MOGeo: Beyond One-to-One Cross-View Object Geo-localization

Bo Lv, Qingwang Zhang, Le Wu, Yuanyuan Li, Yingying Zhu*
College of Computer Science and Software Engineering, Shenzhen University
{2250271009, zhangqingwang2022, 2300271073, 2200271059}@email.szu.edu.cn, zhuyy@szu.edu.cn

Abstract

Cross-View Object Geo-Localization (CVOGL) aims to locate an object of interest in a query image within a corresponding satellite image. Existing methods typically assume that the query image contains only a single object, which does not align with the complex, multi-object geo-localization requirements in real-world applications, making them unsuitable for practical scenarios. To bridge the gap between the realistic setting and existing task, we propose a new task, called Cross-View Multi-Object Geo-Localization (CVMOGL). To advance the CVMOGL task, we first construct a benchmark, CMLocation, which includes two datasets: CMLocation-V1 and CMLocation-V2. Furthermore, we propose a novel cross-view multi-object geo-localization method, MOGeo, and benchmark it against existing state-of-the-art methods. Extensive experiments are conducted under various application scenarios to validate the effectiveness of our method. The results demonstrate that cross-view object geo-localization in the more realistic setting remains a challenging problem, encouraging further research in this area. Our dataset and code will be released at `https://github.com/LV-BO001/MOGeo`.

1. Introduction

Cross-View Geo-Localization (CVGL) aims to determine the geographical location of a query image (e.g., ground or drone image) from geo-tagged reference images (satellite-view images) without relying on GPS or other positioning devices. This technique has broad applications in areas such as autonomous driving [5], urban navigation [10, 17], smart city management [6, 23], and disaster monitoring [3, 9].

In recent years, significant research progress has been made in the CVGL task. The research direction has expanded from center alignment of cross-view images to non-center alignment [38], from coarse-grained alignment to fine-grained alignment [28], and from supervised cross-view geo-localization to unsupervised geo-localization methods [11]. At the same time, research has also shifted from panoramic images to images with limited field of view [22]. The overall trend is gradually moving toward fine-grained localization and closer alignment with real-world scenarios. However, when it comes to localizing objects of interest, such as buildings, within a query image, traditional CVGL methods face significant challenges and limitations, as they are restricted to estimating geo-localization at the image level. To address this problem, the task of Cross-View Object Geo-Localization (CVOGL) is proposed. As illustrated in Fig. 1(a), the CVOGL task enables geo-localization at the object level. However, it idealizes the application scenario to a single-object setting, focusing solely on the geo-localization of a single object, which contrasts with real-world query images that typically contain multiple objects, such as buildings, roads, and bridges. This limitation undermines its ability to meet the fine-grained geo-localization demands of practical applica-

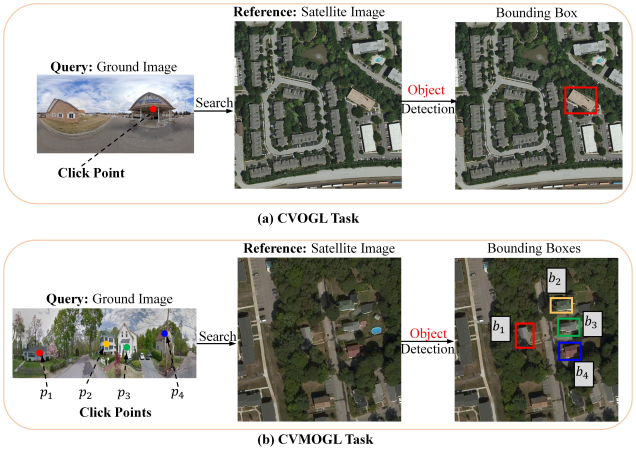

Figure 1. Comparison of cross-view object geo-localization in single-object and multi-object scenarios. Click points represent query objects, while bounding boxes in geo-tagged satellite images indicate location information. Points and bounding boxes of the same color form an object pair, such as p_1 and b_1, where b_1's geographic location is considered the position of p_1.

*Corresponding author.

tions, thereby restricting the CVOGL task's effectiveness in real-world scenarios.

To overcome the limitation of CVOGL, we propose a new task called Cross-View Multi-Object Geo-Localization (**CVMOGL**). As illustrated in Fig. 1(b), given a query image containing multiple query objects (represented by click points), matching objects (represented by bounding boxes) are located in a geo-tagged satellite image. Points and bounding boxes of the same color in the image represent matched object pairs, and the geographic information of the bounding boxes is considered as the geo-localization of the corresponding query objects. For example, b_1 in the satellite image represents the location of the query object p_1. This makes CVMOGL bridge the gap between current research and practical applications, better aligning with practical needs, and enhancing the applicability of CVOGL.

However, the CVMOGL task presents several new challenges beyond existing settings: (1) CVOGL is a special case of CVMOGL, so existing CVOGL methods are not suitable for the CVMOGL task. CVGL methods provide geo-localization at the image level, failing to offer geo-localization at the object level. (2) In addition to localizing multi-objects in satellite images, it is also necessary to establish the correspondence between each query object (represented by point) and the reference object (represented by a bounding box) to ensure the accuracy of the geo-localization results.

To advance the research on the CVMOGL task, we first propose a benchmark, called **CMLocation** (CMLocation-V1 and CMLocation-V2). CMLocation includes a total of 63,888 object instances across 25,520 pairs of query and reference images, providing rich data support for CVMOGL research. Especially, the CMLocation-V2 dataset, with non-center and non-north alignment and varying resolutions, is more representative of real-world scenarios and presents greater challenges.

Moreover, to further advance the research on the CVMOGL task, we propose an end-to-end cross-view multi-object geo-localization method, **MOGeo**. MOGeo includes a dual-branch feature encoding architecture, Multi-Object Position Encoding (MOPE), Cross-View Multi-feature Fusion (CVMF), and training objective function. The dual-branch encoder independently extracts visual features from each view image. As shown in Fig. 2, previous smooth positional encodings (e.g., Gaussian or Euclidean) create diffuse attention maps that hinder precise focus on the query object. To resolve this spatial ambiguity, we propose a sharp, non-smooth delta-like impulse encoding inspired by the Dirac delta function [13]. This method provides a highly discriminative positional information for each query object. We realize this idea with our novel MOPE module, ultimately boosting the model's localization accuracy. Additionally, to fuse cross-view multi-object features, we design a cross-view multi-feature fusion module that emphasizes the importance of attention maps. Furthermore, based on prior knowledge, objects in different images, or different objects within the same image exhibit distinct attention distribution features. To enhance the distinction in attention distribution features, we design a novel optimization function. Finally, the fused features are processed through multiple detection heads to predict the corresponding locations of the query objects within the reference image.

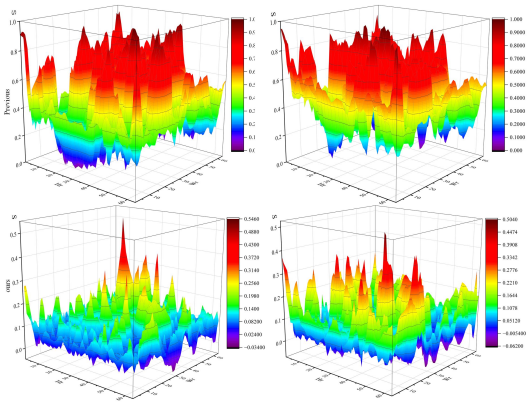

Figure 2. Attention map comparison: previous vs. ours. Previous smooth positional encodings lead to diffuse attention maps, whereas our proposed attention map successfully concentrates on the object region to provide highly discriminative features.

Our main contributions can be summarized as follows:

- To break the limitation of CVOGL, we propose a new task of Cross-View Multi-Object Geo-Localization (CVMOGL). CVMOGL addresses the multi-object geo-localization problem, offering a more challenging and realistic CVOGL setting.
- To bridge the gap between the realistic setting and existing datasets, we first construct a benchmark, called CMLocation(CMLocation-V1 and CMLocation-V2). This benchmark includes a total of 63,888 object instances in 25,520 pairs of query and reference images, providing rich data support for CVMOGL research.
- To promote the research on the CVMOGL task, we propose a cross-view multi-object geo-localization method, called MOGeo. We evaluate MOGeo and other methods on the CMLocation benchmark. The evaluation results show that MOGeo achieves state-of-the-art performance, fully demonstrating its effectiveness.

2. Related work

2.1. Cross-view Image Geo-localization

Cross-view image geo-localization has made significant progress in both methods [1, 2, 12, 16, 21] and datasets [1, 8, 24, 26, 27, 29, 30].

In terms of datasets, the CVUSA [26, 30] and CVACT [14] datasets use ground-view images as queries matched with high-resolution satellite imagery, while University-1652 [36] and SUES-200 [37] datasets support drone geo-localization. However, the aforementioned datasets mainly consider one-to-one retrieval, which may not align well with real-world scenarios. To address this, VIGOR [38] allows non-center-aligned image pairs for more realistic retrieval. However, these datasets provide geo-localization at the image level, but do not offer localization at the object level.

In terms of localization methods, Shi et al. [19] applied polar transformations to align satellite-view and ground images, while Regmi et al. [18] used conditional GANs to synthesize ground-level images from satellite-view inputs. Lu et al. [15] and Toker et al. [25] incorporated geometric priors and multi-task learning to bridge the cross-view domain gap. More recently, methods based on Swin Transformers [16, 20] and unsupervised learning [11] have achieved state-of-the-art results on UAV-view datasets. Chen et al. [1] introduced GeoSSK, which improves geo-localization accuracy between ground and satellite-view images using cross-view semantic similarity learning and knowledge distillation. Zhang et al. [34] proposed GeoDTR+, which enhances cross-region generalization by modeling geometric layouts and generating hard training samples. However, these studies are limited to image-level localization, restricting their real-world applicability.

2.2. Cross-View Object Geo-localization

Cross-view object geo-localization research is still in its early stages, and related studies are relatively limited. The DetGeo method [24] pioneers object-level cross-view geo-localization, and the VAGeo method [12] has further advanced research on CVOGL. More recent works, such as TROGeo [32] and GeoFormer [35], explore this problem from two key aspects: supervision paradigms and object discriminability. However, these methods are primarily restricted to single-object settings, limiting their applicability in real-world scenarios. To bridge the gap between current research and practical applications, we first propose the CVMOGL task.

3. The Proposed Benchmark

To advance research on the cross-view multi-object geo-localization (CVMOGL) task, we construct a cross-view multi-object geo-localization benchmark, named CMLocation. Below, we describe the CMLocation benchmark in detail.

Dataset Collection and Object Annotation. The emphasized properties of the cross-view multi-object geo-localization dataset are two major elements: cross-view and multi-object. Therefore, based on the cross-view image datasets CVUSA, we manually annotate visible objects in both query and reference images using the LabelImg annotation tool. We represent the annotated bounding boxes in the query image with a random point inside the box (as a click point). The corresponding relationships between query points in the query image and objects in the reference image are manually determined and recorded as object correspondence labels. As shown in Fig. 3, points and boxes with the same label number form a matching relationship. Finally, we obtain the CMLocation-V1 dataset. However, because the query and reference images in CMLocation-V1 are arranged in a north-aligned and center-aligned fashion (which facilitates determining object correspondences during manual annotation), this setup differs significantly from real-world application scenarios. Therefore, we applied random cropping, flipping, and scaling to the reference images in CMLocation-V1 so that the query and reference images would no longer share the same center point or fixed orientation. This process resulted in the CMLocation-V2 dataset, which is more realistic and challenging for real-world applications. More details about the dataset can be found in the appendix.

Table 1. Dataset statistics and splits of CMLocation-V1 and CMLocation-V2.

Data name	Split	Query	Reference	Instances
CMLocation-V1	total	12760	12760	31944
	training	8371	8371	20931
	validation	2093	2093	5209
	test	2296	2296	5804
CMLocation-V2	total	12760	12760	31944
	training	8371	8371	20931
	validation	2093	2093	5209
	test	2296	2296	5804

Dataset Split. Table 1 summarizes the dataset scales and splits of CMLocation-V1 and CMLocation-V2. Both versions contain 12,760 query-reference image pairs and share the same set of query objects (31,944 in total), providing rich data support for CVMOGL research.

To analyze object size distributions under different views, we compute the bounding box sizes of query and reference objects in both versions, as shown in Fig. 4. The results reveal a wide range of object sizes in both datasets, with query-side objects generally appearing smaller than their counterparts in the reference images. In CMLocation-V2, the scale discrepancy becomes even more pronounced due to the more extreme viewpoint variations, presenting greater challenges for robust model training.

Evaluation Settings. The acc@t metric in [24] evaluates single-object localization but fails to assess overall multi-object localization accuracy in the CVMOGL task. There-

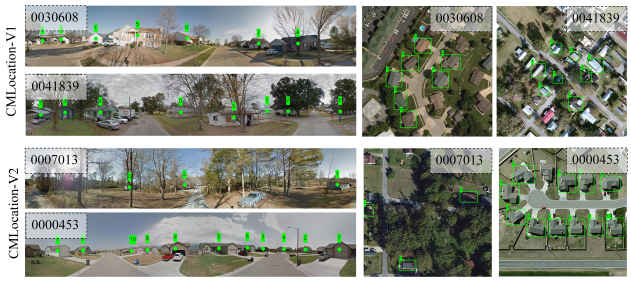

Figure 3. Examples from the CMLocation-V1 and CMLocation-V2 datasets. CMLocation-V1 is constructed with strict alignment constraints, including center alignment and northward orientation, ensuring a consistent spatial distribution across instances. By contrast, CMLocation-V2 does not impose these constraints, resulting in greater variability in object positions and orientations.

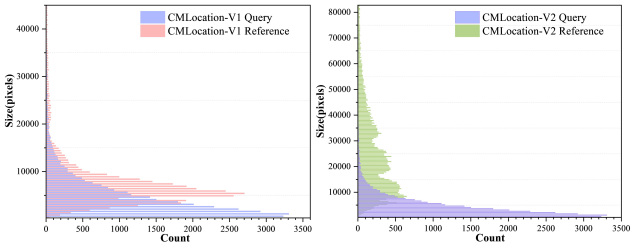

Figure 4. The size distributions of objects in the CMLocation-V1 and CMLocation-V2 datasets.

fore, we introduce the concept of image localization accuracy. The calculation formula is:

$$\text{accI@}t = \frac{1}{n}\sum_{i=1}^{n}\psi_i(t) \tag{1}$$

$$\psi_i(t) = \begin{cases} 0, & \text{if } \exists j \in \{1,2,\ldots,m_i\},\ \text{IoU}(b_j, b_j^*) < t \\ 1, & \text{else} \end{cases} \tag{2}$$

where n is the number of query images, m_i is the number of query objects in the i-th image, b_i is the predicted bounding box, and b_i^* is the ground truth bounding box. In these metrics, accI@t denotes the proportion of images in the dataset where the objects are correctly localized. An image is deemed correctly localized when the Intersection over Union (IoU) between all predicted and ground truth bounding boxes is greater than t. To ensure a fair comparison, we adopt the same experimental setup as in [24], setting the threshold t to 0.25 and 0.5, respectively.

4. The Proposed Method

In this section, we aim to locate objects in cross-view images, typically represented by bounding boxes. To achieve this, we propose a novel detection-based CVMOGL method, called MOGeo, as shown in Fig. 5. MOGeo consists of four main components: a dual-branch image feature extraction network, multi-head query position encoding, a multi-feature fusion module, and an optimization function. In the following, we will provide a detailed explanation of each component.

4.1. Problem Statement

A set of query points and reference objects is represented as $X = \{(q_i, r_i, p_j, b_j)_{j=1}^{m_i}\}_{i=1}^{n}$, where the query image q_i and reference image r_i form a matching image pair, and the query point p_j and bounding box b_j form a matching object pair. m_i represents the number of object pairs contained in the i-th image pair. Given a pair of query and reference images, and the positions of the query points in the query image, our goal is to identify the corresponding bounding boxes in the reference image. Thus, the problem is formulated as $(q_i, r_i, p_j)_{j=1}^{m_i} \rightarrow b_j$.

For convenience, we choose the i-th sample from dataset X as the input data: $x_i = (q_i, r_i, p_j, b_j)_{j=1}^{m_i}$, $q_i \in \mathbb{R}^{H_q \times W_q \times 3}$, $r_i \in \mathbb{R}^{H_r \times W_r \times 3}$. We denote the i-th query image and reference image as I_q and I_r, respectively. $P = \{p_1, p_2, \ldots, p_{m_i}\}$ represents the query points, each corresponding to a ground truth (GT) bounding box, denoted as $\{b_1, b_2, \ldots, b_{m_i}\}$, where $b_j = (x_j^*, y_j^*, w_j^*, h_j^*)$.

Query and Reference Image Representation. To ensure fairness in subsequent experimental comparisons, we choose the same backbone architecture for feature extraction as in the works of [12, 24]. For the query image I_q, we use ResNet18 to extract features, selecting the downsampled 16× result as the feature representation $F_q \in \mathbb{R}^{C \times H_q' \times W_q'}$. For the reference image I_r, we use Darknet53 for feature extraction and select the downsampled 16× result. A fully connected operation is applied to obtain a more expressive feature F_r. We reshape F_r to $\mathbf{V}_r = \{v_{r1}, v_{r2}, \ldots, v_{rc}\}$, where $\mathbf{V}_r \in \mathbb{R}^{4096 \times 512}$, for subsequent feature fusion.

4.2. Multi-Head Query Object Encoding

To support multi-object geo-localization in query images, we propose a multi-head query object position encoding (**MOPE**) module based on feature post-processing. Inspired by the Dirac delta function, we implement our impulse-based positional encoding as a one-hot mask, ensuring a sharp and discriminative spatial prior. Specifically, after obtaining the feature map F_q of the query image, position encoding is performed by creating m binary masks $\{E_1, E_2, \ldots, E_m\}_{j=1}^{m} \subset \{0,1\}^{H_q' \times W_q'}$, each encoding the position of a query point using the one-hot principle. The mathematical representation is as follows:

$$E_j(h,w) = \begin{cases} 1, & \text{if } (h,w) = \lfloor \varphi(p_j) \rfloor \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denotes the transformation that maps the query point coordinates to the feature map coordinates,

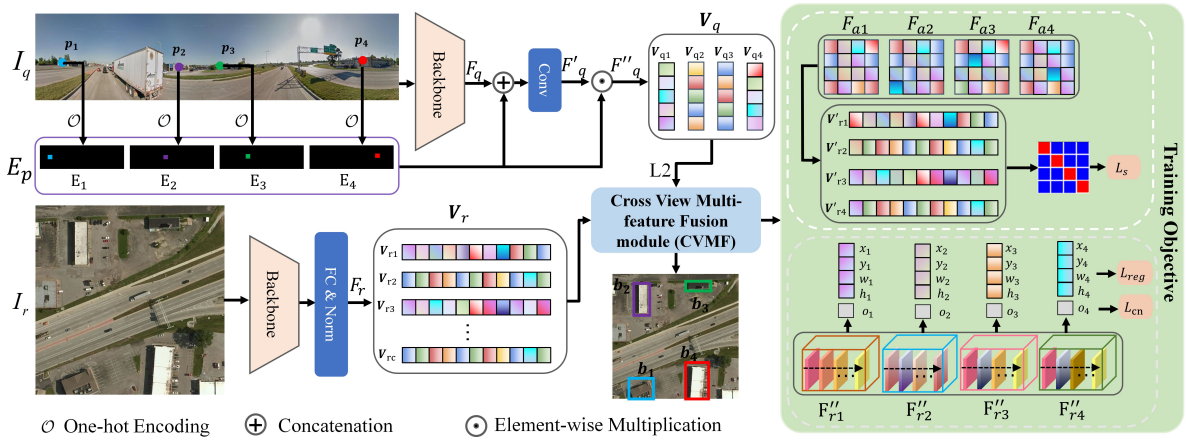

Figure 5. Overview of our proposed MOGeo model. Our method takes as input a query image that contains an arbitrary number of objects of interest (either ground-view images or drone-view images) and a reference image that contains the objects to be localized (satellite image). The input images are processed by our position encoding module, MOPE, to extract feature information for each query object. Then, query object features are fused with the reference image to obtain fused features $\mathbf{F}''_r$, which are fed into the detection head to produce localization results. In the figure, query points and bounding boxes with the same color represent object pairs.

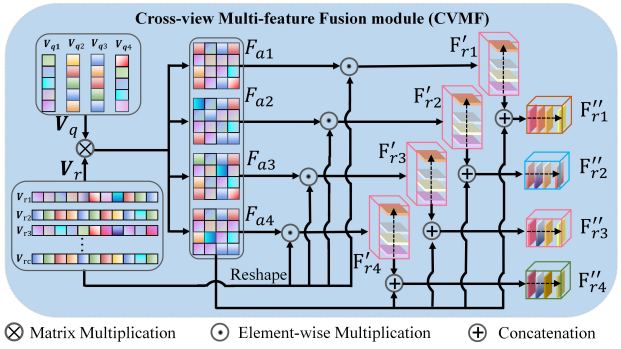

Figure 6. Cross-view multi-object feature fusion module. As shown in the figure, this module sequentially fuses query object features with the reference image to obtain an attention map for each object, and finally combines fused features with the attention map to further enhance localization in the reference image.

and $\lfloor \cdot \rfloor$ represents the floor operation. After obtaining the absolute position information of the query objects, we fuse each position encoding E_j with the visual feature F_q via channel concatenation, followed by a 1×1 convolution to transform the combined feature, resulting in the enhanced feature $F'_{qj} \in \mathbb{R}^{C \times H'_q \times W'_q}$:

$$F'_{qj} = \text{Conv}_{1\times1}\left([F_q \| E_j]\right). \tag{4}$$

Here, $\text{Conv}_{1\times1}$ denotes a convolution with a 1×1 kernel, and $[\cdot\|\cdot]$ represents concatenation along the channel dimension. For simplicity, we denote $\{F'_{q1}, F'_{q2}, \ldots, F'_{qm}\}$ as F'_q. Through the above operations, we obtain the query image information with the initial fusion feature F'_q. However, a significant channel imbalance exists between F_q and E_j, which may lead to the position information being overshadowed or weakened by the visual feature information. To mitigate the potential dilution of position information by semantic features, we propose a position-driven feature enhancement strategy. By expanding the position mask along the channel dimension and performing element-wise multiplication, we obtain the position-enhanced fused feature F''_q:

$$\{F''_{q1}, F''_{q2}, F''_{q3}, \ldots, F''_{qm}\} = F'_q \odot \{E_1, E_2, \ldots, E_m\}. \tag{5}$$

Here, $\odot$ denotes element-wise multiplication. The fused feature $\{F''_{q1}, F''_{q2}, F''_{q3}, \ldots, F''_{qm}\}$ is then transformed and pooled to produce m query feature vectors, denoted as $\mathbf{V}_q = \{v_{q1}, v_{q2}, \ldots, v_{qm}\}$, where $\mathbf{V}_q \in \mathbb{R}^{m \times d}$, with $d = 512$.

4.3. CVMF Module

After obtaining the query object representation vector $\mathbf{V}_q$ and reference image feature vector $\mathbf{V}_r$, we fuse the features from both views for object localization in the satellite image. To ensure robust feature matching, both $\mathbf{V}_q$ and $\mathbf{V}_r$ undergo L_2 normalization. Subsequently, matrix multiplication is applied to generate a cross-view attention map, as defined by:

$$\{\mathbf{V}_p\} = \{v_{p1}, v_{p2}, v_{p3}, \ldots, v_{pm}\} = \{\mathbf{V}_q\} \times \{\mathbf{V}_r\}. \tag{6}$$

Reshape $\mathbf{V}_{pj}$ to a tensor of dimensions $H_r \times W_r$ to obtain the cross-view attention map F_{aj}. The attention map $F_a = \{F_{a1}, F_{a2}, \ldots, F_{am}\}$ captures the potential corresponding regions of the query objects in the reference image. Each attention feature map F_{aj} is then weighted and

Table 2. Performance (accuracy%) comparison of different approaches on our dataset CMLocation.

Method	Test				Validation			
	acc@0.25	acc@0.5	accI@0.25	accI@0.5	acc@0.25	acc@0.5	accI@0.25	accI@0.5
CMLocation-V1								
FRGeo[31]	8.06	1.05	2.74	0.17	7.20	0.92	2.53	0.05
GeoDTR+[34]	7.44	0.45	3.83	0.17	8.62	0.77	4.30	0.53
Sample4Geo[4]	29.19	2.31	16.46	1.44	30.68	2.36	19.10	1.81
DetGeo[24]	56.03	55.00	39.68	38.32	56.29	55.33	39.80	38.93
VAGeo[12]	57.79	57.06	43.03	42.33	57.04	56.50	42.66	42.09
MOGeo	**63.85**	**62.92**	**46.66**	**45.60**	**63.52**	**62.58**	**46.82**	**45.44**
CMLocation-V2								
FRGeo[31]	9.87	1.43	4.05	0.13	9.83	1.27	2.81	0.05
GeoDTR+[34]	11.10	1.72	3.22	0.13	11.69	1.56	3.20	0.19
Sample4Geo[4]	32.43	4.86	18.78	1.22	31.37	4.15	19.10	0.86
DetGeo[24]	32.39	31.79	28.05	27.00	33.10	32.56	30.05	29.14
VAGeo[12]	33.51	33.06	29.18	28.17	34.38	33.98	31.25	30.53
MOGeo	**37.87**	**36.63**	**30.23**	**28.26**	**39.16**	**37.40**	**34.20**	**31.58**

integrated with the satellite image, as defined by the following formula:

$$\{\mathbf{F}'_r\} = \{\mathbf{F}_a\} \odot \{\mathbf{F}_r\}. \tag{7}$$

To further enhance the impact of each attention feature F_{ai} in the fused features, we concatenate the attention features F_{ai} of each query object with the corresponding fused features $\mathbf{F}'_r$. Finally, the concatenated features $\mathbf{F}''_r$ are processed through the prediction heads to generate the bounding box, with the box of highest confidence selected as the final prediction.

4.4. Training Objective

We design a loss function to optimize MOGeo based on [24]. The loss is defined as:

$$\mathcal{L} = L_{\text{cn}} + L_{\text{reg}} + L_s, \tag{8}$$

where L_s is the similarity loss between different objects, L_{reg} is the regression loss, and L_{cn} is the confidence loss. The regression loss minimizes the distance between the predicted bounding box and the GT, while the confidence loss estimates the probability of an object being present in a specific grid cell. The similarity loss computes the differences between attention feature vectors of different query objects. As illustrated in Fig. 2, previous methods suffer from the problem of overly diffuse attention maps. Aiming to further enhance the discriminative power of the attention feature vectors, we utilize this loss to explicitly maximize the distance between non-corresponding query objects, both within a single image and across different images. The calculation formula for L_s is as follows:

$$L_{\text{s}} = \sum_{i=1}^{n} \sum_{k=1}^{m_i} \log\left(1 + e^{(d_{\text{pos}} - d_{\text{neg}})}\right), \tag{9}$$

where d_{pos} represents the Euclidean distance between the attention map of the current query object and itself, while d_{neg} represents the Euclidean distance between the attention map of the current query object and the attention maps of other query objects. Notably, since the positive distance d_{pos} (computed between features of the same object) is already minimized by design, our loss function solely constrains the maximization of distances between different objects.

5. Experiment

5.1. Experimental Settings

Our model is implemented using PyTorch, and the training process is conducted on an NVIDIA V100 GPU. We use the Adam optimizer with an initial learning rate of 1×10^{-4}. The batch size is set to 8, and the model is trained for a total of 24 epochs. We evaluate our method on the CMLocation dataset (CMLocation-V1, CMLocation-V2) and CVOGL(CVOGL-SVI,CVOGL-Drone) dataset. Specifically, CMLocation and CVOGL-SVI are used for the *Ground→Satellite* scenario, while CVOGL-Drone is dedicated to the *Drone→Satellite* scenario. For more experimental results and analysis, please refer to the Appendix.

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

To ensure a fair comparison with existing retrieval-based localization methods, we divide the satellite reference image into 128×128 pixel patches. Each query object is encoded into a feature representation, and the top five most similar patches are retrieved from the reference image. An object is considered successfully localized if at least one retrieved patch has an Intersection over Union (IoU) with the ground truth bounding box exceeding a specified threshold. Otherwise, it is deemed a failure.

Table 3. Performance (accuracy%) comparison of different approaches on CVOGL dataset.

Method	Test		Validation	
	acc@0.25	acc@0.5	acc@0.25	acc@0.5
CVOGL-SVI				
CVM-Net [7]	4.73	0.51	5.09	0.87
TransGeo[39]	21.17	2.88	21.67	3.25
FRGeo [31]	8.12	1.31	7.80	0.87
GeoDTR [33]	32.37	6.06	31.53	5.31
Sample4Geo[4]	8.84	1.23	11.48	1.41
DetGeo [24]	45.43	42.24	46.70	43.99
VAGeo [12]	48.21	45.22	47.56	44.42
Ours	**50.98**	**47.70**	**49.51**	**44.42**
CVOGL-Drone				
CVM-Net[7]	20.14	3.29	20.04	3.47
TransGeo[39]	35.05	6.37	34.78	5.42
FRGeo [31]	11.41	2.67	13.22	2.06
GeoDTR [33]	32.48	5.65	31.13	5.53
Sample4Geo[4]	7.71	0.92	11.16	1.41
DetGeo[24]	61.97	57.66	59.81	55.15
VAGeo [12]	66.19	**61.87**	64.25	**59.59**
Ours	**66.39**	59.71	**65.11**	58.72

Additionally, to compare fairly with existing object detection methods, we convert our dataset into a single-object geo-localization format for evaluation.

(1) Performance of Cross-View Multi-Object Geo-Localization. Table 2 presents a comprehensive comparison of MOGeo with other methods on our CMLocation dataset. MOGeo demonstrates superior performance across both CMLocation-V1 and CMLocation-V2 datasets, particularly in acc@0.25 and accI@0.25 metrics. For example, on the CMLocation-V1 test set, MOGeo surpasses the second-best method, VAGeo, by 6.06% in acc@0.25, and by 3.63% in accI@0.25. On the more challenging CMLocation-V2, although overall performance declines compared to CMLocation-V1, MOGeo consistently outperforms strong baselines. It achieves relative gains of 4.78% and 4.15% over VAGeo and DetGeo in acc@0.25 and accI@0.25, respectively, on the validation set. These results demonstrate the effectiveness and robustness of MOGeo across varying levels of task difficulty, consistently achieving state-of-the-art (SOTA) performance. The substantial improvement in accI@0.25 and accI@0.50 highlights that MOGeo provides fine-grained object geo-localization, rather than merely estimating rough positions. These results also suggest that existing cross-view image geo-localization methods may not be well-suited for the CVMOGL task.

(2) Performance of Cross-View Single-Object Geo-Localization. The CVOGL task is a special case of the CVMOGL task. To further evaluate the effectiveness of our method, we conducted experiments on the CVOGL dataset. The experimental results of different methods are shown in Table 3. As seen from the table, our method achieves SOTA or competitive performance in the *Ground→Satellite* and *Drone→Satellite* tasks. Specifically, in the *Ground→Satellite* task, MOGeo achieves 50.98% in acc@0.25, a 2.77% improvement over VAGeo and a 5.55% improvement over DetGeo. Similarly, in the *Drone → Satellite* task, MOGeo achieves SOTA on both the test and validation sets for acc@0.25, while remaining competitive with the best-performing methods on acc@0.50. These results further confirm MOGeo's superior performance in multi-object geo-localization, while maintaining competitive accuracy in the single-object geo-localization setting.

Table 4. Model performance (accuracy%) under different numbers of query objects on the validation set.

Method	Validation					
	I ($N \leq 3$)		**II** ($3 < N \leq 6$)		**III** ($N > 6$)	
	acc@0.25	acc@0.5	acc@0.25	acc@0.5	acc@0.25	acc@0.5
CMLocation-V1						
DetGeo	61.03	59.90	49.03	48.23	40.58	40.34
VAGeo	62.97	62.30	47.74	47.54	42.27	42.27
Ours	**67.22**	**66.07**	**56.74**	**56.34**	**53.14**	**52.90**
CMLocation-V2						
DetGeo	40.47	39.75	17.90	17.80	11.59	11.59
VAGeo	42.09	41.58	19.09	18.95	10.39	10.39
Ours	**46.37**	**44.05**	**25.86**	**25.31**	**18.84**	**18.36**

(3) Impact of the Query Object Count on Model Performance. To investigate the effect of query object count on model performance, we divided the CMLocation dataset into three categories based on the number of objects (N) in the query image: I $(N \leq 3)$, II $(3 < N \leq 6)$, and III $(N > 6)$. The results in Table 4 show that while all methods experience a performance decline with more query objects, our proposed MOGeo consistently achieves state-of-the-art performance in all settings. Notably, MOGeo exhibits superior robustness, and its performance degradation is less severe than that of the baseline methods as N increases. This leads to a widening performance gap in more challenging scenarios. For example, on CMLocation-V2 (acc@0.25), MOGeo's lead over the second-best method grows from 4.28% in the simplest category (I) to 7.25% in the most complex one (III), demonstrating its enhanced capability in handling dense multi-object localization tasks.

(4) Localization Speed. To further investigate the trade-off between model complexity and inference efficiency, we compare the parameter counts and localization speeds of different methods in Fig. 7. The GeoDTR+ method has the smallest parameter count and achieves fast inference. MOGeo has a comparable number of parameters to GeoDTR+,

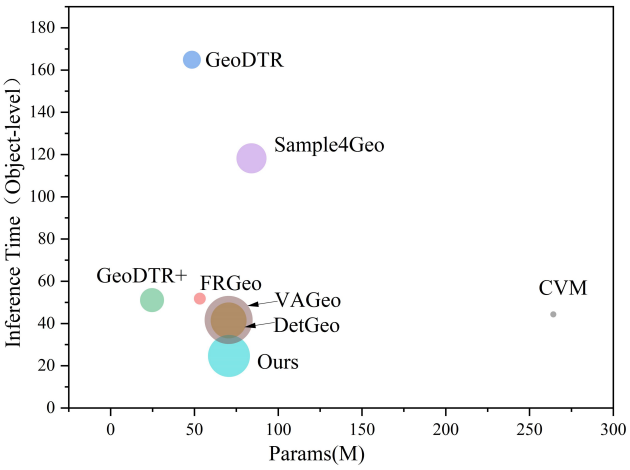

Figure 7. Relationship between model parameters and inference time, measured on the validation set of the CMLocation-V1.

but achieves faster inference. While VAGeo has few parameters, it is designed to localize a single object per inference, leading to increased time when handling multiple objects.

Table 5. Ablation performance (accuracy%) comparison on CVOGL-Drone and CMLocation-V2 datasets.

Method	Test		Validation	
	acc@0.25	acc@0.5	acc@0.25	acc@0.5
CVOGL-Drone				
Ours	66.39	59.71	65.11	58.72
w/o L_s	65.98	59.30	61.21	54.82
w/o CVMF	65.78	59.92	62.62	58.40
w/o MOPE	54.88	50.05	54.60	48.21
CMLocation-V2				
Ours	37.87	36.63	39.16	37.40
w/o L_s	37.56	35.85	38.82	36.74
w/o CVMF	37.19	35.57	37.45	35.73
w/o MOPE	32.74	31.12	33.00	31.54

5.3. Ablation Study

To validate the effectiveness of our proposed components (MOPE, CVMF, L_s), we conducted a series of experiments by sequentially removing each component. We performed these experiments on two challenging datasets: CMLocation-V2 (the *Ground→Satellite* scenario) and CVOGL-Drone (the *Drone→Satellite* scenario), with the results presented in Table 5. As observed, removing any of the three components results in varying degrees of performance degradation across both datasets, which underscores the individual contribution of each component to the final performance. Notably, the removal of the MOPE module leads to the most substantial performance drop. For instance, on the CVOGL-Drone test set, the primary metric

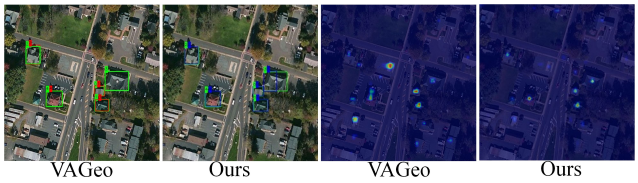

Figure 8. Detection result visualization. Green represents the GT boxes, while the red and blue bounding boxes correspond to the predictions of VAGeo and our method, MOGeo, respectively.

acc@0.25 drops by 11.51%, and on CMLocation-V2 test, it decreases by 5.13%. This significant decline, far exceeding the impact of removing other modules, strongly validates our central hypothesis and demonstrates the critical role of the proposed impulse-based positional encoding in achieving accurate multi-object localization.

5.4. Visualization Analysis

To further analyze the proposed method and compare the attention regions of different models, we present the detection results and corresponding heatmaps in Fig. 8. In the detection results visualization, numbers denote object indices, while green, red, and blue bounding boxes represent GT bounding boxes, VAGeo predictions, and MOGeo (ours) predictions, respectively. MOGeo correctly localizes all query objects, while VAGeo identifies only two (indices 2 and 4). In the heatmap visualizations, MOGeo precisely attends to objects of interest while suppressing background noise. In contrast, VAGeo exhibits spatial misalignment between its attention and objects, with attention partially shifting to non-object regions. This comparison demonstrates superior discriminative ability and robustness of MOGeo in multi-object localization.

6. Conclusion and Limitations

In this work, we address the limitation that existing cross-view geo-localization methods are restricted to single-object settings and fail to meet the practical demands of multi-object localization. To this end, we introduce a novel task of cross-view multi-object geo-localization. To facilitate research research in this direction, we have constructed the CMLocation benchmark, including both V1 and V2 versions, and proposed a MOGeo multi-object localization algorithm. Experimental results demonstrate that the proposed method achieves consistent improvements in multi-object localization. However, limitations remain in accurately aligning query objects with their corresponding objects and localizing them precisely, particularly in complex scenarios. These findings highlight directions for future improvements.

7. Acknowledgement

This work was supported in part Guangdong Basic and Applied Basic Research Foundation under Grant 2026A1515011137, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, and in part by Shenzhen Science and Technology Program under Grant JCYJ20240813142510014 and Grant 20220810142553001.

References

[1] Guanli Chen, Guoheng Huang, Xiaochen Yuan, Xuhang Chen, Guo Zhong, and Chi-Man Pun. Cross-view geo-localization via learning correspondence semantic similarity knowledge. In *International Conference on Multimedia Modeling*, pages 220–233. Springer, 2025. 2, 3

[2] Zhongwei Chen, Zhao-Xu Yang, and Hai-Jun Rong. Multi-level embedding and alignment network with consistency and invariance learning for cross-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 2

[3] Marco Chini, Nazzareno Pierdicca, and William J Emery. Exploiting sar and vhr optical images to quantify damage caused by the 2003 bam earthquake. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1):145–152, 2008. 1

[4] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 6, 7

[5] Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27, 2017. 1

[6] Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021. 1

[7] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 7

[8] Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2

[9] Anuj Kumar, Hiesik Kim, and Gerhard P Hancke. Environmental monitoring systems: A review. *IEEE Sensors Journal*, 13(4):1329–1339, 2012. 1

[10] Ang Li, Huiyi Hu, Piotr Mirowski, and Mehrdad Farajtabar. Cross-view policy learning for street navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8100–8109, 2019. 1

[11] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16719–16729, 2024. 1, 3

[12] Zhongyang Li, Xin Yuan, Wei Liu, and Xin Xu. Vageo: View-specific attention for cross-view object geo-localization. *arXiv preprint arXiv:2501.07194*, 2025. 2, 3, 4, 6, 7

[13] Michael J Lighthill. *An introduction to Fourier analysis and generalised functions*. Cambridge University Press, 1958. 2

[14] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 3

[15] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020. 3

[16] Hongxiang Lv, Hai Zhu, Runzhe Zhu, Fei Wu, Chunyuan Wang, Meiyu Cai, and Kaiyu Zhang. Direction-guided multi-scale feature fusion network for geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2, 3

[17] Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in neural information processing systems*, 31, 2018. 1

[18] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019. 3

[19] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[20] Yujiao Shi, Xin Yu, Liu Liu, Dylan Campbell, Piotr Koniusz, and Hongdong Li. Accurate 3-dof camera geo-localization via ground-to-satellite image matching. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):2682–2697, 2022. 3

[21] Tavis Shore, Oscar Mendez, and Simon Hadfield. Spagbol: Spatial-graph-based orientated localisation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6858–6867. IEEE, 2025. 2

[22] Maxim Shugaev, Ilya Semenov, Kyle Ashley, Michael Klaczynski, Naresh Cuntoor, Mun Wai Lee, and Nathan Jacobs. Arcgeo: Localizing limited field-of-view images using cross-view matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 209–218, 2024. 1

[23] Xihuang Sun, Peng Liu, Yan Ma, Dingsheng Liu, and Yechao Sun. Streaming remote sensing data processing for the future smart cities: state of the art and future challenges. *Environmental Information Systems: Concepts, Methodologies, Tools, and Applications*, pages 1711–1726, 2019. 1

[24] Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2, 3, 4, 6, 7

[25] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura 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*, pages 6488–6497, 2021. 3

[26] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 2, 3

[27] Panwang Xia, Lei Yu, Yi Wan, Qiong Wu, Peiqi Chen, Liheng Zhong, Yongxiang Yao, Dong Wei, Xinyi Liu, Lixiang Ru, et al. Cross-view geo-localization with street-view and vhr satellite imagery in decentrality settings. *arXiv preprint arXiv:2412.11529*, 2024. 2

[28] Zimin Xia and Alexandre Alahi. Fgˆ2: Fine-grained cross-view localization by fine-grained feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6362–6372, 2025. 1

[29] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network. In *European Conference on Computer Vision*, pages 74–90. Springer, 2024. 2

[30] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017. 2, 3

[31] Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7251–7259, 2024. 6, 7

[32] Qingwang Zhang and Yingying Zhu. Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8197–8206, 2025. 3

[33] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3480–3488, 2023. 7

[34] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3, 6

[35] Xiaohan Zhang, Si-Yuan Cao, Zhu Yu, Zhe Wu, Xue Zhang, Xiaokai Bai, Beinan Yu, and Hui-Liang Shen. Geoformer: Boosting object distinguishing and prompt understanding for cross-view object geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025. 3

[36] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020. 3

[37] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo 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*, 33(9):4825–4839, 2023. 3

[38] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 1, 3

[39] Sijie Zhu, Mubarak Shah, and Chen 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*, pages 1162–1171, 2022. 7