Recurrent Cross-View Object Geo-Localization

Xiaohan Zhang[1,2], Si-Yuan Cao[1,3⋆], Xiaokai Bai[2], Yiming Li[2], Zhangkai Shen[2], Zhe Wu[2], Lun Luo[2], Qi Ming[4], Xiaoxi Hu[5], and Hui-Liang Shen[2]

[1] Ningbo Global Innovation Center, Zhejiang University, Ningbo 315100, China
[2] College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China
[3] Jinhua Institute of Zhejiang University, Jinhua 321000, China
[4] College of Computer Science, Beijing University of Technology, Beijing 100124, China
[5] State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China
{zhangxh2023, cao_siyuan}@zju.edu.cn

Abstract. Cross-view object geo-localization (CVOGL) aims to determine the location of a specific object in high-resolution satellite imagery given a query image with a point prompt. Existing approaches treat CVOGL as a one-shot detection process, directly regressing object locations from cross-view information aggregation, but they are vulnerable to feature noise and lack mechanisms for error correction. In this paper, we propose **ReCOT**, a **Re**current **C**ross-view **O**bject geo-localization **T**ransformer, which models CVOGL as a recurrent localization process. ReCOT introduces a set of learnable tokens that encode task-specific intent from the query image and prompt embeddings, and iteratively attend to the reference features to refine the predicted location. To enhance this recurrent process, we incorporate two complementary modules: (1) a SAM-based knowledge distillation strategy that transfers segmentation priors from the Segment Anything Model (SAM) to provide clearer semantic guidance without additional inference cost, and (2) a Reference Feature Enhancement Module (RFEM) that introduces hierarchical attention to emphasize object-relevant regions in the reference features. Extensive experiments on CVOGL benchmarks demonstrate that ReCOT achieves state-of-the-art (SOTA) performance while significantly reducing parameters compared to previous SOTA approaches. Our code is available at https://github.com/Temperature-ai/ReCOT.git.

Keywords: Cross-View Object Geo-Localization · Recurrent Refinement · Learnable Tokens

1 Introduction

Cross-view object geo-localization (CVOGL) aims to determine the geographic location of a specific object indicated by point prompts in a query image on

⋆ Corresponding author.

the reference image [36]. The query images can be captured from devices like phones, autonomous vehicles, robots, and drones, while the reference images are typically high-resolution satellite images. CVOGL is widely used in various applications, such as smart city management [42], disaster monitoring [6, 16], and robot navigation [35, 45]. However, the view gap poses challenges [36].

Recent cross-view image geo-localization (CVIGL) works [11, 19, 31, 41, 52] have demonstrated their superiority in handling view gaps. However, CVIGL approaches are fundamentally designed for camera-level localization using retrieval-based approaches [7,31,49] or fine-grained approaches [28,39]. In contrast, CVOGL aims to localize specific objects (*e.g.*, a building with a red roof) captured in the query image, which demands prompt-guided and object-aware prediction. Therefore, in CVOGL scenarios, CVIGL approaches can only provide a nearby location for the indicated object [36], which is insufficient for precise object-level localization.

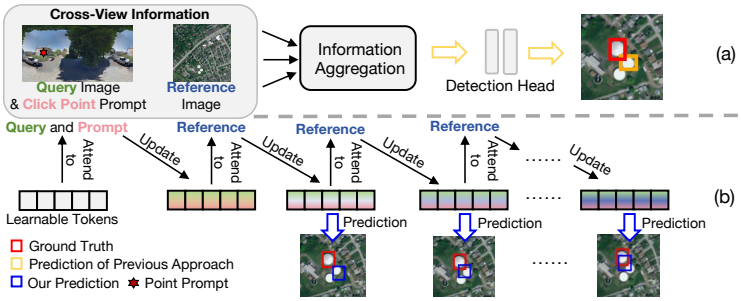

Fig. 1: Comparison between the framework of previous CVOGL approaches and ours. (a) Previous approaches treat the CVOGL as a prompt-based detection process, where the model directly regresses the object location based on information aggregation once. (b) Our framework models the CVOGL as a recurrent localization process, where the model iteratively refines the localization through a set of learnable tokens.

To address this, CVOGL approaches emerge recently. Existing approaches [13, 18, 36, 46, 48, 53] typically treat CVOGL as a one-shot detection paradigm, where the model directly regresses the object location based on prompt-guided information aggregation, as shown in Fig. 1(a). For example, the recent state-of-the-art (SOTA) approach [13] aggregates information from cross-view images and prompts to produce a spatial attention matrix, which is used to enhance the reference image features. The enhanced features are then fed into several convolutional layers to regress the object location. While efficient and architecturally simple, such a framework lacks a correction mechanism for early-stage prediction errors, making it vulnerable to noise in features [3, 36]. Specifically, since large view gaps yield multiple satellite candidates, the one-shot model must commit to a single spatial hypothesis in one pass, and this early commitment conditions the subsequent cross-view localization [36], making errors uncorrectable.

To cope with this, we propose a **Re**current **C**ross-view **O**bject geo-localization **T**ransformer (ReCOT). Motivated by the success of iterative refinement strategies [2,3,43], ReCOT models the CVOGL task as a recurrent localization process, as shown in Fig. 1(b). This serves as the main difference between previous approaches [13,18,36,46,48,53] and our framework. Specifically, ReCOT initializes a set of learnable tokens, which interact with the query image feature and prompt embeddings to extract task-specific intent. These tokens then act as recurrent "questioners" that iteratively attend to the enhanced reference image features, progressively extracting object-relevant information and refining the prediction. The proposed recurrent strategy enables our ReCOT to effectively enhance the performance of CVOGL by iteratively refining the initial prediction, as shown in Fig. 1(b). Compared to one-shot frameworks, recurrent refinement relaxes the one-pass commitment by treating localization as a revisable prediction state, *i.e.*, the model can iteratively revise its spatial hypothesis by re-weighting competing candidates and accumulating object-consistent evidence. Therefore, our recurrent mechanism effectively lifts the performance ceiling, offering greater potential to address more complex scenarios.

Nevertheless, the success of this recurrent strategy in our token-driven framework relies on the semantic clarity of the prompts [18] and the quality of reference features [37]. Specifically, the learnable tokens are first guided by prompt semantics to extract object-relevant intent, then iteratively attend to the reference features to refine the object location in each recurrent step. Therefore, if prompts lack clear semantic intent or reference features are cluttered, the tokens may not accumulate correct task-specific cues across iterations, leading to suboptimal performance. To tackle this, we introduce two complementary methods: (1) a SAM-based knowledge distillation strategy, which injects prior knowledge from a large-scale model into the prompt embeddings to boost prompt understanding while avoiding computational cost during inference, and (2) a Reference Feature Enhancement Module (RFEM), which emphasizes object-relevant reference features through hierarchical attention. These components provide clean visual and semantic cues, enabling the tokens to effectively accumulate task-specific information during iterative refinement.

We evaluate ReCOT on CVOGL benchmarks [13,36]. It achieves state-of-the-art (SOTA) performance while significantly reducing parameter count compared to previous one-shot approaches [36, 48], and runs at a competitive inference speed. In summary, our contributions are as follows:

- We propose ReCOT, a novel framework for CVOGL that models the task as a recurrent localization process, where learnable tokens iteratively attend to reference features to refine object localization.
- We introduce a SAM-based knowledge distillation strategy that transfers prior knowledge from a large foundation model into the prompt embeddings, providing clearer semantic guidance without adding inference cost.
- We design the RFEM, which leverages the proposed hierarchical attention to highlight object-relevant regions in the reference feature, thereby facilitating the recurrent localization process.

2 Related Work

Cross-View Image Geo-Localization (CVIGL). CVIGL is an image-level geo-localization task. It aims to identify the geographic location of the camera by finding the most correlated reference image [7, 31, 49] or the most correlated position [17,28,39] on the reference image. Previous CVIGL works can be roughly divided into direct metric learning-based methods and geometry-based methods. The former [1,10,11,19,24,30,33,34,41,52,54] directly learns viewpoint-invariant features without considering geometric configurations. In contrast, the geometry-based approaches [20,23,27,31,32,38] explicitly reduce the viewpoint discrepancy between ground- and aerial-view images by using geometric configurations such as orientation information. However, existing CVIGL methods determine the approximate location of the query image on the reference image, thus being ineffective for object-level localization [36].

Cross-View Object Geo-Localization (CVOGL). CVOGL is an object-level geo-localization task aiming to determine the geographic location of an object indicated by a point prompt in a query image on the reference image. Considering the vital role of CVOGL and the limitations of current CVIGL methods, DetGeo [36] first proposes the CVOGL task and a corresponding method based on the spatial correlation of cross-view images and an object detection framework [26]. As a pioneering work, DetGeo encodes the point prompt into a Euclidean-distance-based positional map and fuses it with the query image features via element-wise addition. The fused features are then pooled to obtain a channel-only query vector, which interacts with the reference image features through matrix multiplication to produce a spatial attention map. This map is applied to the reference features to highlight the object region, and the enhanced reference features are finally passed through a convolution-based detection head to regress the object location. Subsequent works [13, 18, 46, 53] enhance cross-view feature aggregation and prompt embedding based on the framework of DetGeo [36]. Despite progress, existing CVOGL methods still rely on one-shot detection, which is sensitive to noisy features [2] and lacks mechanisms for error correction. In contrast, we model CVOGL as a recurrent localization process and propose ReCOT to address these limitations.

3 Method

Fig. 2 presents the architecture of ReCOT, which models CVOGL as a recurrent localization process. A set of learnable tokens is initialized to encode task-specific intent from the query image features and prompt embeddings. These tokens act as recurrent "questioners" that iteratively extract object-relevant cues from the reference features and refine the predicted location through cross-attention mechanisms. Additionally, we introduce two complementary methods to enhance this recurrent process: (1) a SAM-based knowledge distillation strategy, which transfers segmentation priors from the Segment Anything Model (SAM) to enhance prompt semantics, and (2) a Reference Feature Enhancement Module (RFEM),

which provides object-relevant reference features through the proposed hierarchical attention for the recurrent stage.

3.1 Recurrent Localization Framework

Motivation. As shown in Fig. 1, existing CVOGL approaches follow a one-shot detection paradigm, directly predicting the object location from enhanced reference features. However, such frameworks are often sensitive to feature noise and lack a mechanism for error correction [2, 3]. Fundamentally, CVOGL can be regarded as a cross-view matching problem, where recurrent strategies have shown superior robustness across domains [2, 37]. Inspired by this, we model CVOGL as a recurrent localization process. Moreover, unlike dense matching tasks [3, 9, 43], CVOGL is prompt-driven and focuses on object-level semantic matching. This calls for a representation that can both encode semantic intent and drive iterative refinement. To this end, we draw inspiration from the class token in vision transformers (ViT) [8], and introduce a set of learnable tokens that absorb task-specific semantics from the query and prompt. Acting as semantic carriers, these tokens can recurrently interact with the reference feature to enable step-wise prediction.

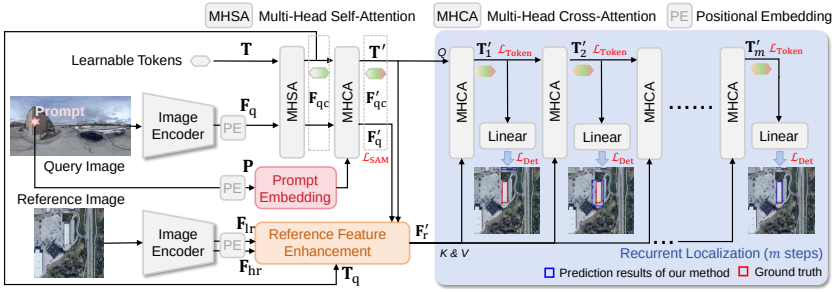

Fig. 2: Architecture of our recurrent cross-view object geo-localization transformer (ReCOT). ReCOT models the CVOGL task as a recurrent localization process, which leverages a set of learnable tokens to extract information from cross-view images and prompts to recurrently refine the prediction. Notably, all "MHCA" and "Linear" components in the recurrent steps share the same weights, following the typical design of recurrent architectures. We perform the positional embedding following DETR [4]

Structure. We initialize a set of learnable tokens $\mathbf{T} \in \mathbb{R}^{n \times c}$, where n and c denote the number of tokens and the feature dimension, respectively. To enable $\mathbf{T}$ to extract object-relevant information from the reference features, $\mathbf{T}$ needs to first acquire task-specific semantics from the query image feature $\mathbf{F}_{\mathrm{q}} \in \mathbb{R}^{h_{\mathrm{q}} w_{\mathrm{q}} \times c}$ and the point prompt embedding $\mathbf{P} \in \mathbb{R}^{c}$. Here, h_{q} and w_{q} denote the height and width of the query feature, respectively. Specifically, we concatenate $\mathbf{T}$ with $\mathbf{F}_{\mathrm{q}}$ along the spatial dimension, yielding $\mathbf{F}_{\mathrm{qc}} \in \mathbb{R}^{(n+h_{\mathrm{q}} w_{\mathrm{q}}) \times c}$. Following standard operations in Vision Transformers (ViT) [8], we apply self-attention to $\mathbf{F}_{\mathrm{qc}}$, allowing

$\mathbf{T}$ to aggregate global context from the query image. The resulting tokens are denoted as $\mathbf{T}_{\mathrm{q}}$. To further inject object-level intent, the embedded point prompt $\mathbf{P}$ interacts with $\mathbf{F}_{\mathrm{qc}}$ through cross-attention. This enables the prompt to semantically guide the tokens, allowing $\mathbf{T}_{\mathrm{q}}$ to further incorporate object-level context. After interaction, we denote the concatenated feature and tokens as $\mathbf{F}'_{\mathrm{qc}}$ and $\mathbf{T}'$, respectively. The $\mathbf{F}'_{\mathrm{qc}}$ and $\mathbf{T}_{\mathrm{q}}$ are further utilized to enhance reference features in RFEM. The enhanced reference feature is denoted as $\mathbf{F}'_{\mathrm{r}} \in \mathbb{R}^{h_{\mathrm{r}} w_{\mathrm{r}} \times c}$, where h_{r} and w_{r} denote the height and width of the reference feature, respectively.

After acquiring task-specific semantics from the query and prompt, we use $\mathbf{T}'$ to perform recurrent localization, as shown in Fig. 2. Let $\mathbf{T}'_i$ denote the token state at the i-th refinement step, where $i \in [0, 1, 2, \ldots, m]$. At each step, $\mathbf{T}'_i$ attends to the enhanced reference feature $\mathbf{F}'_{\mathrm{r}}$ to extract object-relevant cues and update the task-specific intent. Formally, the update process is defined as

$$\mathbf{T}'_{i+1} = \mathrm{MHCA}(\mathbf{T}'_i, \mathbf{F}'_{\mathrm{r}}) \tag{1}$$

where $\mathrm{MHCA}(\cdot, \cdot)$ denotes the multi-head cross-attention module. Here, $\mathbf{T}'_i$ serves as the query, while $\mathbf{F}'_{\mathrm{r}}$ acts as the key and value. This recurrent attention mechanism enables iterative refinement, where the tokens $\mathbf{T}'$ progressively accumulate task-specific semantics and extract increasingly relevant information from the fixed reference feature $\mathbf{F}'_{\mathrm{r}}$. The $\mathbf{T}'_i$ of each step is fed into a linear layer to predict an updated object location, allowing the model to gradually converge toward a precise localization.

In each refinement step i, we introduce a loss $\mathcal{L}_{\mathrm{Token}}$ to guide the generation of $\mathbf{T}'_i$. Specifically, since $\mathbf{T}'_i$ is expected to contain the task-specific intent in cross-view features, it should be able to highlight the required object area on $\mathbf{F}'_{\mathrm{r}}$. Therefore, in each refinement step, we first aggregate $\mathbf{T}'_i$ along the spatial dimension to generate a global embedding $\mathbf{T}''_i$, and use $\mathbf{T}''_i \in \mathbb{R}^{1 \times c}$ and $\mathbf{F}'_{\mathrm{r}}$ to produce an aggregation map $\widehat{\mathbf{m}}_{\mathrm{o}} \in \mathbb{R}^{1 \times h_{\mathrm{r}} \times w_{\mathrm{r}}}$. This can be expressed as

$$\mathbf{T}''_i = \mathrm{Sum}(\mathbf{T}'_i \cdot \mathrm{Softmax}(\mathbf{T}'_i)), \tag{2}$$

$$\widehat{\mathbf{m}}_{\mathrm{o}i} = \sigma(\mathbf{T}''_i \mathbf{F}'^{\mathsf{T}}_{\mathrm{r}}), \tag{3}$$

where $\sigma(\cdot)$ and $\mathrm{Softmax}(\cdot)$ denote the sigmoid and softmax functions, respectively. $\mathrm{Sum}(\cdot)$ denotes summation along the spatial dimension. We then utilize a box-level mask $\mathbf{m}_{\mathrm{o}i}$ produced using the ground truth box to supervise the generation of $\widehat{\mathbf{m}}_{\mathrm{o}}$, which can be expressed as

$$\mathcal{L}_{\mathrm{Token}_i}(\mathbf{m}_{\mathrm{o}}, \widehat{\mathbf{m}}_{\mathrm{o}i}) = \mathcal{L}_{\mathrm{bce}}(\mathbf{m}_{\mathrm{o}}, \widehat{\mathbf{m}}_{\mathrm{o}i}) + \mathcal{L}_{\mathrm{dice}}(\mathbf{m}_{\mathrm{o}}, \widehat{\mathbf{m}}_{\mathrm{o}i}), \tag{4}$$

where $\mathcal{L}_{\mathrm{bce}}(\cdot, \cdot)$ and $\mathcal{L}_{\mathrm{dice}}(\cdot, \cdot)$ denote the binary cross-entropy loss and the Dice Loss [25], respectively.

How ReCOT works. As shown in Fig. 3(a), previous one-shot detection CVOGL approaches [13, 18, 36, 46, 53] rely on a single forward information aggregation and are thus sensitive to noisy features [2, 3]. Our ReCOT adopts a

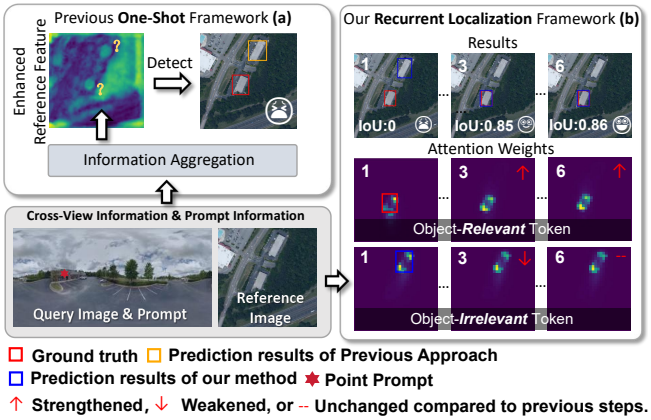

Fig. 3: Comparison between the previous one-shot framework and our recurrent localization framework. (a) Previous approaches [13, 18, 36, 46, 53] rely on single-shot information aggregation, which is sensitive to noisy enhanced features and often leads to incorrect predictions. (b) Our ReCOT iteratively refines predictions through learnable tokens. The attention weight visualizations show that object-relevant tokens progressively focus and strengthen around the target, while object-irrelevant tokens weaken and stabilize to background patterns.

recurrent localization mechanism that iteratively refines predictions. The visualization in Fig. 3(b) of cross-attention weights between tokens and the reference feature reveals the inner dynamics of this refinement process. It can be seen that different tokens focus on different regions of the reference feature, indicating a form of token-level specialization. For object-relevant tokens, their attention gradually concentrates and intensifies around the object region across recurrent steps, reflecting the ability to correct early prediction error and progressively refine the prediction. In contrast, object-irrelevant tokens experience a decrease in their attention responses and eventually stabilize to background patterns once they no longer contribute to the object localization. This behavior highlights the competitive nature of token updates, *i.e.*, multiple tokens initially compete to explain different parts of the reference feature [4], but those correlated with the object receive positive feedback and their attention weights are amplified over recurrent steps, leading to iterative convergence. Such dynamics demonstrate the effectiveness of recurrent refinement mechanism in suppressing irrelevant regions while enhancing object-relevant cues.

3.2 SAM-based Knowledge Distillation

Motivation. Point prompt understanding is essential for CVOGL to correctly locate objects. However, point prompt itself suffers from semantic ambiguity, leading to unsatisfactory performance [15]. To address this, we propose a SAM-based knowledge distillation strategy to boost the prompt understanding of ReCOT. The incorporation of SAM [15] is motivated by an observation that SAM

can give a mask with a clear indication of the required object using point prompts and corresponding images. However, directly applying SAM during inference incurs large computation overhead. Hence, we leverage predictions of SAM as supervision signals, transferring its knowledge through knowledge distillation.

Structure. We extract the query feature $\mathbf{F}'_{\mathrm{q}}$ from the previous concatenated feature $\mathbf{F}'_{\mathrm{qc}}$. Notably, the $\mathbf{F}'_{\mathrm{q}}$ has been interacted with prompt embedding and is supposed to contain the object-level semantics. Therefore, we process it using a lightweight convolutional head followed by a sigmoid activation to generate a segmentation map $\widehat{\mathbf{m}}_{\mathrm{q}}$. Meanwhile, we use SAM to generate a pseudo ground-truth mask $\mathbf{m}_{\mathrm{SAM}}$ from the query image and point prompt. This mask is used to supervise the segmentation out of $\mathbf{F}'_{\mathrm{q}}$ via $\mathcal{L}_{\mathrm{SAM}}$, which can be defined as

$$\mathcal{L}_{\mathrm{SAM}}(\mathbf{m}_{\mathrm{SAM}}, \widehat{\mathbf{m}}_{\mathrm{q}}) = \mathcal{L}_{\mathrm{bce}}(\mathbf{m}_{\mathrm{SAM}}, \widehat{\mathbf{m}}_{\mathrm{q}}) + \mathcal{L}_{\mathrm{dice}}(\mathbf{m}_{\mathrm{SAM}}, \widehat{\mathbf{m}}_{\mathrm{q}}). \tag{5}$$

3.3 Reference Feature Enhancement Module

Motivation. In CVOGL, the reference feature $\mathbf{F}_{\mathrm{r}}$ extracted by the backbone encoder is typically generic and background-dominated, lacking object-level specificity before prompt interaction [18, 36]. In our framework, guiding $\mathbf{F}_{\mathrm{r}}$ to focus on the expected object indicated by the prompt can significantly ease the downstream recurrent localization [3, 37]. To this end, we propose the Reference Feature Enhancement Module (RFEM), which enhances reference features into a more object-aware representation $\mathbf{F}'_{\mathrm{r}}$. Unlike previous approaches [13, 18, 36, 46] that attempt to clearly highlight the object features through one-shot feature enhancement, RFEM serves as a preparatory module to filter irrelevant features and provide more object-relevant information for subsequent recurrent localization. The core of RFEM lies in its hierarchical attention design. We first perform spatial attention to highlight the query-relevant regions in the reference feature, as the expected object is likely confined to these regions. We then apply cross-attention guided by the query and prompt cues to refine these regions with object-level semantics. This spatial-to-cross hierarchy yields a cleaner and more informative reference feature $\mathbf{F}'_{\mathrm{r}}$, which significantly benefits the iterative localization process of ReCOT.

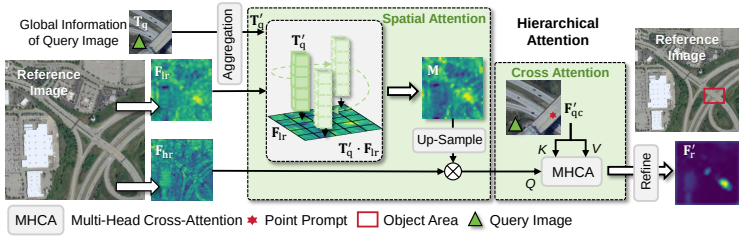

Fig. 4: Architecture of reference feature enhancement module (RFEM). RFEM enhances reference features through a hierarchical attention pipeline to obtain $\mathbf{F}'_{\mathrm{r}}$.

Structure. As illustrated in Fig. 4, we utilize two levels of reference features, *i.e.*, a low-resolution semantic feature $\mathbf{F}_{\mathrm{lr}} \in \mathbb{R}^{h_{\mathrm{r}} \times w_{\mathrm{r}} \times c}$ that captures scene semantics, and a high-resolution detailed feature $\mathbf{F}_{\mathrm{hr}} \in \mathbb{R}^{2h_{\mathrm{r}} \times 2w_{\mathrm{r}} \times c}$ that preserves fine-grained spatial structures [50].

Spatial Attention: We leverage the spatial attention to highlight query-relevant features. Specifically, we first aggregate the query tokens $\mathbf{T}_{\mathrm{q}}$ into a global descriptor $\mathbf{T}'_{\mathrm{q}}$ (Eq. (2)). Notably, the $\mathbf{T}'_{\mathrm{q}}$ contains only the global information of the query image without prompt guidance. This descriptor correlates with $\mathbf{F}_{\mathrm{lr}}$ via a dot-product operation to produce a spatial attention map as

$$\mathbf{M} = \sigma(\mathbf{T}'_{\mathrm{q}} \mathbf{F}_{\mathrm{lr}}^{\mathsf{T}}), \tag{6}$$

which highlights query-relevant regions. The $\sigma(\cdot)$ denotes the sigmoid function. The attention map is then up-sampled and applied to $\mathbf{F}_{\mathrm{hr}}$ by element-wise multiplication, narrowing the focus to potential object areas.

Cross Attention: We leverage the cross-attention to incorporate detailed query features and prompt semantics for further refining the reference feature. The filtered $\mathbf{F}_{\mathrm{hr}}$ is subsequently refined via cross-attention with $\mathbf{F}_{\mathrm{qc}}$, which encodes fine-grained prompt cues and detailed query features. This cross-attention step sharpens the object-relevant responses and injects object-level semantics into the reference representation.

Finally, the updated feature is down-sampled and refined through spatial attention [40] and self-attention [8] to generate $\mathbf{F}'_{\mathrm{r}}$, which is then passed to the recurrent localization stage in ReCOT.

Why multi-resolution reference features? The reference image typically contains both global scene information and fine-grained object details. The low-resolution semantic reference feature $\mathbf{F}_{\mathrm{lr}}$ extracted from deeper layers of the backbone encodes more scene-level context but loses spatial details due to down-sampling. Therefore, we utilize $\mathbf{F}_{\mathrm{lr}}$ to enhance query-relevant regions. Conversely, the high-resolution reference features $\mathbf{F}_{\mathrm{hr}}$ preserve fine structural details but are dominated by background noise. Thus, it needs to be refined by the attention map produced by $\mathbf{F}_{\mathrm{lr}}$, and then RFEM can utilize it to perform object-level enhancement. This design leverages the advantages of multi-resolution features, which are crucial for object-level feature enhancement in CVOGL [13].

3.4 Loss Function

Our training objective $\mathcal{L}$ is defined as a weighted sum of the localization loss $\mathcal{L}_{\mathrm{local}}$ and the auxiliary loss $\mathcal{L}_{\mathrm{aux}}$. It can be defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{local}} + \alpha \mathcal{L}_{\mathrm{aux}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{local}}$ denotes the sum of DETR-style detection losses $\mathcal{L}_{\mathrm{Det}}$ computed at each recurrent localization step (see Fig. 2). The auxiliary loss $\mathcal{L}_{\mathrm{aux}}$ is the sum of $\mathcal{L}_{\mathrm{Token}_i}$ across all recurrent steps and the SAM-based distillation loss $\mathcal{L}_{\mathrm{SAM}}$. The balancing coefficient α is set to 1 in all experiments.

4 Experiments

4.1 Datasets

CVOGL dataset [36] divides the task into "Ground → Satellite" task and "Drone → Satellite" task. It contains 6,239 pairs of "Ground → Satellite" view and 6,239 pairs of "Drone → Satellite" view query and reference images. The ground view query images, drone view query images, and satellite view reference images are sized to 512 × 256, 256 × 256, and 1024 × 1024, respectively. Each cross-view task uses 4,343 pairs for training, 923 pairs for validation, and 973 pairs for testing. Our experiments utilize the training set to train the model and select the best model on the validation set for evaluation on the test set.

CVOGL-few-shot [13] dataset introduces additional object categories and is designed for the "Drone → Satellite" task. It annotates 4 new categories beyond the CVOGL [36] dataset, *e.g.*, lake, parking, slide, and port. The dataset consists of 28 training samples and 24 test samples, with 7 training instances per category across 4 new classes. This dataset is used only for few-shot experiments.

4.2 Evaluation Metrics

Following DetGeo [36], we adopt Acc@0.25 and Acc@0.50 for evaluation. The Acc@0.25 and Acc@0.50 measure the prediction accuracy under a specific intersection over union (IoU) threshold t between the predicted box b_p and ground-truth box b_g as

$$\text{Acc@}t = \frac{1}{n}\sum_{i=1}^{n}\psi(t), \tag{8}$$

where

$$\psi(t) = \begin{cases} 1, \text{IoU}(b_p, b_g) \geq t \\ 0, \text{otherwise} \end{cases}, \tag{9}$$

$$\text{IoU}(b_p, b_g) = \frac{|b_p \cap b_g|}{|b_p \cup b_g|}. \tag{10}$$

For each pair of the query and reference image, we select the box with the highest confidence output by our model as the final prediction box. Additionally, we use the parameter and frames per second (FPS) to show the model efficiency. Higher Acc@0.25, higher Acc@0.50, higher FPS, and fewer parameters denote better performance.

4.3 Implementation Details

We conduct all experiments on 4 NVIDIA GeForce RTX 4090 GPUs. For training, we adopt the AdamW [22] as the optimizer and set the initial learning rate to 2.5×10^{-5}, the weight decay rate to 1×10^{-4}, and the batch size to 16. Since CVOGL is a relatively new task, we select some CVIGL approaches [11, 19, 31, 41, 49, 52] and representative CVOGL approaches [13, 18, 36, 46, 48, 51, 53], as

our comparison methods. We produce the results of CVIGL approaches following previous works [13,18,36,46]. All CVOGL methods are trained to convergence to ensure a fair comparison. Additionally, we adopt Swin Transformer (Swin-t) [21] as the image encoder for its superior performance in various fields [5, 29, 44, 47]. During training, we use a larger value $m = 6$ so that the model learns a stable multi-step refinement process, while at inference we use m that is selected on the validation set to balance accuracy and efficiency.

4.4 Ablation Study

Table 1: Ablation study on the recurrent localization steps m of ReCOT in terms of Average IoU (Avg.IoU)↑, Acc@0.50(%) ↑, and FPS ↑ on the *validation set* of CVOGL dataset. We test the inference speed using one NVIDIA GeForce RTX 4090 GPU. **Bold** and Underline indicate the best and second-best results, respectively. We set $m = 5$ for optimal results.

Steps m	Ground→Satellite			Drone→Satellite		
	Avg.IoU	Acc@0.50	FPS	Avg.IoU	Acc@0.50	FPS
1	0.354	42.25	26.9	0.553	67.06	27.2
2	0.359	42.90	26.2	0.554	67.28	26.5
3	**0.365**	43.55	25.6	0.555	67.17	25.7
4	0.362	43.34	24.8	0.555	67.61	24.9
5	**0.365**	**43.66**	24.1	**0.557**	**67.82**	24.3
6	0.362	43.34	23.4	0.556	67.61	23.5

How to determine m on the dataset? Since recurrent localization constitutes the core mechanism of ReCOT, selecting an appropriate number of refinement steps m is crucial for exploiting its capability. We determine m following a principled and practically applicable strategy that aligns with common practices of existing iterative refinement approaches [2,12,37]. Specifically, m can be chosen by monitoring the performance gain per step on the *validation set* and stopping when the gain becomes negligible. Following this strategy, we investigate the impact of m on ReCOT using the validation set of CVOGL dataset [36]. Increasing m from 1 (one-shot prediction) to 5 yields consistent improvements, with the relatively optimal results achieved at m=5. Further increasing m beyond the optimal point does not bring additional gains and may slightly degrade performance, which can be attributed to over-refinement and overfitting in deeper iterations [3, 14, 43]. Therefore, unless otherwise specified, we set m=5 and keep this setting in other reported experiments.

Explanation of performance saturation with excessive steps. The recurrent mechanism in ReCOT progressively refines localization by predicting residual corrections. As shown in Table 1, performance improves and then gradually saturates with minor fluctuations. In early iterations, residuals correct coarse geometric misalignments, yielding noticeable gains. Once the prediction becomes

sufficiently accurate, subsequent residuals mainly capture subtle perturbations, leading to diminishing returns, as observed in other iterative refinement methods [2,3]. From an optimization perspective, when the objective is already near a local optimum, further iterations bring marginal improvements while increasing computational cost [37]. Therefore, we set $m = 5$ as a balanced trade-off between accuracy and efficiency.

Table 2: Ablation study on the components of ReCOT (m=5) in terms of Acc@0.25(%) ↑ and Acc@0.50(%) ↑ on the *test set* of CVOGL dataset.

Component	Ground→Satellite		Drone→Satellite	
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
w/o RFEM	46.45	42.24	50.15	46.04
w/o $\mathcal{L}_{\text{SAM}}$	49.74	44.81	70.91	65.78
w/o $\mathcal{L}_{\text{Token}}$	50.57	47.58	72.05	66.80
ReCOT (Ours)	52.00	48.10	77.60	72.05

Effect of Complementary Components in ReCOT. Table 2 presents the ablation study of the complementary components in ReCOT on the CVOGL test set. Removing the RFEM module (w/o RFEM) leads to a noticeable performance drop, confirming that early reference feature enhancement is critical for guiding tokens to focus on relevant regions. Similarly, removing the SAM-based distillation loss $\mathcal{L}_{\text{SAM}}$ (w/o $\mathcal{L}_{\text{SAM}}$) results in performance degradation, indicating the importance of accurate prompt semantic understanding. The full ReCOT model achieves the best performance across both scenarios. In addition, removing the token-guidance loss $\mathcal{L}_{\text{Token}}$ (w/o $\mathcal{L}_{\text{Token}}$) also degrades performance, which highlights its role in encouraging the learnable tokens to accurately capture object-relevant areas during recurrent refinement. These results collectively validate that RFEM, $\mathcal{L}_{\text{SAM}}$, and $\mathcal{L}_{\text{Token}}$ complement each other, contributing to robust performance of ReCOT.

Table 3: Ablation study inside the RFEM (m=5) in terms of Acc@0.25(%) ↑ and Acc@0.50(%) ↑ on the *test set* of CVOGL dataset.

RFEM	Ground→Satellite		Drone→Satellite	
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
w/o $\mathbf{M}$	51.18	47.48	75.44	70.30
Replace $\mathbf{F}_{\text{lr}}$ with $\mathbf{F}_{\text{hr}}$	48.00	44.19	75.85	70.09
Replace $\mathbf{F}_{\text{hr}}$ with $\mathbf{F}_{\text{lr}}$	47.49	43.28	75.64	67.11
ReCOT (Ours)	52.00	48.10	77.60	72.05

Effect of Components Within the RFEM. Table 3 investigates the contributions of components within the RFEM. Removing the weight matrix $\mathbf{M}$ (w/o $\mathbf{M}$) leads to a noticeable performance drop. This confirms that the spatial attention stage benefits the reference feature enhancement. Furthermore, replacing the low-resolution semantic feature $\mathbf{F}_{\text{lr}}$ with the high-resolution feature $\mathbf{F}_{\text{hr}}$ and replacing $\mathbf{F}_{\text{hr}}$ with $\mathbf{F}_{\text{lr}}$ both result in performance degradation, highlighting the importance of leveraging multi-resolution reference features in RFEM.

Please refer to the supplementary material for more ablation study results on hyper-parameters.

4.5 Comparison Results

Table 4: Comparisons in terms of Acc@0.25(%) ↑, Acc@0.50(%) ↑, Parameter (M) ↓, and FPS ↑ on the CVOGL dataset. Following Table 1, we set m=5 for final results. We test the inference speed on the Drone→Satellite task using one NVIDIA GeForce RTX 4090 GPU. **Bold** and Underline indicate the best and second-best results, respectively.

Method	Ground→Satellite		Drone→Satellite		Param	FPS
	Test Set		*Test Set*			
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50		
CVM-Net [11]	4.73	0.51	20.14	3.29	–	–
RK-Net [19]	7.40	0.82	19.22	2.67	–	–
L2LTR [41]	10.69	2.16	38.95	6.27	–	–
Polar-SAFA [31]	20.66	3.19	37.41	6.58	–	–
TransGeo [52]	21.17	2.88	35.05	6.47	–	–
SAFA [31]	22.20	3.08	37.41	6.58	–	–
GeoDTR+ [49]	14.19	5.14	16.03	4.73	–	–
DetGeo [36]	45.43	42.24	61.97	57.66	<u>73.8</u>	**29.5**
VAGeo [18]	48.21	45.22	66.19	61.87	–	–
TROGeo [46]	<u>51.08</u>	<u>46.56</u>	76.16	68.96	71.3	13.5
GeoFormer [48]	49.54	45.32	73.79	68.96	739.1	11.1
AttenGeo [53]	50.57	46.15	70.71	62.08	–	–
ReCOT (m=1, one-shot)	49.74	46.25	**78.31**	<u>71.74</u>	**29.9**	<u>27.2</u>
ReCOT (Ours)	**52.00**	**48.10**	<u>77.60</u>	**72.05**	**29.9**	24.3

Quantitative Results. Table 4 compares ReCOT with previous approaches on CVOGL. ReCOT consistently outperforms all competitors, achieving SOTA results on almost all test metrics. These gains are obtained with about 60% fewer parameters, and ReCOT maintains competitive inference speed. Moreover, in scenarios with larger viewpoint gaps (Ground→Satellite), increasing m from 1 to 5 yields substantial performance gains. This clearly highlights the robustness and potential of our recurrent design in addressing challenging scenarios.

To further assess robustness, we evaluate few-shot learning on the CVOGL-few-shot dataset [13] (Table 5). Following OCGNet [13], we initialize ReCOT from the pretrained checkpoints in Table 4 and fine-tune on CVOGL-few-shot.

Table 5: Few-shot learning performance between DetGeo [36], OCGNet [13], and our ReCOT in terms of Acc@0.25(%) ↑ and Acc@0.50(%) ↑ on the *test set* of CVOGL-few-shot dataset [13]. **Bold** and Underline indicate the best and second-best results, respectively.

Method	Drone→Satellite	
	Acc@0.25	Acc@0.50
DetGeo [36]	16.67	16.67
GeoFormer [48]	<u>45.83</u>	<u>29.17</u>
OCGNet [13]	29.17	25.00
ReCOT(m=1)	37.50	25.00
ReCOT(m=2)	41.67	<u>29.17</u>
ReCOT(m=4)	<u>45.83</u>	**33.33**
ReCOT(m=6)	**50.00**	**33.33**

ReCOT clearly surpasses previous CVOGL approaches [13, 36] under few-shot conditions, demonstrating the effectiveness of our design. Besides, increasing the number of recurrent steps m leads to a substantial accuracy gain, highlighting the advantage of our recurrent refinement mechanism under more challenging scenarios (few-shot). Additional comparisons of the backbone can be found in the supplementary material.

Fig. 5: Visualization of some representative results produced by our ReCOT and the representative previous one-shot framework [36].

Qualitative Results. Figure 5 shows representative examples. As recurrent steps proceed, ReCOT progressively refines the bounding boxes and produces more accurate localizations than the single-shot prediction of DetGeo [36]. We set m at the dataset level for simplicity and reproducibility. Although the optimal refinement depth may vary per instance, the refinement process consistently

converges in practice, demonstrating the stability of our recurrent localization mechanism. Future work will investigate adaptive stopping criteria for m.

4.6 Discussion

The Role and Core Advantage of the Recurrent Refinement. Our recurrent localization mechanism fundamentally changes how the model performs CVOGL. Specifically, with m=1, ReCOT functionally behaves as a one-shot predictor performing a single forward prediction without iterative refinement, and achieves performance comparable to existing one-shot approaches in challenging settings (Ground→Satellite, Acc@0.50: TROGeo [46] 46.56, ReCOT(m=1) 46.25; Few-shot, Acc@0.50: OCGNet [13] 25.00, ReCOT(m=1) 25.00). However, by increasing m, ReCOT progressively corrects its initial predictions using the same network and parameters, and gradually achieves the best performance without architectural changes or retraining. This advantage becomes more pronounced in harder few-shot scenarios (Table 5), highlighting its robustness. Therefore, our recurrent mechanism effectively lifts the performance ceiling.

Although certain complementary components yield noticeable performance gains (Table 2), their primary role is to enhance the starting point of localization, whereas the recurrent refinement governs the convergence behavior and determines the final attainable accuracy. Our improvements are therefore complementary, with recurrent localization serving as the core driver of robust performance in difficult conditions. This structural capability is absent in prior one-shot architectures [13, 18, 36, 46, 48, 51, 53], offering greater potential to address more complex scenarios.

Please refer to the supplementary material for more discussions.

5 Conclusions

In this paper, we propose ReCOT, which models the CVOGL as a recurrent localization process. ReCOT introduces learnable tokens to encode task-specific semantics and recurrently refine predictions. We further incorporate a SAM-based knowledge distillation scheme to improve prompt understanding without incurring additional inference costs, and a RFEM to produce object-aware reference features via a hierarchical attention strategy. Extensive experiments on the public benchmarks demonstrate that ReCOT achieves state-of-the-art performance with significantly fewer parameters and competitive inference speed.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62301484, in part by the Jinhua Science and Technology Bureau Project under Grant 2026-1-022.

References

1. Cai, S., Guo, Y., Khan, S., Hu, J., Wen, G.: Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8390–8399 (2019)
2. Cao, S.Y., Hu, J., Sheng, Z., Shen, H.L.: Iterative deep homography estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1869–1878 (2022)
3. Cao, S.Y., Zhang, R., Luo, L., Yu, B., Sheng, Z., Li, J., Shen, H.L.: Recurrent homography estimation using homography-guided image warping and focus transformer. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9833–9842 (2023)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. arXiv e-prints arXiv:2005.12872 (2020)
5. Chi, K., Yuan, Y., Wang, Q.: Trinity-Net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023)
6. Chini, M., Pierdicca, N., Emery, W.J.: 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 (2009)
7. Deuser, F., Habel, K., Oswald, N.: Sample4Geo: Hard negative sampling for cross-view geo-localisation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16847–16856 (October 2023)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for image recognition at scale. arXiv e-prints arXiv:2010.11929 (2021)
9. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching. IEEE Conference on Computer Vision and Pattern Recognition (2024)
10. Guo, Y., Choi, M., Li, K., Boussaid, F., Bennamoun, M.: Soft exemplar highlighting for cross-view image-based geo-localization. IEEE Transactions on Image Processing **31**, 2094–2105 (2022)
11. Hu, S., Feng, M., Nguyen, R.M.H., Lee, G.H.: CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7258–7267 (2018)
12. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: FlowFormer: A transformer architecture for optical flow. ECCV (2022)
13. Huang, Z., Aryal, J., Nahavandi, S., Lu, X., Lim, C.P., Wei, L., Zhou, H.: Object-level cross-view geolocalization with location enhancement and multihead cross attention. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **18**, 22880–22890 (2025)
14. Hur, J., Roth, S.: Iterative residual refinement for joint optical flow and occlusion estimation. In: arXiv e-prints arXiv:1904.05290 (2019)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)

16. Kumar, A., Kim, H., Hancke, G.P.: Environmental monitoring systems: A review. IEEE Sensors Journal **13**(4), 1329–1339 (2013)
17. Lentsch, T., Xia, Z., Caesar, H., Kooij, J.F.P.: SliceMatch: Geometry-guided aggregation for cross-view pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17225–17234 (June 2023)
18. Li, Z., Yuan, X., Liu, W., Xu, X.: VAGeo: View-specific attention for cross-view object geo-localization. ArXiv **2501.07194** (2025)
19. 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 **31**, 3780–3792 (2022)
20. Liu, L., Li, H.: Lending orientation to neural networks for cross-view geo-localization. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5617–5626 (2019)
21. 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 IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
22. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam. arXiv e-prints arXiv:1711.05101 (2017)
23. Lu, X., Li, Z., Cui, Z., Oswald, M.R., Pollefeys, M., Qin, R.: Geometry-aware satellite-to-ground image synthesis for urban areas. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 856–864 (2020)
24. Lu, X., Luo, S., Zhu, Y.: It's Okay to Be Wrong: Cross-view geo-localization with step-adaptive iterative refinement. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–13 (2022)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. arXiv e-print arXiv:1606.04797s (2016)
26. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv e-prints arXiv:1804.02767 (2018)
27. Regmi, K., Shah, M.: Bridging the domain gap for ground-to-aerial image matching. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 470–479 (2019)
28. Sarlin, P.E., DeTone, D., Yang, T.Y., Avetisyan, A., Straub, J., Malisiewicz, T., Bulo, S.R., Newcombe, R., Kontschieder, P., Balntas, V.: OrienterNet: Visual localization in 2d public maps with neural matching. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21632–21642 (2023)
29. Shi, L., Chen, Y., Liu, M., Guo, F.: DuST: Dual swin transformer for multi-modal video and time-series modeling. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4537–4546 (June 2024)
30. Shi, Y., Li, H.: Beyond Cross-view Image Retrieval: Highly accurate vehicle localization using satellite image. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16989–16999 (2022)
31. Shi, Y., Liu, L., Yu, X., Li, H.: Spatial-aware feature aggregation for image based cross-view geo-localization. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
32. Shi, Y., Yu, X., Campbell, D., Li, H.: Where Am I Looking At? joint location and orientation estimation by cross-view matching. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4063–4071 (2020)
33. Shi, Y., Yu, X., Liu, L., Campbell, D., Koniusz, P., Li, H.: Accurate 3-DoF camera geo-localization via ground-to-satellite image matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 2682–2697 (2022)

34. Shi, Y., Yu, X., Liu, L., Zhang, T., Li, H.: Optimal feature transport for cross-view image geo-localization. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 11990–11997 (Apr 2020)
35. Singamaneni, P.T., Bachiller-Burgos, P., Manso, L.J., Garrell, A., Sanfeliu, A., Spalanzani, A., Alami, R.: A survey on socially aware robot navigation: Taxonomy and future challenges. The International Journal of Robotics Research **43**(10), 1533–1572 (2024)
36. 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 **61**, 1–16 (2023)
37. Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: arXiv e-prints arXiv:2003.12039 (2020)
38. Toker, A., Zhou, Q., Maximov, M., Leal-Taixe, L.: 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 (CVPR). pp. 6488–6497 (June 2021)
39. Wang, X., Xu, R., Cui, Z., Wan, Z., Zhang, Y.: Fine-grained cross-view geo-localization using a correlation-aware homography estimator. ArXiv **abs/2308.16906** (2023)
40. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proc. Eur. Conf. Comput. Vis. pp. 3–19 (2018)
41. Yang, H., Lu, X., Zhu, Y.: Cross-view geo-localization with layer-to-layer transformer. In: Advances in Neural Information Processing Systems. vol. 34, pp. 29009–29020. Curran Associates, Inc. (2021)
42. Yao, J., Hong, D., Gao, L., Chanussot, J.: Multimodal remote sensing benchmark datasets for land cover classification. In: 2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). pp. 4807–4810 (2022)
43. Yu, J., Chang, J., He, J., Zhang, T., Yu, J., Wu, F.: Adaptive spot-guided transformer for consistent local feature matching. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21898–21908 (2023)
44. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5728–5739 (June 2022)
45. Zhai, R., Zou, J., Gan, V.J., Han, X., Wang, Y., Zhao, Y.: Semantic enrichment of bim with indoorgml for quadruped robot navigation and automated 3d scanning. Automation in Construction **166**, 105605 (2024)
46. Zhang, Q., Zhu, Y.: Breaking rectangular shackles: Cross-view object segmentation for fine-grained object geo-localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8197–8206 (October 2025)
47. Zhang, T., Zhang, X., Zhu, X., Wang, G., Han, X., Tang, X., Jiao, L.: Multistage enhancement network for tiny object detection in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–12 (2024)
48. Zhang, X., Cao, S.Y., Yu, Z., Wu, Z., Zhang, X., Bai, X., Yu, B., Shen, H.L.: GeoFormer: Boosting object distinguishing and prompt understanding for cross-view object geo-localization. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025)
49. Zhang, X., Li, X., Sultani, W., Chen, C., Wshah, S.: GeoDTR: Toward generic cross-view geolocalization via geometric disentanglement. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10419–10433 (2024)

50. Zhang, X., Zhang, X., Cao, S.Y., Yu, B., Zhang, C., Shen, H.L.: MRF^{3}Net: An infrared small target detection network using multireceptive field perception and effective feature fusion. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)
51. Zhang, Z., Sun, Y., Ding, L., Zhang, H., Wang, Y., Ding, W.: HALO: Hierarchical adaptive feature learning and cross-view interaction for object-level geo-localization. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–15 (2025)
52. Zhu, S., Shah, M., Chen, C.: TransGeo: Transformer is all you need for cross-view image geo-localization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1162–1171 (June 2022)
53. Zhu, X.L.Y.: Improving cross-view object geo-localization: A dual attention approach with cross-view interaction and multi-scale spatial features. arXiv e-prints arXiv:2510.27139 (2025)
54. Zhu, Y., Sun, B., Lu, X., Jia, S.: Geographic semantic network for cross-view image geo-localization. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022)