Multi-branch Joint Representation Learning Based on Information Fusion Strategy for Cross-view Geo-localization

Multi-branch Joint Representation Learning Based on Information Fusion Strategy for Cross-view Geo-localization

Fawei Ge, Yunzhou Zhang*, Yixiu Liu, Guiyuan Wang, Sonya Coleman, Dermot Kerr and Li Wang

***Abstract*—Cross-view geo-localization refers to recognizing images of the same geographic target obtained from different platforms (such as drone-view, satellite-view and ground-view). However, cross-view geo-localization is challenging as image capture using different platforms coupled with extreme viewpoint variations can cause significant changes to the visual image content. Existing methods mainly focus on mining the fine-grained features or the contextual information in neighboring areas, but ignore the complete information of the entire image and the association of contextual information of adjacent regions. Therefore, a multi-branch joint representation learning network model based on information fusion strategies is proposed to solve this cross-view geo-localization problem. Firstly, we obtain feature information from the image through global information fusion branch and local information fusion branch to help the network learn the discernable information in the different images. In addition, a local-guided-global information fusion branch is introduced to make local information assist global features to enhance the learning of potential information in the images. Secondly, we introduced different information fusion strategies in each branch to increase the extraction of contextual information through expanding the global receptive field, thus improving the performance of the model. Finally, a series of experiments is carried out on three prevailing benchmark datasets, namely University-1652, SUES-200, CVUAS and CVACT datasets. The quantitative comparisons from the experiments clearly indicate that the proposed network framework has great performance. For example, compared with some state-of-the-art methods, the quantitative improvements of the R@1 and AP on the University-1652 datasets are 1.91%, 2.18% and 1.55%, 2.99% in both tasks, respectively.**

***Index Terms*—Geo-localization, multi-branch, hybrid information fusion strategies, joint representation learning.**

I. Introduction

CROSS-VIEW geo-localization aims to retrieve the most relevant images of the same geographic target from different platforms, and has been widely used in many fields, such as accurate delivery, autonomous driving, action recognition, change detection, event detection and land cover classification [2]–[8]. In the era of digital maps, it is usually necessary to estimate the geospatial localization of a given object in real-time. This can be done with real-time Kinematic (RTK) GPS, but these sensors are expensive and short time signal interruptions can hinder workflows. In addition, especially in the city, the urban canyon effect will produce a certain deviation. At present, cross-view geo-localization based on image retrieval is an effective method to solve these problems [9]. In practical application, the information obtained from different source data for the same target in different tasks is different and related. Therefore, it is necessary to correlate images from different views [10]. For example, given a drone-view image, it is necessary to retrieve images of the same location from other viewpoints to obtain the geographic information of the location. Given locational information is available from different image sources such as satellite, drone, or ground, studying the cross-view geo-localization problem is extremely important [11]–[14]. However, the scale, viewpoint and imaging modality for images obtained through different platforms can be very different, *e.g.*, the ground-view is almost perpendicular to the horizon, while the satellite-view is almost parallel to the horizon. Therefore, cross-view geo-localization

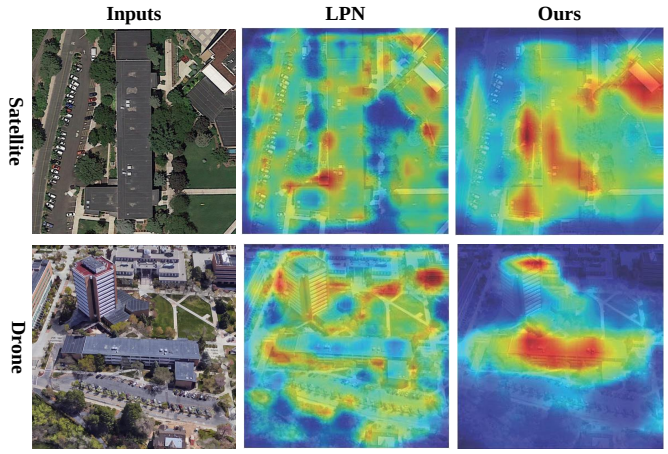

Fig. 1. Difference of the activation maps generated by the LPN method [1] and our method. The images on the left column are the input drone-view and satellite-view images. The images in the middle column are the heatmaps of LPN method. The images ont the right column are the heatmaps of our method. From the visualization result, it can be seen that our method focus on the important information in the image.

Yunzhou Zhang (Corresponding Author) is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China. (e-mail: zhangyunzhou@mail.neu.edu.cn).

Fawei Ge is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China. (e-mail: gefawei0822@163.com).

Yixiu Liu is with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China. (e-mail: liuyixiu@hdu.edu.cn).

Guiyuan Wang is with the Jiangsu Shuguang Opto-electronics co.. LTD, Yangzhou, Jiangsu, China, 225000.

Sonya Coleman and Dermot Kerr are with the Intelligent Systems Research Centre, University of Ulster, Derry BT52 1SA, UK.

Li Wang is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, Liaoning, China.

is a challenging task [15].

In recent years, cross-view geo-localization has made significant progress with the introduction of deep learning. A lot of existing approaches use pre-trained deep learning networks to extract the features from different platforms, and use metric learning to distinguish whether these image feature representations have the same geospatial localization [10], [16], [17]. In this process, the network will learn a feature space to make the image features of the same geographic target closer, and push unmatched pairs far apart to complete cross-view geo-localization tasks [18], [19]. In addition, auxiliary information such as attention mechanisms, orientation information and aligning the spatial layout of features is also widely used in deep learning networks to improve the performance of network models [20]–[22]. However, these geo-location methods only consider the global information of the input images and ignore other contextual information, which can cause images with similar regions to be difficult to differentiate, resulting in mismatching of cross-view images from different platforms.

Due to images obtained by different platforms, such as drone-view or satellite-view, are typically captured at a wide angle. These images also contain other information around the target scene while acquiring the geospatial target, which may also have a significant impact on the results. Most current methods often ignore this contextual information in neighboring areas, which provides critical information for cross-view geo-localization. When there is no obvious landmarks in the scene, the visual difference between similar scenes is not very obvious, and it can be difficult for the human visual system to distinguish such scenes and determine the real target. In view of this, the inclusion of contextual information can effectively reduce the difficulty of the task. Therefore, mining and utilizing the contextual information in images can effectively improve the model performance for cross-view geo-localization [1].

Although mining and utilizing contextual information in images can effectively improve the accuracy of cross-view geo-localization, the use of global information in images is equally important. As shown in Fig.1, a previous work, LPN [1], focuses more on the contextual information for the input image, ignoring other crucial information in the scenes. In order to mine effective information in cross-view images, and inspired by existing methods [1], [23], a multi-branch joint representation learning network model based on information fusion strategies is proposed in this paper to solve the cross-view geo-localization problem. For this problem, we believe that each part of the image has a significant impact on the result of image matching. Therefore, we adapt a multi-branch joint representation learning network model to solve this problem, which is divided into three branches, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In the global information fusion branch, the global features of an image can effectively express the content information in the complete image scene, most existing methods use this concept to solve cross-view scene matching. However, only using the global information may result in the acquired features that are not sufficiently discernible for certain scenes. In order to mine the global information effectively, we introduce the global information fusion strategy into it to increase the global receptive field of the network, which can improve the utilization of global information. However, as it is difficult to distinguish similar scenes completely only using global information, the contextual information in the image can help with feature matching.

Therefore, we design a local information fusion branch to improve the performance of the network model. In the local information fusion branch, the contextual information in the image will have a positive impact on feature matching. In order to better mine the contextual information in the image, we process the global information into blocks. At the same time, in order to better mine and utilize the crucial information of each segmented part, we introduce the local information fusion strategy to expand the receptive field of each part. On this basis, we divide each block feature into several parts in a square-ring partition to obtain the contextual information of each part thus assisting the global features to distinguish similar scenes. In addition, we believe that local features can assist global features to better mine information in the image. In this regard, we also introduce a local-guided-global information fusion branch, which mainly used local features after segmentation to assist global features and introduce a mixing information fusion strategy and attention mechanism to further increase the global receptive field and mine more useful potential information. Through these three branches, the effective information in the cross-view image can be mined and utilized effectively to solve the cross-view geo-localization problem.

The contributions of this paper are summarized as follows:

- A multi-branch joint representation learning network model based on information fusion strategies is proposed to solve the cross-view geo-localization problem, which consists of three parts, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.
- In order to obtain more robust features, based on the use of global information and local information, we utilize the idea of the local-guided-global information to build model branch without introducing additional information and assist network model to further mine latent crucial information in the image, which can further improve the performance network model.
- To further mine and utilize the crucial information in the image, three information fusion strategies (IFS) are designed to the three proposed branches to assist each branch of the network model to increase the global receptive field. On this basis, each branch can more effectively mine and utilize the relevant feature information and improve the discrimination of features.
- A series of experiments is carried out using the University-1652, SUES-200, CVUAS and CVACT datasets, and the experimental results demonstrate that the effectiveness of the proposed network framework.

II. Related Work

A. Deep Cross-view Geo-localization

In recent years, with the emergence of a large number of potential application directions in this field, the cross-view geo-localization problem has attracted more and more attention [24], [25]. Early pioneering work [26]–[28] focussed on addressing the cross-view geo-localization task by extracting hand-crafted features and while some progress was made, however, the features extracted by these methods were not robust in some cross-view scenes. With the development of deep convolutional neural networks (CNNs) and significant success [29], [30], more scholars have studied deep learning to extract robust depth features to complete the cross-view geo-localization task. Workman et al. [31] first used the pre-trained CNN model to extract features from cross-view images to complete the cross-view geo-localization task. Then Workman et al. [32] introduced cross-view training to learn joint semantic feature representations between images. Although these methods have achieved some improvements in the cross-view geo-localization task, focussing on a single type of feature is non-ideal due to the fact that cross-view images are captured from different platforms and have a large content gap.

The idea of aligning the spatial layout of features has been introduced into the cross-view geo-localization task to compensate for the image context changing in different viewpoints and platforms [33]. Shi et al. [10] proposed a regular polar transform to warp an aerial image into a panoramic image close to the ground-view, which can align the features of two-view images in a certain space. In addition, Zhai et al. [16] used an adaptive transformation to map these extracted features into the ground-level perspective. Although the accuracy of the cross-view geo-localization task has been significantly improved by aligning the spatial layout of features, these methods usually only focus on mining the global information but omit contextual information, which can have a significant impact on the final results.

In recent years, methods have focused on the use of contextual information to enhance the accuracy of the cross-view geo-localization task. LPN [1] used contextual information and deployed the square-ring partition strategy to mine additional information in an end-to-end manner. In addition, as a transformer-based model has strong local information mining ability, and with the rapid development of transform models, some methods now include it in the cross-view geo-localization task. FSRA [22] introduced a simple and effective transformer-based structure to enhance the ability of a model to understand contextual information. After considering the contextual information of the image, the overall performance of these methods was improved to some extent.

However, excessively mining and using the contextual information can cause the model to over-consider the peripheral information and ignore other crucial information, which may reduce the performance of the network model. Therefore, a multi-branch joint representation learning network model based on information fusion strategies is proposed to solve the cross-view geo-localization problem. We consider the global information, local information and local-guided-global information to ensure the network model fully mines the crucial information in the image. We then introduce different information fusion strategies to increase the global receptive field of the model and hence increase the model's performance, and effectively complete the cross-view geo-localization task.

B. Part-based Representation Learning

In the design of traditional algorithms, local features have been widely studied and applied in many fields [34]–[36]. Ojala et al. [37] proposed a generalized gray-scale and rotation invariant feature and proved the effectiveness of these features via experiments. Lowe et al. [38] designed a scale-invariant feature transform descriptor (SIFT) for image matching. SIFT is invariant to translations, rotation, and scaling transformations through summarizing the description of the local image structures in a local neighborhood around each interest point, and has been widely applied in many fields due to its excellent performance. Although these methods work well in simple environments, the feature lack robustness in complex environments.

Recently studies have focused on local pattern learning in deep-learning models. Spindle Net [39] is a novel network based on human body region guided multi-stage feature decomposition and tree-structured competitive feature fusion, which improves model performance by extracting semantic features and merging competitive schemes. MSVAN [40] learned the powerful features of each part by stacking multi-scale convolutions to obtain the contextual information from the image. Zheng et al. [41] proposed a novel network for discriminative embedded learning and pedestrian alignment. These methods achieve good results when solving related problems, and demonstrate that global contextual information plays a crucial role in solving some problems.

With the wide application of deep convolutional networks in various fields, attempts have been made to improve the stability of network models by extracting contextual information from images. In a CNN, one of the easiest ways to obtain global contextual information is to use global average pooling [42]. However, this method cannot effectively integrate global contextual information into each pixel representation. In order to solve this problem, attention mechanisms are widely used in deep learning to acquire more critical information in images. Xu et al. [43] proposed an Attention-Aware Compositional Network (AACN) framework, which introduces Pose-guided Part Attention and Attention-aware Feature Composition to enhance the network model performance. Guo et al. [44] applied a human parsing model to extract the binary human part masks and a self-attention mechanism to capture features. Although the attention mechanism is effective in obtaining global contextual information, frequently using it will increase the computational complexity of the network model.

Therefore, we adopt a multi-branch joint representation learning network model based on information fusion strategies to fully mine the crucial information and contextual information in the image. Furthermore, different information fusion strategies is introduced into each branch to provide contextual information to further improve the network model performance. In addition, we introduce the attention mechanism into

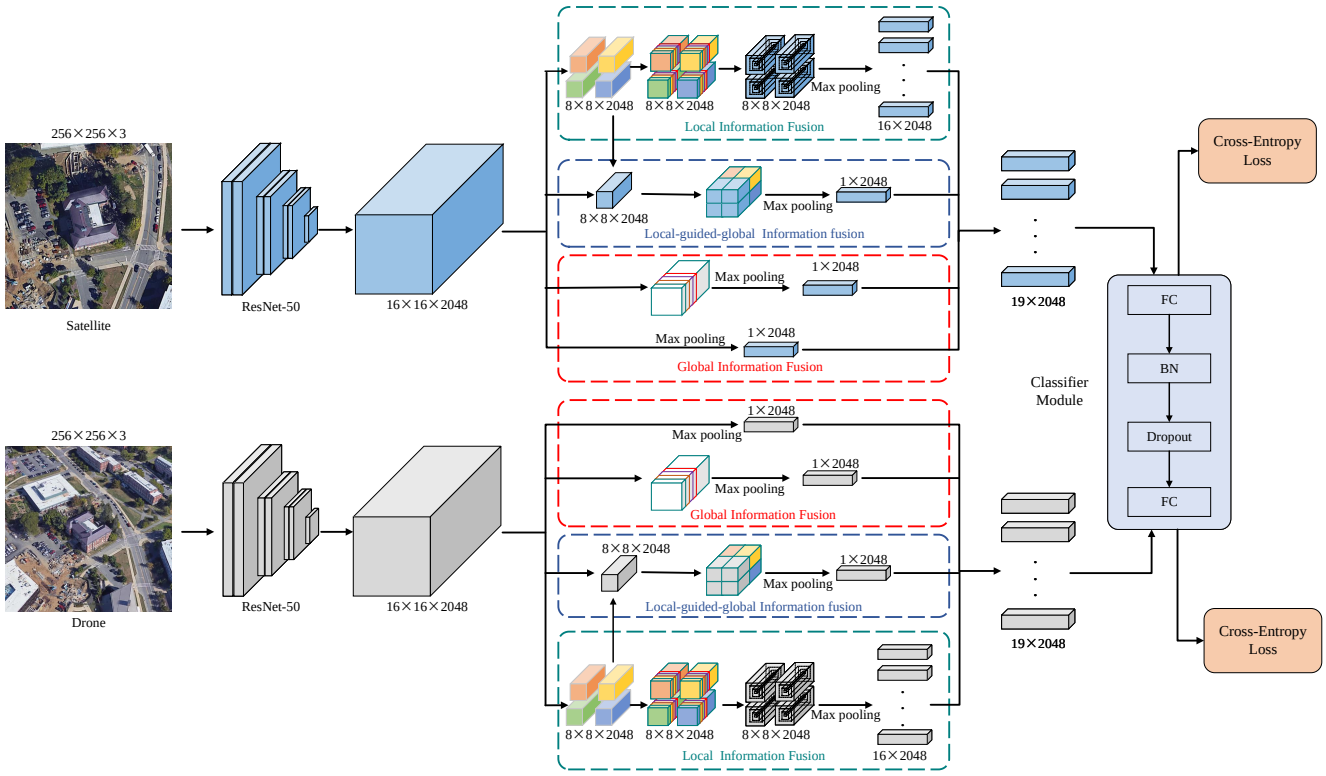

Fig. 2. The overview of the proposed network framework. As the input image may come from different platforms the proposed network model will be set according to different task requirement. It is worth noting that the illustrated network framework shows the drone-view and satellite-view in the University-1652 dataset as example inputs. When the ground-view is needed, network models can be replaced or added as required. In addition, the weights in each models are not shared. During the training period, the network model is optimized mainly through the cross-entropy loss function.

the local-guided-global information fusion branch to further mine the crucial information to enhance the robustness of feature extraction. We only use the attention mechanism in one module, and the network model performance can be improved without significantly increasing the computational complexity.

III. Methodology

In cross-view geo-localization, the contents of different source images will change greatly due to the different obtained views. For the same task object, there will be a lot of relevant information in the different images. Therefore, how to match or mine the relevant information in these different source images is the key to solve this problem. In this regard, by mining the global information and local information of images from different sources, we can associate key information from different views as much as possible and determine the image of the same place, so as to complete the cross-view geo-localization task. In this section, we provide a detailed introduction to the proposed multi-branch joint representation learning network model based on information fusion strategies. The proposed network framework is shown in Fig.2.

Problem formulation. Given a cross-view geo-localization dataset, we represent the input image as x and the input image label as y. In addition, x_i represents the acquisition platform to which the input image data belongs, and $i \in \{1, 2, 3\}$, where x_1 represents the satellite-view image, x_2 represents the drone-view image, and x_3 represents the ground-view image. The label of the sample is $y \in [1, C]$, where C represents the number of sample categories. Suppose that the dataset contains 701 buildings, each building containing multiple images from different perspectives, and then the 701 buildings will be numbered as 701 label indexes, each index representing a category, *i.e.*, the label $y \in [1, 701]$. For the cross-view geo-localization task it is necessary to learn a mapping function for images from different perspectives. Images obtained from different platforms should be projected into a feature space to ensure that the features of images from different perspectives, at the same locations, are similar and the features of images from different locations are not similar.

The proposed network framework is comprised of three branches, the global information fusion branch, the local information fusion branch and the local-guided-global information fusion branch, these branches will be introduced in detail in this section.

A. Global Information Fusion Branch

Firstly, we introduce the global information fusion branch in the network framework. It is worth noting that the input images obtained from different platforms are set as different network branches. The model structure of each model is

consistent, although the weights are not shared among the network models. ResNet50 [45] is chosen as the backbone network to extract global and contextual features. Assuming that this process can be expressed as a function F_{ResNet50}, the process of extracting global features can be expressed as follows:

$$f_i = F_{\mathrm{ResNet50}}(x_i) \tag{1}$$

where x_i is the input image and f_i is the global features extracted from the image x_i.

After obtaining the global features, the proposed global information fusion branch will be activated. We believe that the global information in the image has a significant impact on the performance for the network model. Therefore, it is designed to deeply mine the global information and will provide two global feature descriptors for the final result, namely the global information descriptor and the global feature descriptor, based on the global information fusion strategy. The global information descriptor is obtained through inputting the global feature descriptor into the max pooling layer. The size of the global information description and the global feature descriptors are $16 \times 16 \times 2048$ and 1×2048, respectively. This process can be expressed as follow formula:

$$d_i = Maxpool(f_i) \tag{2}$$

where $Maxpool$ represents the maximum pooling operation, and d_i is the output feature descriptor from global feature f_i.

It is difficult to effectively obtain the complete information from an image by simply manipulating the global features through the pooling layer. Therefore, the global information fusion strategy [23] is introduced to increase the global receptive field of the network and obtain more effective global features. The concrete implementation process of the global information fusion strategy is shown in Fig.3. In this strategy, the features are first divided into $p_h \cdot p_w$ blocks; hence each group of features is segmented based on the channel. The newly generated block feature consists of a large number of features from the original location and a small number of features from other blocks to synthesize a new complete global feature. In this process, the sampling location will be controlled by offsets $\hat{h}$ and $\hat{w}$, and further divided $c = \left\{\frac{C}{2}, \ldots, C-1\right\}$ into $p_h \cdot p_w$ sub-groups, each sub-group has $\left\lfloor \frac{C}{2p_h \cdot p_w} \right\rfloor$ channels. On this basis, the features are divided into spatial dimensions and each part will contain $\left\lfloor \frac{H}{p_h} \right\rfloor \cdot \left\lfloor \frac{W}{p_w} \right\rfloor$ pixels. The offset can be defined as follows:

$$\hat{h} = k \cdot \left\lfloor \frac{H}{p_h} \right\rfloor, \hat{w} = l \cdot \left\lfloor \frac{W}{p_w} \right\rfloor \tag{3}$$

where k and l are the block indexes, $k \in [0,\ p_h - 1]$, $l \in [0,\ p_w - 1]$. The block index is applied to $\hat{h}$ and $\hat{w}$ to generate $\hat{h}_k$ and $\hat{w}_l$. The sampling location can be represented as follows:

$$N_f = \cup_{k,l} \left\{ \left(\hat{h}_k, \hat{w}_l \right) \right\} \tag{4}$$

N_f enumerates all the possible combinations of k, l, which can be formulated as follows:

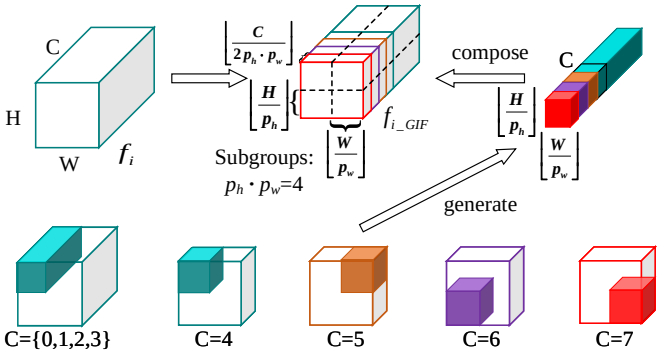

Fig. 3. The concrete implementation process of the global information fusion strategy.

$$\begin{aligned} N_f = \Big\{ &(0,0), \left(0, \left\lfloor \tfrac{W}{p_w} \right\rfloor\right), \cdots, \\ &\left((p_h - 2) \cdot \left\lfloor \tfrac{W}{p_h} \right\rfloor, (p_w - 1) \cdot \left\lfloor \tfrac{W}{p_w} \right\rfloor\right), \\ &\left((p_h - 1) \cdot \left\lfloor \tfrac{W}{p_h} \right\rfloor, (p_w - 1) \cdot \left\lfloor \tfrac{W}{p_w} \right\rfloor\right) \Big\} \end{aligned} \tag{5}$$

N_f includes $p_h \cdot p_w$ offset coordinates in total. It covers almost the entire input feature map.

Each new block feature can be combined in the way shown in Fig.3 to generate a complete global feature set, each newly generated block feature can be formulated as follows:

$$\begin{aligned} f'_{gol,i}(h_k, w_l) = &\sum_{c=0}^{\frac{C}{2}-1} f_i(h_k, w_l, c) + \\ &\sum_{c=\frac{C}{2}}^{C-1} \sum_{(\hat{h},\hat{w}) \in N_f} f_i(h_k + \hat{h}, w_l + \hat{w}, c) \end{aligned} \tag{6}$$

where (h_k, w_l) represent the numbered index of the newly generated feature block.

By partitioning and merging the operations of the global features, each block of the merged features can obtain the information of the original location and the equally important information of the global contextual information, which can enhance the global receptive field of the merged global features. The feature size of the global feature after global information fusion f_{i_GIF} is the same as the original global feature f_i size, which is $16 \times 16 \times 2048$. Finally, a new global feature descriptor is obtained using equation (2) to assist the network model to improve performance.

B. Local Information Fusion Branch

In the cross-view geo-localization task, the content information will change greatly due to the change in the image perspective. Therefore, it is necessary to extract the contextual information to assist the network model. In order to make better use of the contextual information from the image, we adapt the local segmentation approach to mine other information in the image as much as possible, which can improve the stability of the network model. After the extracted global features are partitioned, each part feature contains global information, but the content information is different.

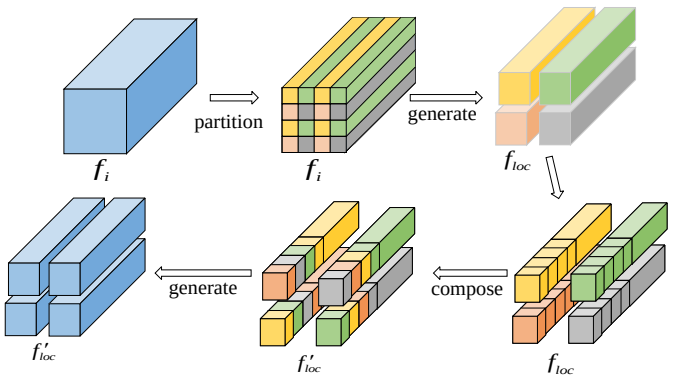

Fig. 4. The local information fusion strategy.

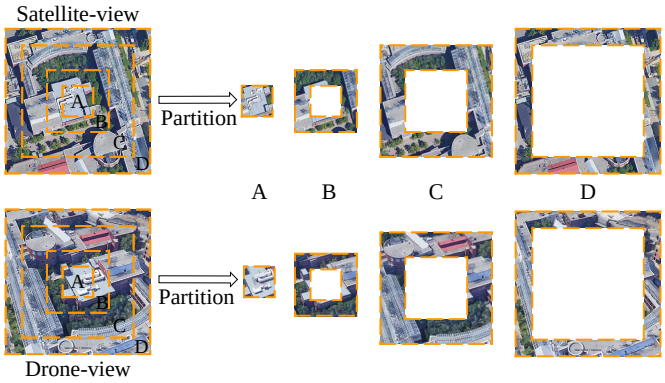

Fig. 5. The square-ring partition strategy.

In order to further increase the global receptive field of each part feature, we introduce the local information fusion strategy to improve the ability to mine contextual information, as illustrated in Fig.4. Since the content information of each part feature is different, to retain the feature contextual information in each block we fully mine the content information from other parts and provide more effective features for the following operations. Similar to equation (5), the process can be formulated as follows:

$$f'_{loc,j}(h,w) = \sum_{c=0}^{\frac{C}{2}-1} f_{loc,j}(h,w,c) + \sum_{c=\frac{C}{2}}^{C-1} \sum_{n=0}^{N} f_{loc,n}(h,w,c) \tag{7}$$

where $f_{loc,j}$ is the j-th block feature before processing, $f'_{loc,j}$ is the j-th block feature after partitioning, and N is the number of blocks of global features. In our work, the size of each part of the local feature $f'_{loc,j}$ is $8 \times 8 \times 2048$.

On this basis, in order to explicitly take advantage of contextual information, we adopt the square-ring partition strategy [1] to process the feature maps after partitioning. For the square-ring partition strategy, the center of the image is approximately aligned with the center of the feature map, and the entire part is partitioned according to the distance from the image center; specific operations are shown in Fig.5. The processed block features are divided into the parts as shown in Fig.5, the geographic target is usually located in the center of the image, and other relevant information is distributed in other locations of the image. It can be seen from the block segmentation that each region after division is also approximately spatially aligned in the cross-view image, which will increase the similarity of the features of each part and provide an effective guarantee for the accuracy of the network model. The process can be formulated as follows:

$$f'^{m}_{loc,j} = F_{slice}\left(f'_{loc,j}, m\right) \tag{8}$$

where $f'^{m}_{loc,j}$ is the j-th block feature after processing, m is the number of divided regions, and F_{slice} represents the square-ring partition processing.

The square-ring partition strategy can not only obtain the geographic target information in the image, but also obtain the contextual information region of the geographic target at different distances. Therefore, it can effectively assist the network model to mine the contextual information in the image. In addition, all partitioned features will generate feature descriptors through equation (2) to improve the accuracy of the final model.

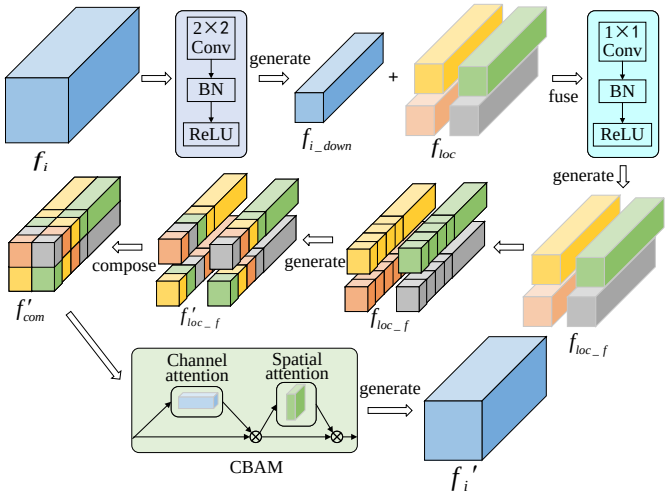

Fig. 6. The specific process of the local-guided-global information fusion branch.

C. Local-guided-global Information Fusion Branch

In order to better utilize the global information and contextual information in the input image, a local-guided-global information fusion branch is designed to assist the model to fully exploit and utilize the potential information in the image. The global features segmented by the local information fusion branch generates some local features, each of them containing a large amount of contextual information. The global features contain more complete and critical information, however, it may be difficult to find this critical information due to the excessive content information in the global features. Therefore, we introduce the local features to assist the global features to potentially obtain more information and improve the performance of the network model. The specific process is shown in Fig.6.

As the dimensions of the global features are different from the partitioned features, it is necessary to downsample the global features to ensure the dimensions of the processed features are consistent with the local partitioned features. The

size of the global feature after downsampling f_{i_down} is the same as that of the local partitioned features f_{loc}, which is $8 \times 8 \times 2048$. Then the processed global features and the local partitioned features are added to generate the new global-local features. On this basis, a mixing information fusion strategy is introduced to recombine the newly generated features. Finally, these features are combined to generate a global feature f'_{com} that is consistent with the starting dimension $16 \times 16 \times 2048$. The newly generated global feature consists of the original global feature and the partitioned local feature, it also contains more contextual information. Since the newly generated global features contain a lot of useful information, in order to make better utilize and mine the crucial parts of these features, we introduced the attention mechanism to solve these problems. In addition, we introduce the information fusion strategy in the local-guided-global information fusion branch, the content of each piece of features has changed. CBAM [46] has two parts: channel attention module and spatial attention module. The combination of the two modules can better mine the crucial information in the integrated features through the information fusion strategy. Therefore, CBAM is introduced into the local-guided-global information fusion branch to mine more useful information from the feature. It is worth noting that the features f'_i generated in the local-guided-global information fusion branch also need to generate a unified form of feature descriptors through equation (2) to improve the accuracy of the final model.

D. Model Optimization

Through the three proposed branches, each model will obtain some features, but they may have different distribution conditions due to the different acquisition platforms, and therefore they cannot be directly used for feature matching. In order to solve this problem, we set up a mapping function that maps all images from different acquisition sources into a shared feature space where the features of the same geo-tags from different platforms are closer together and the feature distances of different geo-tags are separated apart.

The classifier is composed of four parts: full connected layer (FC), batch normalization layer (BN), dropout layer (Dropout) and classification layer (Cls). The classifier module predicts the geo-tag of each part based on the part features. Given the part features d_i^j as the input, the classifier module outputs a column vector z_i^j, and the dimensions of z_i^j are equal to the number of geo-tag categories C. The process can be expressed by the following equation:

$$z_i^j = F_{classifier}(d_i^j) \tag{9}$$

In the training process, the cross-entropy loss function is chosen to optimize the network model and is defined as follows:

$$\hat{p}(y \left| x_i^j \right.) = \frac{\exp(z_i^j(y))}{\sum_{c=1}^{C} \exp(z_i^j(c))} \tag{10}$$

$$Loss = \sum_{i,j} -\log(\hat{p}(y \left| x_i^j \right.)) \tag{11}$$

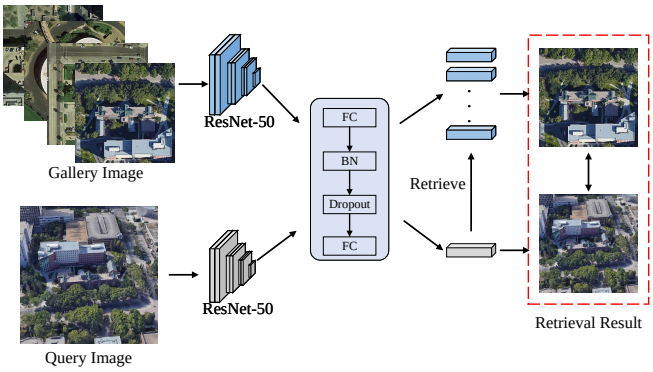

Fig. 7. The retrieval process of proposed network model.

where, $z_i^j(y)$ is the logit score of the ground-truth geo-tag y, the probability score normalized by softmax function in formula (9), and $\hat{p}(y \left| x_i^j \right.)$ is the prediction probability that x_i^j belongs to the geo-tag y.

The cross-entropy loss function is used to accumulate losses on different parts of images from different platforms to optimize the whole network model. In the test stage, we also output the features of various parts of different branches through the classifier module. Hence, we compare the feature similarity of different parts of the input images to determine whether the images from different platforms represent the same geographical target to obtain accurate results. The retrieval process is shown in Fig.7.

IV. Performance Evaluation

Details of the performance evaluation are provided in this section, including experimental datasets, evaluation metrics, experimental details, experimental results, ablation study and qualitative results.

A. Experimental Datasets

Three large-scale geo-localization datasets are chosen to train and evaluate the proposed network model, namely University-1652 [9], CVUSA [16] and CVACT [6]. Table X shows the number of images in the query and gallery sets for testing different tasks using these three datasets.

University-1652 [9] is a multi-view and multi-source dataset, including satellite-view data, drone-view data and ground-view data. It contains 1652 buildings from 72 universities around the world. The training set contains 701 buildings from 33 universities, the testing set contains 951 buildings from 39 universities. There is no overlap between the datasets. Although the University-1652 dataset contains ground-view images, these images are insufficient to fully cover some buildings. Therefore, the dataset also provides additional ground-view images collected from Google Street View with a similar view as the existing ground-view data. As a result, these images can be used as a supplement to the existing ground-view data. For this dataset, the difficulty is the matching and localization task between cross-view images from different platforms, which can be divided into two new tasks, namely drone-view target localization task (Drone→Satellite)

TABLE I. The number of images used in different test datasets for different geo-localization tasks.

Dataset	Task			
	Drone→Satellite		Satellite→Drone	
	Query	Gallery	Query	Gallery
University-1652 [9]	37855	951	701	51355
SUES-200 [47]	16000	200	80	40000
Dataset	Ground→Satellite		Satellite→Ground	
	Query	Gallery	Query	Gallery
CVUSA [16]	8884	8884	8884	8884
CVACT [6]	8884	8884	8884	8884

and drone navigation task (Satellite→Drone). In the training set, 701 buildings and 50218 drone-view images are used. In the drone-view target localization task (Drone→Satellite), there are 37855 drone-view images in the query set, 701 true-matched satellite-view images and 250 satellite-view distractors in the gallery set. In this task, each drone-view image corresponds to only one true-matched satellite-view image. In the drone navigation task (Satellite→Drone), there are 701 satellite-view images in the query set, 37855 true-matched drone-view images and 13500 drone-view distractors in the gallery set. In this task, each satellite-view image will correspond to multiple drone-view images.

SUES-200 [47] is a cross-view geo-localization dataset with multiple sources, multiple scenes, and panoramic views. Specifically, the SUES-200 dataset includes drone-view images at different heights, including school buildings, parks, schools, lakes, and public buildings. The matching and localization tasks are mainly divided into two types: drone-view target localization task (Drone→Satellite) and drone navigation task (Satellite→Drone). The training dataset contains 120 scenarios which has 120 satellite-view and 24000 drone-view images. In the drone-view target localization task (Drone→Satellite), each height in the query set has 4000 drone-view images that are matched with 200 satellite-view images in the gallery set, and includes 120 satellite-view distractors. In the drone navigation task (Satellite→Drone), each height in the query set has 80 satellite-view images are matched with 200 drone-view images in the gallery set, which including 6000 drone-view distractors. In this task, each satellite-view image will correspond to multiple drone-view images.

CVUSA [16] is a cross-view dataset, which includes ground-view data and satellite-view data. Specifically, it contains 35532 sets of ground-and-satellite images used for training and 8884 sets of ground-and-satellite images for testing. All of the ground-view panoramic images are collected from Google Street View and the corresponding satellite-view images are downloaded from the Microsoft Bing Maps.

CVACT [6] is a large-scale cross-view dataset. Similarly to CVUSA it includes ground-view data and satellite-view data. CVACT contains 35532 sets of ground-and-satellite images used for training, and the ground-view images are panoramas. In addition, CVACT provides a validation set with 8884 sets of images and a test set with 92802 sets of images. Each query image only has one true-matched image in the gallery, while each query image may correspond to multiple true-matched images in the gallery for testing.

B. Evaluation Metrics

The Recall@K (R@K) and average accuracy (AP) metrics are selected to evaluate the performance of the network model. R@K represents the proportion of correctly matched images in the top-K of the ranking list, which can be formulated as follows:

$$\mathrm{Recall@}K = \frac{TP@K}{N} \tag{12}$$

where N is the total number of query image.

A higher recall rate demonstrates that the network model has better performance. In addition, we calculate the area under the Precision-Recall curve, called average accuracy (AP), which reflects the precision and recall rate of the retrieval performance. The formula can be shown as follows:

$$AP = \int_0^1 p\,(r)dr \tag{13}$$

C. Implementation Detail

The experiments are completed using a Ubuntu 18.04 system, the model is implemented based on Pytorch, and all experiments are conducted on one NVIDIA GeForce RTX 3090. The ResNet50 [45] is used with pre-training weights using ImageNet [29] to extract visual features. We modify the last layer of ResNet50 and add a new layer, namely the classifier module, and the added layer is initialized using kaiming initialization [48]. During training and testing, the input image is resized to a fixed size of 256×256 pixels for subsequent operations. In the training, random cropping and flipping are used to enhance the input data. For the optimizer, we choose the stochastic gradient descent (SGD) with momentum 0.9 and weight decay 0.0005 with a mini-batch of 32. The initial learning rate of the backbone layer is 0.001, and the initial learning rate of the newly added layer is 0.01. The proposed network model will train in 120 epochs, and the learning rate decreases by 0.1 after 80 epochs. In the test phase, the Cosine distance is used to measure the similarity between the query image and the candidate images in the gallery to complete the cross-view geo-localization task.

D. Experimental Results

1) The Experimental Results using the University-1652 Dataset: The comparison results with the state-of-the-art methods using the University-1652 dataset are given in Table X. The comparison results are mainly divided into three groups, the baseline-related methods, methods harnessing contextual information and Transformer-based methods. The experimental results of the first group of methods are given in the first row to seventh row, these methods pay more attention to the global features and show good results. However, the performance of these methods is not ideal as they may only focus on some global features, these features are difficult to fully effectively identify different scenes due to there are many similar scenes. Moreover, these methods ignore other valid

TABLE II. Comparison with the state-of-the-art methods using the University-1652 dataset. † denotes the input image of size 384×384. For other methods, the image size of the transform-based methods and CNN-based method are 224×224 and 256×256 respectively.

Method	University-1652			
	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
Baseline (Instance Loss) [9]	58.23	62.91	74.47	59.45
Contrastive Loss [30]	52.39	57.44	63.91	52.24
Triplet Loss (M = 0.3) [49]	55.18	59.97	63.62	53.85
Triplet Loss (M = 0.5) [49]	53.58	58.60	64.48	53.15
Soft Margin Triplet Loss [18]	53.21	58.03	65.62	54.47
LCM† [50]	66.65	70.82	79.89	65.38
RK-Net [51]	66.13	70.23	80.17	65.76
LPN [1]	75.93	79.14	86.45	74.49
LPN + USAM [51]	77.60	80.55	86.59	75.96
PCL [15]	79.47	83.63	87.69	78.51
F3-net [52]	78.64	81.60	-	-
Swin-B [53]	84.15	86.62	90.30	83.55
FSRA [22]	84.51	86.71	88.45	83.47
Ours	**86.06**	**88.08**	**91.44**	**85.73**

TABLE III. Comparison with the state-of-the-art methods using the SUES-200 dataset. The input image size for comparison methods is 384×384. For our method, the image size is 256×256.

Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [50]	43.42	49.65	49.42	55.91	57.47	60.31	60.43	65.78
LPN [1]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
Vit [47]	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
Ours	**77.57**	**81.30**	**89.50**	**91.40**	**92.58**	**94.21**	**97.40**	**97.92**
Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LCM [50]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN [1]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
Vit [47]	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
Ours	**93.75**	**79.49**	**97.50**	**90.52**	**97.50**	**96.03**	**100.00**	**97.66**

information in the image that can have a significant impact on the final results. The experimental results of the second group of methods are given in the eighth row to eleventh row, and it can be seen from these results that the performance of these algorithms has been significantly improved after introducing contextual information. From the experimental results, we also see that it is necessary to effectively introduce contextual information into images in the network model. However, these methods only using contextual information while ignoring global information in the image will cause the model to ignore some crucial information in the image, which will have an impact on the final results. The experimental results of the third group of methods are given in the twelfth row and thirteenth row, and it can be seen from the experimental results that the Transformer-based methods have better feature expression ability than the CNN-based algorithm. Therefore, the experimental results for these two methods are significantly better than those for the CNN-based methods. The last row in the table shows the experimental result for the proposed network model. Since the proposed network model fully considers the global information and contextual information in the image, meanwhile, it introduces the idea of local information guiding the global information to improve the ability of the model to discover crucial information. Therefore, the performance of the model has been significantly improved. In the drone-view target localization task (Drone→Satellite), the proposed model achieves 86.06% accuracy for R@1 and 88.08% AP, and in the drone navigation task (Satellite→Drone), the proposed model achieves 91.44% accuracy for R@1 and 85.73% AP. Compared with the LPN method, the R@1 and AP metrics are improved by 10.13% and 8.94% on Drone→Satellite respectively, and by 4.99% and 11.24% on Satellite→Drone respectively. The experimental results also prove the effectiveness of introducing different information fusion branches. In addition, although the Transformer-based method is better than the CNN-based method for feature representation, the performance of the proposed network model is significantly better than these two Transformer-based methods, which also proves the effectiveness of the introduced different branches. For the Drone→Satellite, compared with the Swin-B and SFRA methods, the R@1 and AP are improved by 1.91% and 1.46%, and 1.55% and 1.37% respectively. For the Satellite→Drone, compared with Swin-B and SFRA methods, the R@1 and AP are improved by 1.14% and 2.18%, and 2.99% and 2.26% respectively.

2) The Experimental Results using the SUES-200 Dataset: The comparison results with the state-of-the-art methods using the SUES-200 dataset are given in Table III. The experimental results are mainly divided into three groups, namely the baseline-related methods, the experimental results of methods using contextual information and the experimental results of the Transformer-based method. From the Table III, the the baseline-related methods are given in the first and second rows, the third row shows the experimental results of methods using contextual information, the experimental results of the Transformer-based method are shown in the fourth row and the last row is the experimental result of the proposed network model. It can be seen from the experimental results that the proposed network model achieves the accuracy of R@1 are 77.57, 89.50, 92.58, 97.40 and AP are 81.30, 91.40, 94.21, 97.92 on the drone-view target localization task (Drone→Satellite) at different height, and it can achieve the accuracy of R@1 are 93.75, 97.50, 97.50, 100.00 and AP are 79.49, 90.52, 96.03, 97.66 on the drone navigation task (Satellite→Drone) at different height. Compared with LPN method, the R@1 and AP are improved 10.00%, 8.75%, 5.00%, 7.50% and 12.71%, 15.51%, 14.69%, 11.94% for the Drone→Satellite task at different heights, and the R@1 and AP are improved 15.99%, 18.92%, 12.20%, 15.93% and 14.07%, 15.44%, 10.41%, 13.39% for the Satellite→Drone task at different heights. It can be seen from the experimental results that the proposed network model is effective through introducing global information and local-guided-global information branches on the basis of using contextual information, and the

TABLE IV. Comparison with the state-of-the-art methods using the CVUSA and CVACT datasets. * represents when the method harnesses extra orientation information as input.

Method	Backbone	CVUSA				CVACT			
		R@1	R@5	R@10	R@Top1%	R@1	R@5	R@10	R@Top1%
Zhai [16]	VGG16	-	-	-	43.20	-	-	-	-
Vo [54]	AlexNet	-	-	-	63.70	-	-	-	-
CVM-Net [18]	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87	87.57
Orientation* [6]	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48	92.04
Zheng et al. [9]	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00	85.27
Regmi [17]	X-Fork	48.75	-	81.27	95.98	-	-	-	-
RKNet [51]	USAM	52.50	-	-	96.52	40.53	-	-	89.12
Siam-FCANet [11]	ResNet-34	-	-	-	98.30	-	-	-	-
CVFT [12]	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [1]	ResNet-50	85.79	95.38	96.80	99.41	79.99	90.63	92.56	97.03
GeoNet-II [55]	ResNetX	-	-	-	98.70	58.90	81.80	88.30	97.70
SIRNet [33]	VGG16	81.82	93.39	96.24	99.49	75.37	88.76	91.90	97.42
TransGeo [56]	ViT	94.08	98.36	99.04	**99.77**	-	-	-	-
L2LTR [57]	ViT	91.99	97.68	98.65	99.75	83.14	93.84	95.51	98.40
Polar Transform Methods									
SAFA [10]	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17
DSM [13]	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32
Shi et al. [58]	VGG16	92.69	97.78	98.60	99.61	82.70	92.50	94.42	97.65
LPN [1]	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77
LPN + USAM [51]	ResNet-50	91.22	-	-	99.67	82.02	-	-	98.18
Toker [14]	ResNet-34	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22
SIRNet [33]	VGG16	93.74	98.02	98.85	99.76	86.02	94.45	**96.02**	98.33
L2LTR [57]	ViT	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37
Ours	ResNet-50	**95.09**	**98.85**	**99.34**	**99.77**	**86.64**	**94.61**	95.94	**98.45**

performance of the model has been greatly improved.

3) The Experimental Results using the CVUSA Dataset: The comparison results with the state-of-the-art methods using the CVUSA dataset are given in Table IV. The experimental results are mainly divided into two groups, the method without using polar transform and the method using polar transform. The experimental results for the first group of methods are given in the first row to fourteenth row, these methods show good results for the cross-view geo-localization task. However, the CVUSA dataset is mainly aimed at cross-view image matching between satellite-view and ground-view, due to the huge change of perspective, the content information has changed significantly which presents challenges to the network model. In addition, it is difficult for these methods to spatially align the image features under the changing view, which leads to the model performance is not ideal. Therefore, many methods employ polar transforms to convert satellite-view images. It considers the geometric correspondence of two-platform images and transforms the aerial-view image to approximately align a ground panorama at the pixel level. The experimental results for the second group of methods are given in the last nine rows. From the comparison results of LPN in the two groups, it can be seen that the performance of the method has greatly improved after using a polar transform. From IV, it can be seen that the proposed network model is significantly superior to other methods using the CVUSA dataset after employing a polar transform, and achieves an accuracy of R@1 95.09%, R@5 98.85%, R@10 99.34% and R@Top1% 99.77%. Compared with the LPN method after using a polar transform, the result for R@1 improved 1.31%, which proves the effectiveness of the proposed network model. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.04% and 0.10% on the R@1 and AP, which can proved the effectiveness of the proposed network model.

4) The Experimental Results using the CVACT Dataset: The comparison results with the state-of-the-art methods using the CVACT dataset are given in Table IV. Due to the image perspective structure of CVACT being similar to the CVUSA dataset, the experimental results using this dataset are mainly divided into two groups, the method without using polar transform and the method using polar transform. Similar to the experimental results in CVUSA, the performance of the method is obviously improved after using a polar transform. It can be seen from Table IV that the proposed network model is significantly superior to other methods using the CVACT dataset after employing a polar transform, and achieves accuracies of R@1 86.64%, R@5 94.61%, R@10 95.94% and R@Top1% 98.45%. Compared with the LPN method after using polar transform, the results of R@1, R@5, R@10 and R@Top1% improve by 3.77%, 2.35%, 1.85% and 0.68% respectively. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.75% and 0.08% on the R@1 and AP, which proves the effectiveness of the proposed network model for the cross-view geo-localization task.

E. Ablation Study

In order to prove the validity of each part of the proposed network model, we design several ablation experiments which mainly focus on two tasks, drone-view target localization task (Drone→Satellite) and drone navigation task (Satellite→Ground).

TABLE V. Ablation study on the effect of the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

GIF	LIF	LGGIF	Drone→Satellite		Satellite→Drone	
			R@1	AP	R@1	AP
×	×	×	64.13	68.73	76.32	60.20
✓	×	×	71.78	75.68	80.74	68.69
×	✓	×	68.96	72.42	84.02	66.68
×	×	✓	65.96	70.32	79.32	63.41
✓	✓	×	82.72	84.95	89.44	79.19
✓	×	✓	72.85	76.52	83.17	69.86
×	✓	✓	81.56	83.58	89.73	80.17
✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

1) Effect of the Various Branches: The main contribution of this paper is to design three branches, the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In order to verify the effectiveness of these three proposed branches in the network model, we designed several experiments to test each branch as shown in Table V. It can be seen from Table V that no matter which branch is excluded, the performance of the network model has a certain decline. In addition, the performance of the model will be significantly improved after using any branch, which also proves the effectiveness of the proposed different branches. Intuitively, the performance of the model is greatly improved after combining the local information fusion branch, which also proves that the contextual information in the image has a significant impact on the cross-view geo-localization task. Thus, we fully consider the global information in the image and employ local information to assist the global features to mine the critical information in the image as much as possible. From the experimental results, it can be seen that these three introduced branches are effective and the discriminability of the final feature descriptors can be improved through utilizing the global features, improving the retrieval precision of the network model.

2) Effect of Different Information Fusion Strategies and Attention Mechanism on Different Branches: In order to enhance the limited receptive field in each branch, we introduce different information fusion strategies (IFS) and attention mechanism to improve the global receptive field of each module and enhance the performance of the network model. In order to prove the effectiveness of these strategies, we design some experiments for these strategies, and the experimental results are given in Table VI. Without introducing any information fusion strategy and attention mechanism, the performance of the proposed network model is not very ideal. It can be seen from Table VI that the performance of the model has been clearly improved after introducing these strategies. These strategies are mainly to enable each branch to mine and utilize the critical information from the global or contextual information, and further improve the discriminability of each feature descriptor based on improving the limited global receptive field. It can be seen from the experimental results that the introduced strategies in each branch are essential and have a huge impact on the final result of the network model,

TABLE VI. Ablation study on the effect of the information fusion strategies (IFS) and attention mechanism (CBAM) in the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

IFS in GIF	IFS in LIF	IFS in LGGIF	CBAM	Drone→Satellite		Satellite→Drone	
				R@1	AP	R@1	AP
×	×	×	×	80.35	82.79	87.02	79.49
×	✓	✓	✓	83.59	85.88	89.59	83.15
✓	×	✓	✓	82.20	84.64	88.87	82.19
✓	✓	×	✓	83.24	85.63	90.16	82.10
✓	✓	✓	×	83.97	86.21	89.16	82.75
✓	✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

TABLE VII. Ablation study on the effect of different input sizes on the University-1652 dataset.

Image Size	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
224	77.95	80.84	85.45	76.79
256	86.06	88.08	91.44	85.73
320	86.21	88.38	91.90	85.94
384	86.91	88.83	92.15	**86.50**
512	**87.21**	**88.98**	**92.30**	85.73

which can prove the effectiveness of these designed strategies in the network model.

3) Effect of Different Input Sizes Using the University-1652 Dataset: For model training and testing, the size of the input image will effect the fine-grained information within the image, which will affect the feature representation learning due to the missing information. However, a larger input size introduces more memory costs during training and testing and increases the computational complexity. Therefore, in order to balance the size and performance of the input image, we design some experiments to determine the influence of the input image size on the model performance. In the experiment, we only change the size of the input image and the region covered by the image is not changed; the experimental results are shown in Table VII. We test the size of the input images from 224 to 512 in two studied tasks respectively. It is worth noting that we choose to test the impact of this range of sizes on the final network model performance as the size of the image in the University-1652 dataset is 512×512. It can be seen from the experimental results that the model performance gradually improves with the change of the input image size, which also indicates that the missing image information caused by the reduction of the input image size also has a great impact on the results. When we continue to expand the input size to 512, the improvement is not so clear on Satellite→Drone, which also shows that when the input image size reaches a certain threshold, the impact of the missing information will also be reduced. We hope that this finding can provide effective insights in the case of limited computing resources, in order to choose the size of the input image in real-world applications.

4) Effect of Image Position Shifting Using the University-1652 Dataset: In order to demonstrate the robustness of the proposed network model to position shifting, we design some experiments using different degrees of horizontal shifting to query images; the experimental results are shown in Table

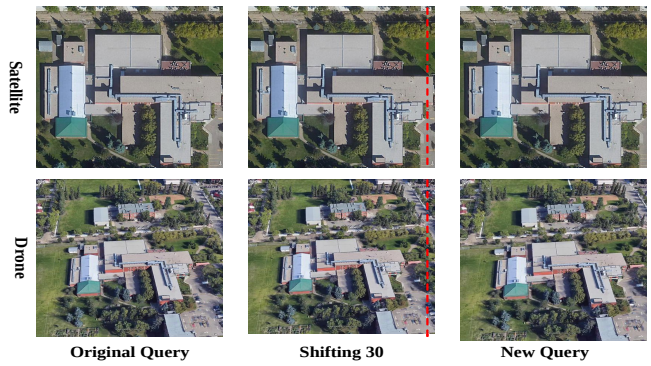

Fig. 8. The image on the left is the original query image, the middle image is the image when translated 30 pixels, and the right image is the new generated query image.

TABLE VIII. Ablation study on the effect of image position shifting on the University-1652 dataset.

Shifted Pixel	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
0	**86.06**	**88.08**	**91.44**	**85.73**
10	86.05	**88.08**	91.28	85.69
20	86.02	88.06	91.01	85.42
30	85.95	87.99	90.87	84.90
40	85.77	87.84	90.44	84.44
50	85.30	87.45	90.37	83.80
60	84.77	87.00	90.34	82.98

VIII. Examples of image translation are shown in Fig.8, where we translate the image by different degrees to generate new query images to test retrieval performance. It can be seen from the experimental results that the model performance gradually decreases with the increase of image translation. However, it does not decrease significantly in the case of slight translation, indicating that the performance of the introduced different branches will not be greatly affected in the case of minimal changes in content, which proves that the proposed network model has a strong robustness to position shifting. Although some information is missing after the image translation, the three designed branches can make the network model better mine and utilize the critical information which is why the proposed network model can resist position shifting.

5) Effect of Matching accuracy of Multiple Queries: In practical applications, it is difficult to fully describe the target location from a single drone-view image. Furthermore, the University-1652 dataset provides images from different perspectives of each scene, which means that we can use multiple drone-view images as queries at the same time to explore whether these multi-view queries can improve the matching accuracy for the Drone→Satellite task. In the test phase, we average the features obtained from multiple images, and take the processed feature as the final query feature. The experimental results are shown in Table IX. For the test phase, we set the number of multi-view images to 1,2,3,9,18,27 and 54 according to the number of drone-view images in the dataset. It can be seen from the experimental results that with the increase of the number of images used, the network model performance has been significantly improved, with an improvement of

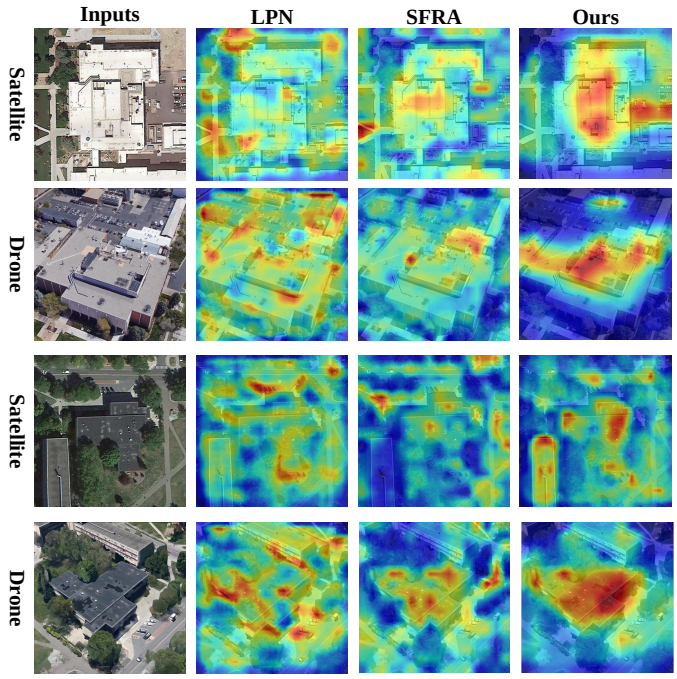

Fig. 9. Heatmaps are produced by the LPN, FSRA and ours on different platforms on the University-1652 dataset.

TABLE IX. Ablation study on the effect of matching accuracy of multiple queries.

Query	Drone→Satellite				
	R@1	R@5	R@10	R@Top1%	AP
54	**91.73**	**97.15**	**97.86**	97.86	**93.00**
27	91.23	97.08	97.79	**97.93**	92.51
18	90.63	96.67	97.75	97.85	91.99
9	89.23	96.34	97.72	97.84	90.84
3	87.03	95.51	97.12	97.30	88.94
2	86.68	95.20	96.97	97.19	88.62
1	86.06	94.95	96.82	97.00	88.08

5.67% and 4.98% for Recall@1 and AP respectively. Multi-view features are effective for Drone→Satellite task and we hope that these experiments can provide an effective solved method for the practical application.

F. Qualitative Results

We illustrate some heatmap visualizations generated using LPN, FSRA and the proposed network model as qualitative results, shown in Fig.10. It can be seen from the heatmaps that the LPN method pays more attention to the contextual information in the image, which may therefore ignore the critical information. FSRA method can pay more attention to the critical information in the image through region alignment strategy, however, it can only focus on a part of the crucial information, and the use of crucial information is still insufficient. Compared with the LPN method and the FSRA method, the proposed network model pays more attention to the critical information in the image after combining the contextual information, which also shows the effectiveness of the designed modules. In addition, some retrieval results of compared methods on different datasets are shown in Fig.11.

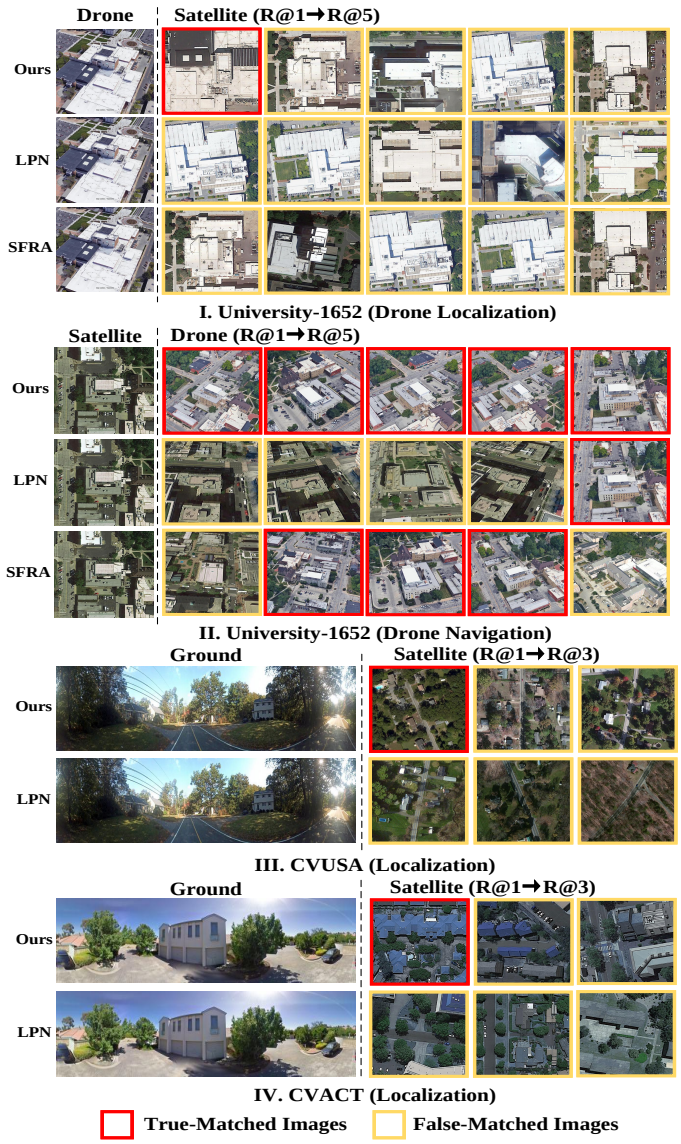

Fig. 10. Qualitative image retrieval results. (I) Top-5 retrieval results for drone-view target localization in different methods using the University-1652 dataset. (II) Top-5 retrieval results for drone navigation in different methods using the University-1652 dataset. (III) Top-3 retrieval results of geographic localization in different methods using the CVUSA dataset. (III) Top-3 retrieval results of geographic localization in different methods using the CVACT dataset. The true matches are in red boxes, while the false matches are displayed in yellow boxes.

The University-1652 dataset is utilized for two tasks, a drone-view target localization task (Drone→Satellite) and a drone navigation task (Satellite→Ground). The retrieval results of these two tasks are given in Fig.11(I) and Fig.11(II). From the compared retrieval results, it can be seen that the proposed method can effectively retrieve the corresponding scenes in both tasks. Although other comparison algorithms can also find the correct scenes, there are some errors in them and they are not the highest feature matching degree. Fig.11(III) and Fig.11(IV) show the retrieval results of the ground→Satellite localization task using the CVUSA and CVACT datasets. It can be seen from these results that given a randomly selected query image, the most relevant image can be retrieved from the candidate gallery through the proposed method, which demonstrates the effectiveness of the proposed network model.

V. Conclusion

We have proposed a multi-branch joint representation learning network model based on information fusion strategies to solve the cross-view geo-localization problem. In order to better focus on the critical information in cross-view images, we introduce three branches into the network model, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch, to extract global information and contextual information in different cross-view images. In addition, we introduce different information fusion strategies into these branches to expand the global receptive field of each module and enhance the discriminability of the each part of the image representation. From the performance comparison experiments with other state-of-the-art methods using the University-1652, CVUSA and CVACT datasets, the proposed framework outperforms these state-of-the-art methods, proving the effectiveness of the proposed network model and improving the retrieval accuracy. Our future work will focus on improving algorithmic performance by introducing semantic or edge information to enhance the ability of the model to obtain discriminable features and improve the model performance.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 61973066) and Major Science and Technology Projects of Liaoning Province(No. 2021JH1/10400049).

References

[1] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 2, pp. 867–879, 2021.

[2] Z. Lv, P. Zhong, W. Wang, Z. You, C. Shi *et al.*, "Novel piecewise distance based on adaptive region key-points extraction for lccd with vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[3] Z. Lv, H. Huang, X. Li, M. Zhao, J. A. Benediktsson, W. Sun, and N. Falco, "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proceedings of the IEEE*, 2022.

[4] Z. Shao, Y. Li, and H. Zhang, "Learning representations from skeletal self-similarities for cross-view action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 160–174, 2020.

[5] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[6] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5624–5633.

[7] Z. Lv, H. Huang, W. Sun, M. Jia, J. A. Benediktsson, and F. Chen, "Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[8] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, J. Li, and W. Wang, "Novel adaptive region spectral-spatial features for land cover classification with high spatial resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[9] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1395–1403.

[10] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 1–11.

[11] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen, "Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8391–8400.

[12] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 990–11 997.

[13] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am i looking at? joint location and orientation estimation by cross-view matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4064–4072.

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

[15] X. Tian, J. Shao, D. Ouyang, and H. T. Shen, "Uav-satellite view synthesis for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4804–4815, 2021.

[16] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 867–875.

[17] K. Regmi and M. Shah, "Bridging the domain gap for ground-to-aerial image matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 470–479.

[18] S. Hu, M. Feng, R. M. Nguyen, and G. H. Lee, "Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 7258–7267.

[19] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 994–1003.

[20] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8287–8294.

[21] Z. Zeng, Z. Wang, F. Yang, and S. Satoh, "Geo-localization via ground-to-satellite cross-view image retrieval," *IEEE Transactions on Multimedia*, vol. 25, pp. 2176–2188, 2022.

[22] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4376–4389, 2021.

[23] Z. Deng, X. Ren, J. Ye, J. He, and Y. Qiao, "Fcn+: Global receptive convolution makes fcn great again," *arXiv preprint arXiv:2303.04589*, 2023.

[24] Y. Zhu, S. Chen, X. Lu, and J. Chen, "Cross-view image synthesis from a single image with progressive parallel gan," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[25] Y. Sun, Y. Ye, J. Kang, R. Fernandez-Beltran, S. Feng, X. Li, C. Luo, P. Zhang, and A. Plaza, "Cross-view object geo-localization in a local region with satellite imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–16, 2023.

[26] F. Castaldo, A. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 9–17.

[27] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2013, pp. 891–898.

[28] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1125–1128.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[30] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2015, pp. 5007–5015.

[31] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 70–78.

[32] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3961–3969.

[33] X. Lu, S. Luo, and Y. Zhu, "It¡¯s okay to be wrong: Cross-view geo-localization with step-adaptive iterative refinement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[34] Y. Amit and A. Trouvé, "Pop: Patchwork of parts models for object recognition," *International Journal of Computer Vision*, vol. 75, pp. 267–282, 2007.

[35] D. Crandall, P. Felzenszwalb, and D. Huttenlocher, "Spatial priors for part-based recognition using statistical models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, vol. 1, 2005, pp. 10–17.

[36] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International journal of computer vision*, vol. 77, pp. 259–289, 2008.

[37] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[38] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, 1999, pp. 1150–1157.

[39] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 1077–1085.

[40] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 384–393.

[41] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 3037–3045, 2018.

[42] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[43] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 2119–2128.

[44] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3642–3651.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision*, 2018, pp. 3–19.

[47] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu, "Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2015, pp. 1026–1034.

[49] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking." *Journal of Machine Learning Research*, vol. 11, no. 3, pp. 1109–1135, 2010.

[50] L. Ding, J. Zhou, L. Meng, and Z. Long, "A practical cross-view image matching method between uav and satellite for uav-based geo-localization," *Remote Sensing*, vol. 13, no. 1, pp. 47–67, 2020.
[51] J. Lin, Z. Zheng, Z. Zhong, Z. Luo, S. Li, Y. Yang, and N. Sebe, "Joint representation learning and keypoint detection for cross-view geo-localization," *IEEE Transactions on Image Processing*, vol. 31, pp. 3780–3792, 2022.
[52] B. Sun, G. Liu, and Y. Yuan, "F3-net: Multi-view scene matching for drone-based geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–11, 2023.
[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
[54] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proceedings of the European conference on computer vision*, 2016, pp. 494–509.
[55] Y. Zhu, B. Sun, X. Lu, and S. Jia, "Geographic semantic network for cross-view image geo-localization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.
[56] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1162–1171.
[57] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 009–29 020, 2021.
[58] Y. Shi, X. Yu, L. Liu, D. Campbell, P. Koniusz, and H. Li, "Accurate 3-dof camera geo-localization via ground-to-satellite image matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 2682–2697, 2022.

Response Letter

Author Response to Reviews of

Multi-branch Joint Representation Learning Based on Information Fusion Strategy for Cross-view Geo-localization

IEEE Transactions on Geoscience and Remote Sensing, Number: TGRS-2023-05070

Response to Associate Editor

Comments:
The manuscript should be further revised to address following comments:

***Response*:** Thanks very much for your help! The comments of reviewers are very valuable, we have made detailed revisions based on the reviewer comments. After that, the paper has been significantly improved. Thanks again!

Comments 1: The contribution and novelty should be summarized and clarified in the manuscript.

***Response*:** Thank you very much for your comments. We have summarized and clarified the contribution and novelty of the manuscript.

Comment 2: The experimental scene should be expanded.

***Response*:** This is a valuable suggestion. We have added experiments on the new dataset to prove the effectiveness of the proposed network model. In addition, we have added some necessary experiments to the ablation study.

Comment 3: Visual comparison of the experimental results should be presented.

***Response*:** Thank you for the reminder. In the experimental section, we have added visual comparison results to prove the effectiveness of the network model proposed in this manuscript, which include heatmap visual results and image retrieval result.

Comment 4: More experimental analysis needs to be added.

***Response*:** This is a valuable opinion. We have rewritten the experimental analysis section in this manuscript to analyze and explain the experimental results in more detail. In addition, we also added some new experimental and comparative methods in the experimental section.

Responses to the reviewers

Reviewer #1

Comments to Authors:

This paper proposed a multi-branch joint representation learning network model based on information fusion strategies. There are some problems in this paper that need to be improved. Before publication, I had some comments. Details of my comments are presented below:

***Response*:** I am very glad to receive your valuable comments and give a positive response to this paper. I actively responded to your comments one by one, and the revised paper is more complete. I hope I can answer your questions and get your positive response. The modified contents are all highlighted in red in the manuscript.

Comment 1: The contribution and novelty of this study should be summarized and clarified in the manuscript.

***Response*:** Thanks greatly for your suggestion! We have rewritten and summarized the contribution and novelty of the manuscript to make them clarified and more explicit.

Revision: Although mining and utilizing contextual information in images can effectively improve the accuracy of cross-view geo-localization, the use of global information in images is equally important. As shown in Fig.1, a previous work, LPN [1], focuses more on the contextual information for the input image, ignoring other crucial information in the scenes. In order to mine effective information in cross-view images, and inspired by existing methods [1], [23], a multi-branch joint representation learning network model based on information fusion strategies is proposed in this paper to solve the cross- view geo-localization problem. For this problem, we believe that each part of the image has a significant impact on the result of image matching. Therefore, we adapt a multi-branch joint representation learning network model to solve this problem, which is divided into three branches, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In the global information fusion branch, the global features of an image can effectively express the content information in the complete image scene, most existing methods use this concept to solve cross-view scene matching. However, only using the global information may result in the acquired features that are not sufficiently discernible for certain scenes. In order to mine the global information effectively, we introduce the global information fusion strategy into it to increase the global receptive field of the network, which can improve the utilization of global information. However, as it is difficult to distinguish similar scenes completely only using global information, the contextual information in the image can help with feature matching.

Therefore, we design a local information fusion branch to improve the performance of the network model. In the local information fusion branch, the contextual information in the image will have a positive impact on feature matching. In order to better mine the contextual information in the image, we process the global information into blocks. At the same time, in order to better mine and utilize the crucial information of each segmented part, we introduce the local information fusion strategy to expand the receptive field of each part. On this basis, we divide each block feature into several parts in a square-ring partition to obtain the contextual information of each part thus assisting the global features to distinguish similar scenes. In addition, we believe that local features can assist global features to better mine information in the image. In this regard, we also introduce a local-guided-global information fusion branch, which mainly used local features after segmentation to assist global features and introduce a mixing information fusion strategy and attention mechanism to further increase the global receptive field and mine more useful potential information. Through these three branches, the effective information in the cross-view image can be mined and utilized effectively to solve the cross-view geo-localization problem. The contributions of this paper are summarized as follows:

- A multi-branch joint representation learning network model based on information fusion strategies is proposed to solve the cross-view geo-localization problem, which consists of three parts, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.
- In order to obtain more robust features, based on the use of global information and local information, we utilize the idea of the local-guided-global information to build model branch without introducing additional information and assist network model to further mine latent crucial information in the image, which can further improve the performance network model.
- To further mine and utilize the crucial information in the image, three information fusion strategies (IFS) are designed to the three proposed branches to assist each branch of the network model to increase the global receptive field. On this basis, each branch can more effectively mine and utilize the relevant feature information and improve the discrimination of features.

- A series of experiments is carried out using the University-1652, SUES-200, CVUAS and CVACT datasets, and the experimental results demonstrate that the effectiveness of the proposed network framework. (see the fourth and fifth paragraph of INTRODUCTION, page 2)

Comment 2: Although some related works such as land cover change detection (DOI:10.1109/TNNLS.2023.3282935, doi:10.1109/LGRS.2023.3267879, doi:10.1109/JPROC.2022.3219376) and land cover classification (10.1109/TGRS.2023.3268038, doi: 10.1109/TGRS.2023.3275753) with remote sensing images have been applied successfully in practical engineering. Image fusion is required when more than one image used for applications.—— This should be cleared to enhance the significant of this work.

***Response*:** Thank you for the reminder. We have revised the section of the application background and application scenarios in the manuscript to highlight the importance of the studied work. In addition, we also cite some new literature, as shown below.

[2] Novel piecewise distance based on adaptive region key-points extraction for lccd with vhr remote sensing images, IEEE Transactions on Geoscience and Remote Sensing, 2023.
[3] Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective, Proceedings of the IEEE, 2022.
[5] Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images, IEEE Geoscience and Remote Sensing Letters, 2023.
[7] Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network, IEEE Transactions on Neural Networks and Learning Systems, 2023.
[8] Novel adaptive region spectral-spatial features for land cover classification with high spatial resolution remotely sensed imagery, IEEE Transactions on Geoscience and Remote Sensing, 2023.

Revision: Cross-view geo-localization aims to retrieve the most relevant images of the same geographic target from different platforms, and has been widely used in many fields, such as accurate delivery, autonomous driving, action recognition, change detection, event detection and land cover classification [2]-[8]. In the era of digital maps, it is usually necessary to estimate the geospatial localization of a given object in real-time. This can be done with real-time Kinematic (RTK) GPS, but these sensors are expensive and short time signal interruptions can hinder workflows. In addition, especially in the city, the urban canyon effect will produce a certain deviation. At present, cross-view geo-localization based on image retrieval is an effective method to solve these problems [9]. In practical application, the information obtained from different source data for the same target in different tasks is different and related. Therefore, it is necessary to correlate images from different views [10]. For example, given a drone-view image, it is necessary to retrieve images of the same location from other viewpoints to obtain the geographic information of the location. Given locational information is available from different image sources such as satellite, drone, or ground, studying the cross-view geo-localization problem is extremely important [11]-[14]. However, the scale, viewpoint and imaging modality for images obtained through different platforms can be very different, e.g., the ground-view is almost perpendicular to the horizon, while the satellite-view is almost parallel to the horizon. Therefore, cross-view geo-localization is a challenging task [15]. (see the first paragraph of INTRODUCTION, pages 1-2)

Comment 3: The abstract section should quantify the benefits of improved accuracy on the dataset.

***Response*:** This is a valuable comment. We have revised the summary of the experimental results in the abstract to quantify the benefits of improved the accuracy of the dataset.

Revision: Finally, a series of experiments is carried out on three prevailing benchmark datasets, namely University-1652, SUES-200, CVUAS and CVACT datasets. The quantitative comparisons from the experiments clearly indicate that the proposed network framework has great performance. For example, compared with some state-of-the-art methods, the quantitative improvements of the R@1 and AP on the University-1652 datasets are 1.91%, 2.18% and 1.55%, 2.99% in both tasks, respectively. (see the Abstract, page 1)

Comment 4: In the introduction section, the research significance of cross-view is not well described, but the application fields of cross-view are simply described, which is not conducive to readers' understanding.

***Response*:** Thank you very much for your valuable comments. In the introduction, we have revised and sort out the research significance of cross-view, which can help readers better understand the research significance of this manuscript.

Revision: Cross-view geo-localization aims to retrieve the most relevant images of the same geographic target from different platforms, and has been widely used in many fields, such as accurate delivery, autonomous driving, action recognition, change detection, event detection and land cover classification [2]-[8]. In the era of digital maps, it is usually necessary to estimate

the geospatial localization of a given object in real-time. This can be done with real-time Kinematic (RTK) GPS, but these sensors are expensive and short time signal interruptions can hinder workflows. In addition, especially in the city, the urban canyon effect will produce a certain deviation. At present, cross-view geo-localization based on image retrieval is an effective method to solve these problems [9]. In practical application, the information obtained from different source data for the same target in different tasks is different and related. Therefore, it is necessary to correlate images from different views [10]. For example, given a drone-view image, it is necessary to retrieve images of the same location from other viewpoints to obtain the geographic information of the location. Given locational information is available from different image sources such as satellite, drone, or ground, studying the cross-view geo-localization problem is extremely important [11]-[14]. However, the scale, viewpoint and imaging modality for images obtained through different platforms can be very different, e.g., the ground-view is almost perpendicular to the horizon, while the satellite-view is almost parallel to the horizon. Therefore, cross-view geo-localization is a challenging task [15]. (see the first paragraph of INTRODUCTION, pages 1-2)

Comment 5: In the proposed method, may I ask why cross-view can realize geo-positioning by fusing global information and local information? Or the principle of cross-view localization should be explained clearly.

***Response*:** Thanks a lot for your suggestion. In the methods section, we have added the explanation of the principle of cross-view geo-localization. In addition, in the introduction and methods section, we also explain in detail why the proposed network model can solve the cross-view geo-localization problem.

Revision: In cross-view geo-localization, the contents of different source images will change greatly due to the different obtained views. For the same task object, there will be a lot of relevant information in the different images. Therefore, how to match or mine the relevant information in these different source images is the key to solve this problem. In this regard, by mining the global information and local information of images from different sources, we can associate key information from different views as much as possible and determine the image of the same place, so as to complete the cross-view geo-localization task. In this section, we provide a detailed introduction to the proposed multi-branch joint representation learning network model based on information fusion strategies. The proposed network framework is shown in Fig.2. (see the METHODOLOGY A-C, pages 4-7)

Although mining and utilizing contextual information in images can effectively improve the accuracy of cross-view geo-localization, the use of global information in images is equally important. As shown in Fig.1, a previous work, LPN [1], focuses more on the contextual information for the input image, ignoring other crucial information in the scenes. In order to mine effective information in cross-view images, and inspired by existing methods [1], [23], a multi-branch joint representation learning network model based on information fusion strategies is proposed in this paper to solve the cross-view geo-localization problem. For this problem, we believe that each part of the image has a significant impact on the result of image matching. Therefore, we adapt a multi-branch joint representation learning network model to solve this problem, which is divided into three branches, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In the global information fusion branch, the global features of an image can effectively express the content information in the complete image scene, most existing methods use this concept to solve cross-view scene matching. However, only using the global information may result in the acquired features that are not sufficiently discernible for certain scenes. In order to mine the global information effectively, we introduce the global information fusion strategy into it to increase the global receptive field of the network, which can improve the utilization of global information. However, as it is difficult to distinguish similar scenes completely only using global information, the contextual information in the image can help with feature matching.

Therefore, we design a local information fusion branch to improve the performance of the network model. In the local information fusion branch, the contextual information in the image will have a positive impact on feature matching. In order to better mine the contextual information in the image, we process the global information into blocks. At the same time, in order to better mine and utilize the crucial information of each segmented part, we introduce the local information fusion strategy to expand the receptive field of each part. On this basis, we divide each block feature into several parts in a square-ring partition to obtain the contextual information of each part thus assisting the global features to distinguish similar scenes. In addition, we believe that local features can assist global features to better mine information in the image. In this regard, we also introduce a local-guided-global information fusion branch, which mainly used local features after segmentation to assist global features and introduce a mixing information fusion strategy and attention mechanism to further increase the global receptive field and mine more useful potential information. Through these three branches, the effective information in the cross-view image can be mined and utilized effectively to solve the cross-view geo-localization problem. (see the fourth and fifth paragraph of INTRODUCTION, page 2)

Comment 6: The backbone of the network is ResNet-50, and the innovation of the proposed method in the neural network structure should be elaborated.

***Response*:** Thank you very much for reminding. In our work, the main innovation of this paper is mainly reflected in the three proposed branches and the information fusion strategy used in each branches. Based on the three proposed branches, namely namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch, we use different information fusion strategies in these branch, which can obtain deep latent information of the features in each branch. In addition, we believe that the content of the image will changing greatly due to the huge changes in the view of the cross-view image. Therefore, we introduce the idea of the local information to guide the global information, so that the network model can pay more attention to the crucial information in the different view image through local features, which can improve the performance of the proposed model. In addition, in the local-guided-global information fusion branch we also design a simple network structure to handle the global features, which is convenient for our operations on the feature level. Moreover, the innovation of each proposed branch is also introduced in detail in the introduction and methods. (see the fourth and fifth paragraph of INTRODUCTION and METHODOLOGY A-C, pages 2 and 4-7)

Revision: Although mining and utilizing contextual information in images can effectively improve the accuracy of cross-view geo-localization, the use of global information in images is equally important. As shown in Fig.1, a previous work, LPN [1], focuses more on the contextual information for the input image, ignoring other crucial information in the scenes. In order to mine effective information in cross-view images, and inspired by existing methods [1], [23], a multi-branch joint representation learning network model based on information fusion strategies is proposed in this paper to solve the cross- view geo-localization problem. For this problem, we believe that each part of the image has a significant impact on the result of image matching. Therefore, we adapt a multi-branch joint representation learning network model to solve this problem, which is divided into three branches, namely the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In the global information fusion branch, the global features of an image can effectively express the content information in the complete image scene, most existing methods use this concept to solve cross-view scene matching. However, only using the global information may result in the acquired features that are not sufficiently discernible for certain scenes. In order to mine the global information effectively, we introduce the global information fusion strategy into it to increase the global receptive field of the network, which can improve the utilization of global information. However, as it is difficult to distinguish similar scenes completely only using global information, the contextual information in the image can help with feature matching.

Therefore, we design a local information fusion branch to improve the performance of the network model. In the local information fusion branch, the contextual information in the image will have a positive impact on feature matching. In order to better mine the contextual information in the image, we process the global information into blocks. At the same time, in order to better mine and utilize the crucial information of each segmented part, we introduce the local information fusion strategy to expand the receptive field of each part. On this basis, we divide each block feature into several parts in a square-ring partition to obtain the contextual information of each part thus assisting the global features to distinguish similar scenes. In addition, we believe that local features can assist global features to better mine information in the image. In this regard, we also introduce a local-guided-global information fusion branch, which mainly used local features after segmentation to assist global features and introduce a mixing information fusion strategy and attention mechanism to further increase the global receptive field and mine more useful potential information. Through these three branches, the effective information in the cross-view image can be mined and utilized effectively to solve the cross-view geo-localization problem. (see the fourth and fifth paragraph of INTRODUCTION and METHODOLOGY A-C, pages 2 and 4-7)

Comment 7: In the proposed local-guided-global information fusion branch, the CBAM attention module is used. Attention mechanism has developed rapidly in recent years. May I ask why CBAM is used here, or whether CBAM module can aggregate information? I don't see the motivation of choosing CBAM in the method.

***Response*:** This is a valuable suggestion. We have explained why we chose the CBAM attention module in the local-guided-global information fusion branch. In addition, we have tested the CBAM attention module and demonstrated that it can indeed assist the network model to focus on the crucial information, although we did not include these results in the manuscript. It can also be seen from the figure that the use of CBAM attention module is effective.

Revision: The newly generated global feature consists of the original global feature and the partitioned local feature, it also contains more contextual information. Since the newly generated global features contain a lot of useful information, in order to make better utilize and mine the crucial parts of these features, we introduced the attention mechanism to solve these problems. In addition, we introduce the information fusion strategy in the local-guided-global information fusion branch, the content of each piece of features has changed. CBAM [46] has two parts: channel attention module and spatial attention module. The combination of the two modules can better mine the crucial information in the integrated features through the information fusion strategy. Therefore, CBAM is introduced into the local-guided-global information fusion branch to mine more useful information from the feature. (see the METHODOLOGY C, page 7)

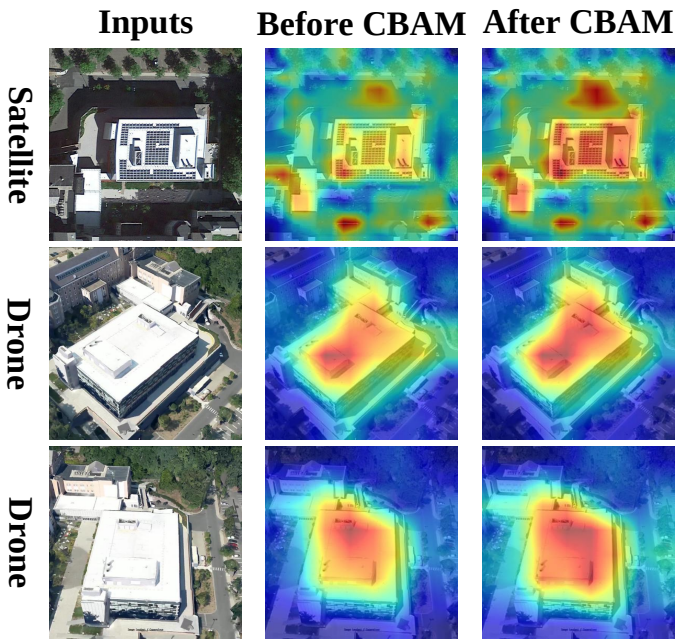

Fig. 0. The visualization results of CBAM.

TABLE X. The number of images used in different test datasets for different geo-localization tasks.

Dataset	Task			
	Drone→Satellite		Satellite→Drone	
	Query	Gallery	Query	Gallery
University-1652 [9]	37855	951	701	51355
SUES-200 [47]	16000	200	80	40000
Dataset	Ground→Satellite		Satellite→Ground	
	Query	Gallery	Query	Gallery
CVUSA [16]	8884	8884	8884	8884
CVACT [6]	8884	8884	8884	8884

Comment 8: The experimental dataset is too small to prove the effectiveness of the proposed method, and the experimental scene should be expanded and the relevant geographical location should be labeled.

***Response*:** Thanks for your question! In order to prove the effectiveness of the proposed network model, we have supplemented the experimental results of the new dataset in the experimental section. In addition, some new experiments were added to the ablation study to analyze the proposed network model.

Revision: SUES-200 [47] is a cross-view geo-localization dataset with multiple sources, multiple scenes, and panoramic views. Specifically, the SUES-200 dataset includes drone-view images at different heights, including school buildings, parks, schools, lakes, and public buildings. The matching and localization tasks are mainly divided into two types: drone-view target localization task (Drone→Satellite) and drone navigation task (Satellite→Drone). The training dataset contains 120 scenarios which has 120 satellite-view and 24000 drone-view images. In the drone-view target localization task (Drone→Satellite), each height in the query set has 4000 drone-view images that are matched with 200 satellite-view images in the gallery set, and includes 120 satellite-view distractors. In the drone navigation task (Satellite→Drone), each height in the query set has 80 satellite-view images are matched with 200 drone-view images in the gallery set, which including 6000 drone-view distractors. In this task, each satellite-view image will correspond to multiple drone-view images. (see the PERFORMANCE EVALUATION A, page 8)

2) The Experimental Results using the SUES-200 Dataset: The comparison results with the state-of-the-art methods using the SUES-200 dataset are given in Table III. The experimental results are mainly divided into three groups, namely the baseline-related methods, the experimental results of methods using contextual information and the experimental results of the Transformer-based method. From the Table III, the the baseline-related methods are given in the first and second rows, the third row shows the experimental results of methods using contextual information, the experimental results of the Transformer-based method are shown in the fourth row and the last row is the experimental result of the proposed network model. It

TABLE III. Comparison with the state-of-the-art methods using the SUES-200 dataset. The input image size for comparison methods is 384×384. For our method, the image size is 256×256.

Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [50]	43.42	49.65	49.42	55.91	57.47	60.31	60.43	65.78
LPN [1]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
Vit [47]	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
Ours	**77.57**	**81.30**	**89.50**	**91.40**	**92.58**	**94.21**	**97.40**	**97.92**
Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LCM [50]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN [1]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
Vit [47]	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
Ours	**93.75**	**79.49**	**97.50**	**90.52**	**97.50**	**96.03**	**100.00**	**97.66**

TABLE V. Ablation study on the effect of the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

GIF	LIF	LGGIF	Drone→Satellite		Satellite→Drone	
			R@1	AP	R@1	AP
×	×	×	64.13	68.73	76.32	60.20
✓	×	×	71.78	75.68	80.74	68.69
×	✓	×	68.96	72.42	84.02	66.68
×	×	✓	65.96	70.32	79.32	63.41
✓	✓	×	82.72	84.95	89.44	79.19
✓	×	✓	72.85	76.52	83.17	69.86
×	✓	✓	81.56	83.58	89.73	80.17
✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

can be seen from the experimental results that the proposed network model achieves the accuracy of R@1 are 77.57, 89.50, 92.58, 97.40 and AP are 81.30, 91.40, 94.21, 97.92 on the drone-view target localization task (Drone→Satellite) at different height, and it can achieve the accuracy of R@1 are 93.75, 97.50, 97.50, 100.00 and AP are 79.49, 90.52, 96.03, 97.66 on the drone navigation task (Satellite→Drone) at different height. Compared with LPN method, the R@1 and AP are improved 10.00%, 8.75%, 5.00%, 7.50% and 12.71%, 15.51%, 14.69%, 11.94% for the Drone→Satellite task at different heights, and the R@1 and AP are improved 15.99%, 18.92%, 12.20%, 15.93% and 14.07%, 15.44%, 10.41%, 13.39% for the Satellite→Drone task at different heights. It can be seen from the experimental results that the proposed network model is effective through introducing global information and local-guided-global information branches on the basis of using contextual information, and the performance of the model has been greatly improved. (see the PERFORMANCE EVALUATION D 2), pages 9-10)

1) Effect of the Various Branches: The main contribution of this paper is to design three branches, the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In order to verify the effectiveness of these three proposed branches in the network model, we designed several experiments to test each branch as shown in Table V. It can be seen from Table V that no matter which branch is excluded, the performance of the network model has a certain decline. In addition, the performance of the model will be significantly improved after using any branch, which also proves the effectiveness of the proposed different branches. Intuitively, the performance of the model is greatly improved after combining the local information fusion branch, which also proves that the contextual information in the image has a significant impact on the cross-view geo-localization task. Thus, we fully consider the global information in the image and employ local information to assist the global features to mine the critical information in the image as much as possible. From the experimental results, it can be seen that these three introduced branches are effective and the discriminability of the final feature descriptors can be improved through utilizing the global features, improving the retrieval precision of the network model. (see the PERFORMANCE EVALUATION E 1), page 11)

5) Effect of Matching accuracy of Multiple Queries: In practical applications, it is difficult to fully describe the target location from a single drone-view image. Furthermore, the University-1652 dataset provides images from different perspectives of each scene, which means that we can use multiple drone-view images as queries at the same time to explore whether these multi-view queries can improve the matching accuracy for the Drone→Satellite task. In the test phase, we average the features obtained from multiple images, and take the processed feature as the final query feature. The experimental results are shown in Table IX. For the test phase, we set the number of multi-view images to 1,2,3,9,18,27 and 54 according to the number of drone-view images in the dataset. It can be seen from the experimental results that with the increase of the

TABLE IX. Ablation study on the effect of matching accuracy of multiple queries.

Query	Drone→Satellite				
	R@1	R@5	R@10	R@Top1%	AP
54	**91.73**	**97.15**	**97.86**	97.86	**93.00**
27	91.23	97.08	97.79	**97.93**	92.51
18	90.63	96.67	97.75	97.85	91.99
9	89.23	96.34	97.72	97.84	90.84
3	87.03	95.51	97.12	97.30	88.94
2	86.68	95.20	96.97	97.19	88.62
1	86.06	94.95	96.82	97.00	88.08

number of images used, the network model performance has been significantly improved, with an improvement of 5.67% and 4.98% for Recall@1 and AP respectively. Multi-view features are effective for Drone→Satellite task and we hope that these experiments can provide an effective solved method for the practical application. (see the PERFORMANCE EVALUATION E 5), page 12)

Comment 9: It is suggested to give the calculation formula of Evaluation Metrics.

***Response*:** Thanks a lot for your suggestion. We have supplemented the calculation formula of evaluation metrics in the experimental section.

Revision: B. Evaluation Metrics

The Recall@K (R@K) and average accuracy (AP) metrics are selected to evaluate the performance of the network model. R@K represents the proportion of correctly matched images in the top-K of the ranking list, which can be formulated as follows:

$$\mathrm{Recall@}K = \frac{TP@K}{N} \tag{12}$$

where N is the total number of query image.

A higher recall rate demonstrates that the network model has better performance. In addition, we calculate the area under the Precision-Recall curve, called average accuracy (AP), which reflects the precision and recall rate of the retrieval performance. The formula can be shown as follows:

$$AP = \int_0^1 p\left(r\right)dr \tag{13}$$

(see the PERFORMANCE EVALUATION B, page 8)

Comment 10: Visual comparison of the experimental results should be presented in the paper.

***Response*:** Thank you very much for your valuable opinions. We have modified the qualitative results section, adding the visual comparison of heatmap results with other methods and the visual comparison of Image retrieval results.

Revision: F. Qualitative Results

We illustrate some heatmap visualizations generated using LPN, FSRA and the proposed network model as qualitative results, shown in Fig.10. It can be seen from the heatmaps that the LPN method pays more attention to the contextual information in the image, which may therefore ignore the critical information. FSRA method can pay more attention to the critical information in the image through region alignment strategy, however, it can only focus on a part of the crucial information, and the use of crucial information is still insufficient. Compared with the LPN method and the FSRA method, the proposed network model pays more attention to the critical information in the image after combining the contextual information, which also shows the effectiveness of the designed modules. In addition, some retrieval results of compared methods on different datasets are shown in Fig.11. The University-1652 dataset is utilized for two tasks, a drone-view target localization task (Drone→Satellite) and a drone navigation task (Satellite→Ground). The retrieval results of these two tasks are given in Fig.11(I) and Fig.11(II). From the compared retrieval results, it can be seen that the proposed method can effectively retrieve the corresponding scenes in both tasks. Although other comparison algorithms can also find the correct scenes, there are some errors in them and they are not the highest feature matching degree. Fig.11(III) and Fig.11(IV) show the retrieval results of the ground→Satellite localization task using the CVUSA and CVACT datasets. It can be seen from these results that given a randomly selected query image, the most relevant image can be retrieved from the candidate gallery through the proposed method, which demonstrates the effectiveness of the proposed network model. (see the PERFORMANCE EVALUATION F, pages 12-13)

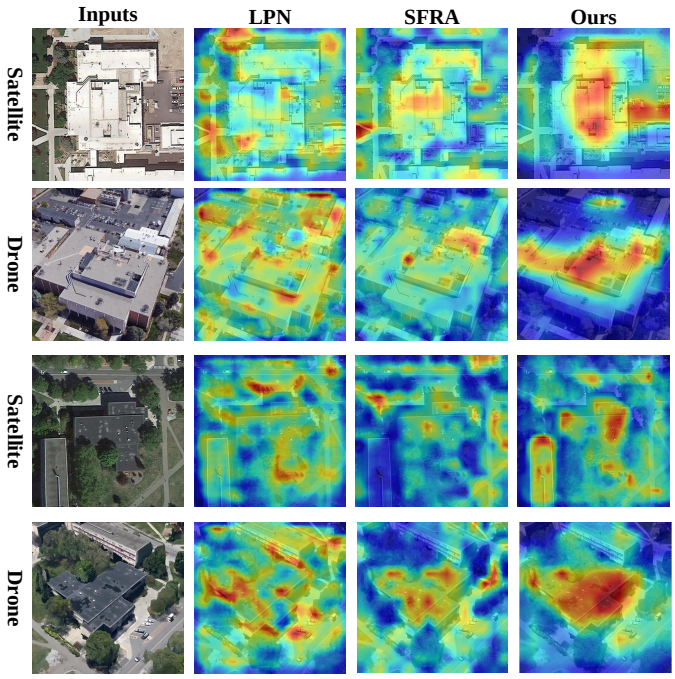

Fig. 10. Heatmaps are produced by the LPN, FSRA and ours on different platforms on the University-1652 dataset.

Comment 11: The experimental analysis is insufficient, more experimental analysis needs to be added in this paper.

***Response*:** Thanks greatly for your suggestion! We have revised the experimental results section and discussed and analyzed the experimental results of each part more deeply.

Revision: D. Experimental Results

TABLE II. Comparison with the state-of-the-art methods using the University-1652 dataset. † denotes the input image of size 384×384. For other methods, the image size of the transform-based methods and CNN-based method are 224×224 and 256×256 respectively.

Method	University-1652			
	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
Baseline (Instance Loss) [9]	58.23	62.91	74.47	59.45
Contrastive Loss [30]	52.39	57.44	63.91	52.24
Triplet Loss (M = 0.3) [49]	55.18	59.97	63.62	53.85
Triplet Loss (M = 0.5) [49]	53.58	58.60	64.48	53.15
Soft Margin Triplet Loss [18]	53.21	58.03	65.62	54.47
LCM† [50]	66.65	70.82	79.89	65.38
RK-Net [51]	66.13	70.23	80.17	65.76
LPN [1]	75.93	79.14	86.45	74.49
LPN + USAM [51]	77.60	80.55	86.59	75.96
PCL [15]	79.47	83.63	87.69	78.51
F3-net [52]	78.64	81.60	-	-
Swin-B [53]	84.15	86.62	90.30	83.55
FSRA [22]	84.51	86.71	88.45	83.47
Ours	**86.06**	**88.08**	**91.44**	**85.73**

1) The Experimental Results using the University-1652 Dataset: The comparison results with the state-of-the-art methods using the University-1652 dataset are given in Table X. The comparison results are mainly divided into three groups, the baseline-related methods, methods harnessing contextual information and Transformer-based methods. The experimental results of the first group of methods are given in the first row to seventh row, these methods pay more attention to the global features and show good results. However, the performance of these methods is not ideal as they may only focus on some global features, these features are difficult to fully effectively identify different scenes due to there are many similar scenes. Moreover, these

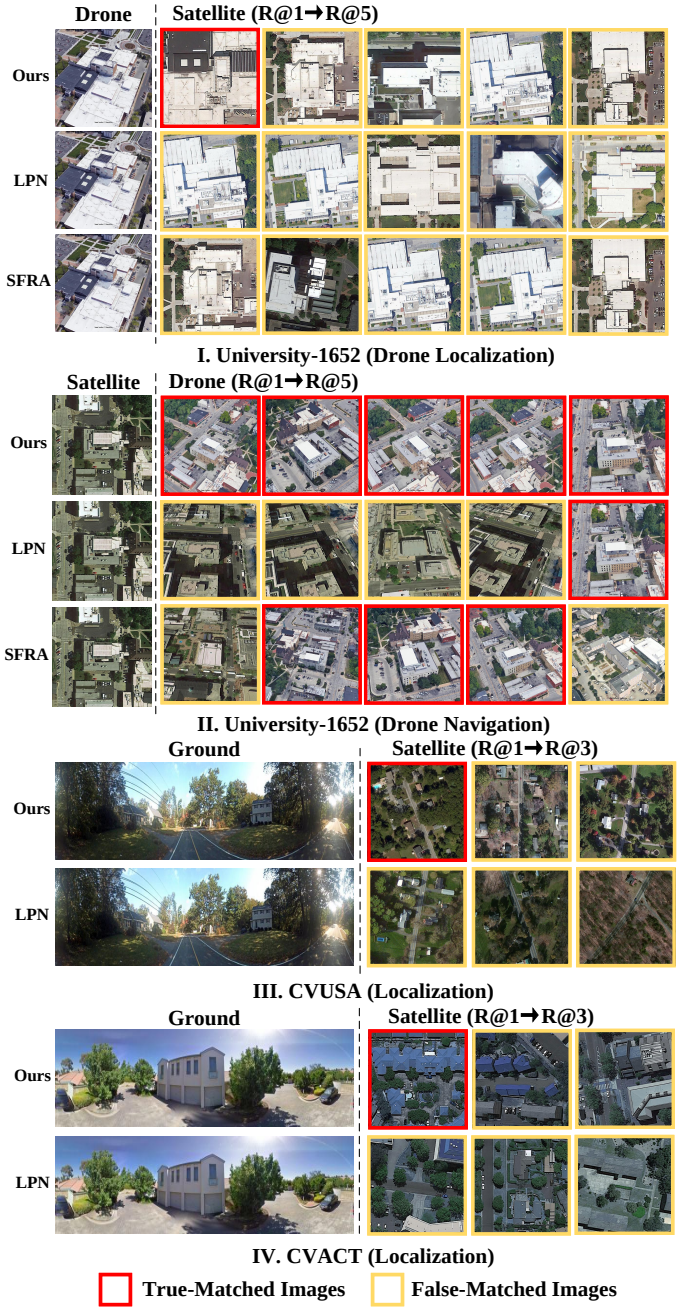

Fig. 11. Qualitative image retrieval results. (I) Top-5 retrieval results for drone-view target localization in different methods using the University-1652 dataset. (II) Top-5 retrieval results for drone navigation in different methods using the University-1652 dataset. (III) Top-3 retrieval results of geographic localization in different methods using the CVUSA dataset. (III) Top-3 retrieval results of geographic localization in different methods using the CVACT dataset. The true matches are in red boxes, while the false matches are displayed in yellow boxes.

TABLE III. Comparison with the state-of-the-art methods using the SUES-200 dataset. The input image size for comparison methods is 384×384. For our method, the image size is 256×256.

Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [50]	43.42	49.65	49.42	55.91	57.47	60.31	60.43	65.78
LPN [1]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
Vit [47]	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
Ours	**77.57**	**81.30**	**89.50**	**91.40**	**92.58**	**94.21**	**97.40**	**97.92**
Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LCM [50]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN [1]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
Vit [47]	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
Ours	**93.75**	**79.49**	**97.50**	**90.52**	**97.50**	**96.03**	**100.00**	**97.66**

methods ignore other valid information in the image that can have a significant impact on the final results. The experimental results of the second group of methods are given in the eighth row to eleventh row, and it can be seen from these results that the performance of these algorithms has been significantly improved after introducing contextual information. From the experimental results, we also see that it is necessary to effectively introduce contextual information into images in the network model. However, these methods only using contextual information while ignoring global information in the image will cause the model to ignore some crucial information in the image, which will have an impact on the final results. The experimental results of the third group of methods are given in the twelfth row and thirteenth row, and it can be seen from the experimental results that the Transformer-based methods have better feature expression ability than the CNN-based algorithm. Therefore, the experimental results for these two methods are significantly better than those for the CNN-based methods. The last row in the table shows the experimental result for the proposed network model. Since the proposed network model fully considers the global information and contextual information in the image, meanwhile, it introduces the idea of local information guiding the global information to improve the ability of the model to discover crucial information. Therefore, the performance of the model has been significantly improved. In the drone-view target localization task (Drone→Satellite), the proposed model achieves 86.06% accuracy for R@1 and 88.08% AP, and in the drone navigation task (Satellite→Drone), the proposed model achieves 91.44% accuracy for R@1 and 85.73% AP. Compared with the LPN method, the R@1 and AP metrics are improved by 10.13% and 8.94% on Drone→Satellite respectively, and by 4.99% and 11.24% on Satellite→Drone respectively. The experimental results also prove the effectiveness of introducing different information fusion branches. In addition, although the Transformer-based method is better than the CNN-based method for feature representation, the performance of the proposed network model is significantly better than these two Transformer-based methods, which also proves the effectiveness of the introduced different branches. For the Drone→Satellite, compared with the Swin-B and SFRA methods, the R@1 and AP are improved by 1.91% and 1.46%, and 1.55% and 1.37% respectively. For the Satellite→Drone, compared with Swin-B and SFRA methods, the R@1 and AP are improved by 1.14% and 2.18%, and 2.99% and 2.26% respectively.

2) The Experimental Results using the SUES-200 Dataset: The comparison results with the state-of-the-art methods using the SUES-200 dataset are given in Table III. The experimental results are mainly divided into three groups, namely the baseline-related methods, the experimental results of methods using contextual information and the experimental results of the Transformer-based method. From the Table III, the the baseline-related methods are given in the first and second rows, the third row shows the experimental results of methods using contextual information, the experimental results of the Transformer-based method are shown in the fourth row and the last row is the experimental result of the proposed network model. It can be seen from the experimental results that the proposed network model achieves the accuracy of R@1 are 77.57, 89.50, 92.58, 97.40 and AP are 81.30, 91.40, 94.21, 97.92 on the drone-view target localization task (Drone→Satellite) at different height, and it can achieve the accuracy of R@1 are 93.75, 97.50, 97.50, 100.00 and AP are 79.49, 90.52, 96.03, 97.66 on the drone navigation task (Satellite→Drone) at different height. Compared with LPN method, the R@1 and AP are improved 10.00%, 8.75%, 5.00%, 7.50% and 12.71%, 15.51%, 14.69%, 11.94% for the Drone→Satellite task at different heights, and the R@1 and AP are improved 15.99%, 18.92%, 12.20%, 15.93% and 14.07%, 15.44%, 10.41%, 13.39% for the Satellite→Drone task at different heights. It can be seen from the experimental results that the proposed network model is effective through introducing global information and local-guided-global information branches on the basis of using contextual information, and the performance of the model has been greatly improved.

3) The Experimental Results using the CVUSA Dataset: The comparison results with the state-of-the-art methods using the CVUSA dataset are given in Table IV. The experimental results are mainly divided into two groups, the method without using polar transform and the method using polar transform. The experimental results for the first group of methods are given in the

TABLE IV. Comparison with the state-of-the-art methods using the CVUSA and CVACT datasets. * represents when the method harnesses extra orientation information as input.

Method	Backbone	CVUSA				CVACT			
		R@1	R@5	R@10	R@Top1%	R@1	R@5	R@10	R@Top1%
Zhai [16]	VGG16	-	-	-	43.20	-	-	-	-
Vo [54]	AlexNet	-	-	-	63.70	-	-	-	-
CVM-Net [18]	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87	87.57
Orientation* [6]	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48	92.04
Zheng et al. [9]	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00	85.27
Regmi [57]	X-Fork	48.75	-	81.27	95.98	-	-	-	-
RKNet [51]	USAM	52.50	-	-	96.52	40.53	-	-	89.12
Siam-FCANet [11]	ResNet-34	-	-	-	98.30	-	-	-	-
CVFT [12]	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [1]	ResNet-50	85.79	95.38	96.80	99.41	79.99	90.63	92.56	97.03
GeoNet-II [55]	ResNetX	-	-	-	98.70	58.90	81.80	88.30	97.70
SIRNet [33]	VGG16	81.82	93.39	96.24	99.49	75.37	88.76	91.90	97.42
TransGeo [56]	ViT	94.08	98.36	99.04	**99.77**	-	-	-	-
L2LTR [57]	ViT	91.99	97.68	98.65	99.75	83.14	93.84	95.51	98.40
Polar Transform Methods									
SAFA [10]	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17
DSM [13]	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32
Shi et al. [58]	VGG16	92.69	97.78	98.60	99.61	82.70	92.50	94.42	97.65
LPN [1]	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77
LPN + USAM [51]	ResNet-50	91.22	-	-	99.67	82.02	-	-	98.18
Toker [14]	ResNet-34	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22
SIRNet [33]	VGG16	93.74	98.02	98.85	99.76	86.02	94.45	**96.02**	98.33
L2LTR [57]	ViT	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37
Ours	ResNet-50	**95.09**	**98.85**	**99.34**	**99.77**	**86.64**	**94.61**	95.94	**98.45**

first row to fourteenth row, these methods show good results for the cross-view geo-localization task. However, the CVUSA dataset is mainly aimed at cross-view image matching between satellite-view and ground-view, due to the huge change of perspective, the content information has changed significantly which presents challenges to the network model. In addition, it is difficult for these methods to spatially align the image features under the changing view, which leads to the model performance is not ideal. Therefore, many methods employ polar transforms to convert satellite-view images. It considers the geometric correspondence of two-platform images and transforms the aerial-view image to approximately align a ground panorama at the pixel level. The experimental results for the second group of methods are given in the last nine rows. From the comparison results of LPN in the two groups, it can be seen that the performance of the method has greatly improved after using a polar transform. From IV, it can be seen that the proposed network model is significantly superior to other methods using the CVUSA dataset after employing a polar transform, and achieves an accuracy of R@1 95.09%, R@5 98.85%, R@10 99.34% and R@Top1% 99.77%. Compared with the LPN method after using a polar transform, the result for R@1 improved 1.31%, which proves the effectiveness of the proposed network model. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.04% and 0.10% on the R@1 and AP, which can proved the effectiveness of the proposed network model.

4) The Experimental Results using the CVACT Dataset: The comparison results with the state-of-the-art methods using the CVACT dataset are given in Table IV. Due to the image perspective structure of CVACT being similar to the CVUSA dataset, the experimental results using this dataset are mainly divided into two groups, the method without using polar transform and the method using polar transform. Similar to the experimental results in CVUSA, the performance of the method is obviously improved after using a polar transform. It can be seen from Table IV that the proposed network model is significantly superior to other methods using the CVACT dataset after employing a polar transform, and achieves accuracies of R@1 86.64%, R@5 94.61%, R@10 95.94% and R@Top1% 98.45%. Compared with the LPN method after using polar transform, the results of R@1, R@5, R@10 and R@Top1% improve by 3.77%, 2.35%, 1.85% and 0.68% respectively. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.75% and 0.08% on the R@1 and AP, which proves the effectiveness of the proposed network model for the cross-view geo-localization task.

E. Ablation Study

In order to prove the validity of each part of the proposed network model, we design several ablation experiments which mainly focus on two tasks, drone-view target localization task (Drone→Satellite) and drone navigation task (Satellite→Ground).

1) Effect of the Various Branches: The main contribution of this paper is to design three branches, the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In order to verify the effectiveness of these three proposed branches in the network model, we designed several experiments to test each branch as shown in Table V. It can be seen from Table V that no matter which branch is excluded, the performance of the network model has a certain decline. In addition, the performance of the model will be significantly improved after using any branch, which also proves the effectiveness of the proposed different branches. Intuitively, the performance of the model

TABLE V. Ablation study on the effect of the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

GIF	LIF	LGGIF	Drone→Satellite		Satellite→Drone	
			R@1	AP	R@1	AP
×	×	×	64.13	68.73	76.32	60.20
✓	×	×	71.78	75.68	80.74	68.69
×	✓	×	68.96	72.42	84.02	66.68
×	×	✓	65.96	70.32	79.32	63.41
✓	✓	×	82.72	84.95	89.44	79.19
✓	×	✓	72.85	76.52	83.17	69.86
×	✓	✓	81.56	83.58	89.73	80.17
✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

is greatly improved after combining the local information fusion branch, which also proves that the contextual information in the image has a significant impact on the cross-view geo-localization task. Thus, we fully consider the global information in the image and employ local information to assist the global features to mine the critical information in the image as much as possible. From the experimental results, it can be seen that these three introduced branches are effective and the discriminability of the final feature descriptors can be improved through utilizing the global features, improving the retrieval precision of the network model.

2) Effect of Different Information Fusion Strategies and Attention Mechanism on Different Branches: In order to enhance the limited receptive field in each branch, we introduce different information fusion strategies (IFS) and attention mechanism to improve the global receptive field of each module and enhance the performance of the network model. In order to prove the effectiveness of these strategies, we design some experiments for these strategies, and the experimental results are given in Table VI. Without introducing any information fusion strategy and attention mechanism, the performance of the proposed network model is not very ideal. It can be seen from Table VI that the performance of the model has been clearly improved after introducing these strategies. These strategies are mainly to enable each branch to mine and utilize the critical information from the global or contextual information, and further improve the discriminability of each feature descriptor based on improving the limited global receptive field. It can be seen from the experimental results that the introduced strategies in each branch are essential and have a huge impact on the final result of the network model, which can prove the effectiveness of these designed strategies in the network model.

TABLE VI. Ablation study on the effect of the information fusion strategies (IFS) and attention mechanism (CBAM) in the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

IFS in GIF	IFS in LIF	IFS in LGGIF	CBAM	Drone→Satellite		Satellite→Drone	
				R@1	AP	R@1	AP
×	×	×	×	80.35	82.79	87.02	79.49
×	✓	✓	✓	83.59	85.88	89.59	83.15
✓	×	✓	✓	82.20	84.64	88.87	82.19
✓	✓	×	✓	83.24	85.63	90.16	82.10
✓	✓	✓	×	83.97	86.21	89.16	82.75
✓	✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

TABLE VII. Ablation study on the effect of different input sizes on the University-1652 dataset.

Image Size	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
224	77.95	80.84	85.45	76.79
256	86.06	88.08	91.44	85.73
320	86.21	88.38	91.90	85.94
384	86.91	88.83	92.15	**86.50**
512	**87.21**	**88.98**	**92.30**	85.73

3) Effect of Different Input Sizes Using the University-1652 Dataset: For model training and testing, the size of the input image will effect the fine-grained information within the image, which will affect the feature representation learning due to the missing information. However, a larger input size introduces more memory costs during training and testing and increases the computational complexity. Therefore, in order to balance the size and performance of the input image, we design some experiments to determine the influence of the input image size on the model performance. In the experiment, we only change the size of the input image and the region covered by the image is not changed; the experimental results are shown in Table VII. We test the size of the input images from 224 to 512 in two studied tasks respectively. It is worth noting that we choose to test the impact of this range of sizes on the final network model performance as the size of the image in the University-1652

dataset is 512×512. It can be seen from the experimental results that the model performance gradually improves with the change of the input image size, which also indicates that the missing image information caused by the reduction of the input image size also has a great impact on the results. When we continue to expand the input size to 512, the improvement is not so clear on Satellite→Drone, which also shows that when the input image size reaches a certain threshold, the impact of the missing information will also be reduced. We hope that this finding can provide effective insights in the case of limited computing resources, in order to choose the size of the input image in real-world applications.

4) Effect of Image Position Shifting Using the University-1652 Dataset: In order to demonstrate the robustness of the proposed network model to position shifting, we design some experiments using different degrees of horizontal shifting to query images; the experimental results are shown in Table VIII. Examples of image translation are shown in Fig.8, where we translate the image by different degrees to generate new query images to test retrieval performance. It can be seen from the experimental results that the model performance gradually decreases with the increase of image translation. However, it does not decrease significantly in the case of slight translation, indicating that the performance of the introduced different branches will not be greatly affected in the case of minimal changes in content, which proves that the proposed network model has a strong robustness to position shifting. Although some information is missing after the image translation, the three designed branches can make the network model better mine and utilize the critical information which is why the proposed network model can resist position shifting.

Fig. 8. The image on the left is the original query image, the middle image is the image when translated 30 pixels, and the right image is the new generated query image.

TABLE VIII. Ablation study on the effect of image position shifting on the University-1652 dataset.

Shifted Pixel	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
0	**86.06**	**88.08**	**91.44**	**85.73**
10	86.05	**88.08**	91.28	85.69
20	86.02	88.06	91.01	85.42
30	85.95	87.99	90.87	84.90
40	85.77	87.84	90.44	84.44
50	85.30	87.45	90.37	83.80
60	84.77	87.00	90.34	82.98

TABLE IX. Ablation study on the effect of matching accuracy of multiple queries.

Query	Drone→Satellite				
	R@1	R@5	R@10	R@Top1%	AP
54	**91.73**	**97.15**	**97.86**	97.86	**93.00**
27	91.23	97.08	97.79	**97.93**	92.51
18	90.63	96.67	97.75	97.85	91.99
9	89.23	96.34	97.72	97.84	90.84
3	87.03	95.51	97.12	97.30	88.94
2	86.68	95.20	96.97	97.19	88.62
1	86.06	94.95	96.82	97.00	88.08

5) Effect of Matching accuracy of Multiple Queries: In practical applications, it is difficult to fully describe the target location from a single drone-view image. Furthermore, the University-1652 dataset provides images from different perspectives

of each scene, which means that we can use multiple drone-view images as queries at the same time to explore whether these multi-view queries can improve the matching accuracy for the Drone→Satellite task. In the test phase, we average the features obtained from multiple images, and take the processed feature as the final query feature. The experimental results are shown in Table IX. For the test phase, we set the number of multi-view images to 1,2,3,9,18,27 and 54 according to the number of drone-view images in the dataset. It can be seen from the experimental results that with the increase of the number of images used, the network model performance has been significantly improved, with an improvement of 5.67% and 4.98% for Recall@1 and AP respectively. Multi-view features are effective for Drone→Satellite task and we hope that these experiments can provide an effective solved method for the practical application. (see the PERFORMANCE EVALUATION D-E, pages 8-12)

Reviewer #2

Comments to Authors:

The author designs a multi-branch joint representation learning model based on information fusion strategies, which includes a global information fusion (GIF) branch, a local information fusion (LIF) branch and a local-guided-global information fusion (LGGIF) branch, to extract global information and contextual information in different cross-view images. Furthermore, this paper introduces different information fusion strategies into these branches to expand the global receptive field of each module and enhance the discriminability of each part of the image representation. The experiment results verify the effectiveness of the proposed model.

I only have some minor concerns with details.

***Response*:** Thank you very much for your valuable suggestions for this paper. We have revised the paper according to your suggestions, which has greatly improved the quality of this paper. Thank you again for your valuable suggestions for the paper. The modified contents are all highlighted in blue.

Comment 1: Instance loss is proposed in another paper [a]. It would be great to clarify.
[a] Dual-Path Convolutional Image-Text Embeddings with Instance Loss. TOMM

***Response*:** This is a valuable suggestion. We want to refer to in the paper is the paper of the builder of the University-1652 dataset, which uses the instance function to solve this problem, and we have revised this section.

[9] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in Proceedings of the 28th ACM international conference on Multimedia, 2020, pp. 1395–1403.

Revision: (see the PERFORMANCE EVALUATION D, page 9)

TABLE X. Comparison with the state-of-the-art methods using the University-1652 dataset. † denotes the input image of size 384×384. For other methods, the image size of the transform-based methods and CNN-based method are 224×224 and 256×256 respectively.

Method	University-1652			
	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
Baseline (Instance Loss) [9]	58.23	62.91	74.47	59.45
Contrastive Loss [30]	52.39	57.44	63.91	52.24
Triplet Loss (M = 0.3) [49]	55.18	59.97	63.62	53.85
Triplet Loss (M = 0.5) [49]	53.58	58.60	64.48	53.15
Soft Margin Triplet Loss [18]	53.21	58.03	65.62	54.47
LCM† [50]	66.65	70.82	79.89	65.38
RK-Net [51]	66.13	70.23	80.17	65.76
LPN [1]	75.93	79.14	86.45	74.49
LPN + USAM [51]	77.60	80.55	86.59	75.96
PCL [15]	79.47	83.63	87.69	78.51
F3-net [52]	78.64	81.60	-	-
Swin-B [53]	84.15	86.62	90.30	83.55
FSRA [22]	84.51	86.71	88.45	83.47
Ours	**86.06**	**88.08**	**91.44**	**85.73**

Comment 2: Figure 1 is not introduced in the paper text. Moreover, it is better to arrange the figures/tables above the text content rather than inserting them in the middle of the content.

***Response*:** Thank you very much for reminding. We have added the illustration of Figure 1 to the introduction. In addition, we modified the position of the tables and figures to make them appearing above the text content.

Revision: Although mining and utilizing contextual information in images can effectively improve the accuracy of cross-view geo-localization, the use of global information in images is equally important. As shown in Fig.1, a previous work, LPN [1], focuses more on the contextual information for the input image, ignoring other crucial information in the scenes. (see the fourth paragraph of INTRODUCTION, page 2)

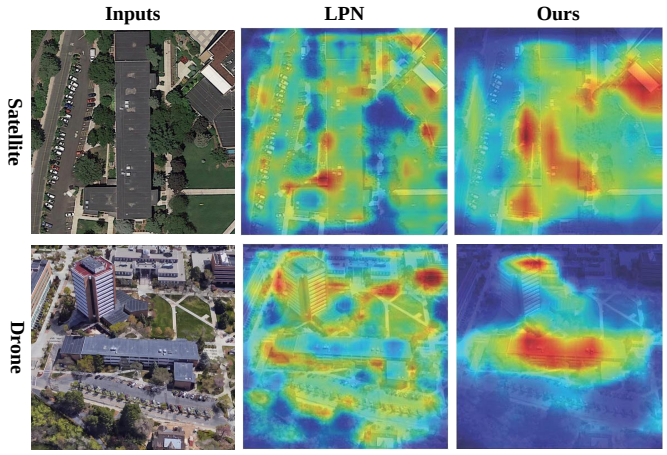

Fig. 1. Difference of the activation maps generated by the LPN method [1] and our method. The images on the left column are the input drone-view and satellite-view images. The images in the middle column are the heatmaps of LPN method. The images ont the right column are the heatmaps of our method. From the visualization result, it can be seen that our method focus on the important information in the image.

Comment 3: In the Section III.B lines 56, "specific operations are shown in Fig.4"→"specific operations are shown in Fig.5".
In the Section III.D lines 48, "full connection layer"→"fully connected layer".
In the Section IV.D.3) lines 38, "CVUSA"→"CVACT".

***Response*:** Thanks greatly for your suggestion! We have carefully reviewed the full text and made changes to the incorrect parts of the full text to avoid such errors.

Revision: 1) For the square-ring partition strategy, the center of the image is approximately aligned with the center of the feature map, and the entire part is partitioned according to the distance from the image center; specific operations are shown in Fig.5. (see the METHODOLOGY B, page 6)

2) The classifier is composed of four parts: full connected layer (FC), batch normalization layer (BN), dropout layer (Dropout) and classification layer (Cls). (see the METHODOLOGY D, page 7)

3) It can be seen from Table IV that the proposed network model is significantly superior to other methods using the CVACT dataset after employing a polar transform, and achieves accuracies of R@1 86.64%, R@5 94.61%, R@10 95.94% and R@Top1% 98.45%. (see the PERFORMANCE EVALUATION D 4), page 10)

Comment 4: Does the proposed network include a branch of ground-view image for the University1652 dataset?

***Response*:** This is a valuable opinion. While training the University-1652 dataset, we only used satellite-view and drone-view images, but did not use ground-view image. Therefore, the proposed network model only has two branches, which includes the satellite-view and drone-view images.

Comment 5: What is the difference between the global information descriptor and the global feature descriptor? It is better to mark out the mentioned features in the figures and describe the dimensions of features in content for readers to understand.

***Response*:** Thanks greatly for your suggestion! In our work, the global information descriptor is different from global feature descriptor. The global information descriptors. The global information descriptor is the feature obtained from the backbone network ResNet50, with the size of $16 \times 16 \times 2048$. The global feature descriptor is the feature processed by the maximum pooling layer and has a size of 1×2048. In addition, we have revised some descriptions and figures in the manuscript, highlighted the mentioned features in the figures, and described the dimensions of the features in the content for readers to read.

Revision: The global information descriptor is obtained through inputting the global feature descriptor into the max pooling layer. The size of the global information description and the global feature descriptors are $16 \times 16 \times 2048$ and 1×2048, respectively. (see the METHODOLOGY A, page 5)

The feature size of the global feature after global information fusion f_{i_GIF} is the same as the original global feature f_i size, which is $16 \times 16 \times 2048$. (see the METHODOLOGY A, page 5)

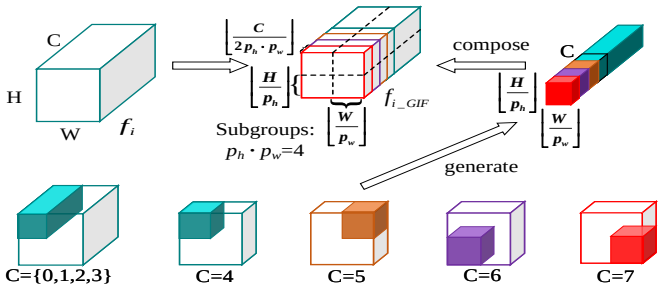

Fig. 3. The concrete implementation process of the global information fusion strategy.

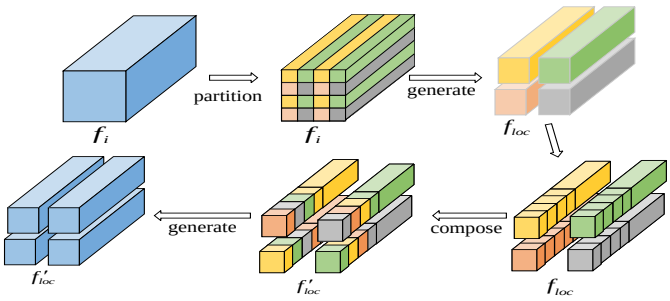

Fig. 4. The local information fusion strategy.

where $f_{loc,j}$ is the j-th block feature before processing, $f'_{loc,j}$ is the j-th block feature after partitioning, and N is the number of blocks of global features. In our work, the size of each part of the local feature $f'_{loc,j}$ is $8 \times 8 \times 2048$. (see the METHODOLOGY B, page 6)

The size of the global feature after downsampling f_{i_down} is the same as that of the local partitioned features f_{loc}, which is $8 \times 8 \times 2048$. Then the processed global features and the local partitioned features are added to generate the new global-local features. On this basis, a mixing information fusion strategy is introduced to recombine the newly generated features. Finally, these features are combined to generate a global feature f'_{com} that is consistent with the starting dimension $16 \times 16 \times 2048$. (see the METHODOLOGY C, page 7)

It is worth noting that the features f'_i generated in the local-guided-global information fusion branch also need to generate a unified form of feature descriptors through equation (2) to improve the accuracy of the final model. (see the METHODOLOGY C, page 7)

Comment 6: What is the difference between the proposed global information fusion strategy and the reference(19) cited in the paper?

(19)Z. Deng, X. Ren, J. Ye, J. He, and Y. Qiao, "Fcn+: Global receptive convolution makes fcn great again," arXiv preprint arXiv:2303.04589, 2023.

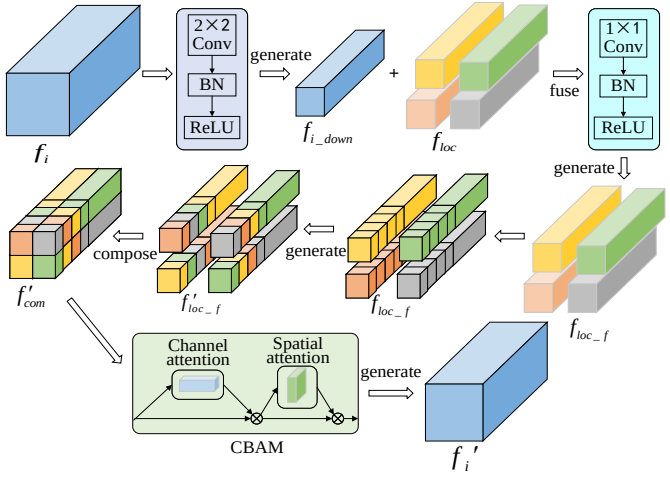

Fig. 6. The specific process of the local-guided-global information fusion branch.

***Response*:** This is a valuable opinion. In the global information fusion branch, the global information fusion strategy used in our work is roughly the same as that used in reference (19), but the subsequent processing and utilizing are different. In reference (19), they fused this process into a single block to solve other problem, as shown in the figure, while we used this information fusion strategy alone. In our work, we use the idea of multi-branch joint learning to solve the problem of cross-perspective positioning. This information fusion strategy is only a part of our overall network model and is trained in conjunction with other branches to make the model perform well. Therefore, there are still differences between us and reference (19) in the way of utilizing this part of information fusion strategy. In addition, we also illustrate in the manuscript that we are introducing this part of the mechanism.

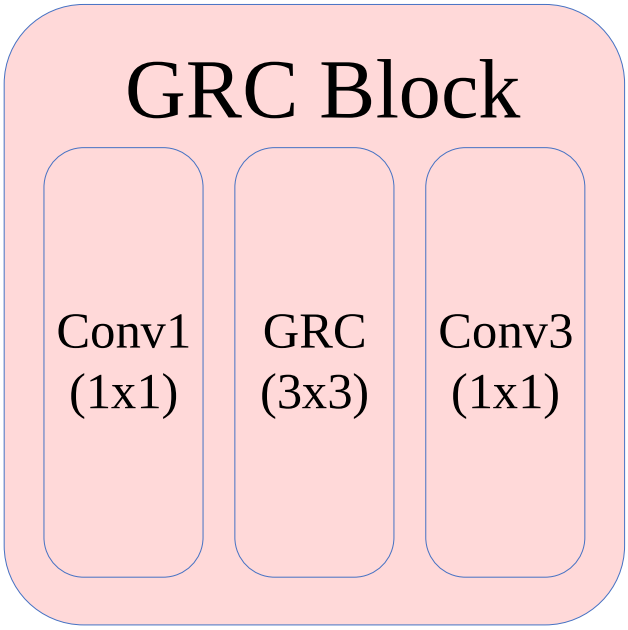

Fig. 0. The GRC Block.

Therefore, the global information fusion strategy [23] is introduced to increase the global receptive field of the network and obtain more effective global features. (see the METHODOLOGY A, page 5)

7: In the Global Information Fusion Branch, the two output global feature descriptors after max pooling seem to be the same. Because the proposed global information fusion strategy only changes the spatial position of features without changing the distribution of feature channels, which will not affect the results after max pooling.

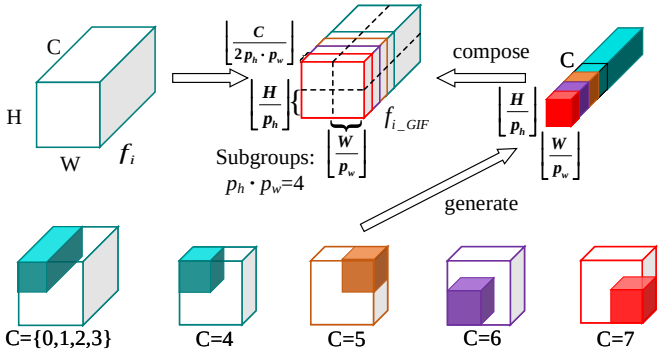

Fig. 3. The concrete implementation process of the global information fusion strategy.

***Response*:** Thank you for the reminder. The results of these two features are different after passing through the maximum

pooling layer. In the global information fusion strategy, each position of the feature obtains a part from other features, and although most of the features are retained, other features are also obtained. When the features of other parts are acquired, the channel distribution of the features of this part will also be collected to a new location, thus affecting the spatial position of the features of the original part and the distribution of the channels of the features. Therefore, the spatial position and channel distribution of the features in the newly generated part will change. Finally, the features obtained after passing through the maximum pooling layer are also different. The specific operation process is shown in the figure.

Comment 8: Can you explain equation 4 in more detail? For the equation 7, what does f_l mean?

***Response*:** This is a valuable suggestion. We have added an explanation of equation 4 to make it more detailed. f_l is the feature generated after the local information fusion strategy. In our work, this part of the feature is divided into four parts, and we need to continue to operate on this part of the feature. In addition, we revised the representation and formula of these features to correspond to the features in the image, so that each feature is clearer and easier to understand for readers.

Revision: The sampling location can be represented as follows:

$$N_f = \cup_{k,l} \left\{ \left(\hat{h}_k, \hat{w}_l \right) \right\} \tag{4}$$

N_f enumerates all the possible combinations of k, l, which can be formulated as follows:

$$\begin{aligned} N_f = \Big\{ &(0,0), \left(0, \left\lfloor \frac{W}{p_w} \right\rfloor \right), \cdots, \\ &\left((p_h - 2) \cdot \left\lfloor \frac{W}{p_h} \right\rfloor, (p_w - 1) \cdot \left\lfloor \frac{W}{p_w} \right\rfloor \right), \\ &\left((p_h - 1) \cdot \left\lfloor \frac{W}{p_h} \right\rfloor, (p_w - 1) \cdot \left\lfloor \frac{W}{p_w} \right\rfloor \right) \Big\} \end{aligned} \tag{5}$$

N_f includes $p_h \cdot p_w$ offset coordinates in total. It covers almost the entire input feature map.

The process can be formulated as follows:

$$f'^{m}_{loc,j} = F_{slice}\left(f'_{loc,j}, m\right) \tag{8}$$

where $f'^{m}_{loc,j}$ is the j-th block feature after processing, m is the number of divided regions, and F_{slice} represents the square-ring partition processing. (see the METHODOLOGY B, page 6)

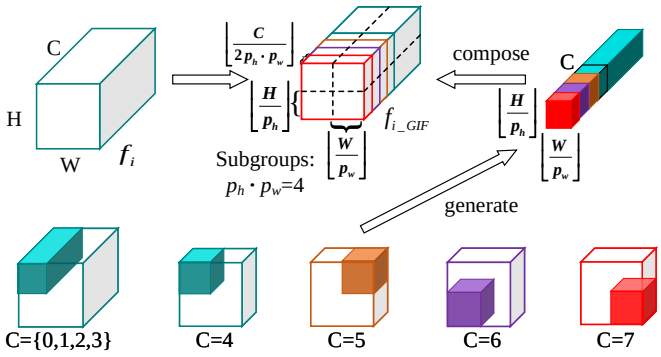

Fig. 3. The concrete implementation process of the global information fusion strategy.

The global information descriptor is obtained through inputting the global feature descriptor into the max pooling layer. The size of the global information description and the global feature descriptors are $16 \times 16 \times 2048$ and 1×2048, respectively. (see the METHODOLOGY A, page 5)

The feature size of the global feature after global information fusion f_{i_GIF} is the same as the original global feature f_i size, which is $16 \times 16 \times 2048$. (see the METHODOLOGY A, page 5)

where $f_{loc,j}$ is the j-th block feature before processing, $f'_{loc,j}$ is the j-th block feature after partitioning, and N is the number of blocks of global features. In our work, the size of each part of the local feature $f'_{loc,j}$ is $8 \times 8 \times 2048$. (see the METHODOLOGY B, page 6)

The size of the global feature after downsampling f_{i_down} is the same as that of the local partitioned features f_{loc}, which is $8 \times 8 \times 2048$. Then the processed global features and the local partitioned features are added to generate the new global-local features. On this basis, a mixing information fusion strategy is introduced to recombine the newly generated features. Finally, these features are combined to generate a global feature f'_{com} that is consistent with the starting dimension $16 \times 16 \times 2048$. (see the METHODOLOGY C, page 7)

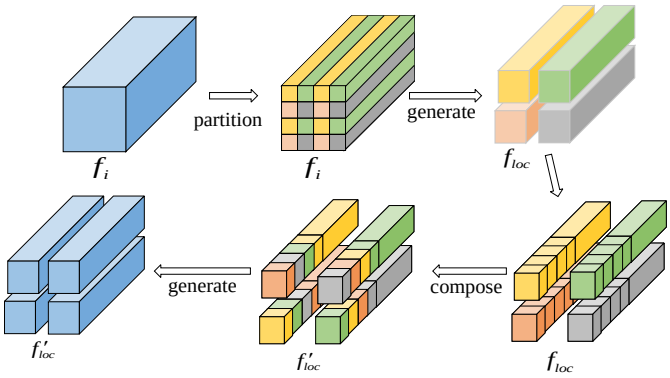

Fig. 4. The local information fusion strategy.

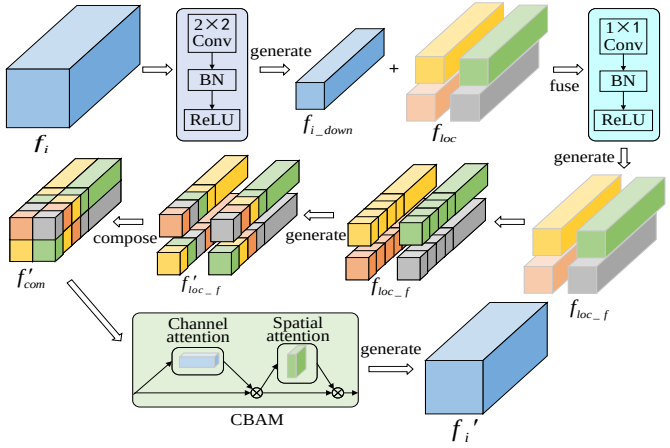

Fig. 6. The specific process of the local-guided-global information fusion branch.

It is worth noting that the features f_i' generated in the local-guided-global information fusion branch also need to generate a unified form of feature descriptors through equation (2) to improve the accuracy of the final model. (see the METHODOLOGY C, page 7)

Comment 9: In the Section III.C, the paper says "Then the processed global features and the local partitioned features are combined to generate new global-local features", how to combine global features and local partitioned features? Concat or addition?

***Response*:** Thank you very much for your comments. In our work, the processed global features and the local partitioned features are added to generate the new global-local features. We have revised and explained it in the manuscript.

Revision: Then the processed global features and the local partitioned features are added to generate the new global-local features. On this basis, a mixing information fusion strategy is introduced to recombine the newly generated features. Finally, these features are combined to generate a global feature f_{com}' that is consistent with the starting dimension $16 \times 16 \times 2048$. (see the METHODOLOGY C, page 7)

Comment 10: For the Table III, the result of "shi et al." in CVACT dataset is the best, which should be in bold font.
***Response*:** Thank you very much for your valuable comments. We reviewed the relevant literature again and found that it was our problem that led to the error of the numerical index, which was not the best. We have modified the result of "shi et al." in CVACT dataset.

Revision:
(see the PERFORMANCE EVALUATION D, page 10)

Comment 11: Lack of some competitive methods, such as SIRNet[1], L2LTR[2], TransGeo[3] and so on.
[1] Xiufan Lu, Siqi Luo, Yingying Zhu. It's Okay to Be Wrong: Cross-View Geo-Localization With Step-Adaptive Iterative Refinement. IEEE Transactions on Geoscience and Remote Sensing 2022

TABLE IV. Comparison with the state-of-the-art methods using the CVUSA and CVACT datasets. * represents when the method harnesses extra orientation information as input.

Method	Backbone	CVUSA				CVACT			
		R@1	R@5	R@10	R@Top1%	R@1	R@5	R@10	R@Top1%
Zhai [16]	VGG16	-	-	-	43.20	-	-	-	-
Vo [54]	AlexNet	-	-	-	63.70	-	-	-	-
CVM-Net [18]	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87	87.57
Orientation* [6]	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48	92.04
Zheng et al. [9]	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00	85.27
Regmi [57]	X-Fork	48.75	-	81.27	95.98	-	-	-	-
RKNet [51]	USAM	52.50	-	-	96.52	40.53	-	-	89.12
Siam-FCANet [11]	ResNet-34	-	-	-	98.30	-	-	-	-
CVFT [12]	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [1]	ResNet-50	85.79	95.38	96.80	99.41	79.99	90.63	92.56	97.03
GeoNet-II [55]	ResNetX	-	-	-	98.70	58.90	81.80	88.30	97.70
SIRNet [33]	VGG16	81.82	93.39	96.24	99.49	75.37	88.76	91.90	97.42
TransGeo [56]	ViT	94.08	98.36	99.04	**99.77**	-	-	-	-
L2LTR [57]	ViT	91.99	97.68	98.65	99.75	83.14	93.84	95.51	98.40
Polar Transform Methods									
SAFA [10]	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17
DSM [13]	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32
Shi et al. [58]	VGG16	92.69	97.78	98.60	99.61	82.70	92.50	94.42	97.65
LPN [1]	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77
LPN + USAM [51]	ResNet-50	91.22	-	-	99.67	82.02	-	-	98.18
Toker [14]	ResNet-34	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22
SIRNet [33]	VGG16	93.74	98.02	98.85	99.76	86.02	94.45	**96.02**	98.33
L2LTR [57]	ViT	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37
Ours	ResNet-50	**95.09**	**98.85**	**99.34**	**99.77**	**86.64**	**94.61**	95.94	**98.45**

[2] Hongji Yang, Xiufan Lu, Yingying Zhu. Cross-view Geo-localization with Layer-to-Layer Transformer. NeurIPS 2021

[3] Sijie Zhu, Mubarak Shah, Chen Chen. TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization. CVPR 2022

***Response*:** Thank you very much for reminding. We have added competitive methods to the experimental comparison on the CVUSA and CVACT datasets. In addition, we also perform experiments on new datasets to prove the effectiveness of the proposed network model.

Revision: 3) The Experimental Results using the CVUSA Dataset

The comparison results with the state-of-the-art methods using the CVUSA dataset are given in Table IV. The experimental results are mainly divided into two groups, the method without using polar transform and the method using polar transform. The experimental results for the first group of methods are given in the first row to fourteenth row, these methods show good results for the cross-view geo-localization task. However, the CVUSA dataset is mainly aimed at cross-view image matching between satellite-view and ground-view, due to the huge change of perspective, the content information has changed significantly which presents challenges to the network model. In addition, it is difficult for these methods to spatially align the image features under the changing view, which leads to the model performance is not ideal. Therefore, many methods employ polar transforms to convert satellite-view images. It considers the geometric correspondence of two-platform images and transforms the aerial-view image to approximately align a ground panorama at the pixel level. The experimental results for the second group of methods are given in the last nine rows. From the comparison results of LPN in the two groups, it can be seen that the performance of the method has greatly improved after using a polar transform. From IV, it can be seen that the proposed network model is significantly superior to other methods using the CVUSA dataset after employing a polar transform, and achieves an accuracy of R@1 95.09%, R@5 98.85%, R@10 99.34% and R@Top1% 99.77%. Compared with the LPN method after using a polar transform, the result for R@1 improved 1.31%, which proves the effectiveness of the proposed network model. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.04% and 0.10% on the R@1 and AP, which can proved the effectiveness of the proposed network model.

4) The Experimental Results using the CVACT Dataset

The comparison results with the state-of-the-art methods using the CVACT dataset are given in Table IV. Due to the image perspective structure of CVACT being similar to the CVUSA dataset, the experimental results using this dataset are mainly divided into two groups, the method without using polar transform and the method using polar transform. Similar to the experimental results in CVUSA, the performance of the method is obviously improved after using a polar transform. It can be seen from Table IV that the proposed network model is significantly superior to other methods using the CVACT dataset after employing a polar transform, and achieves accuracies of R@1 86.64%, R@5 94.61%, R@10 95.94% and R@Top1%

TABLE IV. Comparison with the state-of-the-art methods using the CVUSA and CVACT datasets. * represents when the method harnesses extra orientation information as input.

Method	Backbone	CVUSA				CVACT			
		R@1	R@5	R@10	R@Top1%	R@1	R@5	R@10	R@Top1%
Zhai [16]	VGG16	-	-	-	43.20	-	-	-	-
Vo [54]	AlexNet	-	-	-	63.70	-	-	-	-
CVM-Net [18]	VGG16	18.80	44.42	57.47	91.54	20.15	45.00	56.87	87.57
Orientation* [6]	VGG16	27.15	54.66	67.54	93.91	46.96	68.28	75.48	92.04
Zheng et al. [9]	VGG16	43.91	66.38	74.58	91.78	31.20	53.64	63.00	85.27
Regmi [57]	X-Fork	48.75	-	81.27	95.98	-	-	-	-
RKNet [51]	USAM	52.50	-	-	96.52	40.53	-	-	89.12
Siam-FCANet [11]	ResNet-34	-	-	-	98.30	-	-	-	-
CVFT [12]	VGG16	61.43	84.69	90.94	99.02	61.05	81.33	86.52	95.93
LPN [1]	ResNet-50	85.79	95.38	96.80	99.41	79.99	90.63	92.56	97.03
GeoNet-II [55]	ResNetX	-	-	-	98.70	58.90	81.80	88.30	97.70
SIRNet [33]	VGG16	81.82	93.39	96.24	99.49	75.37	88.76	91.90	97.42
TransGeo [56]	ViT	94.08	98.36	99.04	**99.77**	-	-	-	-
L2LTR [57]	ViT	91.99	97.68	98.65	99.75	83.14	93.84	95.51	98.40
Polar Transform Methods									
SAFA [10]	VGG16	89.84	96.93	98.14	99.64	81.03	92.80	94.84	98.17
DSM [13]	VGG16	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32
Shi et al. [58]	VGG16	92.69	97.78	98.60	99.61	82.70	92.50	94.42	97.65
LPN [1]	ResNet-50	93.78	98.50	99.03	99.72	82.87	92.26	94.09	97.77
LPN + USAM [51]	ResNet-50	91.22	-	-	99.67	82.02	-	-	98.18
Toker [14]	ResNet-34	92.56	97.55	98.33	99.57	83.28	93.57	95.42	98.22
SIRNet [33]	VGG16	93.74	98.02	98.85	99.76	86.02	94.45	**96.02**	98.33
L2LTR [57]	ViT	94.05	98.27	98.99	99.67	84.89	94.59	95.96	98.37
Ours	ResNet-50	**95.09**	**98.85**	**99.34**	**99.77**	**86.64**	**94.61**	95.94	**98.45**

98.45%. Compared with the LPN method after using polar transform, the results of R@1, R@5, R@10 and R@Top1% improve by 3.77%, 2.35%, 1.85% and 0.68% respectively. In addition, compared with the Transformer-based method (L2LTR), the proposed network model can improve 1.75% and 0.08% on the R@1 and AP, which proves the effectiveness of the proposed network model for the cross-view geo-localization task. (see the PERFORMANCE EVALUATION D 3) and 4), page 10)

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

[33] X. Lu, S. Luo, and Y. Zhu, "Its okay to be wrong: Cross-view geo-localization with step-adaptive iterative refinement," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-13, 2022.

[56] S. Zhu, M. Shah, and C. Chen, "Transgeo: Transformer is all you need for cross-view image geo-localization," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1162-1171.

[57] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," Advances in Neural Information Processing Systems, vol. 34, pp. 29009-29020, 2021.

TABLE III. Comparison with the state-of-the-art methods using the SUES-200 dataset. The input image size for comparison methods is 384×384. For our method, the image size is 256×256.

Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [50]	43.42	49.65	49.42	55.91	57.47	60.31	60.43	65.78
LPN [1]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
Vit [47]	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
Ours	**77.57**	**81.30**	**89.50**	**91.40**	**92.58**	**94.21**	**97.40**	**97.92**
Method	Drone→Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Baseline [47]	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LCM [50]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN [1]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
Vit [47]	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
Ours	**93.75**	**79.49**	**97.50**	**90.52**	**97.50**	**96.03**	**100.00**	**97.66**

2) The Experimental Results using the SUES-200 Dataset: The comparison results with the state-of-the-art methods using the SUES-200 dataset are given in Table III. The experimental results are mainly divided into three groups, namely the baseline-related methods, the experimental results of methods using contextual information and the experimental results of the Transformer-based method. From the Table III, the the baseline-related methods are given in the first and second rows, the third row shows the experimental results of methods using contextual information, the experimental results of the Transformer-based method are shown in the fourth row and the last row is the experimental result of the proposed network model. It can be seen from the experimental results that the proposed network model achieves the accuracy of R@1 are 77.57, 89.50, 92.58, 97.40 and AP are 81.30, 91.40, 94.21, 97.92 on the drone-view target localization task (Drone→Satellite) at different height, and it can achieve the accuracy of R@1 are 93.75, 97.50, 97.50, 100.00 and AP are 79.49, 90.52, 96.03, 97.66 on the drone navigation task (Satellite→Drone) at different height. Compared with LPN method, the R@1 and AP are improved 10.00%, 8.75%, 5.00%, 7.50% and 12.71%, 15.51%, 14.69%, 11.94% for the Drone→Satellite task at different heights, and the R@1 and AP are improved 15.99%, 18.92%, 12.20%, 15.93% and 14.07%, 15.44%, 10.41%, 13.39% for the Satellite→Drone task at different heights. It can be seen from the experimental results that the proposed network model is effective through introducing global information and local-guided-global information branches on the basis of using contextual information, and the performance of the model has been greatly improved. (see the PERFORMANCE EVALUATION D 2), pages 9-10)

Comment 12: For the Table V, it is better to explain the abbreviations in the caption.

***Response*:** This is a valuable suggestion. We have modified Table V and Table VI to explain the abbreviation in the caption.

Revision: (see the PERFORMANCE EVALUATION E 1) and 2), page 11)

TABLE V. Ablation study on the effect of the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

GIF	LIF	LGGIF	Drone→Satellite		Satellite→Drone	
			R@1	AP	R@1	AP
×	×	×	64.13	68.73	76.32	60.20
✓	×	×	71.78	75.68	80.74	68.69
×	✓	×	68.96	72.42	84.02	66.68
×	×	✓	65.96	70.32	79.32	63.41
✓	✓	×	82.72	84.95	89.44	79.19
✓	×	✓	72.85	76.52	83.17	69.86
×	✓	✓	81.56	83.58	89.73	80.17
✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

TABLE VI. Ablation study on the effect of the information fusion strategies (IFS) and attention mechanism (CBAM) in the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

IFS in GIF	IFS in LIF	IFS in LGGIF	CBAM	Drone→Satellite		Satellite→Drone	
				R@1	AP	R@1	AP
×	×	×	×	80.35	82.79	87.02	79.49
×	✓	✓	✓	83.59	85.88	89.59	83.15
✓	×	✓	✓	82.20	84.64	88.87	82.19
✓	✓	×	✓	83.24	85.63	90.16	82.10
✓	✓	✓	×	83.97	86.21	89.16	82.75
✓	✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**

Comment 13: For the ablation study on the effect of the various branches (section VI.E.1), what are the result of baseline and results of using each branch separately?

***Response*:** Thanks for your question! We have supplemented the experimental results of baseline and separate use of each branch in TABLE V, and conducted in-depth analysis of the results.

Revision: 1) Effect of the Various Branches

The main contribution of this paper is to design three branches, the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In order to verify the effectiveness of these three proposed branches in the network model, we designed several experiments to test each branch as shown in Table V. It can be seen from Table V that no matter which branch is excluded, the performance of the network

model has a certain decline. In addition, the performance of the model will be significantly improved after using any branch, which also proves the effectiveness of the proposed different branches. Intuitively, the performance of the model is greatly improved after combining the local information fusion branch, which also proves that the contextual information in the image has a significant impact on the cross-view geo-localization task. Thus, we fully consider the global information in the image and employ local information to assist the global features to mine the critical information in the image as much as possible. From the experimental results, it can be seen that these three introduced branches are effective and the discriminability of the final feature descriptors can be improved through utilizing the global features, improving the retrieval precision of the network model. The main contribution of this paper is to design three branches, the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch. In order to verify the effectiveness of these three proposed branches in the network model, we designed several experiments to test each branch as shown in Table V. It can be seen from Table V that no matter which branch is excluded, the performance of the network model has a certain decline. In addition, the performance of the model will be significantly improved after using any branch, which also proves the effectiveness of the proposed different branches. Intuitively, the performance of the model is greatly improved after combining the local information fusion branch, which also proves that the contextual information in the image has a significant impact on the cross-view geo-localization task. Thus, we fully consider the global information in the image and employ local information to assist the global features to mine the critical information in the image as much as possible. From the experimental results, it can be seen that these three introduced branches are effective and the discriminability of the final feature descriptors can be improved through utilizing the global features, improving the retrieval precision of the network model. (see the PERFORMANCE EVALUATION E 1), page 11)

TABLE V. Ablation study on the effect of the global information fusion (GIF) branch, the local information fusion (LIF) branch and the local-guided-global information fusion (LGGIF) branch.

GIF	LIF	LGGIF	Drone→Satellite		Satellite→Drone	
			R@1	AP	R@1	AP
×	×	×	64.13	68.73	76.32	60.20
✓	×	×	71.78	75.68	80.74	68.69
×	✓	×	68.96	72.42	84.02	66.68
×	×	✓	65.96	70.32	79.32	63.41
✓	✓	×	82.72	84.95	89.44	79.19
✓	×	✓	72.85	76.52	83.17	69.86
×	✓	✓	81.56	83.58	89.73	80.17
✓	✓	✓	**86.06**	**88.08**	**91.44**	**85.73**