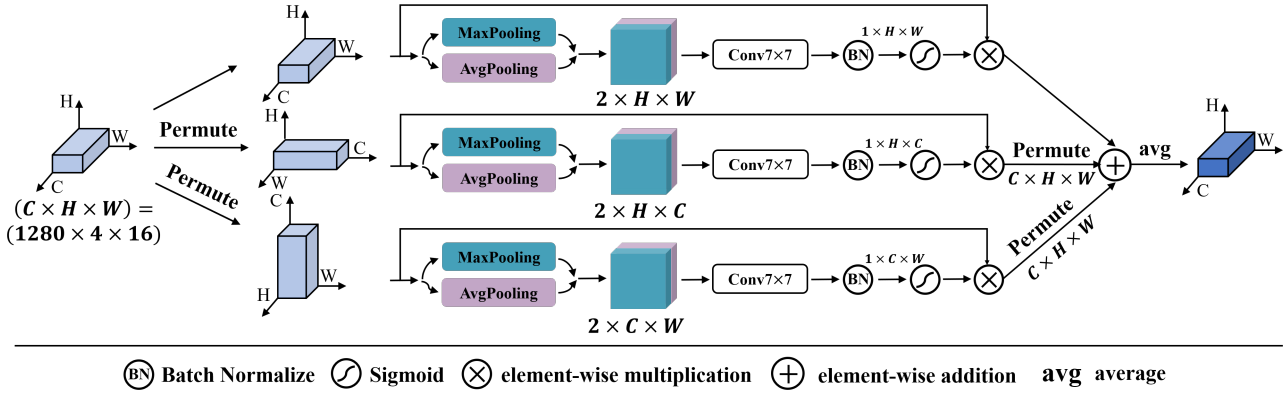


Figure 7. TA module structure.

3.3. Loss function based on MHNW strategy

As shown in Figure 8(a), if a ground image a is regarded as the anchor image, the corresponding reference satellite image p (as shown in Figure 8(b)) is called the positive sample of this anchor image whose Euclidean distance from the anchor image a is $d_{a,p}$, and the satellite images n taken at different geographic locations are called the negative samples of this anchor image. According to the distance to the anchor image, the negative samples can be further classified into the following three categories.

(a) Ground Image (b) Satellite Image

(c) Easy Negative (d) Semi-hard Negative (e) Hard Negative

Figure 8. Anchor, positive and negative sample examples.

The sample in the first category is called an easy negative sample, whose Euclidean distance $d_{a,n}$ from the anchor image a is much larger than the Euclidean distance $d_{a,p}$ between the anchor a and the positive sample p. Namely, this negative sample satellite image c should be obviously different to the anchor ground image a, such as the satellite image in Figure 8(c). Such samples are easily distinguished by the network.

The sample in the second category is called a semi-hard negative sample, whose Euclidean distance $d_{a,n}$ from the anchor image a is very close to $d_{a,p}$, but still larger than $d_{a,p}$. Namely, this satellite image sample is similar to the anchor image. For example, the negative sample in Figure 7(d) and the anchor image a both consist of multiple house buildings. Such negative samples can not be distinguished by

the network easily.

The sample in the third category is called a hard negative sample, whose Euclidean distance $d_{a,n}$ is smaller than $d_{a,p}$. Namely, this satellite image sample is more similar to the anchor image than the corresponding reference satellite image. For example, the negative sample in Figure 8(e) is extremely similar to the ground image a. It is hard to distinguish this positive sample image by the network.

Intuitively, it is expected that the total loss is minimum when the loss between the anchor image and the positive sample is minimum, and the loss between the anchor image and the hard negative sample is maximum. Therefore, Hu et al. [20] used the hard sample mining strategy proposed by Hermans et al. [38] to find the hard negative sample, and proposed the following weighted soft-margin ranking loss function by synthesising the distances of the positive sample and negative samples to the anchor image.

$$L_{\mathrm{Hu}} = \ln\left(1 + e^{\alpha\left(d_{a,p} - \min\limits_{n\in B} d_{a,n}\right)}\right) \tag{3.1}$$

where $d_{a,p}$ is the Euclidean distance between the features of each ground image a and its positive sample p in the current batch B and α is a weighting parameter used to improve the convergence speed of the network.

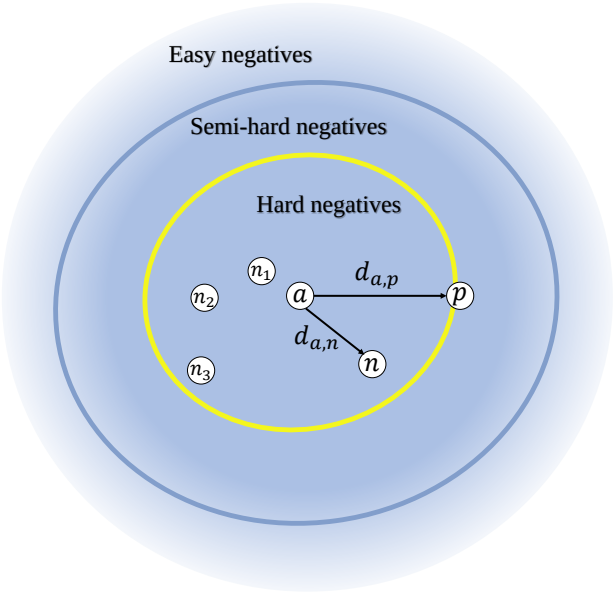


Figure 9. Illustration of the anchor image with its corresponding three negative samples.

During the training process of the network, we found that, for an anchor image, there may be N hard negative samples $n_1, n_2, \cdots, n_N$ in each batch, as shown in Figure 9. However, the hard sample mining strategy proposed by Hermans et al. [38] only selects the negative sample which is closest to the anchor image and ignores other hard negative samples. Thus, in order to make the network learn more adequately for hard negative samples, we designed a loss function based on multiple hard negative samples weighting (MHNW) strategy, i.e., when there are several hard negative samples in a batch, the losses of all these hard negative samples are emphasized. Specifically, for an anchor image a in a

batch B, first calculate the Euclidean distance between a and all negative samples and take the N hard negative samples n_i ($i \in N$) of the anchor, whose distance from the anchor is $d_{a,n_i} < d_{a,p}$. Then, the respective difficulty measure value of each hard negative sample is measured by $D_i = d_{a,p} - d_{a,n_i}$, where the smaller d_{a,n_i}, the higher the difficulty. Finally, according to the difficulty of each hard negative sample n_i, compute the weight w_i as follows,

$$w_i = \frac{D_i}{\max(D_i)} (i \in N) \tag{3.2}$$

Finally, the corresponding loss of each hard negative sample is multiplied by the weights w_i. The obtained results of all samples are summed as the final loss for the batch. Therefore, our loss is defined as

$$L_{\mathrm{MHNW}} = \frac{1}{N} \sum_{a \in batch} w_i * \ln\left(1 + e^{\alpha\left(d_{a,p} - d_{a,n_i}\right)}\right), \quad (i \in N) \tag{3.3}$$

4. Experimental results and analysis

4.1. Experimental setup

The performance of the proposed method was tested in the experimental setup as shown in Table 2.

Datasets. The experiments were conducted on two standard benchmark datasets: CVUSA and CVACT_val. The original CVUSA dataset is a large-scale dataset constructed by Workman and Jacobs [18] which consists of ground and satellite images from all over the U.S.. Zhai et al. [47] selected 35,532 pairs of cross-view images from the original CVUSA dataset for training and 8,884 pairs of cross-view images for testing. The CVUSA dataset constructed by Zhai et al. has been widely used in research on cross-view image geo-localization, and thus, in this section, CVUSA denotes the CVUSA dataset constructed by Zhai et al. CVACT_val is a new city-scale cross-view image dataset contructed by Liu and Li [36], which densely covers the city of Canberra. This dataset provides 35,532 pairs of cross-view images as the training set and 8884 cross-view image pairs as the validation set. The size of all input ground and satellite images were resized to 128×512.

Network training. The proposed method AENet was implemented in a PyTorch environment and used a TITAN RTX GPU with 24 GB of memory. The network was initialized with pre-trained parameters in ImageNet, then updated by the AdamW optimizer. During the training process, the batch size was set to 24, the learning rate was set to 0.00001 and the weight decay was chosen to be 0.00005.

Evaluation metric. The top K recall accuracy proposed by Vo and Hays [35] was used as an evaluation metric. When K is a integer, the top K is the set of the K satellite images whose descriptors are closest to that of a query ground image. When K is a percentage, the top K is the set of the $K \times T$ (T is the total number of satellite images in the reference satellite image set) satellite images whose descriptors are closest to that of a query ground image. The top K recall accuracy denotes the ratio of query images whose corresponding satellite image in top K, and is denoted R@K. In this section, R@1, R@5, R@10, and R@1% were used to evaluate the performance of cross-view image geo-localization.

Table 2. Experimental setting.

Operation	Setup
Input Image Size	128×512
Dataset	CVUSA [47] and CVACT_val [36]
Training Strategies	Batch size = 24, AdamW Optimizer, learning rate = 0.00001, weight decay = 0.00005
Experimental Platform	24GB TITAN RTX GPU, PyTorch 1.7.1.
Evaluation Protocol	Recall accuracy at top K (K $\in$ 1, 5, 10, 1%)

4.2. Comparison to the existing methods

The proposed method was compared with several state-of-the-art methods on two standard datasets, CVUSA and CVACT_val, and the experimental results are shown in Tables 3 and 4, respectively. It can be seen from Tables 3 and 4 that our method has higher recall accuracy than other methods, and the recall accuracy of our method is significantly improved on the key evaluation metric R@1. On the CVUSA dataset, our method achieves a recall accuracy of 95.97%, which is 1.92% higher than the second-best method. On the CVACT_val dataset, our method achieves a recall accuracy of 91.78%, which is 6.89% higher than the second-best method. From the experimental results, it is evident that our method suppresses the interference of irrelevant features on the extracted feature descriptors, thus improving the recall accuracy.

Table 3. Comparisons with state-of-the-art methods on the CVUSA [47] dataset.

Model	CVUSA			
	R@1	R@5	R@10	R@1%
Workman and Jacobs [18]	-	-	-	34.30
Zhai et al. [47]	-	-	-	43.20
Vo and Hays [35]	-	-	-	63.70
CVM-Net [20]	22.53	50.01	63.19	93.52
Regmi and Shah [26]	48.75	-	81.27	95.98
GeoCapsNet [21]	-	-	-	98.07
Siam-FCANet34 [22]	-	-	-	98.30
Liu and Li [36]	40.79	66.82	76.36	96.08
CVFT [23]	61.43	84.69	90.49	99.02
SAFA [28]	89.84	96.93	98.14	99.64
DSM [29]	91.96	97.50	98.54	99.67
Toker et al. [32]	92.56	97.55	98.33	99.57
Polar-L2LTR [30]	94.05	98.27	98.99	99.67
Ours	95.97	98.80	99.11	99.84

Table 4. Comparisons with state-of-the-art methods on the CVACT_val [36] dataset.

Model	CVACT_val			
	R@1	R@5	R@10	R@1%
CVM-Net [20]	20.15	45.00	56.87	87.57
Liu and Li [36]	46.96	68.28	75.48	92.01
CVFT [23]	61.05	81.33	86.52	95.93
SAFA [28]	81.03	92.80	94.84	98.17
DSM [29]	82.49	92.44	93.99	97.32
Toker et al. [32]	83.28	93.57	95.42	98.22
Polar-L2LTR [30]	84.89	94.59	95.96	98.37
Ours	91.78	96.28	97.29	99.29

4.3. Ablation experiments

TA module. To evaluate the effectiveness of the TA module, we removed the TA module from the AENet to obtain a Baseline containing only the EfficientNetV2 network, and trained the Baseline. We also added BAM and CBAM after Baseline and trained these two networks respectively, denoted as Baseline+BAM and Baseline+CBAM. The comparison results on CVUSA and CVACT_val datasets are shown in Table 5. From the experimental results, it can be seen that AENet combined with the TA module achieves the highest recall accuracy on both datasets, reaching R@1 of 95.97% and 91.78% on CVUSA and CVACT_val respectively. This is because the TA module can filter features on spatial location, allowing the network to focus more on the region of the image that are relevant to the cross-view image geo-localization task, thus improving the recall accuracy.

Table 5. Ablation experiment of TA module

Model	CVUSA				CVACT_val			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
Baseline	94.06	98.21	98.85	99.80	91.14	96.06	97.02	99.11
Baseline+BAM	89.80	96.58	97.90	99.52	82.21	91.90	94.12	98.01
Baseline+CBAM	90.89	96.93	98.05	99.53	83.13	92.19	94.24	98.21
AENet	95.97	98.80	99.11	99.84	91.78	96.28	97.29	99.29

Loss function based on MHNW strategy. To test the effectiveness of the MHNW strategy proposed in this paper, we conducted ablation experiments on CVUSA and CVACT_val datasets according to whether the MHNW strategy was used or not. The experimental results are shown in Table 6, where "without MHNW" means that we use the weighted soft-margin ranking loss function based on the hard sample mining strategy proposed by Hermans et al. [38]. "with MHNW" means that we use the weighted soft-margin ranking loss function based on the MHNW Strategy. It can be seen from the results that after using the MHNW strategy, all of four evaluation metrics R@1, R@5, R@10, R@1% on CVUSA and CVACT_val dataset were improved. This shows that the MHNW strategy can enhance the learning ability of the network by emphasizing multiple hard samples in the training process, and obtain more discriminative image features.

Table 6. Ablation experiment of MHNW strategy.

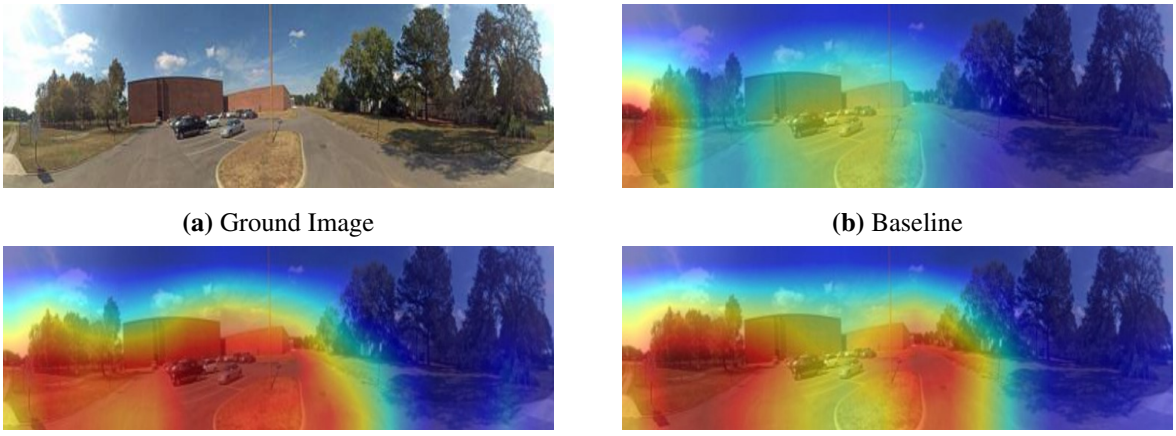
Model	CVUSA				CVACT_val			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
without MHNW	95.56	98.66	99.06	99.83	91.69	96.09	97.26	99.18
with MHNW	95.97	98.80	99.11	99.84	91.78	96.28	97.29	99.29

Complexity and computation cost. In order to compare the complexity and computation cost of the networks, we provide the number of parameters, GFLOPs (Giga Floating Point Operations per Second) of SAFA [28], DSM [29] and Polar-L2LTR [30] in Table 7. It can be seen from the results that the proposed method has lower GFLOPs than the other three networks. In terms of the number of parameters, the proposed method has many fewer than the Polar-L2LTR method, which has the second-best performance on recall accuracy, but still needs to be improved compared to SAFA and DSM methods.

Table 7. Comparison with previous works in terms of parameters and GFLOPs.

Model	Param (M)	GFLOPs
SAFA [28]	29.50	15.64
DSM [29]	17.90	7.25
Polar-L2LTR [30]	195.90	44
Ours	40.34	7.14

Visualization analysis. To more intuitively observe the effects of the proposed AENet, we visualized some heat maps of the extracted features. In order to test the superiority of the TA module, we replaced the TA module in AENet with the classical channel and spatial attention mechanism module CBAM [46], and then made a comparison. Figure 10 shows the heat maps of features extracted from ground image by Baseline, Baseline+CBAM and AENet. The darker red the color is, the more attention the network pays to this part. Figure 10 demonstrates that the AENet can successfully ignore the transient cars and sky in the image, and pay more attention to the region relevant to cross view image geo-localization task.



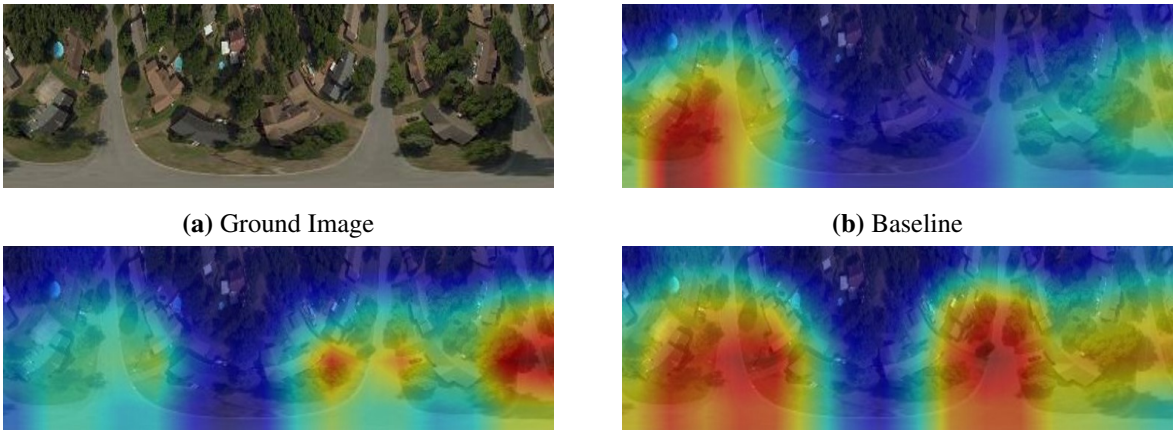
(a) Ground Image **(b)** Baseline

(c) Baseline+CBAM **(d)** AENet

Figure 10. Heat map of ground image features.

Figure 11 shows the heat maps of features extracted from a satellite image after a polar transform by Baseline, Baseline+CBAM and AENet. It can be clearly seen that the AENet can filter out the redundant content covered by satellite images and focus on the content of common region between satellite image and ground image.

Moreover, by comparing Figure 10(c) and 10(d), we can see that when processing ground images, the combination of the EfficientNetV2 and the CBAM can not effectively filter out the moving vehicles. By comparing Figure 11(c) and 11(d), we can see that, although the combination of the EfficientNetV2 and the CBAM can filter out the redundant content in satellite image, it pays less attention to the region where useful features are located (green area in Figure 11(c) and red area in Figure 11(d)) than AENet.



(a) Ground Image **(b)** Baseline

(c) Baseline+CBAM **(d)** AENet

Figure 11. Heat map of satellite image features.

5. Conclusions

In this paper, we propose a novel AENet for cross-view image geo-localization, aiming to address the interference of irrelevant features in the feature extraction process. The proposed AENet can reduce the interference of irrelevant features by making the network focus more on the useful features through attention enhancement. In addition, this paper also proposes a MHNW strategy, which can effectively improve the retrieval accuracy. We tested our method on two existing benchmark datasets, and the experimental results show that our method significantly improves the cross-view image geo-localization accuracy. Moreover, one major limitation of the AENet is that it is not applicable to the scenario when cross-view image pairs' orientation are not consistent. Therefore, we intend to increase the scenario applicability of AENet in future work.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Nos. 61872448, U1804263, 62272163), and the Science and Technology Research Project of Henan Province (No. 222102210075), China.

Conflict of interest

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this paper.

References

1. J. Brejcha, M. Čadík, State-of-the-art in visual geo-localization, *Pattern Anal. Appl.*, **20** (2017), 613–637. https://doi.org/10.1007/s10044-017-0611-1

2. C. McManus, W. Churchill, W. Maddern, A. D. Stewart, P. Newman, Shady dealings: robust, long-term visual localisation using illumination invariance, in *IEEE International Conference on Robotics and Automation (ICRA)*, (2014), 901–906. https://doi.org/10.1109/icra.2014.6906961

3. S. Middelberg, T. Sattler, O. Untzelmann, L. Kobbelt, Scalable 6-dof localization on mobile devices, in *European Conference on Computer Vision (ECCV)*, **8690** (2014), 268–283. https://doi.org/10.1007/978-3-319-10605-2_18

4. N. Suenderhauf, S. Shirazi, A. Jacobson, F. Dayoub, E. Pepperell, B. Upcroft, et al., Place recognition with ConvNet landmarks: viewpoint-robust, condition-robust, training-free, *Rob. Sci. Syst.*, **2015** (2015), 1–10. https://doi.org/10.15607/RSS.2015.XI.022

5. J. Hays, A. A. Efros, Im2gps: estimating geographic information from a single image, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2008), 1–8. https://doi.org/10.1109/CVPR.2008.4587784

6. A. R. Zamir, M. Shah, Image geo-localization based on multiplenearest neighbor feature matching usinggeneralized graphs, *IEEE Trans. Pattern Anal. Mach. Intell.*, **36** (2014), 1546–1558. https://doi.org/10.1109/TPAMI.2014.2299799

7. T. Sattler, M. Havlena, K. Schindler, M. Pollefeys, Large-scale location recognition and the geometric burstiness problem, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 1582–1590. https://doi.org/10.1109/CVPR.2016.175

8. N. N. Vo, N. Jacobs, J. Hays, Revisiting im2gps in the deep learning era, in *IEEE International Conference on Computer Vision (ICCV)*, (2017), 2621–2630. https://doi.org/10.1109/ICCV.2017.286

9. M. Noda, T. Takahashi, D. Deguchi, I. Ide, H. Murase, Y. Kojima, et al., Vehicle ego-localization by matching in-vehicle camera images to an aerial image, in *Asian Conference on Computer Vision Workshops*, **6469** (2010), 163–173. https://doi.org/10.1007/978-3-642-22819-3_17

10. T. Senlet, A. Elgammal, 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)*, (2011), 2034–2041. https://doi.org/10.1109/ICCVW.2011.6130498

11. M. Bansal, H. S. Sawhney, H. Cheng, K. Daniilidis, Geo-localization of street views with aerial image databases, in *19th ACM international conference on Multimedia*, (2011), 1125–1128. https://doi.org/10.1145/2072298.2071954

12. T. Y. Lin, S. Belongie, J. Hays, Cross-view image geolocalization, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2013), 891–898. https://doi.org/10.1109/CVPR.2013.120

13. D. G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision*, **60** (2004), 91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

14. H. Bay, T. Tuytelaars, L. van Gool, Surf: speeded up robust features, in *European Conference on Computer Vision*, **3951** (2006), 404–417. https://doi.org/10.1007/11744023_32

15. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, (2005), 886–893. https://doi.org/10.1109/CVPR.2005.177

16. E. Shechtman, M. Irani, Matching local selfsimilarities across images and videos, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2007), 1–8. https://doi.org/10.1109/CVPR.2007.383198

17. A. Oliva, A. Torralba, Modeling the shape of the scene: a holistic representation of the spatial envelope, *Int. J. Comput. Vision*, **42** (2017), 145–175. https://doi.org/10.1023/A:1011139631724

18. S. Workman, N. Jacobs, Wide-area image geolocalization with aerial reference imagery, in *IEEE International Conference on Computer Vision (ICCV)*, (2015), 3961–3969. https://doi.org/10.1109/ICCV.2015.451

19. A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Commun. ACM*, **60** (2017), 84–90. https://doi.org/10.1145/3065386

20. S. Hu, M. Feng, R. M. H. Nguyen, G. H. Lee, CVM-Net: cross-view matching network for image-based ground-to-aerial geo-localization, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7258–7267. https://doi.org/10.1109/CVPR.2018.00758

21. B. Sun, C. Chen, Y. Zhu, J. Jiang, GEOCAPSNET: ground to aerial view image geo-localization using capsule network, in *IEEE International Conference on Multimedia and Expo (ICME)*, (2019), 742–747. https://doi.org/10.1109/ICME.2019.00133

22. S. Cai, Y. Guo, S. Khan, J. Hu, G. Wen, Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss, in *IEEE International Conference on Computer Vision (ICCV)*, (2019), 8391–8400. https://doi.org/10.1109/ICCV.2019.00848

23. Y. Shi, X. Yu, L. Liu, T. Zhang, H. Li, Optimal feature transport for cross-view image geo-localization, in *AAAI Conference on Artificial Intelligence*, **34** (2019), 11990–11997. https://doi.org/10.1609/aaai.v34i07.6875

24. S. Karen, A. Zisserman, Very deep convolutional networks for large-scale image recognition, preprint, arXiv:1409.1556.

25. S. Xie, R. Girshick, P. Dollar, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 1492–1500. https://doi.org/10.1109/CVPR.2017.634

26. K. Regmi, M. Shah, Bridging the domain gap for ground-to-aerial image matching, in *IEEE International Conference on Computer Vision (ICCV)*, (2019), 470–479. https://doi.org/10.1109/ICCV.2019.00056

27. M. Mehdi, S. Osindero, Conditional generative adversarial nets, preprint, arXiv:1411.1784.

28. Y. Shi, L. Liu, X. Yu, H. Li, Spatial-aware feature aggregation for image based cross-view geo-localization, in *33rd Conference on Neural Information and Processing Systems (NIPS)*, (2019), 10090–10100. Available from: https://proceedings.neurips.cc/paper/2019/file/ba2f0015122a5955f8b3a50240fb91b2-Paper.pdf.

29. Y. Shi, X. Yu, D. Campbell, H. Li, Where am i looking at? Joint location and orientation estimation by cross-view matching, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 4064–4072. https://doi.org/10.1109/CVPR42600.2020.00412

30. H. Yang, X. Lu, Y. Zhu, Cross-view geo-localization with layer-to-layer transformer, in *35th Conference on Neural Information and Processing Systems (NIPS)*, (2021), 29009–29020. Available from: https://papers.nips.cc/paper_files/paper/2021/file/f31b20466ae89669f9741e047487eb37-Paper.pdf.

31. T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, et al., Each part matters: local patterns facilitate cross-view geo-localization, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 867–879. https://doi.org/10.1109/TCSVT.2021.3061265

32. A. Toker, Q. Zhou, M. Maximov, L. Leal-Taixé, Coming down to earth: satellite-to-street view synthesis for geo-localization, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 6484–6493. https://doi.org/10.1109/CVPR46437.2021.00642

33. M. Tan, Q. Le, Efficientnetv2: smaller models and faster training, in *38th International Conference on Machine Learning (ICML)*, preprint, arXiv:2104.00298.

34. D. Misra, T. Nalamada, A. U. Arasanipalai, Q. Hou, Rotate to attend: convolutional triplet attention module, in *IEEE/CVF Winter Conference on Applications of Computer Vision*, (2021), 3139–3148. https://doi.org/10.1109/WACV48630.2021.00318

35. N. N. Vo, J. Hays, Localizing and orienting street views using overhead imagery, in *European Conference on Computer Vision (ECCV)*, **9905** (2016), 494–509. https://doi.org/10.1007/978-3-319-46448-0_30

36. L. Liu, H. Li, Lending orientation to neural networks for cross-view geo-localization, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 5624–5633. Available from: https://openaccess.thecvf.com/content_CVPR_2019/papers/Liu_Lending_Orientation_to_Neural_Networks_for_Cross-View_Geo-Localization_CVPR_2019_paper.pdf.

37. R. Rodrigues, M. Tani, Are these from the same place? Seeing the unseen in cross-view image geo-localization, in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, (2021), 3753–3761. https://doi.org/10.1109/WACV48630.2021.00380

38. A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, preprint, arXiv:1703.07737.

39. M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, *International conference on machine learning*, preprint, arXiv:1905.11946.

40. B. Zoph, Q. V. Le, Neural architecture search with reinforcement learning, preprint, arXiv:1611.01578.

41. M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, et al., MnasNet: platform-aware neural architecture search for mobile, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019), 2820–2828. https://doi.org/10.1109/CVPR.2019.00293

42. S. Gupta, M. Tan, Efficientnet-edgetpu: creating accelerator-optimized neural networks with automl, Google AI Blog, 2019. Available from: https://torontoai.org/2019/08/05/efficientnet-edgetpu-creating-accelerator-optimized-neural-networks-with-automl/.

43. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

44. J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and excitation networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **42** (2020), 2011–2023. https://doi.org/10.1109/TPAMI.2019.2913372

45. J. Park, S. Woo, J. Y. Lee, I. S. Kweon, BAM: bottleneck attention module, preprint, arXiv:1807.06514.

46. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: convolutional block attention module, in *European Conference on Computer Vision (ECCV)*, **11211** (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

47. M. Zhai, Z. Bessinger, S. Workman, N. Jacobs, Predicting ground-level scene layout from aerial imagery, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2017), 867–875. https://doi.org/10.1109/CVPR.2017.440

AIMS Press