Enhancing Cross-View Geo-Localization Through Global-Local Quadrant Interaction Network

Jin Xu , Junping Yin * , Juan Zhang , Tianyan Gao

Posted Date: 17 March 2025

doi: 10.20944/preprints202503.1203.v1

Keywords: geo-localization; Quadrant Insight; Integrated Global-Local Attention; cross-view

Article

Enhancing Cross-View Geo-Localization Through Global-Local Quadrant Interaction Network

Jin Xu [1,2,†], Junping Yin [1,2,3], Juan Zhang [2,4] and Tianyan Gao [2,3]

[1] Institute of Applied Physics and Computational Mathematics
[2] Shanghai Zhangjiang Institute of Mathematics
[3] Northeast Normal University
[4] Beihang University
* Correspondence: yinjp829829@126.com
† Current address: Affiliation.

Abstract: Cross-view geo-localization aims to match images of the same location captured from different perspectives, such as drone and satellite views. This task is inherently challenging due to significant visual discrepancies caused by viewpoint variations. Existing approaches often rely on global descriptors or limited directional cues, failing to effectively integrate diverse spatial information and global-local interactions. To address these limitations, we propose the Global-Local Quadrant Interaction Network (GLQINet), which enhances feature representation through two key components: the Quadrant Insight Module (QIM) and the Integrated Global-Local Attention Module (IGLAM). QIM partitions feature maps into directional quadrants, refining multi-scale spatial representations while preserving intra-class consistency. Meanwhile, IGLAM bridges global and local features by aggregating high-association feature stripes, reinforcing semantic coherence and spatial correlations. Extensive experiments on the University-1652 and SUES-200 benchmarks demonstrate that GLQINet significantly improves geo-localization accuracy, achieving state-of-the-art performance and effectively mitigating cross-view discrepancies.

Keywords: geo-localization; Quadrant Insight; Integrated Global-Local Attention; cross-view

1. Introduction

Cross-view object geo-localization [1–5] aims to determine the precise geographic location of an object in a reference image using its coordinates from a query image. This task is crucial in remote sensing applications, including autonomous navigation, 3D scene reconstruction [6], and precision delivery [7]. As a complementary technique to the Global Positioning System (GPS), image-based geo-localization enhances localization accuracy and provides a robust alternative in GPS-denied environments.

Despite its significance, cross-view geo-localization remains a challenging problem due to inherent visual discrepancies across images captured from different platforms, such as drones and satellites. Unlike traditional geo-localization methods [8–10], which operate on images from similar viewpoints, cross-view geo-localization must overcome extreme viewpoint differences, scale variations, and partial occlusions between ground-level and aerial perspectives. The key to success lies in extracting salient scene features that effectively bridge the domain gap, enabling robust cross-view matching and accurate localization.

In the early stages of cross-view geo-localization, traditional image processing techniques were commonly employed, such as gradient-based methods [11] and hand-crafted features [12], to obtain effectively matched image pairs. With the rapid development of deep learning, there has been significant progress in this field. A notable method [13] uses pre-trained convolutional neural networks (CNNs) as the backbone to extract image descriptors, followed by a classifier that aggregates correctly matched pairs of drone and satellite images within the feature space. As a result of the strong global

 doi:10.20944/preprints202503.1203.v1

modeling capabilities, Transformer has gained considerable attention in visual tasks [14,15], positioning Transformer-based architectures as the preferred choice [16–18] for geo-localization challenges. Building on this approach, several methodologies [19–21] explicitly utilize contextual information to enhance feature representation. Nevertheless, the critical aspect of addressing cross-view geo-localization lies in identifying relevant information between images and thoroughly comprehending both global and local contextual information. This task proves quite challenging for CNNs, as many CNN architectures primarily emphasize small, discriminative features rather than capturing broader contextual cues.

On the other hand, attention mechanisms [14,22,23], renowned for focusing on essential components while suppressing irrelevant ones within feature maps, are increasingly employed to help networks learn critical cues from images and improve the overall contextual information. Furthermore, as the self-attention mechanism has evolved in natural language processing, Vision Transformer—a self-attention-based architecture for visual tasks—has been incorporated into several studies related to cross-view geo-localization [16–18,24], yielding impressive results. However, these approaches typically concentrate on extracting fine-grained information from the final feature map, often overlooking low-level cues such as texture and edge details, which are crucial for a comprehensive understanding of the images.

To address the limitations of the previously mentioned methods in cross-view geo-localization, we propose a Global-Local Quadrant Interaction Network (GLQINet), as shown in Figure 1. GLQINet focuses on extracting fine-grained information from feature maps oriented in different directions, thereby enhancing the accuracy of cross-view geo-localization. Specifically, GLQINet is built upon a two-stream architecture and comprises two main components: the Quadrant Insight Module (QIM) and the Integrated Global-Local Attention Module (IGLAM). Instead of utilizing the feature maps produced by the backbone directly, our network employs the QIM to capture multiple local pattern stripes for a more comprehensive feature representation.QIM segments the feature maps from four directions into equal parts of contextual information stripes. This division, coupled with varying receptive fields, allows the model to effectively fuse information across different stripes, facilitating a better understanding of the correspondence between aerial and ground images. Additionally, we introduce IGLAM to adaptively enhance the discriminative regions and semantic representations while preserving global cues. IGLAM utilizes an attention-based mechanism, specifically a Cross Attention block, which significantly improves the model's capacity to capture local features of the query object while minimizing the representation of non-target features. Finally, drawing inspiration from previous work on dataset sampling [25], we construct batches that aid in extracting meaningful statistics to guide the training process and improve convergence. The primary contributions of this paper can be summarized as follows:

- We propose the Quadrant Insight Module (QIM), which is specifically designed to capture multi-scale spatial features and preserve intra-class consistency, adaptively emphasizing discriminative regions within the features and enhancing the overall representation of the object.
- We employ the Integrated Global-Local Attention Module (IGLAM), which bridges overall and partial information by aggregating high-association feature stripes, enhancing both broad contextual information and specific details while minimizing unnecessary background elements.
- Through extensive experiments conducted on public datasets, namely University-1652 and SUES-200, we prove that our GLQINet achieves superior performance in cross-view geo-localization compared to competing methods.

The rest of this paper is organized as follows. Section 2 introduces the related work. Section 3 elaborates our method. Section 4 presents the experimental results to show our method's superiority. Section 5 concludes this paper.

 doi:10.20944/preprints202503.1203.v1

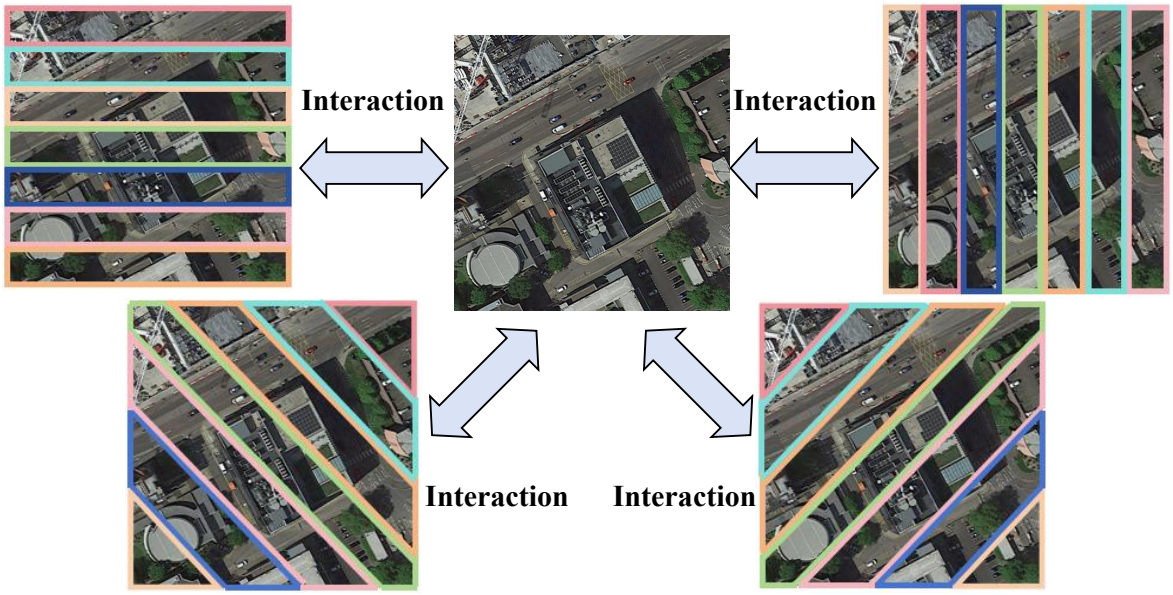


Figure 1. An illustration of the motivation behind our work. Our proposed GLQINet generates diverse patterns to encourage the network to learn informative feature representations by focusing on discriminative aspects of the input. In addition, the model employs an attention-based mechanism in an interactive manner to effectively learn both global and local features, enabling a comprehensive understanding of the geographic context across different views.

2. Related Work

2.1. Cross-View Geo-Localization

Cross-view geo-localization has become a significant research focus due to its broad applications in areas like urban planning and autonomous navigation. This section reviews the evolution of datasets and techniques for geo-localization across different viewpoints. Early studies [26,27] organized datasets into image pairs to address challenges in ground-to-aerial localization. Key datasets, such as CVUSA [28] and CVACT [29], use ground-view panoramic images as queries and satellite images as references. The Vigor dataset [30] offers a more practical testbed, bridging research and real-world applications. The University-1652 dataset [13] includes images from drones, satellites, and ground cameras to focus on university buildings. While the CVOGL dataset [31] addresses "Ground $\rightarrow$ Satellite" and "Drone $\rightarrow$ Satellite" localization, the SUES-200 dataset [32] provides aerial imagery at various altitudes for better context in drone-based geo-localization.

One of the key challenges in this task is extracting viewpoint-invariant features. Early methods transformed ground images to bird's-eye view (BEV) [12]. With the rise of deep learning, CNNs became popular for feature extraction [33], improving cross-view localization. LONN [29] incorporated orientation information for better accuracy, while polar transforms [34] helped mitigate viewpoint differences but introduced distortions. Recent work, such as LPN [19], uses feature partition strategies to improve model performance, while FSRA [16] and TransGeo [17] adopt Transformer-based approaches, reducing the need for geometric preprocessing and offering improved results. While significant progress has been made, issues such as background clutter, occlusion, and misalignment of buildings remain.

2.2. Based on Part Feature Methods

Part-based representations divide image features into smaller regions to capture fine-grained details and emphasize salient regions, which is particularly beneficial in retrieval tasks such as vehicle and person re-identification [35–37]. PCB [38] and MGN [39] enhance recognition accuracy by aggregating global and local features, allowing for more robust feature discrimination. AlignedReID [40] and DMLI [41] improve localization by aligning local features without additional supervision, ef-

 doi:10.20944/preprints202503.1203.v1

fectively reducing misalignment issues. Meanwhile, MSBA [1] integrates self-attention mechanisms to optimize feature extraction efficiency, while FSRA [16] and SGM [24] achieve pixel-level feature division, enhancing retrieval performance by capturing finer spatial relationships. These methods have proven effective in handling complex scenarios, particularly when addressing variations in pose, background clutter, and viewpoint shifts. In cross-view geo-localization, the integration of global context and local feature alignment plays a crucial role in bridging perspective differences, ensuring robust and reliable matching. Inspired by these advancements, we introduce a novel quadrant-based strategy that refines spatial representations and strengthens multi-scale interactions, significantly improving feature compactness and geo-localization accuracy.

2.3. Based on Transformer Methods

The Transformer model [22] has revolutionized natural language processing (NLP) and computer vision by effectively capturing long-range dependencies through self-attention. In geo-localization, attention mechanisms have been integrated into feature extraction pipelines, significantly improving performance in aerial-view-based tasks [16,17]. Notably, models such as TransGeo [17] and EgoTR [18] leverage Transformer-based attention to encode global dependencies, effectively mitigating cross-view visual ambiguities. Transformer models typically require extensive datasets and are computationally intensive. The limited scale of existing cross-view geo-localization datasets [13,28,29] poses a challenge for training these models effectively. Recent research has further extended Transformer-based approaches to pose-guided generation [42] and customizable virtual dressing [43], demonstrating their adaptability across diverse applications. Moreover, advancements in progressive conditional diffusion models [44,45] and long-term video generation [46] highlight the growing role of attention mechanisms in tackling complex, dynamic tasks beyond static image matching. To address this, recent research has explored structured feature decomposition and global-local attention mechanisms to better capture view-invariant representations for geo-localization. Inspired by these advancements, our work integrates direction-aware quadrant partitioning and multi-scale global-local interaction to enhance feature compactness and improve cross-view matching accuracy.

3. Proposed Method

In this section, we present our proposed GLQINet. The overall network architecture is illustrated in Figure 2. GLQINet primarily consists of two components. The first component is the Quadrant Insight Module (QIM), which divides the feature maps obtained from the drone and satellite view—after processing through the backbone network—into equal parts of contextual information stripes. The second component is the Integrated Global-Local Attention Module (IGLAM), where the feature vectors generated by the QIM are fed into the Merged Attention block, adaptively enhancing the discriminative regions and semantic representations while maintaining global cues. Finally, following a global average pooling operation, the feature vectors output by the IGLAM are directed to separate classifier blocks, all sharing the same structure. This allows for the acquisition of multiple feature representations, which are then used to calculate the loss, ultimately minimizing it to enhance the utilization of semantic information.

 doi:10.20944/preprints202503.1203.v1

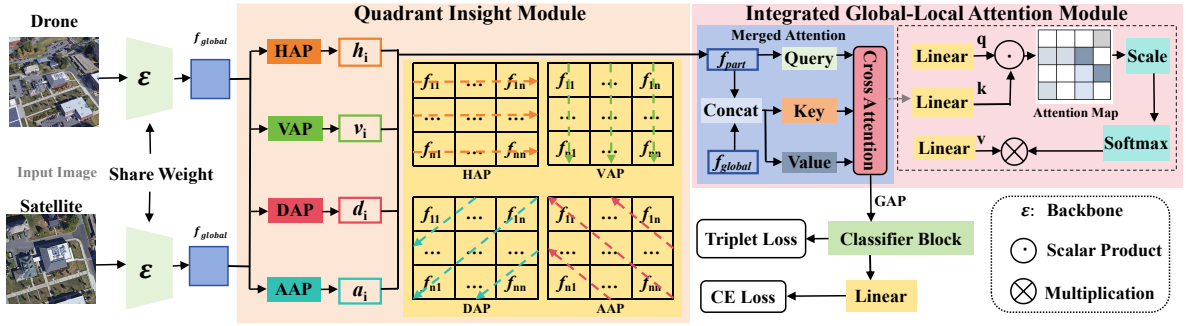


Figure 2. The proposed network's architecture comprises a dual-stream feature extraction backbone, the Quadrant Insight Module (QIM), and the Integrated Global-Local Attention Module (IGLAM). QIM leverages fine-grained details to generate four-directional local representations of the feature. IGLAM integrates both global and local embeddings, enabling simultaneous attention to different perspectives within various feature spaces and incorporating additional key features into comprehensive final representations.

3.1. Problem Formulation

In this section, we examine a geo-localization dataset, denoting the input image and its corresponding label as x_j and y_j, respectively. Here, j indicates the platform from which x_j is obtained, with $j \in \{1, 2\}$. Specifically, x_1 represents the satellite-view image, while x_2 corresponds to the drone-view image. The label y_j falls within the range $[1, R]$, where R is the total number of categories. Our GLQINet is structured with two branches: the drone-view and the satellite-view branch. Following the approach outlined in [13], we employ shared weights between the two branches, as the aerial views exhibit similar patterns. For feature extraction, we utilize the pre-trained ConvNeXt [47] as the backbone embedding. ConvNeXt is a standard CNN-based architecture, which achieves performance comparable to that of Vision Transformer networks [14,15], offering a balance of processing speed and accuracy while maintaining a simpler design.

The subsequent sections provide a detailed overview of the Quadrant Insight Module in Section 3.2. Following that, Section 3.3 delves into the Integrated Global-Local Attention Module. Finally, Section 3.4 presents the model's optimization module, including the formulation of the loss function.

3.2. Quadrant Insight Module (QIM)

Cross-view geo-localization requires robust feature representations to mitigate variations in perspective, scale, and structural alignment. Existing methods often struggle with feature fragmentation when handling aerial and ground-level images, making it difficult to establish meaningful correspondences. To address this challenge, we introduce the Quadrant Insight Module (QIM) , which decomposes feature maps into directional components, capturing multi-scale spatial relationships across different quadrants. This structured partitioning enhances the model's ability to distinguish fine-grained details , improving feature compactness and intra-class semantic consistency .

As illustrated in the middle part of Figure 2, the QIM operates on the global feature map output by the ConvNeXt backbone $\mathcal{F}_{backbone}$. Given an input image x_j, the extracted feature representation $f_{global} \in \mathbb{R}^{H \times W \times C}$, where H and W denote the height and width, and C is the number of channels, is defined as:

$$f_{global} = \mathcal{F}_{backbone}(x_j). \tag{1}$$

To effectively capture directional dependencies in different orientations, we introduce four specialized quadrant-based pooling layers : Horizontal Average Pooling (HAP), Vertical Average Pooling (VAP), Diagonal Average Pooling (DAP), and Anti-diagonal Average Pooling (AAP) . Each layer extracts structured features along a distinct axis, allowing the network to encode spatial hierarchies critical for accurate cross-view matching.

Horizontal Average Pooling (HAP) Layer: The HAP layer averages feature values along horizontal stripes , capturing spatial dependencies across rows. The output feature map is denoted as

 doi:10.20944/preprints202503.1203.v1

$h_i \in \mathbb{R}^{\frac{H}{t} \times W \times C}$, where t represents the number of partitioned stripes. For instance, when $t = n$, the first horizontal feature is computed as:

$$h_1 = \frac{1}{n} \sum_{i=1}^{n} f_{1i}. \tag{2}$$

Vertical Average Pooling (VAP) Layer: The VAP layer aggregates feature responses across vertical stripes , facilitating structural alignment between aerial and ground viewpoints. The feature map $v_i \in \mathbb{R}^{H \times \frac{W}{t} \times C}$ is transposed to $v_i^T \in \mathbb{R}^{\frac{H}{t} \times W \times C}$ to maintain dimensional consistency with h_i. When $t = n$, an example computation is:

$$v_2 = \frac{1}{n} \sum_{i=1}^{n} f_{i2}. \tag{3}$$

Diagonal Average Pooling (DAP) Layer: The DAP layer enhances feature continuity along diagonal axes , extracting oblique structural patterns. When $t = n$, the first diagonal feature is computed as:

$$d_1 = \frac{1}{n} \sum_{k=1}^{n} f_{p_k q_k}, \tag{4}$$

where (p_k, q_k) indexes diagonal elements.

Anti-diagonal Average Pooling (AAP) Layer: Similarly, the AAP layer captures information across anti-diagonal directions , reinforcing spatial consistency. Given $t = n$, the first anti-diagonal feature is:

$$a_1 = \frac{1}{n} \sum_{k=1}^{n} f_{r_k s_k}, \tag{5}$$

where (r_k, s_k) indexes anti-diagonal elements.

These directional pooling operations enhance spatial feature encoding by leveraging multiple receptive fields, allowing the model to effectively integrate multi-scale patterns . The aggregated outputs from QIM are formulated as:

$$\begin{cases} h_i = HAP(f_{global}), \\ v_i = VAP(f_{global}), \\ d_i = DAP(f_{global}), \\ a_i = AAP(f_{global}). \end{cases} \tag{6}$$

By capturing multi-directional interactions , QIM enables more compact and informative feature representations, bridging the gap between aerial and ground views . The extracted quadrant-based descriptors are subsequently refined within our Global-Local Attention Module , further enhancing geo-localization robustness .

3.3. Integrated Global-Local Attention Module (IGLAM)

Cross-view geo-localization requires effectively integrating both global and local visual cues to robustly match corresponding features across aerial and ground perspectives. However, existing attention mechanisms either focus on global feature extraction while neglecting local spatial structures or rely on local descriptors without incorporating high-level semantic understanding. To bridge this gap, we introduce an Integrated Global-Local Attention Module (IGLAM) that facilitates efficient cross-scale feature interaction, capturing both high-level contextual information and fine-grained local details. Compared to existing interaction modules [48,49], IGLAM achieves superior computational efficiency while preserving discriminative feature representation, as illustrated in Figure 3.

 doi:10.20944/preprints202503.1203.v1

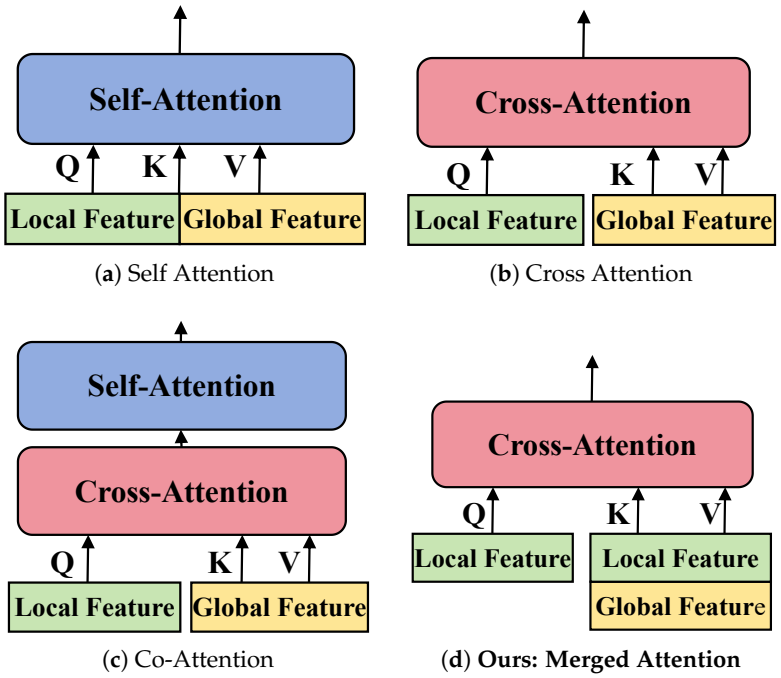


Figure 3. Comparison of our Integrated Global-Local Attention Module (IGLAM) with existing interaction mechanisms. (a) Self-Attention concatenates local and global features before passing them through a self-attention block. (b) Cross-Attention fuses features via a cross-attention layer. (c) Co-Attention applies a Cross-Attention layer followed by a Self-Attention block. (d) **Our Merged Attention** first concatenates global and local features, then processes them through a single Cross-Attention block, enabling effective cross-view interaction.

As shown in the upper right corner of Figure 2, IGLAM first concatenates the global feature representation f_{global} with the local features extracted from different directional stripes (horizontal, vertical, diagonal, and anti-diagonal) obtained from the Quadrant Insight Module (QIM), as defined in Eq. 7. The concatenation operation can be expressed as:

$$f_{cat} = concatenate(f_{part}, f_{global}), \tag{7}$$

where f_{part} represents one of the local feature representations (h_i, v_i, d_i, a_i).

Next, we process f_{cat} through a single Cross-Attention block. Specifically, as shown in the dashed rectangular box of the right end of Figure 2, three linear layers are applied to map the concatenated feature into the query, key, and value representations:

$$Q = W^Q f_{part}, \quad K = W^K f_{cat}, \quad V = W^V f_{cat}, \tag{8}$$

where W^Q, W^K, and W^V are learnable transformation matrices. The attention operation is then formulated as:

$$Attention(Q, K, V) = softmax(QK^T)V. \tag{9}$$

This mechanism ensures that local feature representations retrieve relevant contextual information from the global embeddings, reinforcing discriminative features across different views.

After the attention mechanism is applied, the output f^j_{local} is further processed using a Global Average Pooling (GAP) operation:

$$g_j = GAP(f_{local}), \tag{10}$$

where g_j represents the final feature vector for the input image x_j. This operation condenses spatial features into a compact descriptor while preserving global semantics.

To demonstrate the efficiency of IGLAM, we compare it against three widely used interaction modules in Table 1. As shown in Figure 3, (a) Self-Attention treats local and global features as a single

 doi:10.20944/preprints202503.1203.v1

entity, potentially leading to redundant computations. (b) Cross-Attention fuses them in a one-way manner, limiting bidirectional interaction. (c) Co-Attention employs sequential fusion, increasing computational overhead. In contrast, our Merged Attention method in (d) allows direct bidirectional interaction, significantly reducing computational complexity while preserving robust feature fusion. Through extensive evaluation, we find that IGLAM not only enhances Rank-k performance across multiple benchmarks but also significantly improves computational efficiency. This balance between accuracy and efficiency makes it particularly well-suited for large-scale cross-view geo-localization tasks.

Table 1. Comparison of different attention mechanisms for cross-view tasks on University-1652.

Method	Drone→ Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
Self Attention [22]	91.48	92.39	93.72	91.11
Cross Attention [22]	91.02	92.42	94.58	91.04
Co-Attention [48]	91.12	92.54	94.29	90.30
Merged Attention (Ours) [48]	**91.66**	**92.94**	**94.58**	**91.11**

3.4. Loss Function

Individual feature vectors are generated after feature extraction and subsequently fed into the classifier block. This block comprises a linear layer, followed by a batch normalization layer, a ReLU activation layer and a dropout layer. During the training phase, each feature vector g_j is processed through a classifier block and then a linear layer, transforming it into a category vector that is used to calculate the cross-entropy loss. In the testing phase, only the classifier block is utilized to process g_j, resulting in a 16×512 feature representation.

The cross-entropy loss optimizes the network's parameters and is defined as follows:

$$L_{CE} = -\sum_{s=1}^{R} p(x_s) \log q(x_s) \tag{11}$$

where $p(x_s)$ represents the ground truth probability, and $q(x_s)$ denotes the estimated probability. If x_s corresponds to the feature of the target (where the label value is 1), then $p(x_s) = 1$. Otherwise, $p(x_s) = 0$.

To minimize the distance between feature vectors of the same category across different views, we adopt the triplet loss, leveraging the conventional Euclidean distance as referenced in previous works [16,50,51]. The triplet loss is formulated as:

$$L_{\text{triplet}} = \max(||F_a - F_b||_2 - ||F_a - F_c||_2 + M, 0), \tag{12}$$

where $||\cdot||_2$ denotes the L_2-norm, F_a is the feature vector of the anchor image a (which can be either a satellite or drone-view image), F_b is the feature vector of an image from the same category as a, and F_c is the feature vector of an image from a different category. M is the triplet margin and we set to 0.3.

The total loss integrates both cross-entropy and triplet losses across multiple feature representations, computed as:

$$L_{\text{total}} = \frac{1}{3} \sum_{k=1}^{3} \sum_{j=1}^{2} (L_{\text{CE}}^{jk} + L_{\text{triplet}}^{jk}), \tag{13}$$

where k indexes the feature representations, and L_{CE}^{jk} and L_{triplet}^{jk} denote the cross-entropy and triplet losses for the k^{th} feature of x_j, respectively.

In summary, we employ cross-entropy loss for classification and triplet loss to enhance feature consistency across different domains.

 doi:10.20944/preprints202503.1203.v1

4. Experiment and Analysis

To validate the proposed GLQINet's superiority, it is compared with multiple state-of-the-art cross-view geo-location approaches on two large-scale datasets, namely, University-1652 and SUES-200.

4.1. Datasets

University-1652. The University-1652 dataset serves as a benchmark for drone-based geolocalization, featuring multi-view and multi-source imagery, which includes ground-view, drone-view and satellite-view images collected from 1,652 buildings across 72 universities. The dataset is divided into a training set comprising images from 701 buildings at 33 universities and a testing set consisting of images from the remaining 39 universities. Notably, the training set contains an average of 71.64 images per location, in contrast to existing datasets, which typically include only two images per location.

SUES-200. SUES-200 is a cross-view matching dataset distinguished by its varied sources, multiple scenes and panoramic perspectives, which consists of images captured from drone and satellite viewpoints. What differentiates SUES-200 from earlier datasets is its incorporation of tilted images from the drone perspective, complete with labeled flight heights of 150 meters, 200 meters, 250 meters and 300 meters, making it more representative of real-world conditions. Additionally, this dataset encompasses a wide array of scene types, extending beyond campus buildings to include parks, schools, lakes and other public structures.

4.2. Evaluation Metrics

In cross-view geo-localization, Recall@k (R@K) and Average Precision (AP) serve as standard evaluation metrics. R@K quantifies how likely a correct match will appear within the top k rankings. A higher R@K signifies superior network performance, which is defined as follows:

$$\text{R@K} = \begin{cases} 1, & if\ order_{true} < K+1, \\ 0, & otherwise. \end{cases} \tag{14}$$

AP measures the area under the precision-recall curve, accounting for all true matches, which is formulated as follows:

$$\text{AP} = \frac{1}{N} \sum_{s=1}^{N} \frac{p_{s-1} + p_s}{2}, \tag{15}$$

where $p_0 = 1$ and

$$p_s = \frac{T_s + 1}{T_s + F_s}. \tag{16}$$

Here, N represents the number of true matches for a query, while T_s and F_s denote the counts of true and false matches before the $(i+1)$-th true match.

4.3. Implementation Details

We select ConvNeXt-Tiny as the baseline model and utilize the pre-trained weights on ImageNet to extract visual features. To accommodate the requirements of the pre-trained model, we resize the input images to a fixed size of 256×256 pixels for both training and testing. The parameter n is set to be 8.

For parameter initialization, we employ the Kaiming initialization method [60] specifically for the classifier block. For the optimization process, we use stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005, operating on a batch size of 32. The learning rate is set to be 0.003 for the backbone parameters and 0.01 for the remaining layers. We train the model for a total of 200 epochs, applying a reduction of the learning rate by a factor of 0.1 after the 80th and 120th epochs. To measure similarity between query and gallery images, we utilize the cosine distance of the extracted features.

 doi:10.20944/preprints202503.1203.v1

4.4. Comparison with State-of-the-Art Methods

The comparison results presented in Table 2 and Table 3, demonstrate that the proposed GLQINet is superior to many state-of-the-art cross-view geo-localization methods.

Table 2. Comparison with the state-of-the-art results reported on University-1652, where the **BOLD** indicate the best results.

Method	University-1652			
	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
U-baseline [13]	58.49	63.31	71.18	58.74
DWDR [52]	69.77	73.73	81.46	70.45
MuSe-Net [53]	74.48	77.83	88.02	75.10
LPN [19]	75.93	79.14	86.45	74.49
SAIG [54]	78.85	81.62	86.45	78.48
LDRVSD [55]	78.66	81.55	89.30	79.17
PCL [56]	79.47	83.63	87.69	78.51
FSRA [16]	82.25	84.82	87.87	81.53
SGM [24]	82.14	84.72	88.16	81.80
PAAN [57]	84.51	86.78	91.01	82.28
MSBA [1]	86.61	88.55	92.15	84.45
MBF [58]	89.05	90.61	93.15	88.17
MCCG [59]	89.64	91.32	94.30	89.39
Ours	**91.66**	**92.94**	**94.58**	**91.11**

Table 3. Comparison with the state-of-the-art results reported on SUES-200, where the **BOLD** indicate the best results.

Method	Drone → Satellite							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LCM [3]	43.42	49.65	49.42	55.91	57.47	60.31	60.43	65.78
Vit [32]	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67
LPN [19]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53
SwinV2-T [61]	66.40	71.64	77.63	81.91	84.62	87.73	90.01	92.27
FSRA [16]	68.25	73.45	83.00	85.99	90.68	92.27	91.95	93.46
Ours	**82.07**	**85.55**	**91.50**	**93.33**	**96.72**	**97.49**	**96.82**	**97.42**
Method	Satellite → Drone							
	150m		200m		250m		300m	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LCM [3]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
Vit [32]	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10
LPN [19]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72
SwinV2-T [61]	82.51	71.18	90.03	82.20	93.23	92.11	97.52	92.09
FSRA [16]	83.75	76.67	90.00	85.34	93.75	90.17	95.00	92.03
Ours	**95.00**	**85.44**	**98.75**	**95.26**	**98.75**	**97.14**	**98.75**	**97.98**

4.4.1. Comparisons on University-1652

To evaluate the performance of GLQINet against state-of-the-art methods on the University-1652 dataset, we compare it with both CNN-based and Transformer-based techniques. The CNN-based methods include MCCG [59], MBF [58], PANN [57], MSBA [1] and LPN [19], while the Transformer-based methods consist of SAIG [54], SGM [24] and FSRA [16]. Additionally, we reference the baseline

 doi:10.20944/preprints202503.1203.v1

model provided for the University-1652 dataset [13] (U-baseline). The quantitative results are presented in Table 2, where the **BOLD** indicate the superior performance. The University-1652 dataset comprises two primary tasks: Drone-to-Satellite target localization and Satellite-to-Drone navigation. In the Drone $\rightarrow$ Satellite task, GLQINet achieves a R@1 of 91.66% and an AP of 92.94%. For the Satellite $\rightarrow$ Drone navigation task, it attains a R@1 of 94.58% and an AP of 91.11%. These results underscore GLQINet's superior performance compared to existing state-of-the-art models. Notably, GLQINet outperforms MCCG [59] by 2.02% R@1 in the Drone $\rightarrow$ Satellite task, demonstrating its effectiveness and advancements in cross-domain tasks within drone applications. This comparison highlights GLQINet's capability to deliver enhanced results across both localization and navigation tasks, solidifying its position as an advanced method in this domain.

4.4.2. Comparisons on SUES-200

To further assess the performance of GLQINet, we present the results on the SUES-200 dataset [32] in Table 3. In the drone-view target localization task (Drone $\rightarrow$ Satellite), GLQINet achieves R@1 scores of 82.07%, 91.50%, 96.72% and 96.82%, along with AP values of 85.55%, 93.33%, 97.49% and 97.42% across four different heights. These significant improvements substantially surpass the performance of the LPN model [19], which reports R@1 ranging from 61.58% to 81.47% at the same altitudinal intervals. In the drone navigation task (Satellite $\rightarrow$ Drone), GLQINet records R@1 scores of 95.00%, 98.75%, 98.75% and 98.75%, and achieves AP values of 85.44%, 95.26%, 97.14% and 97.98% at four heights. This demonstrates that GLQINet exhibits excellent and consistent performance across varying flight altitudes. Moreover, when compared to other Transformer-based models, our proposed approach outperforms SwinV2-T and FSRA by nearly 7% and 5%, respectively.

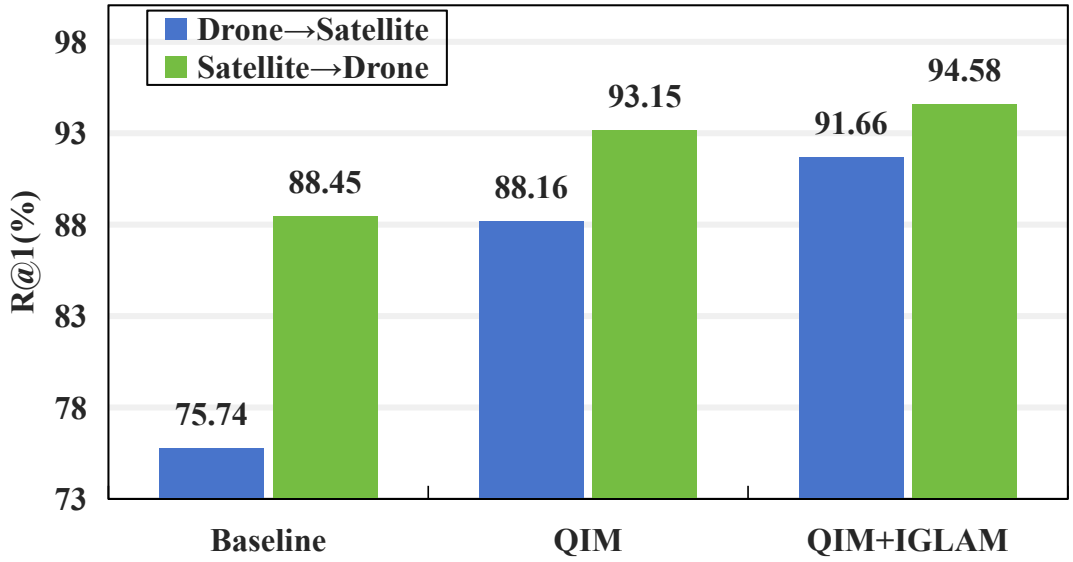
4.5. Ablation Studies and Analysis

In this section, the proposed GLQINet is comprehensively analyzed from four aspects to investigate the logic behind its superiority. Specifically, we conduct ablation experiments to evaluate (1) the role of our proposed QIM and IGLAM, (2) the impact of the number of stripes, (3) analysis of the attention mechanism in IGLAM and (4) the effect of the utilized baseline network.

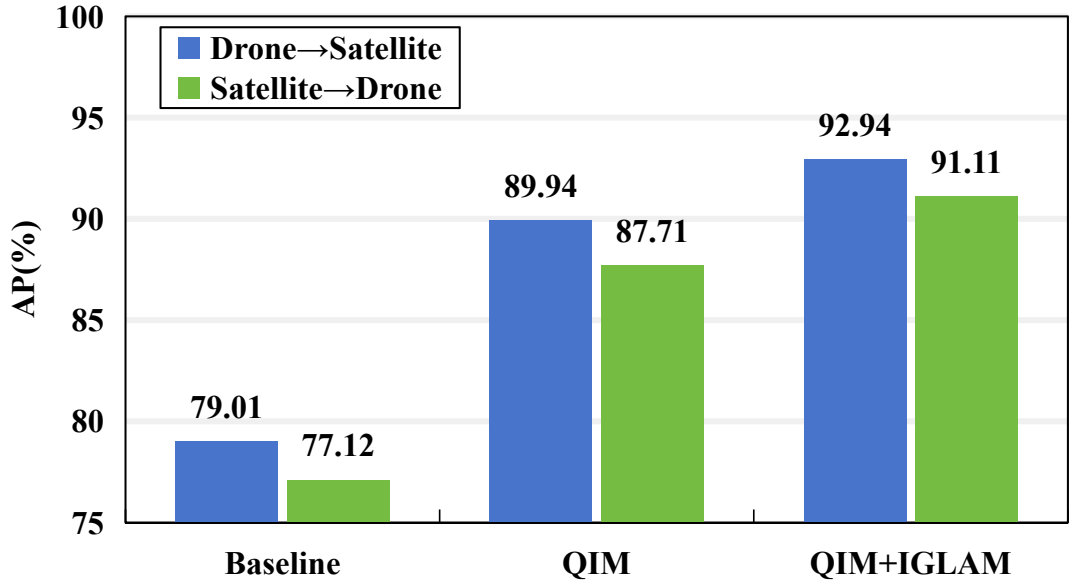
(1) Role of QIM and IGLAM. First, we evaluate the effectiveness of the proposed components, namely QIM and IGLAM, as shown in Figure 4, on both the University-1652 and SUES-200 datasets. The results reveal that the simultaneous introduction of QIM and IGLAM leads to significant performance improvements compared to the baseline. Specifically, for the Drone $\rightarrow$ Satellite task, the complete model enhances R@1 from 76.74% to 91.66% (an increase of 14.92%), while for the Satellite $\rightarrow$ Drone task, it improves R@1 from 88.45% to 94.58% (an increase of 6.13%). From these results, we can draw the following conclusions: 1) The removal of all components results in notably poorer matching performance across both datasets. 2) Utilizing any single proposed module independently results in significant performance enhancements, with QIM contributing the most, highlighting the effectiveness of leveraging diverse features to improve model performance. 3) When all components are used together, the model achieves optimal performance across all metrics, which shows that our proposed modules work collaboratively to enhance the overall effectiveness of the model. All these above findings underscore the efficacy of the proposed QIM and IGLAM.

(2) Impact of the number of stripes. In the following experiments, we only vary the parameter t, which denotes the number of stripes along each direction. This parameter is a crucial factor in the stripe partitioning strategy, and by default, we set $t = 4$. To assess the impact of t on the accuracy of R@1 and AP, we performed several experiments using different t values ranging from 1 to 8. The results are illustrated in Figure 5. For the drone navigation task, both R@1 and AP accuracy improve as t increases, reaching their peak values of 91.66% and 91.11% when $t = 4$, after which there is a slight decline. On the other hand, in the drone-view target localization task, R@1 and AP accuracy exhibit an upward trend when t is less than 4, followed by a gradual decrease as t continues to increase. This indicates that the proposed method demonstrates resilience to variations in the number of stripes.

 doi:10.20944/preprints202503.1203.v1



(a) R@1

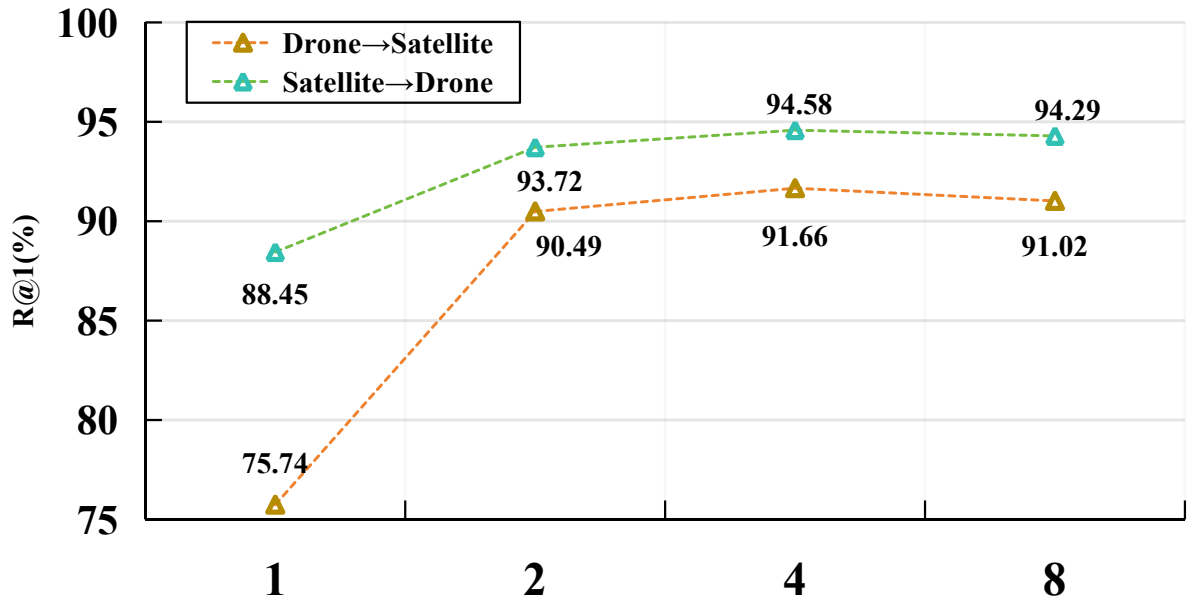


(b) AP

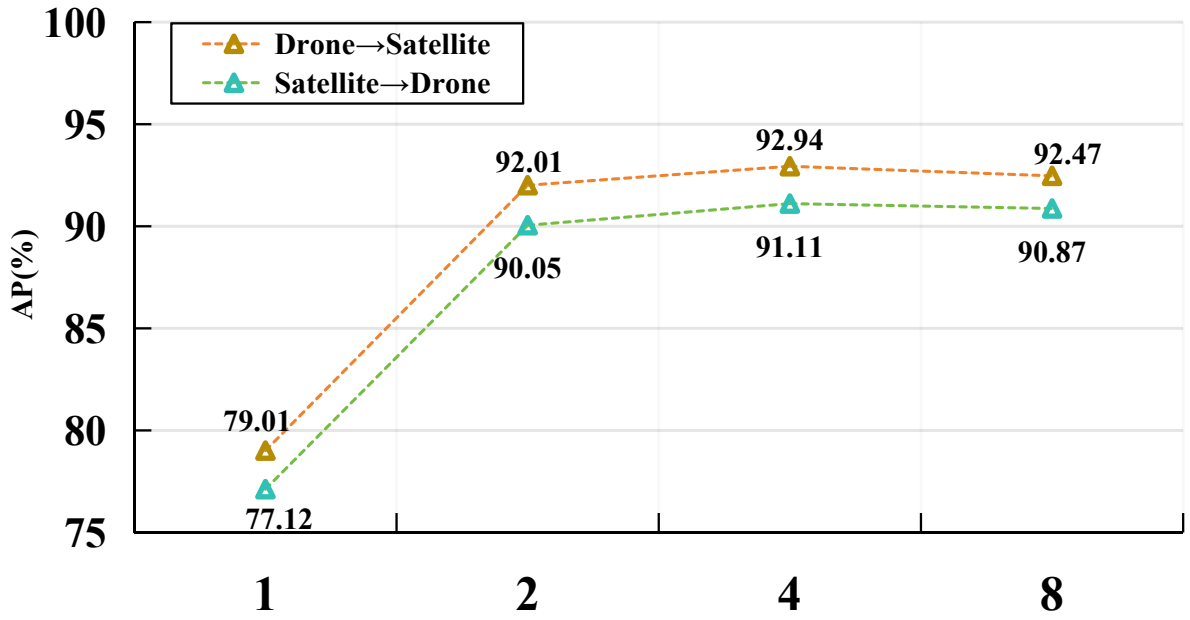
Figure 4. Ablation study on comparison with different components.

(3) Analysis of the attention mechanism in IGLAM. Among the various attention mechanisms discussed in the literature, we selected Merged Attention for its efficiency in enhancing feature representation and incurring minimal additional computational costs. We conducted several experiments comparing different attention mechanisms, which contain Self Attention, Cross Attention, and Co-Attention, in order to assess the effectiveness of the attention mechanism utilized in the IGLAM on the final results. Notably, all these attention mechanisms demonstrated inferior performance compared to Merged Attention within the IGLAM framework. The proposed model was trained using different attention-based IGLAM configurations under the same training conditions. As detailed in Table 1, the performance metrics for the Merged Attention-based IGLAM surpassed those of the Co-Attention-based IGLAM, with R@1 and AP improving by 0.29% (0.54%) and 0.81% (0.40%) for the Satellite → Drone (and Drone → Satellite) tasks, respectively. Merged Attention consistently achieved superior performance compared to other attention mechanisms used in the IGLAM. We hypothesize that the attention weights generated by Merged Attention effectively emphasize crucial components of the overall feature maps, significantly influencing the final outcomes.

 doi:10.20944/preprints202503.1203.v1



(**a**) R@1



(**b**) AP

Figure 5. Ablation study on comparison with different values of t.

(4) Effect of the utilized baseline network. The methods compared in Subsection 4.4 employ various backbones for feature extraction. As indicated in Table 4, to enable a fair comparison and assess the generalization capability of the proposed QIM and IGLAM, we report the performance of our model using different backbone architectures. Specifically, we utilize ConvNeXt-Tiny [47], ConvNeXt-Small [47], ResNet-50 [62], and ResNet-101 [62] as backbones. The results reveal that employing GLQINet across these four backbones leads to a significant improvement in average Recall@1, with a 17.16% increase for the Drone → Satellite task and a 7.13% increase for the Satellite → Drone task. Notably, in the Drone → Satellite task, our method enhances the Recall@1 from 59.56% to 83.01% (a 23.45% increase) for ResNet-101, and from 82.43% to 92.40% (a 9.97% increase) for ConvNeXt-Small. These findings underscore the generalization capability of GLQINet across different backbones. Furthermore, the ConvNeXt-Tiny network outperforms the ResNet-50 network by at least 5.99% in both Recall@1 and AP for the two cross-view tasks. In our comparisons of ConvNeXt-Small with other backbones, we observe that a deeper ConvNeXt model shows a slight increase in performance,

 doi:10.20944/preprints202503.1203.v1

likely due to its more complex architecture. Consequently, to achieve a balance between accuracy and complexity, we select ConvNeXt-Tiny as the baseline model for the remainder of this paper.

Table 4. Comparison of ResNet and ConvNeXt networks for cross-view tasks on University-1652.

Method	Backbone	Drone → Satellite		Satellite → Drone	
		R@1	AP	R@1	AP
Baseline	ResNet-50 [62]	58.50	63.28	79.17	57.81
GLQINet	ResNet-50 [62]	**78.78**	**81.86**	**88.59**	**78.49**
Baseline	ResNet-101 [62]	59.56	64.00	81.17	60.56
GLQINet	ResNet-101 [62]	**83.01**	**85.64**	**89.87**	**82.93**
Baseline	ConvNeXt-Tiny [47]	76.74	79.01	88.45	77.12
GLQINet	ConvNeXt-Tiny [47]	**91.66**	**92.94**	**94.58**	**91.11**
Baseline	ConvNeXt-Small [47]	82.43	85.03	91.30	82.89
GLQINet	ConvNeXt-Small [47]	**92.40**	**93.65**	**95.58**	**91.90**

4.6. Visualization

As a key qualitative assessment, we present the visualization of the retrieval results of our GLQINet on the University-1652 dataset. Figure 6 illustrates the visualization outcomes for various tasks in the performance from R@1 to R@5, where blue boxes indicate correct matches and red boxes represent incorrect matches. It is evident that our model demonstrates remarkable matching results, even with randomly selected images that differ in domain and viewpoint. In the Drone → Satellite task, our model successfully distinguishes between negative samples that closely resemble positive ones. This visual evaluation confirms the effectiveness of GLQINet in discerning and aligning complex imagery, highlighting its potential as a valuable tool for advanced geo-localization applications.

To better recognize the model's limitations and shortcomings, we analyze the scenarios of the error cases carefully. As shown in Figure 6, GLQINet clearly outperforms the baseline in terms of discriminative ability. However, failures occur when there is a high degree of similarity among drone-view images. The imprecise texture mapping of front-layer features leads to difficulties in providing the model with more accurate retrievals.

 doi:10.20944/preprints202503.1203.v1

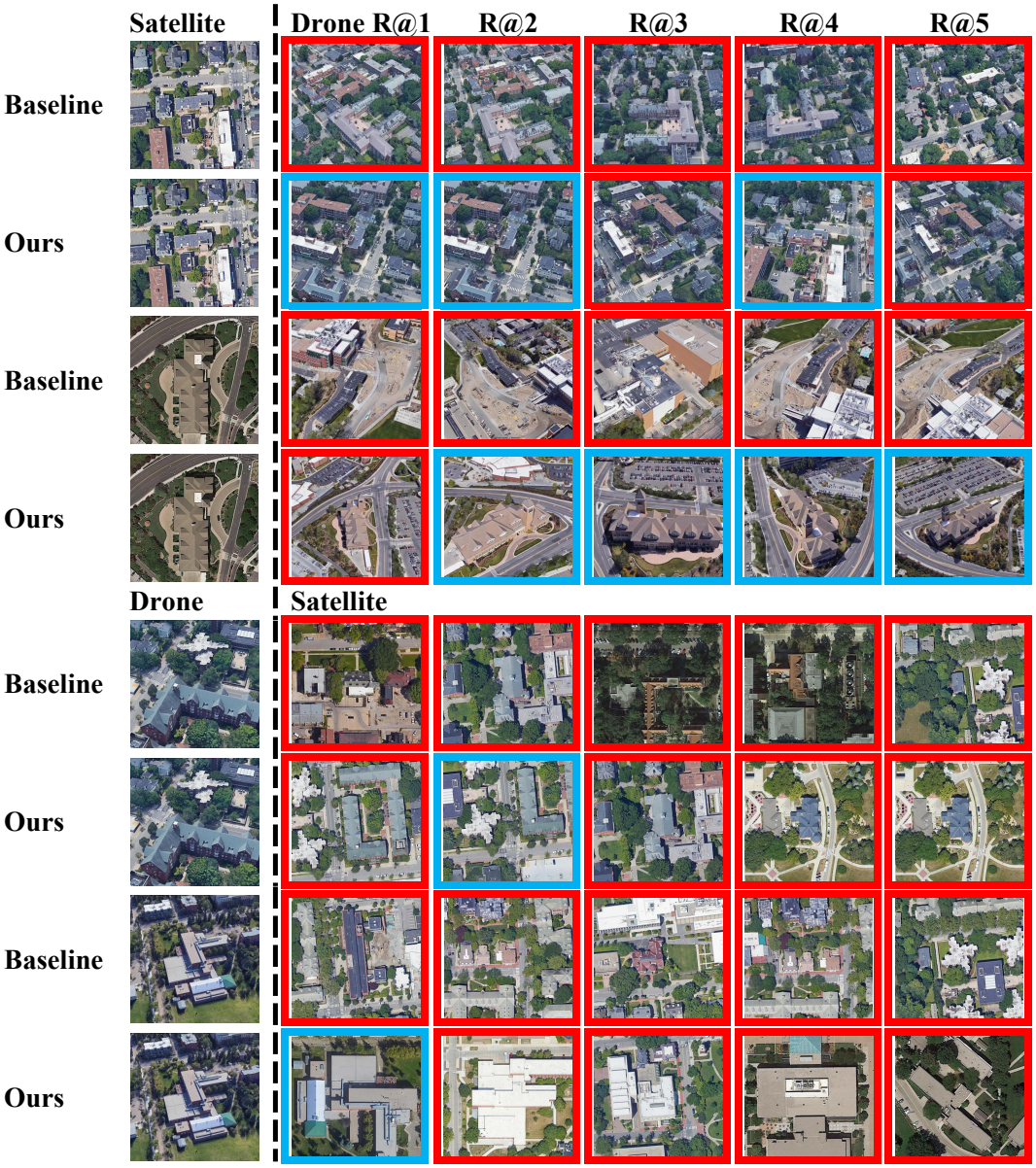


Figure 6. The error cases visualization of our method and baseline, with blue boxes denoting correct matching and red boxes signifying false matching.

5. Conclusions

With the swift advancement of UAV technology, there is an increasing demand for autonomous control of UAVs, particularly in navigating without GPS signals. Image-based localization has emerged as a crucial solution to this challenge. This paper addresses the problem of cross-view image matching for geo-localization. We identified the limitations of current CNN and Transformer-based approaches and introduced an innovative method that ensured a comprehensive feature representation, incorporating a Quadrant Insight Module (QIM) and an Integrated Global-Local Attention Module (IGLAM). The effectiveness of our proposed method was validated on two benchmark datasets, University-1652 and SUES-200, with experimental results showcasing its superior performance compared to existing state-of-the-art methods. Significantly, to utilize spatial correlations across the image, our GLQINet effectively captured diverse local pattern stripes and fostered a mutual reinforcement between global and local semantics. This paves the way for promising applications in future Vision Transformer-related research. In the next stages, we will focus on investigating ways to further enhance matching accuracy, especially with real UAV image data, and work towards optimizing the Transformer model for practical UAV applications.

 doi:10.20944/preprints202503.1203.v1

Author Contributions: Conceptualization, Jin Xu; Data curation, Jin Xu; Formal analysis, Jin Xu; Investigation, Jin Xu; Methodology, Jin Xu; Resources, Jin Xu; Software, Jin Xu and Tianyan Gao; Validation, Jin Xu; Writing – original draft, Jin Xu; Writing – review & editing, Junping Yin and Juan Zhang. All authors have read and agreed to the published version of manuscript.

Funding: This research is partially funded by the Major Program of National Natural Science Foundation of China (No.12292980,12292984,NSFC12031016 and NSFC12426529), National Key R&D Program of China (2023YFA1009000, 2023YFA1009004, 2020YFA0712203 and 2020YFA0712201), Beijing Natural Science Foundation (BNSF-Z210003), and the Department of Science, Technology and Information of the Ministry of Education (No. 8091B042240).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets source are listed in the paper.

Acknowledgments: The authors would like to thank Dr. Shunan Mao for his technical assistance.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GLQINet	Global-Local Quadrant Interaction Network
QIM	Quadrant Insight Module
IGLAM	Integrated Global-Local Attention Module
GPS	Global Positioning System
CNNs	Convolutional Neural Networks
BEV	Bird's-Eye View
NLP	Natural Language Processing
HAP	Horizontal Average Pooling
VAP	Vertical Average Pooling
DAP	Diagonal Average Pooling
AAP	Anti-diagonal Average Pooling
R@K	Recall@K
AP	Average Precision
SGD	Stochastic Gradient Descent

References

1. Zhuang, J.; Dai, M.; Chen, X.; Zheng, E. A faster and more effective cross-view matching method of UAV and satellite images for UAV geolocalization. *Remote Sensing* **2021**, *13*, 3979.
2. Cui, Z.; Zhou, P.; Wang, X.; Zhang, Z.; Li, Y.; Li, H.; Zhang, Y. A novel geo-localization method for UAV and satellite images using cross-view consistent attention. *Remote Sensing* **2023**, *15*, 4667.
3. Ding, L.; Zhou, J.; Meng, L.; Long, Z. A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization. *Remote Sensing* **2020**, *13*, 47.
4. Hou, Q.; Lu, J.; Guo, H.; Liu, X.; Gong, Z.; Zhu, K.; Ping, Y. Feature relation guided cross-view image based geo-localization. *Remote Sensing* **2023**, *15*, 5029.
5. Yan, Y.; Wang, M.; Su, N.; Hou, W.; Zhao, C.; Wang, W. IML-Net: A Framework for Cross-View Geo-Localization with Multi-Domain Remote Sensing Data. *Remote Sensing* **2024**, *16*, 1249.
6. Middelberg, S.; Sattler, T.; Untzelmann, O.; Kobbelt, L. Scalable 6-dof localization on mobile devices. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13. Springer, 2014, pp. 268–283.
7. An, Z.; Wang, X.; Li, B.; Xiang, Z.; Zhang, B. Robust visual tracking for UAVs with dynamic feature weight selection. *Applied Intelligence* **2023**, *53*, 3836–3849.
8. Krylov, V.A.; Kenny, E.; Dahyot, R. Automatic discovery and geotagging of objects from street view imagery. *Remote Sensing* **2018**, *10*, 661.

 doi:10.20944/preprints202503.1203.v1

9. Nassar, A.S.; Lefèvre, S.; Wegner, J.D. Simultaneous multi-view instance detection with learned geometric soft-constraints. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6559–6568.
10. Chaabane, M.; Gueguen, L.; Trabelsi, A.; Beveridge, R.; O'Hara, S. End-to-end learning improves static object geo-localization from video. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2063–2072.
11. Lin, T.Y.; Belongie, S.; Hays, J. Cross-view image geolocalization. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 891–898.
12. Castaldo, F.; Zamir, A.; Angst, R.; Palmieri, F.; Savarese, S. Semantic cross-view matching. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 9–17.
13. Zheng, Z.; Wei, Y.; Yang, Y. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In Proceedings of the Proceedings of the 28th ACM international conference on Multimedia, 2020, pp. 1395–1403.
14. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012–10022.
15. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
16. Dai, M.; Hu, J.; Zhuang, J.; Zheng, E. A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 4376–4389.
17. Zhu, S.; Shah, M.; Chen, C. Transgeo: Transformer is all you need for cross-view image geo-localization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 1162–1171.
18. Yang, H.; Lu, X.; Zhu, Y. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems* **2021**, *34*, 29009–29020.
19. Wang, T.; Zheng, Z.; Yan, C.; Zhang, J.; Sun, Y.; Zheng, B.; Yang, Y. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 867–879.
20. Lin, J.; Zheng, Z.; Zhong, Z.; Luo, Z.; Li, S.; Yang, Y.; Sebe, N. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing* **2022**, *31*, 3780–3792.
21. Li, H.; Chen, Q.; Yang, Z.; Yin, J. Drone Satellite Matching based on Multi-scale Local Pattern Network. In Proceedings of the Proceedings of the 2023 Workshop on UAVs in Multimedia: Capturing the World from a New Perspective, 2023, pp. 51–55.
22. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
23. Peng, J.; Wang, H.; Xu, F.; Fu, X. Cross domain knowledge learning with dual-branch adversarial network for vehicle re-identification. *Neurocomputing* **2020**, *401*, 133–144.
24. Zhuang, J.; Chen, X.; Dai, M.; Lan, W.; Cai, Y.; Zheng, E. A semantic guidance and transformer-based matching method for UAVs and satellite images for UAV geo-localization. *Ieee Access* **2022**, *10*, 34277–34287.
25. Kuma, R.; Weill, E.; Aghdasi, F.; Sriram, P. Vehicle re-identification: an efficient baseline using triplet embedding. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019, pp. 1–9.
26. Tian, Y.; Chen, C.; Shah, M. Cross-view image matching for geo-localization in urban environments. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3608–3616.
27. Lin, T.Y.; Cui, Y.; Belongie, S.; Hays, J. Learning deep representations for ground-to-aerial geolocalization. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5007–5015.
28. Workman, S.; Souvenir, R.; Jacobs, N. Wide-area image geolocalization with aerial reference imagery. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3961–3969.
29. Liu, L.; Li, H. Lending orientation to neural networks for cross-view geo-localization. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5624–5633.
30. Zhu, S.; Yang, T.; Chen, C. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3640–3649.

 doi:10.20944/preprints202503.1203.v1

31. Sun, Y.; Ye, Y.; Kang, J.; Fernandez-Beltran, R.; Feng, S.; Li, X.; Luo, C.; Zhang, P.; Plaza, A. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
32. Zhu, R.; Yin, L.; Yang, M.; Wu, F.; Yang, Y.; Hu, W. 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**, *33*, 4825–4839.
33. Workman, S.; Jacobs, N. On the location dependence of convolutional neural network features. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 70–78.
34. Shi, Y.; Liu, L.; Yu, X.; Li, H. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems* **2019**, *32*.
35. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In Proceedings of the Proceedings of the 31th ACM International Conference on Multimedia, 2023.
36. Shen, F.; Du, X.; Zhang, L.; Tang, J. Triplet Contrastive Learning for Unsupervised Vehicle Re-identification. *arXiv preprint arXiv:2301.09498* **2023**.
37. Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Transactions on Image Processing* **2023**.
38. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 480–496.
39. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the Proceedings of the 26th ACM international conference on Multimedia, 2018, pp. 274–282.
40. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184* **2017**.
41. Luo, H.; Jiang, W.; Zhang, X.; Fan, X.; Qian, J.; Zhang, C. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition* **2019**, *94*, 53–61.
42. Shen, F.; Tang, J. IMAGPose: A Unified Conditional Framework for Pose-Guided Person Generation. In Proceedings of the The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
43. Shen, F.; Jiang, X.; He, X.; Ye, H.; Wang, C.; Du, X.; Li, Z.; Tang, J. IMAGDressing-v1: Customizable Virtual Dressing. *arXiv preprint arXiv:2407.12705* **2024**.
44. Shen, F.; Ye, H.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Advancing pose-guided image synthesis with progressive conditional diffusion models. *arXiv preprint arXiv:2310.06313* **2023**.
45. Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; Yang, W. Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models. *arXiv preprint arXiv:2407.02482* **2024**.
46. Shen, F.; Wang, C.; Gao, J.; Guo, Q.; Dang, J.; Tang, J.; Chua, T.S. Long-Term TalkingFace Generation via Motion-Prior Conditional Diffusion Model. *arXiv preprint arXiv:2502.09533* **2025**.
47. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
48. Dou, Z.Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. An empirical study of training end-to-end vision-and-language transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18166–18176.
49. Hendricks, L.A.; Mellor, J.; Schneider, R.; Alayrac, J.B.; Nematzadeh, A. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 570–585.
50. Zeng, W.; Wang, T.; Cao, J.; Wang, J.; Zeng, H. Clustering-guided pairwise metric triplet loss for person reidentification. *IEEE Internet of Things Journal* **2022**, *9*, 15150–15160.
51. Chen, K.; Lei, W.; Zhao, S.; Zheng, W.S.; Wang, R. PCCT: Progressive class-center triplet loss for imbalanced medical image classification. *IEEE Journal of Biomedical and Health Informatics* **2023**, *27*, 2026–2036.
52. Wang, T.; Zheng, Z.; Zhu, Z.; Gao, Y.; Yang, Y.; Yan, C. Learning cross-view geo-localization embeddings via dynamic weighted decorrelation regularization. *arXiv preprint arXiv:2211.05296* **2022**.
53. Wang, T.; Zheng, Z.; Sun, Y.; Yan, C.; Yang, Y.; Chua, T.S. Multiple-environment Self-adaptive Network for Aerial-view Geo-localization. *Pattern Recognition* **2024**, *152*, 110363.
54. Zhu, Y.; Yang, H.; Lu, Y.; Huang, Q. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572* **2023**.

 doi:10.20944/preprints202503.1203.v1

55. Hu, Q.; Li, W.; Xu, X.; Liu, N.; Wang, L. Learning discriminative representations via variational self-distillation for cross-view geo-localization. *Computers and Electrical Engineering* **2022**, *103*, 108335.
56. Tian, X.; Shao, J.; Ouyang, D.; Shen, H.T. UAV-satellite view synthesis for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *32*, 4804–4815.
57. Bui, D.V.; Kubo, M.; Sato, H. A part-aware attention neural network for cross-view geo-localization between UAV and satellite. *Journal of Robotics, Networking and Artificial Life* **2022**, *9*, 275–284.
58. Zhu, R.; Yang, M.; Yin, L.; Wu, F.; Yang, Y. Uav's status is worth considering: A fusion representations matching method for geo-localization. *Sensors* **2023**, *23*, 720.
59. Shen, T.; Wei, Y.; Kang, L.; Wan, S.; Yang, Y.H. MCCG: A ConvNeXt-based multiple-classifier method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* **2023**.
60. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
61. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 12009–12019.
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.