ITC 3/54

**Information Technology and Control**
**Vol. 54 / No. 3/ 2025**
**pp. 778-795**
**DOI 10.5755/j01.itc.54.3.41802**

**Cross-View Image Geo-localization Based on Attention Weight Masks**

Received 2025/06/05 | Accepted after revision 2025/07/20

**HOW TO CITE:** Zhu M., Sun, S., Xu, J. (2025). Cross-View Image Geo-localization Based on Attention Weight Masks. *Information Technology and Control,* 54(3), 778-795. https://doi.org/10.5755/j01.itc.54.3.41802

# Cross-View Image Geo-localization Based on Attention Weight Masks

**Ma Zhu**

Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou, Information Engineering University, Zhengzhou, 450001, China; e-mail: qingling800@163.com

**Siyue Sun**

Key Laboratory of Cyberspace Situation Awareness of Henan Province, Zhengzhou, Information Engineering University, Zhengzhou, 450001, China; e-mail: sunsy12@163.com

**Jingqian Xu***

School of Computer Science and Information Engineering, Anyang Institute of Technology, Anyang, 455000, China; e-mail: zzuxjq@126.com

**Corresponding author:** zzuxjq@126.com

Cross-view Image Geo-localization is the process of determining the geographic location of a ground-view query image by matching it with geotagged satellite or unmanned aerial vehicle (UAV) captured images. In the context of ground images characterized by a constrained field of view (FoV), the query image exhibits a reduced coverage area, limited scene content, and an unknown imaging direction. Furthermore, reference satellite images from the same location may contain significant feature redundancy. These issues lead to low localization accuracy when existing methods are applied to ground images with a limited FoV. We propose a cross-view image geo-localization method based on attention weight mask alignment. The Coordinate Attention (CA) mechanism, embedded in a lightweight ResNet18 network, generates weight masks to enable precise alignment of limited FoV ground images with satellite image feature maps. This process eliminates redundant areas in satellite images, thereby enhancing localization accuracy. Since feature maps at various levels capture images at different granularities, we introduce a multi-scale feature fusion strategy. It generates more representative image descriptors by combining features from different convolutional layers. Experimental results on the CVUSA and CVACT_val benchmark datasets demonstrate that when the FoV of ground images to be located is 70° and 90° with a random imaging direction, the proposed method significantly improves location accuracy.

KEYWORDS: Cross-view image geo-localization, Field of view, Attention mechanism, Feature alignment.

## 1. Introduction

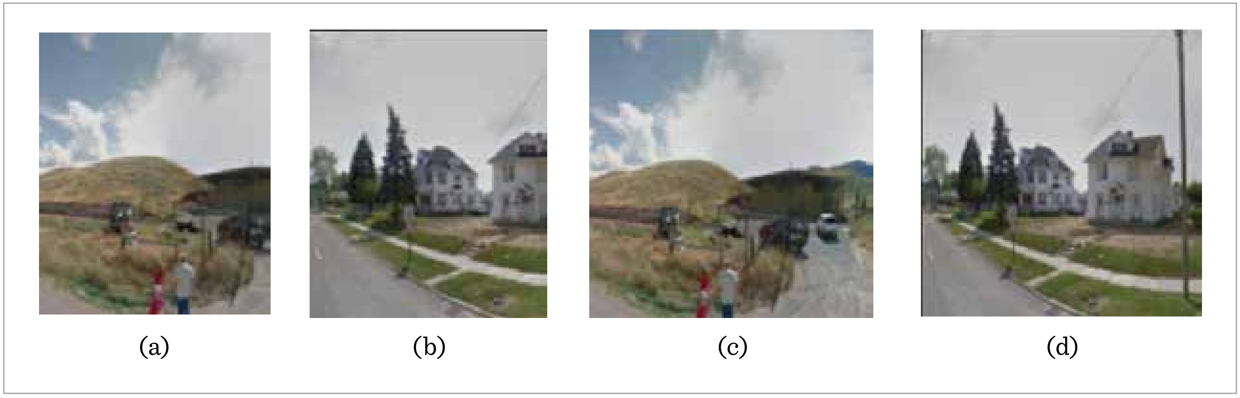
Image geo-localization is a technology of determining geographical locations of captured scenes or photographers using visual content. It offers significant potential across various domains, such as target tracking [30], autonomous driving systems [4], robotic navigation [13], and virtual/augmented reality (VR/AR) [9], [15]. Cross-view image geo-localization, a key focus in this field, matches ground-view query images with geotagged satellite or UAV images to pinpoint locations. Due to the wide coverage, easy acquisition, and inherent geotags of aerial images, cross-view image geo-localization has become a research hot point.

A major challenge in cross-view image geo-localization is the feature discrepancy between ground and satellite images. Researchers address this by extracting viewpoint-invariant features from images taken from different perspectives. Traditional handcrafted feature methods struggle to overcome visual and spatial differences between these views, often yielding lower localization accuracy. Consequently, these methods often result in lower location accuracy. With the rapid development of deep learning, methods for cross-view geo-localization based on deep neural networks [12], [21], [24] have become the mainstream method. Lin et al. [11] first applied Convolution Neural Network (CNN) to cross-view image geo-localization, presenting a Siamese-like Network for feature extraction of ground and satellite images. Since then, AlexNet [27], VGG [8], [31], ResNet [32] and their variants have been incorporated into two-branch architectures. Researchers are also exploring new neural networks with better structural modeling capabilities. Sun et al. [23] proposed GeoCapsNet, a capsule-network-based approach. The capsule layer was used to encode CNN-extracted features, enhancing representation by modeling spatial hierarchy. Zhu et al. [38] proposed GeoNet, an end-to-end network with ResNetX and GeoCaps modules. The ResNetX learns intermediate feature maps, while the GeoCaps converts them into capsules. Capsule length and orientation denote the existence probability and spatial information of scene objects, modeling the relationships between them. Wang et al. [26] proposed a Local Pattern Network (LPN) that uses a square-ring feature partitioning strategy to learn spatial features based on distance from the image center, thus better adapting to rotation variations. Dai et al. [3] designed a Transformer-based deep neural network architecture named Feature Segmentation and Region Alignment (FSRA). By incorporating a saliency-heatmap-guided feature segmentation mechanism and a region alignment module, it effectively enhances cross-view image matching performance under challenges like position shift and scale change.

Existing cross-view geo-localization studies primarily rely on matching panoramic ground images with satellite images. Panoramic images, offering a

**Figure 1**

Examples of 70° and 90° Ground Images (a) sample of 70° FoV ground image (b) sample of 70° FoV ground image (c) sample of 90° FoV ground image (d) sample of 90° FoV ground image.



full 360° FoV, capture abundant contextual information, which significantly alleviates perceptual discrepancies in cross-view matching. However, acquiring panoramic images is often challenging in practical scenarios. Instead, common ground images—captured by pedestrians, vehicle cameras, or mobile devices—typically have a limited FoV. The FOV of a typical mobile phone's rear camera is approximately 70°-90°, while that of the front camera is usually 60°-80°, as shown in Figure 1(a)-(b). Some high-end mobile phone cameras, equipped with ultra-wide-angle lenses, may have a much wider FOV of up to 100° or more, as shown in Figure 1(c)-(d).

Limited FoV ground images, due to the restricted perspective, incomplete information, and severe scale variations, cannot provide extensive geographical reference information. This causes feature redundancy in cross-view image geo-localization due to coverage discrepancies between these ground images and satellite images. Consequently, traditional methods relying on global consistency often fail in such scenarios. To address cross-view image geo-localization with limited FoV ground images, we propose a method based on attention-weighted masks. By generating this mask to align features between ground-level and satellite images, the precision of geo-localization can be enhanced effectively. The main contributions of this work are as follows.

1 A feature alignment method based on attention-weighted masks is proposed. Coordinate attention (CA) mechanism is embedded into the lightweight ResNet18 network to capture position info of feature maps in vertical and horizontal directions, enhancing the network's perception of key-feature regions.

2 A feature cropping module based on attention-weighted masks is proposed. It prunes redundant regions in satellite images that are irrelevant to ground images, retaining only key matching areas. This improves the efficiency and accuracy of feature matching.

3 A multi-scale feature concatenate strategy is introduced. It integrates features from different convolutional layers of the ResNet18 network. This generates more representative image descriptors by combining both high-level semantic and low-level detail features, thereby improving the accuracy of cross-view image geo-localization.

4 Comprehensive experiments were conducted on the CVUSA and CVACT_val benchmark datasets. Results indicate that, compared to existing advanced methods, the proposed method significantly improves recall rates (R@1, R@5, R@10, R@1%) for ground images with 70° and 90° FoV.

The structure of this paper is organized as follows: Section 2 provides a summary and overview of existing related research; Section 3 details the proposed methodological framework, including the feature alignment method based on attention-weighted masks and the multi-scale feature fusion strategy; Section 4 describes the experimental setup, the datasets used, and the evaluation metrics, and presents a comprehensive analysis of the experimental results and ablation studies; Section 5 discusses the potential limitations of the method and outlines future research directions; Section 6 summarizes the entire paper.

## 2. Related Work

### 2.1. Applications of Attention Mechanisms in Cross-View Image Geo-localization

Attention mechanisms aid feature extraction operators in extracting more representative features. The core function of feature-processing modules based on attention mechanisms lies in leveraging attention mechanisms to deeply explore key features within images. Cai et al. [1] successfully introduced a Feature Context-based Attention Module (FCAM) into the ResNet network, enabling deeper learning of multi-scale contextual semantic information of features and more representative feature representation. Shi et al. [19] designed a Spatial-Aware Feature Aggregation (SAFA) module to address image distortion caused by rectangular - polar transformation. Based on the VGG network, the module introduces spatial attention mechanisms to enhance the focus on features from different spatial content. Rodrigues et al. [17] innovatively proposed a Multi-scale Attention Module and successfully embedded it into the ResNet network to accurately capture salient image features. Zhu et al. [36] introduced an Attention-guided Non-uniform Cropping strategy. Using the self-attention mechanism of Transformers, it removes redundant areas in satellite images, cuts computation, and redirects saved resources to high-resolution, information-rich regions, improving model performance. Wang et al. [25] proposed a multi-scale window Transformer module based on self-attention. By combining global and local scale features, it enables precise image feature extraction and improves feature distinctiveness. Zhao et al. [35] designed a cross-attention Transformer module, which establishes interaction between the ground image and satellite image network branches, enabling effective cross-view information learning and enhancing the network's overall learning ability. Yang et al. [28] and Zhu et al. [36], [37] both used Vision Transformers (ViTs) to cross-view image geo-localization, leveraging their multi-scale attention to capture global object position relationships. The latter research also used ViT's patching mechanism to discard less important satellite image patches, enhancing training efficiency. Zhang et al. [33] designed a SubSpace Attention (SSA) module to highlight salient corresponding layout features across different scales. The encoded features represent different objects and reflect their relative positions, enabling the learning of more distinctive deep features.

### 2.2. Cross-View Image Geo-localization for Limited FoV Ground Images

The limited FoV restricts available visual information, causing insufficient features in ground images and reducing cross-view matching accuracy. To address this, Regmi et al. [16] attempted to synthesize panoramic images for limited FoV ground images. They proposed two novel cGAN-based architectures. With multi-objective training, these architectures enhance the model's ability to understand and transfer semantic information. Zhang et al. [34] proposed a cross-view sequential localization framework, which combines temporal modeling with adaptive sequence enhancement. By introducing a multi-head self-attention mechanism to construct a Temporal Feature Aggregation Module (TFAM), the framework captures potential spatiotemporal structural information in small FoV image sequences. Additionally, a sequence random dropout strategy is employed to enhance the model's adaptability to variable-length inputs. Rodrigues et al. [18] proposed a novel representation method that integrates global and local features of satellite images. Based on a Siamese network architecture and combined with data augmentation strategies such as sky region removal and rotation-robustness, this method effectively improves the cross-view matching capability of limited FoV ground images. Cheng et al. [2] proposed the Window-to-Window Bird's Eye View (W2W-BEV) representation learning method. Using a context-aware window matching strategy and depth information initialization, it improves matching accuracy under limited FoV and unknown shooting direction conditions. Mi et al. [14] proposed ConGeo, a feature learning framework integrating cross-view appearance consistency modeling with local perception enhancement. By aligning features of images from the same location under varying appearances in latent space and simulating appearance changes to enhance model adaptability, it effectively reduces performance degradation of conventional methods in scenarios with significant appearance variation. Shugaev et al. [22] devised an angle-constrained ArcGeo loss function, combined large-scale pre-training and FoV transformation data augmentation, optimized the feature embedding space structure, and significantly improved geo-localization under limited FoV conditions. Li et al. [10] proposed the Automatic Progressive Learning (AMPLE) method. Combining an improved ConvNeXt network and two progressive training strategies, it effectively improves geo-localization under unknown orientations and limited FoV conditions.

The above researches have somewhat improved cross-view matching of limited FoV ground images, but mainly focus on feature extraction and alignment, neglecting cropping/filtering of redundant areas in satellite images. Most also use whole-image or local-block feature modeling, ignoring "key matching regions" like important local details. Moreover, how to integrate different-scale feature info (fine-grained textures and high-level semantics) remains an issue needing attention.

## 3. Method

### 3.1. Problem Analysis

The primary challenge in cross-view image geo-localization lies in the significant viewpoint discrepancy and substantial feature redundancy between ground-level and satellite imagery, particularly under FoV-constrained scenarios. Shi et al. [20] transformed the cross-view geo-localization of limited FoV images into a feature alignment problem. They proposed a Dynamic Similarity Matching (DSM)-based cross-view image geo-localization method. First, a polar transform converts the satellite image to a ground-view pseudo-panorama. Then, an improved VGG16 backbone extracts feature from the limited FoV ground image and the pseudo-panorama. Next, using the feature map size of the ground image to be located as the sliding window size, it slides horizontally across the pseudo-panoramic feature map. The similarity between the sliding window's portion and the limited FoV ground image feature map is then calculated. Finally, the direction corresponding to the highest-similarity sliding window location is estimated as the shooting direction of the ground image to be located, as shown in Figure 2. Based on the similarity value in this direction, it is determined whether the ground image and the reference satellite image are taken at the same location.

This method conducts similarity matching on the feature maps extracted by the final convolutional layer of the feature extraction backbone network. It aims to determine the shooting direction of the ground image to be located. The feature maps from the final convolutional layer of the backbone network have low resolution, and the varying importance of different semantic objects for geo-localization is ignored. Consequently, the method has significant misalignment in the shooting direction of ground images, causing low positioning accuracy.

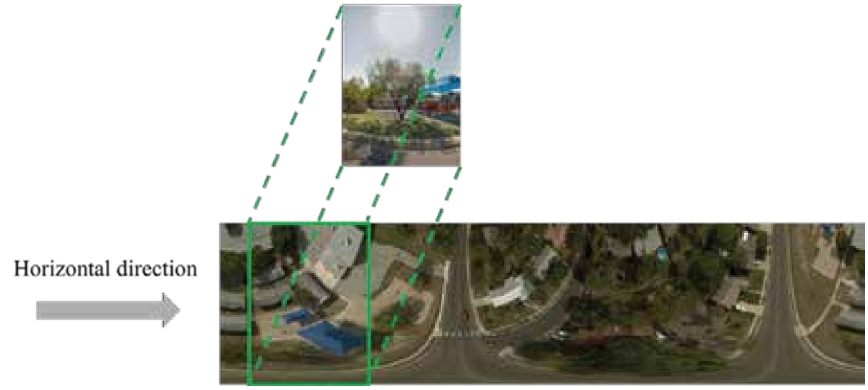
## 3.2. Method Overview

We propose an attention-weighted mask-based feature alignment method for cross-view localization of non-panoramic ground images. The attention-weighted mask, a two-dimensional weight map produced by the CA module integrated into the ResNet18 residual blocks. It is designed to characterize the spatial distribution of importance across different regions of the input image. Each weight in the mask reflects the contribution of the corresponding feature location to the downstream task. By independently generating these masks in the ground and satellite image branches, the network learns to identify pivotal regions within each respective view during training, thereby offering spatial guidance for subsequent feature alignment and region cropping. The overall framework of our method is shown in Figure 3 and consists of three main components: a Feature Extraction module, a Feature Cropping and Alignment module, and a Multi-scale Feature Map Concatenation operation.

To reduce the perspective-structure representation gap between ground and satellite images, we perform polar transformations on the original satellite image $I_{sat}$ using its center as the pole, generating a pseudo-ground image $I_s$. First, non-panoramic ground images ($I_g$) with limited FoV and pseudo-ground-view panoramic images ($I_s$) transformed via polar coordinate transformation are input into a network with independent dual branches. The network consists of two branches: one for ground images and another for

**Figure 2**

Dense Search Pattern of DSM.



satellite images. Pseudo-ground images are fed into the satellite-image branch for feature extraction instead of original satellite images. Each branch contains an identical lightweight ResNet18 network with a CA module for feature extraction. Attention masks $M_g$ for ground images and $M_s$ for pseudo-ground-view images are obtained through CA module processing. Second, in the satellite image branch, an attention-weighted mask-based feature cropping method is introduced to crop the pseudo-ground-view panoramic image. This cropping retains only the pseudo-ground-view image areas corresponding to the ground image, achieving feature alignment between the two perspectives. Finally, the aligned ground features ($F_g$) and pseudo-ground-view features ($F_s$) are input into a Triplet Loss function for similarity metric learning. Minimizing the distance between same-location cross-view image pairs and maximizing the distance between different-location pairs to determine whether the query non-panoramic image and reference satellite image whether the same location or not.

Though polar transformation may cause nonlinear stretching in peripheral regions, the ResNet18_CA network used in this work, with its integrated coordinate attention mechanism. It has strong directional and positional perception capabilities due to its structural, which effectively reduces interference from geometric deformation. Moreover, the pseudo-ground images resulting from the transformation are structurally more consistent with ground images, which helps enhance the learning stability of subsequent region-cropping and alignment modules.

## 3.3. Feature Alignment Based on Attention-Weighted Masking

In this section, we detail the feature alignment process using attention weight masks. It includes two key steps: generating the masks and performing feature cropping and alignment.

### 3.3.1. Attention-Weighted Mask Generation

Non-panoramic ground images, due to their limited FoV, only contain partial content of the reference image from the same location, or show only part of a complete object. If the redundant areas in satellite images can be cropped out, and matching is done using only the regions corresponding to the non-panoramic ground images, the localization accuracy can be significantly improved. Hou et al. [7] proposed a CA module that captures cross-channel information and encodes vertical and horizontal position information of image feature maps, embedding it into channel attention to help networks focus on task relevant regions. Given this, we incorporate the CA module into our work. The module integrates cross-channel information and encodes the vertical and horizontal position information of image features. This information is embedded into channel attention, enabling the network to precisely localize key feature areas in images and capture direction-aware and position-sensitive information.

Compared with conventional attention mechanisms such as SE and CBAM, the CA mechanism demonstrates a distinctive advantage by introducing positional awareness of spatial coordinates during the feature encoding process. Its core innovation lies in the direction-decoupled modeling of spatial features and channel dependencies, which achieves more fine-grained spatial-position information encoding. Let the input feature map be denoted as $X \epsilon R^{C\times H\times W}$, where $C$ represents the number of channels, and $H$ and $W$ denote the height and width of the feature map, respectively. CA mechanism performs average pooling along the height and width directions separately, yielding two one-dimensional directional features. The operation along the height direction is formulated in Equation (1): for the $h$-th row in the $c$-th channel, average pooling is applied across the horizontal (column-wise) direction. This operation generates a feature representation denoted as $F^h \epsilon R^{C\times H\times 1}$.

$$f^h(c,h) = \frac{1}{W}\sum_{w=1}^{W} X(c,h,w). \tag{1}$$

Similarly, in the width direction, as shown in Equation (2), average pooling is performed along the horizontal (column) direction for the $w$-th column in the $c$-th channel, resulting in a one-dimensional feature $F^w \epsilon R^{C\times 1\times w}$.

$$f^w(c,w) = \frac{1}{H}\sum_{h=1}^{H} X(c,h,w). \tag{2}$$

Subsequently, the directional features $F^h$ and $F^w$ are concatenated. Shared convolution and nonlinear transformations are then applied to generate two 1D attention weight maps. These maps are utilized for the height and width dimensions, respectively.

By multiplying the original input feature map $X$ with these two directional attention weight maps respectively, bidirectional spatial attention weighting for each position point $(h, w)$ in $X$ is achieved. This process constructs the final attention weight mask.

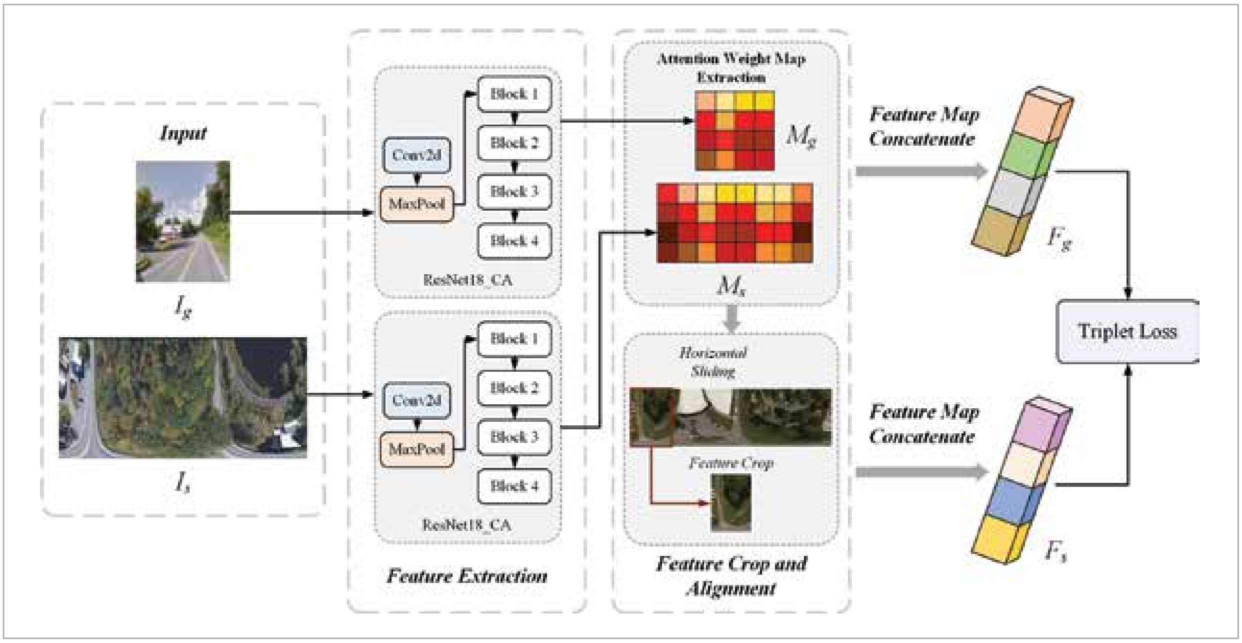
ResNet18 is adopted as the backbone network architecture in this work. Within the ResNet18 architecture, each residual block comprises two consecutive 3×3 convolutional layers designed to progressively extract multi-level feature information from images. Specifically, the first 3×3 convolutional layer primarily performs initial local feature extraction, generating intermediate feature maps. The second 3×3 convolutional layer subsequently integrates and refines these intermediate features to produce more discriminative feature representations. To enhance the network's capability to focus on critical regions and channel features, we embed the CA module into the residual blocks of the ResNet18 network, between the two 3×3 convolutional layers. The improved network is termed ResNet18_CA, as shown in Figure 4. First, 70°FoV ground images are input into the network. These images have 3 channels ($C$), a height ($H$) of 128, and a width ($W$) of 99. The input images sequentially pass through a convolutional layer with a 7×7 kernel, followed by a batch normalization (BN) layer, a ReLU activation layer, and a max-pooling layer. Then, the feature maps, reduced in spatial dimension by the MaxPool layer, are sequentially input into four residual blocks (Block 1~Block 4). Finally, the output feature maps, sized $C$ = 512, $H$ = 4, $W$ = 4, are used as input for subsequent feature cropping and alignment. Each residual block's internal structure is shown in the right half of Figure 4. The input feature maps first passing through a 3×3 convolutional layer for feature extraction and then through a CA layer to introduce spatial coordinate information. This process enhances the model's ability to identify key feature locations in images.

Then, the feature maps pass through another 3×3 convolutional layer for further feature extraction. Finally, a skip connection adds the original input to the convolutional output to form the residual block's output. This structure not only retains the advantages of deep feature propagation from residual networks but also enhances the network's sensitivity to spatial information through the CA module, improving its feature representation ability.

The internal structure of the CA module is shown in Figure 5. Generating attention-weighted masks and feature maps are obtained from the following steps for the input feature map $f_{C\times H\times W}$.

**Figure 3**

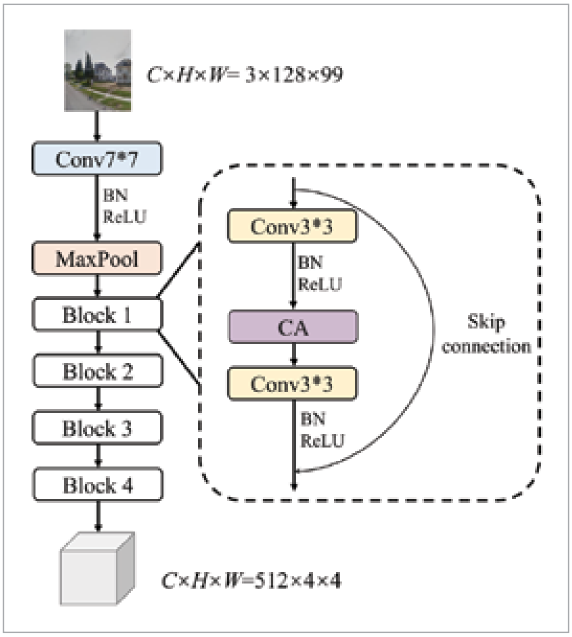
Framework of Feature Alignment Based on Attention-Weighted Masking.



1 For the input feature map $f_{C\times H\times W}$, two 1D average pooling layers are used to process. $X$ AvgPool performs average pooling horizontally (along the feature map's width W) to get $f_{C\times H\times 1}$. $Y$ AvgPool performs average pooling vertically (along the feature map's height $H$) to get $f_{C\times 1\times W}$. This aims to aggregate feature information horizontally and vertically while maintaining the spatial dimensions of the feature map, thereby generating direction-aware feature representations.

**Figure 4**

Structure of ResNet18 with Integrated CA Module (ResNet18_CA).



2 The feature maps $f_{C\times H\times 1}$ and $f_{C\times 1\times W}$ obtained from the two average pooling operations are concatenated along the spatial dimension (dim=2). A feature map that integrates information from both horizontal and vertical directions is obtained.

3 The concatenated feature map sequentially passes through a 1×1 Conv2d layer, a BN layer, and a nonlinear activation layer. This process further compresses the channel information and yields the feature representation $f_{C/r\times 1\times (H+W)}$.

4 To further refine the feature representation, we split $f_{C/r\times 1\times (H+W)}$ into two directional feature vectors through a feature separation operation, namely $f_{C/r\times 1\times (H+W)}$ and $f_{C/r\times 1\times W}$.

5 The two feature vectors are each input into a 1×1 Conv2d layer for channel adjustment, yielding $f_{C\times H\times 1}$ and $f_{C\times 1\times W}$.

6 The Sigmoid activation function generates weight masks $Mask_{C\times H\times 1}$ and $Mask_{C\times 1\times W}$ for the $H$ and $W$ directions of the feature map. These masks help the network focus on important features in each row and column, capturing long-range dependencies along the spatial dimensions (H or W) of the input feature map $f_{C\times H\times W}$.

7 By applying multiplication with the two attention masks, the network can more accurately identify the coordinate positions of important features in the input feature $f_{C\times H\times W}$.

### 3.3.2. Feature Cropping and Alignment

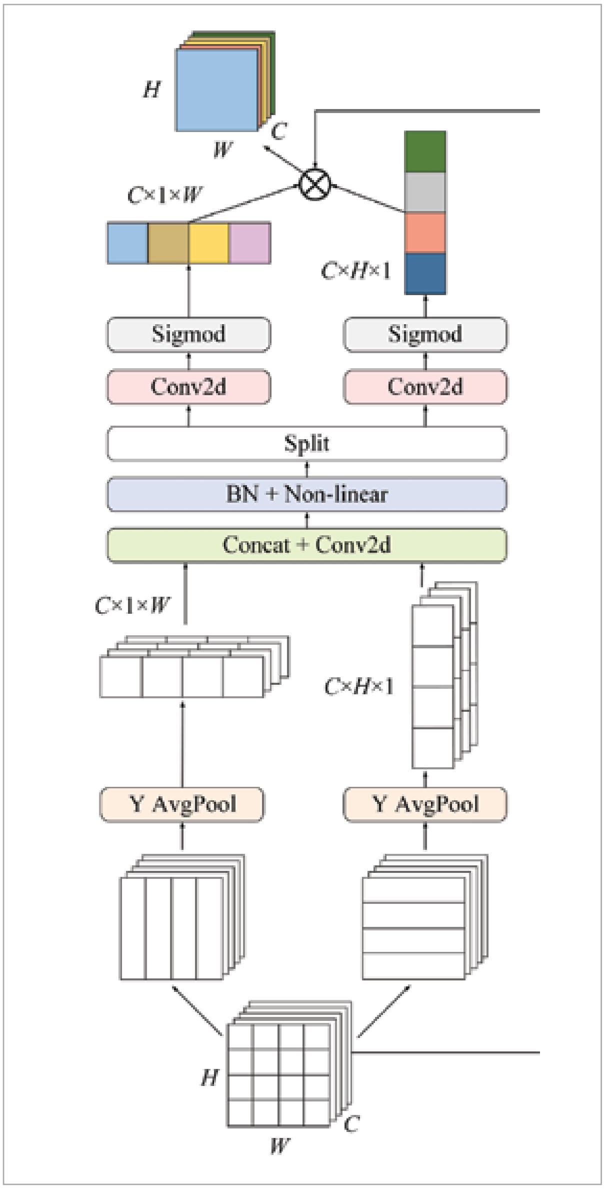
After generating the mask, we propose an attention-based weighted mask method to achieve feature-map cropping and alignment through Feature Cropping and Alignment (FCA). We perform cropping and alignment on the satellite-image feature map using the satellite-attention mask $M_{s.}$ The sliding window searches for areas corresponding to the ground-attention mask $M_g$, removing redundant features and keeping key matching regions. The satellite image referred to here are pseudo-ground image generated by polar coordinate transformation. To distinguish them from ground images, we collectively refer to pseudo-ground images as satellite images in the subsequent description.

As shown in Figure 6(a) illustrates the non-panoramic ground image, Figure 6(b) presents the polar-transformed reference satellite image, and Figure 6(c)-(d) respectively show the visualization heatmaps of the ground image feature attention-weighted mask $M_g$ and the satellite image feature attention-weighted mask $M_s$. During training, the attention mechanism generates optimized weight masks to indicate the importance of different regions in the network. The weight values in the attention masks reflect the significance of corresponding regions in the feature maps. Darker colors in the masks signify larger weight values, highlighting the greater importance of those regions.

Using the weight masks learned from the CA layer in ResNet18_CA, a dense search identifies the regions with the highest weights. These regions, corresponding to ground image features, guide the removal of redundant areas in satellite image feature maps

**Figure 5**

Internal Structure of the CA Module.



($F_{s\_block\_i}$) and enhance the key areas related to ground images. During feature alignment cropping, the vertical ($y$-axis) range of ground image features remain fixed, while the cropping area's start and end positions on the horizontal ($x$-axis) are determined. Once the optimal $x$-coordinate interval is determined, satellite image feature maps are cropped using this interval.

We introduce a sliding window matching strategy to locate candidate regions on the satellite attention mask $M_s$ that exhibit maximal relevance to the ground image feature areas. During feature cropping and alignment, the sliding window dimensions are dynamically adapted to the spatial resolution of ground feature maps generated by distinct convolutional blocks in ResNet18_CA—rather than employing a fixed size. At each network hierarchy, the corresponding ground-image attention mask is leveraged as a complete sliding window. This window performs horizontal sliding matching on the satellite image attention weight mask. The goal is to find regions in the satellite image that are most relevant to the ground image. The attention mask of the ground image accurately represents the spatial location of salient regions in the current view. It guides the sliding window to accurately identify corresponding key regions in the satellite image.

The specific implementation process is as described below.

1 In the satellite image network branch, the weight masks of satellite image feature maps $M_s$, generated by the CA module in each Block of the ResNet18_CA network, are extracted.

2 In the ground image network branch, ground image feature maps are extracted from each Block of the ResNet18_CA network, and their height ($H$) and width ($W$) are used as the sliding window size.

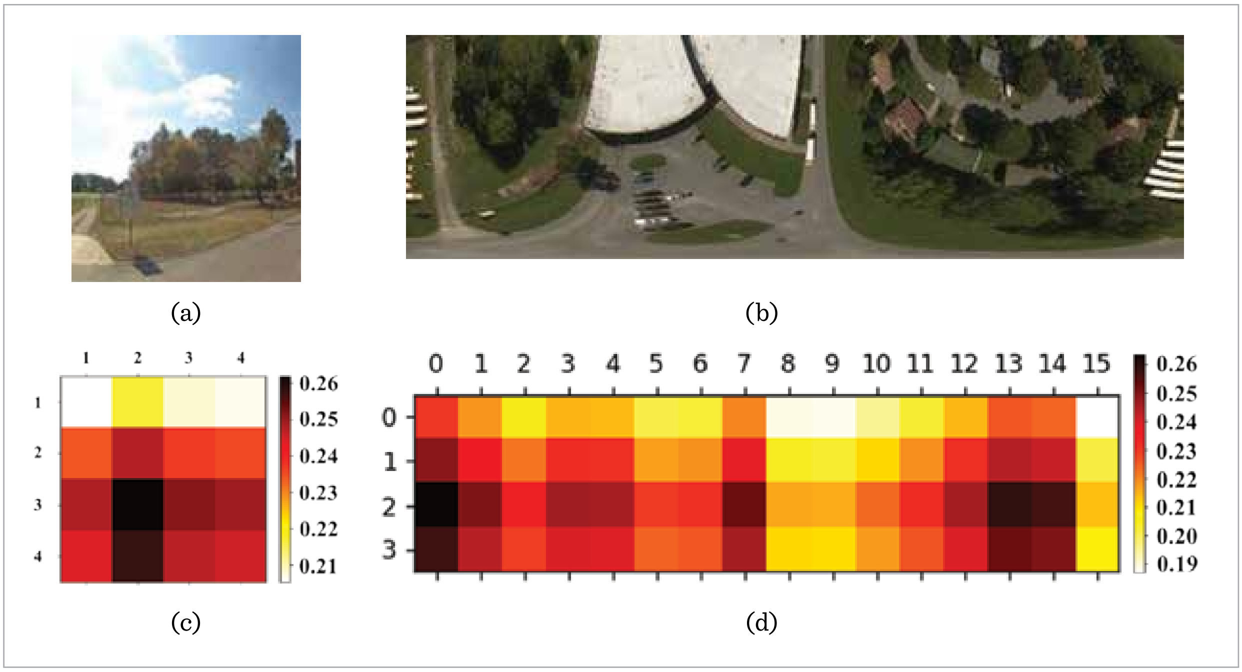
3 Using the sliding window set in step (2), slide it across the satellite image feature weight mask $M_s$. At each step, sum the total weight values within each window.

4 Compare the weights of sliding windows and identify the window with the maximum total weight. The corresponding region of this window in the satellite image feature map is considered aligned with the ground image feature map. This region is cropped to retain key features, while other irrelevant areas are discarded. The resulting feature map is the final satellite image feature map ($F_{sc\_block\_i}$).

The feature cropping and alignment process is shown in Figure 7. The satellite image ($I_s$) is transformed into the cropped satellite image ($I_{sc}$) by this stage. Although the sliding window covers multiple regions, the final window is selected by traversing all positions and comprehensively evaluating the summed attention weights within each region. It ensures the extracted satellite features maximally correspond to ground image features, enhancing cropping accu-

**Figure 6**

Visualization of Attention-Weighted Masks in Block 4 of ResNet18 Network (a) 70° FoV Ground Image (b) Matching Satellite Panoramic Image (c) Heatmap of Ground Image Feature Attention Weights (d) Heatmap of Satellite Image Feature Attention Weights.

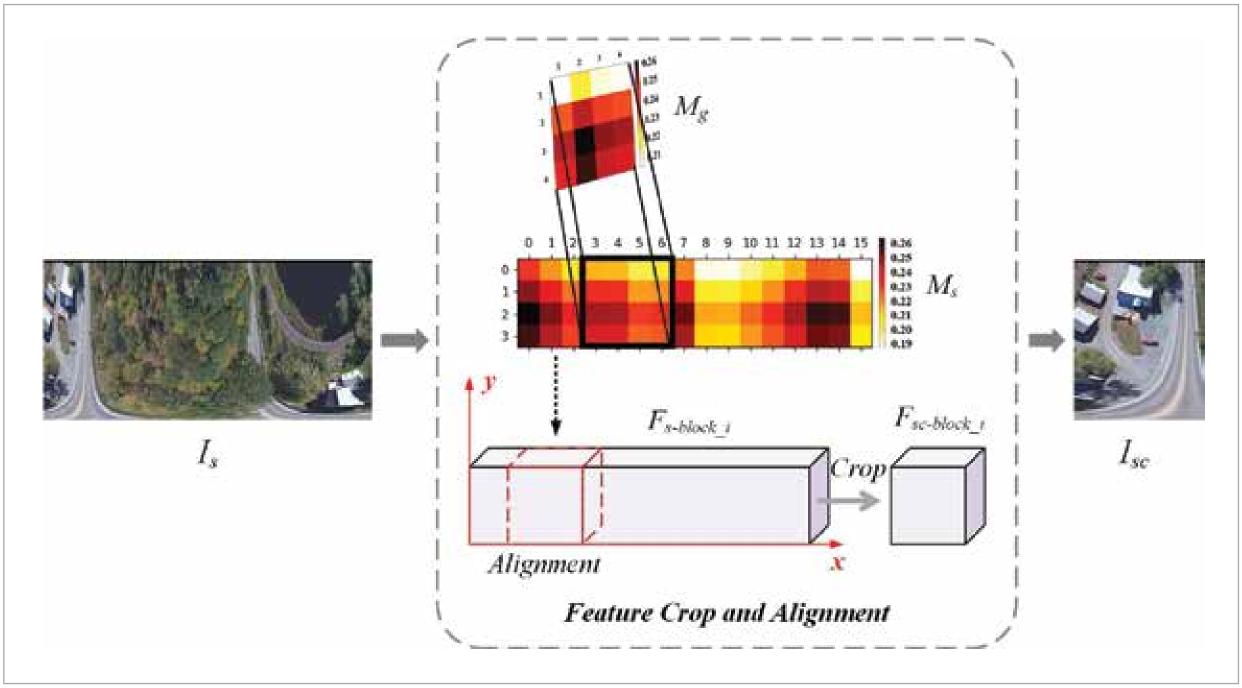


(a) (b)

(c) (d)

**Figure 7**

Feature Cropping and Alignment Process of Satellite Images.



racy and matching effectiveness, rather than simply choosing the first window.

In feature cropping, sliding matching is only done horizontally. The heights of all ground and satellite images in the dataset are uniformly scaled to 128 pixels, ensuring vertical (y-axis) consistency. After undergoing polar coordinate transformation, satellite images generate pseudo-ground panoramic images. These images visually present a geographical scene distribution that is expanded along the horizontal direction, aligning with the horizontal-perspective characteristics of ground images. This implies that the viewpoint differences between ground and satellite images mainly lie in the horizontal direction. The vertical direction shows smaller semantic-structural changes. Thus, we directly use the feature map of ground images as the sliding window. When the height is aligned, sliding-window matching in the vertical direction is unnecessary.

### 3.4. Multi-scale Feature Fusion

When learning similarity measurement between satellite and non-panoramic ground images, feature maps at different levels describe the image at different granularities. When determining if the query and reference satellite images depict the same location, people consider both coarse-grained high-level semantic info (like object outlines and spatial layouts) and fine-grained low-level features (such as colors and textures) from the two view images. Given this, we introduce a multi-scale feature fusion strategy: the Multi-Convolutional Layer Feature Map Concatenate (FMC) method. By concatenating features extracted from different convolutional layers, both high-level semantic information and low-level visual details are effectively integrated to construct image descriptors that capture rich hierarchical representations.

Due to the structural characteristics of the ResNet18 network, the spatial size and number of channels of the feature maps change after passing through each residual block. The satellite-image feature maps cropped from Block 1~Block 4 differ in spatial resolution and channel dimensions, so they cannot be directly concatenated. To effectively fuse multi-scale features, we first flatten each cropped feature map from the four blocks, converting their spatial and channel information into one-dimensional vectors. Then, the four vectors are concatenated on the feature dimension to form the final multi-scale image-feature descriptor.

Specifically, first, in the satellite image network branch, the satellite image features learned from Block 1 to Block 4 of the ResNet18_CA backbone network are each subjected to feature alignment and cropping as described in Section 3.3.2. Second, the cropped satellite image features from four scales are concatenated to form the final multi-scale satellite image feature descriptor $F_s$, as shown in Figure 8(a). Then, for the ground image network branch, as the image is a limited FoV one with a concentrated target area, the ground image features extracted from Block 1 to Block 4 are concatenated. This forms the multi-scale feature descriptor $F_g$ of the ground image, as shown in Figure 8(b). Finally, the concatenated $F_s$ and $F_g$ are input into the loss function module to perform similarity-metric learning, which comprehensively evaluates the impact of multi-scale features on matching results.

### 3.5. Loss Function

The design of the loss function is pivotal for optimizing the cross-view image geo-localization model. We have adopted a weighted soft-margin triplet loss function, as proposed by Hu et al. [8]. The mathematical formulation of this loss is:

$$L = \ln(1 + e^{\alpha(d_{a,p} - d_{a,n})}). \tag{3}$$

Here, $d_{a,p}$ denotes the Euclidean distance between the feature representations of the anchor (a ground image) and the positive sample (a satellite image from the same geographical location), computed as shown in Equation (4). $d_{a,n}$ represents the Euclidean distance between the anchor and the negative sample (a satellite image from a different location), computed as shown in Equation (5). The hyperparameter $\alpha$ controls the margin between positive and negative pairs, which not only accelerates the convergence of the network but also enhances the model's discriminative ability. $f(\cdot)$ represents the feature extraction function, implemented by the ResNet18_CA network.

$$d_{a,p} = \left\| f(a) - f(p) \right\|_2 \tag{4}$$

$$d_{a,n} = \left\| f(a) - f(n) \right\|_2. \tag{5}$$

# 4. Experiment

## 4.1. Datasets and Evaluation Protocols

1 **Dataset**

In this section, the ground and satellite images from the CVUSA dataset [29] and CVACT_val [12] are resized. For ground images with a field of view of 70°, the normalized dimensions are set to 128×99; when the field of view is 90°, the dimensions are adjusted to 128×128. Satellite images are uniformly resized to 128×512. Our proposed model's performance is evaluated on two cross-view datasets. We use the image preprocessing method from [20] to obtain non-panoramic ground images with random shooting directions and FoV angles of 70° and 90°. In model training, the shooting directions of these ground images were randomly selected for each training round.

2 **Training Configuration**

Our proposed method is implemented using PyTorch 1.4.1. Experiments are performed on a system equipped with a TITAN RTX GPU with 24 GB of memory. The network is trained for a total of 100 epochs. For the optimizer, we adopt AdamW with a batch size of 24. The backbone network is initialized with ImageNet pre-trained weights to accelerate convergence through transfer learning. During each training epoch, both the training and validation sets are randomly shuffled to ensure

**Figure 8**

Multi-scale Feature Concatenation of Two view Images: (a) Feature Concatenation of Satellite Image, and (b) Feature Concatenation of Ground Image.

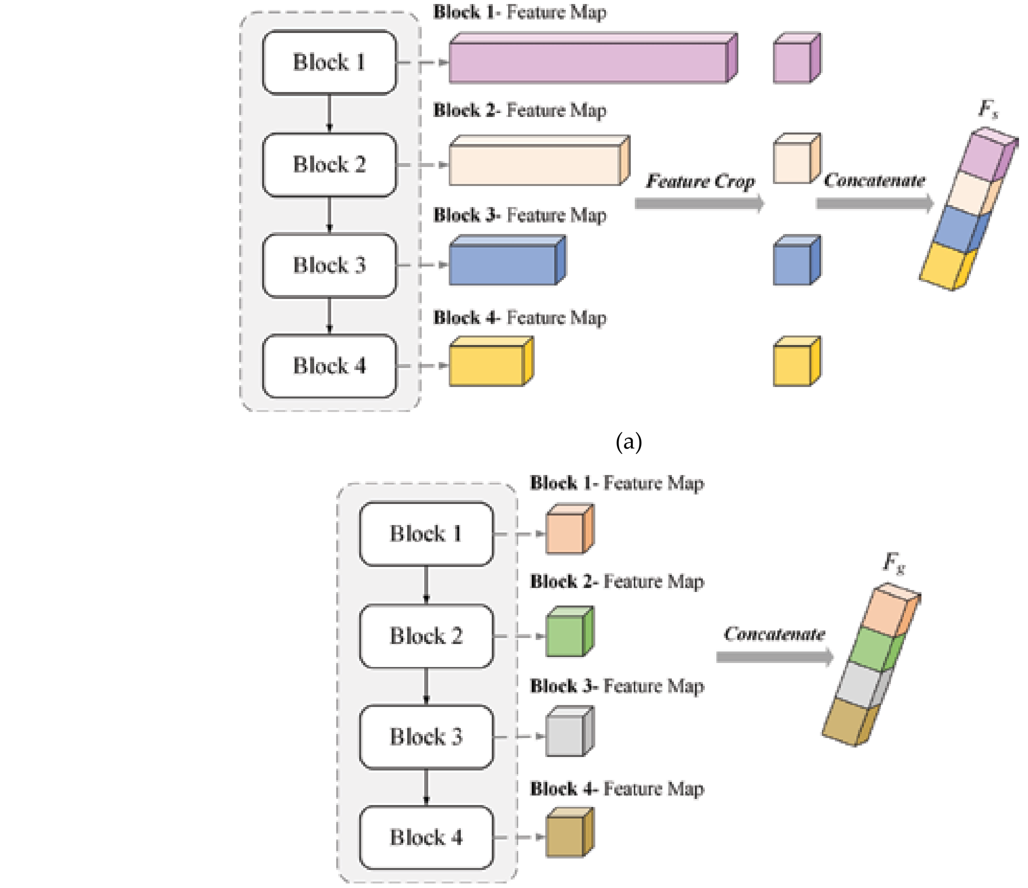


**Table 1**

Comparison with Existing Methods on the CVUSA.

| Method | CVUSA 70° FoV | | | | CVUSA 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| CVM-Net [8] | 2.62 | 9.30 | 15.06 | 21.77 | 2.76 | 10.11 | 16.74 | 55.49 |
| CVFT [21] | 3.79 | 12.44 | 19.33 | 55.56 | 4.80 | 14.84 | 23.18 | 61.23 |
| DSM [20] | 8.78 | 19.90 | 27.30 | 61.20 | 16.19 | 31.44 | 39.85 | 71.13 |
| SEH [5] | 8.39 | 19.54 | 26.04 | / | 17.54 | 32.91 | 40.54 | / |
| **Ours** | **16.51** | **32.67** | **42.88** | **75.78** | **22.45** | **41.07** | **52.15** | **82.09** |

**Table 2**

Comparison with Existing Methods on the CVACT_val.

| Method | CVACT_val 70° FoV | | | | CVACT_val 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| CVM-Net [8] | 1.24 | 4.98 | 8.42 | 34.74 | 1.47 | 5.70 | 9.64 | 38.05 |
| CVFT [21] | 1.49 | 4.13 | 8.19 | 34.59 | 1.85 | 6.28 | 10.54 | 39.25 |
| DSM [20] | 6.91 | 16.46 | 22.28 | / | 13.85 | 28.39 | 36.24 | / |
| SEH [5] | 8.29 | 20.72 | 27.13 | 57.08 | 18.11 | 33.34 | 40.94 | 68.65 |
| **Ours** | **9.81** | **22.73** | **29.29** | **58.91** | **19.56** | **34.02** | **41.50** | **69.24** |

data diversity. Additionally, we employ the hard negative mining strategy proposed by Hermans et al. [6], which focuses on repeatedly learning the most similar but mismatched satellite images to enhance the network's generalization ability. The initial learning rate is set to 0.00005 and is gradually reduced during training to ensure stable model convergence. The weight decay is set to 0.0005. We conducted L2 regularization on the feature descriptors of ground and satellite images.

### 3 Evaluation Protocols

We use the top-$k$ (R@$k$, $k \in \{1,5,10,1\%\}$) recall rate as the evaluation metric, which is the probability of correct results appearing among the top-$k$ retrieved images. When $W$ is an integer, the top-$k$ set refers to the collection of the k satellite images whose feature descriptors are closest to that of a queried ground image. When k is a percentage, the top-$k$ set consists of the first $k \times N$ satellite images closest in feature descriptor to the queried ground image, where $N$ denotes the number of reference satellite images. This section uses four metrics: R@1, R@5, R@10, and R@1%.

## 4.2. Experimental Results

### 1 Comparison with other Methods

We compared our method with several advanced ones on CVUSA and CVACT_val. The results are shown in Tables 1-2. As shown in Table 1, on the CVUSA dataset, our method significantly improves recall accuracy. For ground images with a 70° FoV, our R@1 metric improves by 8.12%, R@5 by 13.13%, R@10 by 16.24%, and R@1% surpasses DSM by 24.68%. For ground images with a 90° FoV, our method achieves notable improvements: a 5% increase in R@1, an 8% rise in R@5, an 11.6% gain in R@10, and an 11% improvement in R@1% over DSM.

As shown in Table 2, on the CVACT_val dataset, for ground images with a 70° field of view, we improved R@1 from 8.29% to 9.81%, R@5 from 20.72% to 22.73%, R@10 from 27.13% to 29.29%, and increased R@1% by 11.83% from its original value. For ground images with a 90° field of view, we improved R@1 from 18.11% to 19.56%, R@5 from 33.34% to 34.02%, R@10 from 40.94% to 41.50%, and R@1% from 68.65% to 69.24%. This indicates that the proposed method can better handle cross-view geo-localization tasks on non-panoramic ground images taken in real-world settings.

2 **Experimental Results Analysis of Two Datasets**

The performance difference observed between the two datasets primarily stems from their inherent variations in scene distribution and visual content. The CVUSA dataset mainly comprises suburban scenes, such as forests, deserts, and open roads, which are relatively simple and contain fewer occlusions. These characteristics facilitate the establishment of stable visual correspondences between different viewpoints. In contrast, the CVACT dataset is collected in urban environments and includes a wide range of complex semantic elements, such as vehicles, pedestrians, and densely packed buildings. These elements often lack clear spatial localization cues and are susceptible to dynamic changes and occlusions, making it challenging to provide stable references for image matching. Furthermore, urban scenes exhibit greater content variations across viewpoints, resulting in reduced consistency between ground-level and satellite images. Consequently, the proposed model demonstrates a comparatively smaller performance improvement on the CVACT dataset.

## 4.3. Ablation Studies

1 **Effect of the Utilized CA module**

To verify the effectiveness of the CA module in cross-view image geo-localization, we conducted ablation studies on the CVUSA and CVACT_val benchmark datasets. The results are presented in Tables 3-4. First, we use the baseline network ResNet18 for cross-view image matching and record its performance metrics. Then, we conduct experiments with the improved ResNet18_CA network incorporating the CA module under the same settings, and compare their recall performance at different FoV angles (70° and 90°).

As the results show, on both the CVUSA and CVACT_val datasets, the ResNet18_CA network with the CA module significantly improves recall accuracy at all FoV angles compared to the baseline ResNet18 model. It demonstrates that the CA module, by encoding spatial positions in feature maps, enhances the model's sensitivity to spatial information, and enables more precise capture of key areas in ground images. It also shows the CA module's effectiveness and robustness in better aligning cross-view image features.

2 **Effect of the Utilized CA Module, FCA Module, and FMC Module**

To systematically evaluate the combined impact of the CA module, FCA module, and FMC module on cross-view image geo-localization performance, we conducted joint ablation studies on the CVUSA and CVACT_val benchmark datasets. The experiments used various module combination configurations to assess each module's contribution to model performance. The results are shown in Tables 5-6, where "√" indicates a module was used and "-" indicates it was not.

The results indicate that for non-panoramic images with 70° and 90° FoV, combining the CA and FCA modules enhances recall accuracy compared to using only the CA module. It shows that feature cropping is effective for removing redundant features and strengthening key-area matching.

Moreover, the model achieves optimal R@1, R@5, R@10, and R@1% metrics when the FMC module is further introduced. This validates the effective synergy of the CA, FCA, and FMC modules. By fusing multi-scale features, the model's representation ability is enhanced. It improves capture multi-granularity semantic information in cross-view images, boosting overall cross-view geo-localization performance.

**Table 3**

Ablation Study of the CA Module on the CVUSA.

| Model | CVUSA 70° FoV | | | | CVUSA 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| ResNet18 | 12.31 | 27.47 | 36.02 | 67.73 | 15.92 | 34.53 | 44.95 | 76.45 |
| ResNet18_CA | **13.36** | **29.68** | **38.73** | **72.18** | **18.33** | **37.47** | **47.26** | **79.27** |

**Table 4**
Ablation Study of the CA Module on the CVACT_val.

| Model | CVACT_val 70° FoV | | | | CVACT_val 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| ResNet18 | 7.65 | 17.63 | 24.86 | 55.64 | 11.15 | 24.65 | 32.28 | 64.06 |
| ResNet18_CA | **9.04** | **21.69** | **28.12** | **58.20** | **15.42** | **27.21** | **36.69** | **67.18** |

**Table 5**
Ablation Study of Combined Three Modules on CVUSA.

| Module | | | CVUSA 70° FoV | | | | CVUSA 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CA | FCA | FMC | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| √ | - | - | 13.36 | 29.68 | 38.73 | 72.18 | 18.33 | 37.47 | 47.26 | 79.27 |
| √ | √ | - | 15.09 | 31.72 | 41.64 | 74.36 | 21.18 | 39.26 | 49.42 | 81.40 |
| √ | √ | √ | **16.51** | **32.67** | **42.88** | **75.78** | **22.45** | **41.07** | **52.15** | **82.09** |

**Table 6**
Ablation Study of Combined Three Modules on CVACT_val.

| Module | | | CVACT_val 70° FoV | | | | CVACT_val 90° FoV | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CA | FCA | FMC | R@1 | R@5 | R@10 | R@1% | R@1 | R@5 | R@10 | R@1% |
| √ | - | - | 9.04 | 21.69 | 28.12 | 58.20 | 15.42 | 27.21 | 36.69 | 67.18 |
| √ | √ | - | 9.34 | 22.09 | 28.46 | 58.83 | 18.92 | 31.60 | 39.09 | 67.57 |
| √ | √ | √ | **9.81** | **22.72** | **29.29** | **58.91** | **19.56** | **34.02** | **41.50** | **69.24** |

## 5. Discussion

In this work, we primarily focus on analyzing the impact of ground images with limited F FoV on localization accuracy. The current research is centered on the examination of FoV, without incorporating the potential influence of camera parameters—such as focal length and exposure—on the content of ground image scenes. However, we recognize that these camera parameters constitute critical factors in image analysis. They may significantly affect the accuracy of cross-view matching by altering the image's perspective range, brightness distribution, and the quality of feature extraction. For instance, variations in focal length can influence the scaling and detail presentation of the scene, thereby impacting the effectiveness of feature extraction. Similarly, adjustments in exposure may affect the image's contrast and feature discernibility, particularly in scenes with varying lighting conditions. In future research, we intend to prioritize this aspect as a key focus area, aiming to analyzing the impact of factors in the practical applications on geo-localization performance.

## 6. Conclusion

In this paper, we address the challenge of cross-view image geo-localization for non-panoramic ground images with limited field of view. We propose a novel attention-weighted mask alignment approach, leveraging a lightweight ResNet18 network embedded with the Coordinate Attention mechanism. By generating attention weight masks to prioritize task-relevant regions that enable precise feature alignment between ground and satellite images. Additionally, we incorporate a feature cropping module to eliminate redundant areas in satellite images, retaining only the key regions corresponding to the ground images, thereby enhancing both the efficiency and accuracy of feature matching. Furthermore, a multi-scale feature fusion strategy is employed to combine features from different convolutional layers, generating more representative image descriptors that improve the overall localization accuracy. Experimental results show that our method achieves remarkable performance improvements on the CVUSA and CVACT_val datasets. For non-panoramic ground images with 70° and 90° FoV angles, our method outperforms existing ones in R@1, R@5, R@10, and R@1% metrics. This indicates it effectively overcomes limitations of traditional methods in feature alignment and direction perception, offering a new solution for cross-view geo-localization of non-panoramic ground images.

## References

1. Cai, S. D., Guo, Y. L., Khan, S., Hu, J. W., Wen, G. J. Ground-to-Aerial Image Geo-Localization with a Hard Exemplar Reweighting Triplet Loss. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 8391-8400. https://doi.org/10.1109/ICCV.2019.00848
2. Cheng, L., Wang, T., Meng, L. Q., Sun, C. Y. Window-to-Window BEV Representation Learning for Limited FoV Cross-View Geo-Localization. arXiv Preprint, 2024, arXiv:2407.06861. https://doi.org/10.48550/arXiv.2407.06861
3. Dai, M., Hu, J. H., Zhuang, J. D., Zheng, E. H. A Transformer-Based Feature Segmentation and Region Alignment Method for UAV-View Geo-Localization. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(7), 4376-4389. https://doi.org/10.1109/TCSVT.2021.3135013
4. Fervers, F., Bullinger, S., Bodensteiner, C., Arens, M., Stiefelhagen, R. Uncertainty-Aware Vision-Based Metric Cross-View Geo-Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, 21621-21631. https://doi.org/10.1109/CVPR52729.2023.02071
5. Guo, Y. L., Choi, M., Li, K. H., Boussaid, F., Bennamoun, M. Soft Exemplar Highlighting for Cross-View Image-Based Geo-Localization. IEEE Transactions on Image Processing, 2022, 31, 2094-2105. https://doi.org/10.1109/TIP.2022.3152046
6. Hermans, A., Beyer, L., Leibe, B. In Defense of the Triplet Loss for Person Re-Identification. arXiv Preprint, 2017, arXiv:1703.07737. https://doi.org/10.48550/arXiv.1703.07737
7. Hou, Q. B., Zhou, D. Q., Feng, J. S. Coordinate Attention for Efficient Mobile Network Design. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350
8. Hu, S. X., Feng, M. D., Nguyen, R. M. H., Lee, G. H. CVM-Net: Cross-View Matching Network for Image-Based Ground-to-Aerial Geo-Localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 7258-7267. https://doi.org/10.1109/CVPR.2018.00758
9. Li, A., Hu, H. Y., Mirowski, P., Farajtabar, M. Cross-View Policy Learning for Street Navigation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, 8100-8109. https://doi.org/10.1109/ICCV.2019.00819
10. Li, C. R., Yan, C., Xiang, X. J., Lai, J., Zhou, H., Tang, D. Q. AMPLE: Automatic Progressive Learning for Orientation Unknown Ground-to-Aerial Geo-Localization. IEEE Transactions on Geoscience and Remote Sensing, 2024, 63, 5800115. https://doi.org/10.1109/TGRS.2024.3517654
11. Lin, T. Y., Cui, Y., Belongie, S., Hays, J. Learning Deep Representations for Ground-to-Aerial Geo-Localization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, 5007-5015. https://doi.org/10.1109/CVPR.2015.7299135
12. Liu, L., Li, H. Lending Orientation to Neural Networks for Cross-View Geo-Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, 5624-5633. https://doi.org/10.1109/CVPR.2019.00577
13. Meng, X. Y., Wang, W., Leong, B. SkyStitch: A Cooperative Multi-UAV-Based Real-Time Video Surveillance System with Stitching. Proceedings of the 23rd ACM International Conference on Multimedia, 2015, 261-270. https://doi.org/10.1145/2733373.2806225
14. Mi, L., Xu, C., Castillo-Navarro, J., Montariol, S., Yang, W., Bosselut, A., Tuia, D. ConGeo: Robust Cross-View Geo-Localization Across Ground View Variations. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024, 214-230. https://doi.org/10.1007/978-3-031-72630-9_13

15. Mithun, N. C., Minhas, K. S., Chiu, H. P., Oskiper, T., Sizintsev, M., Samarasekera, S. Cross-View Visual Geo-Localization for Outdoor Augmented Reality. 2023 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), IEEE, 2023, 493-502. https://doi.org/10.1109/VR55154.2023.00064

16. Regmi, K., Borji, A. Cross-View Image Synthesis Using Conditional GANs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, 3501-3510. https://doi.org/10.1109/CVPR.2018.00369

17. Rodrigues, R., Tani, M. Are These from the Same Place? Seeing the Unseen in Cross-View Image Geo-Localization. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, 3753-3761. https://doi.org/10.1109/WACV48630.2021.00380

18. Rodrigues, R., Tani, M. Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, 3871-3879. https://doi.org/10.1109/WACV51458.2022.00275

19. Shi, Y. J., Liu, L., Yu, X., Li, H. D. Spatial-Aware Feature Aggregation for Image Based Cross-View Geo-Localization. Advances in Neural Information Processing Systems, 2019, 905, 10090-10100.

20. Shi, Y. J., Yu, X., Campbell, D., Li, H. D. Where Am I Looking At? Joint Location and Orientation Estimation by Cross-View Matching. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 4064-4072. https://doi.org/10.1109/CVPR42600.2020.00412

21. Shi, Y. J., Yu, X., Liu, L., Zhang, T., Li, H. D. Optimal Feature Transport for Cross-View Image Geo-Localization. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(07), 11990-11997. https://doi.org/10.1609/aaai.v34i07.6875

22. Shugaev, M., Semenov, I., Ashley, K., Klaczynski, M., Cuntoor, N., Lee, M. W. ArcGeo: Localizing Limited Field-of-View Images Using Cross-View Matching. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, 209-218. https://doi.org/10.1109/WACV57701.2024.00028

23. Sun, B., Chen, C., Zhu, Y. Y., Jiang, J. M. GeoCapsNet: Ground to Aerial View Image Geo-Localization Using Capsule Network. 2019 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2019, 742-747. https://doi.org/10.1109/ICME.2019.00133

24. Toker, A., Zhou, Q., Maximov, M., Leal-Taixe, L. Coming Down to Earth: Satellite-to-Street View Synthesis for Geo-Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, 6488-6497. https://doi.org/10.1109/CVPR46437.2021.00642

25. Wang, T., Fan, S. J., Liu, D. K., Sun, C. Y. Transformer-Guided Convolutional Neural Network for Cross-View Geo-Localization. arXiv Preprint, 2022, arXiv:2204.09967. https://doi.org/10.48550/arXiv.2204.09967

26. Wang, T. Y., Zheng, Z. D., Yan, C. G., Zhang, J. Y., Sun, Y. Q., Zheng, B. L. Each Part Matters: Local Patterns Facilitate Cross-View Geo-Localization. IEEE Transactions on Circuits and Systems for Video Technology, 2021, 32(2), 867-879. https://doi.org/10.1109/TCSVT.2021.3061265

27. Workman, S., Souvenir, R., Jacobs, N. Wide-Area Image Geolocalization with Aerial Reference Imagery. Proceedings of the IEEE International Conference on Computer Vision, 2015, 3961-3969. https://doi.org/10.1109/ICCV.2015.451

28. Yang, H. J., Lu, X. F., Zhu, Y. Y. Cross-View Geo-Localization with Layer-To-Layer Transformer. Advances in Neural Information Processing Systems, 2021, 29009-29020.

29. Zhai, M., Bessinger, Z., Workman, S., Jacobs, N. Predicting Ground-Level Scene Layout from Aerial Imagery. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 867-875. https://doi.org/10.1109/CVPR.2017.440

30. Zhang, C. H., Ge, S. M., Zhang, K. K., Zeng, D. Accurate UAV Tracking with Distance-Injected Overlap Maximization. Proceedings of the 28th ACM International Conference on Multimedia, 2020, 565-573. https://doi.org/10.1145/3394171.3413959

31. Zhang, C., Lam, K.-M., Wang, Q. Cof-Net: A Progressive Coarse-to-Fine Framework for Object Detection in Remote-Sensing Imagery. IEEE Transactions on Geoscience and Remote Sensing, 2023, 61, 1-17. https://doi.org/10.1109/TGRS.2022.3233881

32. Zhang, X. H., Li, X. Y., Sultani, W., Chen, C., Wshah, S. GeoDTR+: Toward Generic Cross-View Geo-Localization via Geometric Disentanglement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12), 10419-10433. https://doi.org/10.1109/TPAMI.2024.3443652

33. Zhang, X. W., Meng, X. C., Yin, H. L., Wang, Y. X., Yue, Y. Z., Xing, Y. H. SSA-Net: Spatial Scale Attention Network for Image-Based Geo-Localization. IEEE Geoscience and Remote Sensing Letters, 2022, 19, 1-5. https://doi.org/10.1109/LGRS.2021.3120658

34. Zhang, X. H., Sultani, W., Wshah, S. Cross-View Image Sequence Geo-Localization. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, 2914-2923. https://doi.org/10.1109/WACV56688.2023.00293

35. Zhao, J. W., Zhai, Q., Huang, R., Cheng, H. Mutual Generative Transformer Learning for Cross-View Geo-Localization. arXiv e-prints, 2022, arXiv:2203.09135. https://doi.org/10.3390/rs15092221

36. Zhu, S. J., Shah, M., Chen, C. TransGeo: Transformer Is All You Need for Cross-View Image Geo-Localization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 1162-1171. https://doi.org/10.1109/CVPR52688.2022.00123

37. Zhu, S. J., Yang, T. J. N., Chen, C. ViGOR: Cross-View Image Geo-Localization Beyond One-To-One Retrieval. IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE, 2021, 3640-3649. https://doi.org/10.1109/CVPR46437.2021.00364

38. Zhu, Y. Y., Sun, B., Lu, X. F., Jia, S. Geographic Semantic Network for Cross-View Image Geo-Localization. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60, 1-15. https://doi.org/10.1109/TGRS.2021.3121337