RECURRENT CROSS-VIEW OBJECT GEO-LOCALIZATION

Anonymous authors
Paper under double-blind review

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 task, 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 reformulates CVOGL as a recurrent localization task. 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 a hierarchical attention to emphasize object-relevant regions in the reference features. Extensive experiments on standard CVOGL benchmarks demonstrate that ReCOT achieves state-of-the-art (SOTA) performance while reducing parameters by 60% compared to previous SOTA approaches. Our code will be made available upon acceptance.

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 the reference image Sun et al. (2023). 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 Yao et al. (2022), disaster monitoring Chini et al. (2009); Kumar et al. (2013), and robot navigation Singamaneni et al. (2024); Zhai et al. (2024). However, the view gap poses challenges for CVOGL Sun et al. (2023).

Recent cross-view image geo-localization (CVIGL) works Hu et al. (2018); Shi et al. (2019); Zhu et al. (2022a); Yang et al. (2021); Lin et al. (2022) have demonstrated their superiority in handling view gaps. However, CVIGL approaches are fundamentally designed for camera-level localization using retrieval-based approaches Deuser et al. (2023); Zhang et al. (2024b); Shi et al. (2019) or fine-grained approaches Sarlin et al. (2023); Wang et al. (2023). However, 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 Sun et al. (2023), which is insufficient for precise object-level localization.

To address this, CVOGL approaches emerge recently. Existing approaches Sun et al. (2023); Li et al. (2025); Huang et al. (2025) 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 Huang et al. (2025) 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 is sensitive to the quality of the enhanced feature Cao et al. (2022). It

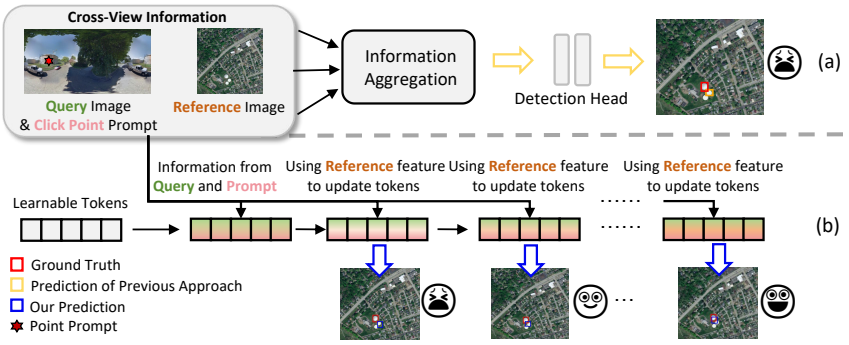

Figure 1: Comparison between the framework of previous CVOGL approaches and ours. (a) Previous approaches treat the CVOGL as a prompt-based detection task, where the model directly regresses the object location based on information aggregation once. (b) Our framework reformulates the CVOGL as a recurrent localization problem, where the model iteratively refines the localization through a set of learnable tokens. Please refer to the zoomed-in view for better visualization.

lacks a correction mechanism for early-stage prediction errors, making them vulnerable to noise in features, $i.e.$, the model cannot give correct localization once the enhanced feature leads to a wrong prediction Cao et al. (2023); Sun et al. (2023).

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 Yu et al. (2023); Cao et al. (2022; 2023), ReCOT reformulates the CVOGL task as a recurrent localization problem, as shown in Fig. 1(b). This serves as the main difference between previous approaches Sun et al. (2023); Li et al. (2025); Huang et al. (2025) 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).

Nevertheless, the success of this recurrent strategy in our token-driven framework relies on the semantic clarity of the prompts Li et al. (2025) and the quality of reference features Teed & Deng (2020). 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 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 the standard CVOGL benchmark Sun et al. (2023). It achieves state-of-the-art (SOTA) performance while reducing parameter count by 60% compared to the previous SOTA approach Huang et al. (2025) (29.9M vs. 74.8M), and runs at a competitive inference speed. In summary, our contributions are as follows:

- We propose ReCOT, a novel framework for CVOGL that reformulates the task as a recurrent localization problem, 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 a 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 aims to determine the camera's geographic location by matching a ground-view query image with the most correlated reference image Deuser et al. (2023); Zhang et al. (2024b); Shi et al. (2019) or position Sarlin et al. (2023); Wang et al. (2023); Lentsch et al. (2023). Existing CVIGL approaches can be grouped into metric learning-based methods Lu et al. (2022); Zhu et al. (2022b); Cai et al. (2019); Shi et al. (2020b); Guo et al. (2022); Shi & Li (2022); Shi et al. (2022); Hu et al. (2018); Yang et al. (2021); Lin et al. (2022); Zhu et al. (2022a), which learn viewpoint-invariant features, and geometry-based methods Shi et al. (2020a); Lu et al. (2020); Toker et al. (2021); Liu & Li (2019); Regmi & Shah (2019); Shi et al. (2019), which exploit orientation or structural cues to reduce viewpoint gaps. However, CVIGL methods only provide camera-level localization and cannot accurately pinpoint object-level targets Sun et al. (2023).

Cross-View Object Geo-Localization (CVOGL). CVOGL focuses on locating a specific object indicated by prompts in a query image. DetGeo Sun et al. (2023) first formalized this task and proposed a detection-based framework. Subsequent works, such as VaGeo Li et al. (2025) and OCGNet Huang et al. (2025), enhanced cross-view feature aggregation and prompt embedding. Despite progress, existing CVOGL methods still rely on one-shot detection, which is sensitive to noisy features and lacks mechanisms for error correction Cao et al. (2022). In contrast, we reformulate CVOGL as a recurrent localization problem and propose ReCOT to address these limitations.

3 Methodology

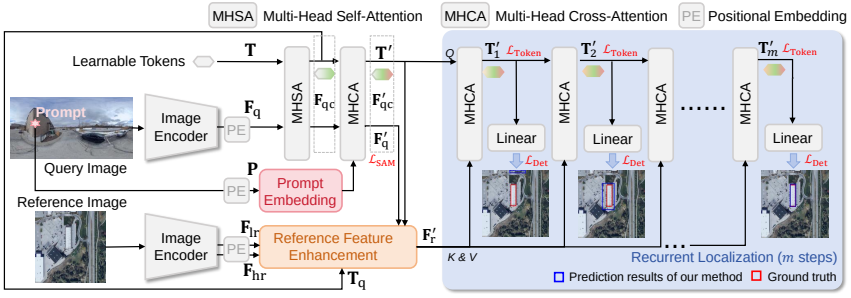

Figure 2: Architecture of our recurrent cross-view object geo-localization transformer (ReCOT). ReCOT reformulates the CVOGL task as a recurrent localization task, which leverages a set of learnable tokens to extract information from cross-view images and prompts to recurrently refine the prediction. Notably, all recurrent steps in ReCOT share the same "MHCA" and "Linear" component.

Fig. 2 presents the architecture of ReCOT, which reformulates CVOGL as a recurrent localization task. 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 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 Cao et al. (2022; 2023). Fundamentally, CVOGL can be regarded as a cross-view matching problem, where recurrent strategies have shown superior robustness across domains Cao et al. (2022); Teed & Deng (2020). Inspired by this, we reformulate CVOGL as a recurrent localization process. Moreover, unlike dense

matching tasks Cao et al. (2023); Yu et al. (2023); Edstedt et al. (2024), 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) Dosovitskiy et al. (2021), 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.

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) Dosovitskiy et al. (2021), 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 point prompt $\mathbf{P}$ embedded 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 are 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 $i \in [0, 1, 2, \ldots, m]$. We set m to 6 in our work experimentally. 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.

At each refinement step i, we introduce a loss $\mathcal{L}_{\text{Token}}$ 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 function, respectively. $\mathrm{Sum}(\cdot)$ is the summing 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

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

where $\mathcal{L}_{\text{bce}}(\cdot,\cdot)$ and $\mathcal{L}_{\text{dice}}(\cdot,\cdot)$ denote the binary cross-entropy loss and the Dice Loss Milletari et al. (2016), respectively.

How ReCOT works. As shown in Fig. 3(a), previous one-shot detection CVOGL approaches Sun et al. (2023); Li et al. (2025); Huang et al. (2025) rely on a single forward information aggregation and are thus sensitive to noisy features Cao et al. (2022; 2023). Our ReCOT adopts a 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

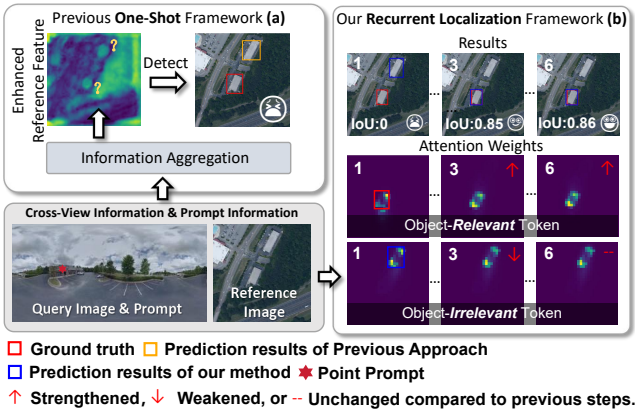

Figure 3: Comparison between the previous one-shot framework and our recurrent localization framework. (a) Previous approaches Li et al. (2025); Sun et al. (2023); Huang et al. (2025) rely on single-shot information aggregation, which is sensitive to noisy enhanced features and often leads to incorrect 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. Please refer to the zoomed-in view for better visualization.

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 Carion et al. (2020), 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 Kirillov et al. (2023). To address this, we propose a SAM-based knowledge distillation strategy to boost the prompt understanding of ReCOT. The incorporation of SAM Kirillov et al. (2023) 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 a 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 semantic. 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 Sun et al. (2023); Li et al. (2025). 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 Cao et al. (2023); Teed & Deng (2020). To this end, we propose the Reference Feature Enhancement Module (RFEM),

which enhances reference features into a more object-aware representation $\mathbf{F}_r'$. Unlike previous approaches Sun et al. (2023); Li et al. (2025); Huang et al. (2025) 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 key 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 $\mathbf{F}_r'$, which significantly benefits the iterative localization process of ReCOT.

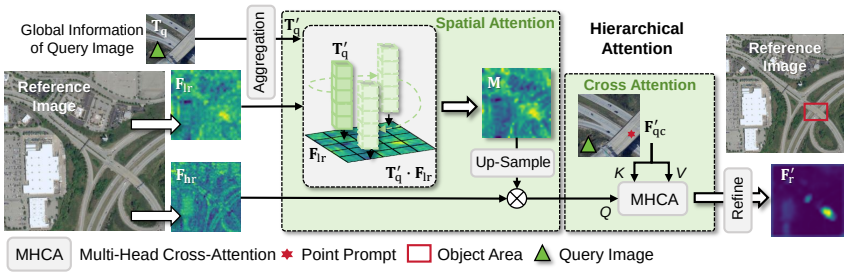

Figure 4: Architecture of reference feature enhancement module (RFEM). RFEM enhances reference features through a hierarchical attention pipeline to obtain $\mathbf{F}_r'$.

Structure. Specifically, as illustrated in Fig. 4, we utilize two levels of reference features, *i.e.*. a low-resolution semantic feature $\mathbf{F}_{lr} \in \mathbb{R}^{h_r \times w_r \times c}$ that captures scene semantics, and a high-resolution detailed feature $\mathbf{F}_{hr} \in \mathbb{R}^{2h_r \times 2w_r \times c}$ that preserves fine-grained spatial structures Zhang et al. (2024c).

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

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

which highlights query-relevant regions. The attention map is then up-sampled and applied to $\mathbf{F}_{hr}$ by element-wise multiplication, suppressing irrelevant background and 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 $\mathbf{F}_{hr}$ is subsequently refined via cross-attention with $\mathbf{F}_{qc}$, 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 Woo et al. (2018) and self-attention Dosovitskiy et al. (2021) to generate the $\mathbf{F}_r'$, 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 $\mathbf{F}_{lr}$ extracted from deeper layers of the backbone encodes more scene-level context but lose spatial details due to down-sampling. Therefore, we utilize $\mathbf{F}_{lr}$ to enhance query-relevant regions. Conversely, the high-resolution reference features $\mathbf{F}_{hr}$ preserve fine structural details but are dominated by background noise. Thus, it needs to be refined by the attention map produced by $\mathbf{F}_{lr}$, and then RFEM can utilize it to perform object-level enhancement. This design leverages advantages of multi-resolution features, which is crucial for object-level feature enhancement in CVOGL Huang et al. (2025).

3.4 Loss Function

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

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

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

4 Experiment

4.1 Datasets

CVOGL-DetGeo dataset Sun et al. (2023) divides the task into "Ground $\rightarrow$ Satellite" task and "Drone $\rightarrow$ Satellite" task. This dataset is the only public dataset for the relatively new task CVOGL. It contains 6,239 pairs of "Ground $\rightarrow$ Satellite" view and 6,239 pairs of "Drone $\rightarrow$ Satellite" view query and reference images. The ground view query images, drone view query images, and satellite view reference images are sized to 512 $\times$ 256, 256 $\times$ 256, and 1024 $\times$ 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.

4.2 Evaluation Metrics

Following the pioneering work DetGeo Sun et al. (2023), we adopt Acc@0.25 and Acc@0.50 for evaluation. 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 parameter denote better performance. Please refer to the appendix for more detailed introduction of evaluation metrics.

4.3 Implementation Details

Please refer to the appendix for detailed introduction of implementation details.

4.4 Ablation Study

Effect of Total Recurrent Localization Steps. The role of the recurrent localization lies in refining and correcting early predictions. Table 1 investigates the impact of varying the total number of recurrent steps m on ReCOT performance. For the Ground$\rightarrow$Satellite task, increasing m from 1 (one-shot prediction) to 5 yields consistent improvements, with the best performance achieved at $m = 5$. In contrast, for the Drone$\rightarrow$Satellite task, the Drone and Satellite images share similar top-down geometry and exhibit cleaner alignment, a single forward pass already produces a strong estimate. Its performance saturates earlier, with $m = 3$ giving the relatively optimal results.This discrepancy indicates that the optimal number of recurrent steps is task-dependent due to differences in viewpoint variations and feature alignment difficulty across scenarios. Additionally, 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 Hur & Roth (2019); Cao et al. (2023); Yu et al. (2023). Based on the results, we set $m = 5$ for the Ground$\rightarrow$Satellite task, while $m = 3$ for the Drone$\rightarrow$Satellite task, and keep this setting in other experiments.

Effect of Components in ReCOT. Table 2 presents the ablation study of the key components in ReCOT on the CVOGL-DetGeo 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 semantics 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 the robust performance of ReCOT.

Table 1: Ablation study on the recurrent localization steps m of ReCOT in terms of Acc@$0.25(\%)$ $\uparrow$ and Acc@$0.50(\%)$ $\uparrow$ on the test set of CVOGL-DetGeo dataset. Bold and Underline indicate the best and second-best results, respectively.

Steps m	Ground→Satellite			Drone→Satellite		
	Acc@0.25	Acc@0.50	FPS	Acc@0.25	Acc@0.50	FPS
1	49.74	46.25	26.9	**78.31**	71.74	27.2
2	51.08	47.17	26.2	<u>78.21</u>	72.15	26.5
3	51.28	47.58	25.6	<u>78.21</u>	<u>72.35</u>	25.7
4	51.39	<u>47.89</u>	24.8	77.90	**72.56**	24.9
5	**52.00**	**48.10**	24.1	77.60	72.05	24.3
6	<u>51.70</u>	**48.10**	23.4	77.49	71.84	23.5

Table 2: Ablation study on the components of ReCOT in terms of Acc@$0.25(\%)$ $\uparrow$, Acc@$0.50(\%)$ $\uparrow$, and FPS $\uparrow$ on the test set of CVOGL-DetGeo 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	78.21	72.35

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, particularly in the Drone→Satellite setting. This confirms that the spatial attention stage benefit 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 appendix for more ablation study results.

Table 3: Ablation study inside the RFEM in terms of Acc@$0.25(\%)$ $\uparrow$ and Acc@$0.50(\%)$ $\uparrow$ on the test set of CVOGL-DetGeo 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	78.21	72.35

4.5 Comparison Results

Quantitative Results. Table 4 compares ReCOT with existing SOTA cross-view localization approaches on the CVOGL-DetGeo dataset. ReCOT consistently outperforms all competitors in both the Ground→Satellite and Drone→Satellite settings, achieving new SOTA performance across almost all metrics in the test set. Despite relatively lower performance on the validation set for Ground→Satellite, it maintains the best performance on the test set, indicating stronger generalization ability. Notably, as shown in Table 4, these performance gains are achieved with significantly fewer parameters, representing a 60% reduction in model size. ReCOT is a little slower in inference speed compared to the previous CVOGL works Sun et al. (2023); Huang et al. (2025) due to iterative framework Cao et al. (2023). However, it still achieves a competitive inference speed, making ReCOT both efficient and scalable for real-world applications. Compared to the Drone→Satellite setting, the improvement of ReCOT on Ground→Satellite is relatively smaller. This is mainly due to the larger viewpoint gap and severe occlusions in ground-view images Sun et al. (2023), where

objects are often partially visible or obstructed by surrounding structures. Moreover, ground images typically contain more background clutter, making recurrent localization harder. We believe integrating geometric priors ($e.g.$, camera pose estimation) or multi-view fusion could further boost Ground→Satellite performance. Please refer to the appendix for more comparison results.

Table 4: Comparisons in terms of Acc@0.25(%) ↑, Acc@0.50(%) ↑, Parameter (M) ↓, and FPS ↑ on the CVOGL-DetGeo dataset. Following Table 1, we set $m = 5$ and $m = 3$ for the Ground→Satellite and Drone→Satellite task, respectively. The inference speed is test on the Drone→Satellite ($m = 3$) 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		Validation Set		Test Set		Validation Set			
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50		
CVM-Net Hu et al. (2018)	4.73	0.51	5.09	0.87	20.14	3.29	20.04	3.47	–	–
RK-Net Lin et al. (2022)	7.40	0.82	8.67	0.98	19.22	2.67	19.94	3.03	–	–
L2LTR Yang et al. (2021)	10.69	2.16	12.24	1.84	38.95	6.27	38.68	5.96	–	–
Polar-SAFA Shi et al. (2019)	20.66	3.19	19.18	2.71	37.41	6.58	36.19	6.39	–	–
TransGeo Zhu et al. (2022a)	21.17	2.88	21.67	3.25	35.05	6.47	34.78	5.42	–	–
SAFA Shi et al. (2019)	22.20	3.08	20.59	3.25	37.41	6.58	36.19	6.39	–	–
GeoDTR+ Zhang et al. (2024b)	14.19	5.14	14.08	1.95	16.03	4.73	15.71	3.68	–	–
DetGeo Sun et al. (2023)	45.43	42.24	46.70	43.99	61.97	57.66	59.81	55.15	<u>73.8</u>	**29.5**
VaGeo Li et al. (2025)	48.21	45.22	47.56	**44.42**	66.19	61.87	64.25	59.59	–	–
OCGNet Huang et al. (2025)	<u>51.49</u>	<u>47.69</u>	**48.54**	<u>44.20</u>	<u>68.39</u>	<u>63.93</u>	<u>66.52</u>	<u>61.86</u>	74.8	<u>27.7</u>
ReCOT (Ours)	**52.00**	**48.10**	<u>48.43</u>	43.66	**78.21**	**72.35**	**74.00**	**67.17**	**29.9**	25.7

Figure 5: Visualization of some representative results produced by our ReCOT and the previous work Sun et al. (2023). Please refer to the zoomed-in view for better visualization.

Qualitative Results. Fig. 5 visualizes some representative results. As the recurrent steps proceed, our model progressively refines the bounding boxes, leading to higher-quality localization compared to the single-shot prediction of the previous approach Sun et al. (2023). However, the optimal number of recurrent steps varies across different scenarios, and excessive iterations may cause over-refinement Hur & Roth (2019), resulting in a slight performance drop. Therefore, based on the ablation results in Table 1, we set the number of recurrent steps to 5 for Ground→Satellite and 4 for Drone→Satellite, which provides the best trade-off between accuracy and stability.

5 Conclusion

In this paper, we propose ReCOT, which reformulates the CVOGL task as a recurrent localization problem. 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 CVOGL-DetGeo benchmark demonstrated that ReCOT achieves state-of-the-art performance with significantly fewer parameters and competitive inference speed.

REFERENCES

Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. 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.

Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(5):1483–1498, 2021.

Si-Yuan Cao, Jianxin Hu, Zehua Sheng, and Hui-Liang Shen. Iterative deep homography estimation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1869–1878, 2022.

Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, and Hui-Liang Shen. 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.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv e-prints arXiv:2005.12872*, 2020.

Kaichen Chi, Yuan Yuan, and Qi Wang. Trinity-Net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

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, 2009.

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 (ICCV)*, pp. 16847–16856, October 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for image recognition at scale. *arXiv e-prints arXiv:2010.11929*, 2021.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Yulan Guo, Michael Choi, Kunhong Li, Farid Boussaid, and Mohammed Bennamoun. Soft exemplar highlighting for cross-view image-based geo-localization. *IEEE Transactions on Image Processing*, 31:2094–2105, 2022.

Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. 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.

Zheyang Huang, Jagannath Aryal, Saeid Nahavandi, Xuequan Lu, Chee Peng Lim, Lei Wei, and Hailing Zhou. Object-level cross-view geo-localization with location enhancement and multi-head cross attention, 2025.

Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Anuj Kumar, Hiesik Kim, and Gerhard P. Hancke. Environmental monitoring systems: A review. *IEEE Sensors Journal*, 13(4):1329–1339, 2013.

Ted Lentsch, Zimin Xia, Holger Caesar, and Julian F. P. Kooij. 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.

Zhongyang Li, Xin Yuan, Wei Liu, and Xin Xu. VAGeo: View-specific attention for cross-view object geo-localization. *ArXiv*, 2501.07194, 2025.

Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing*, 31:3780–3792, 2022.

Liu Liu and Hongdong Li. 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.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv e-prints arXiv:1711.05101*, 2017.

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 *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 856–864, 2020.

Xiufan Lu, Siqi Luo, and Yingying Zhu. 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.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv e-print arXiv:1606.04797s*, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, 2019.

Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 470–479, 2019.

Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. 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.

Liang Shi, Yixin Chen, Meimei Liu, and Feng Guo. 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.

Yujiao Shi and Hongdong Li. 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.

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

Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. 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, 2020a.

Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (07):11990–11997, Apr. 2020b.

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*, 45:2682–2697, 2022.

Phani Teja Singamaneni, Pilar Bachiller-Burgos, Luis J. Manso, Anaís Garrell, Alberto Sanfeliu, Anne Spalanzani, and Rachid Alami. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research*, 43(10):1533–1572, 2024.

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*, 61:1–16, 2023.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixe. 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.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers &; distillation through attention. In *International Conference on Machine Learning*, volume 139, pp. 10347–10357, July 2021.

Xiaolong Wang, Runsen Xu, Zuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator. *ArXiv*, abs/2308.16906, 2023.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Proc. Eur. Conf. Comput. Vis.*, pp. 3–19, Cham, September 2018. Springer Nature Switzerland.

Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29009–29020. Curran Associates, Inc., 2021.

Jing Yao, Danfeng Hong, Lianru Gao, and Jocelyn Chanussot. Multimodal remote sensing benchmark datasets for land cover classification. In *2022 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 4807–4810, 2022.

Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, and Feng Wu. 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.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 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.

Ruoming Zhai, Jingui Zou, Vincent J.L. Gan, Xianquan Han, Yushuo Wang, and Yinzhi Zhao. Semantic enrichment of bim with indoorgml for quadruped robot navigation and automated 3d scanning. *Automation in Construction*, 166:105605, 2024. ISSN 0926-5805.

Tianyang Zhang, Xiangrong Zhang, Xiaoqian Zhu, Guanchun Wang, Xiao Han, Xu Tang, and Licheng Jiao. Multistage enhancement network for tiny object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024a.

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*, 46(12):10419–10433, 2024b.

Xiaohan Zhang, Xue Zhang, Si-Yuan Cao, Beinan Yu, Chenghao Zhang, and Hui-Liang Shen. 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, 2024c.

Sijie Zhu, Mubarak Shah, and Chen Chen. 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 2022a.

Yingying Zhu, Bin Sun, Xiufan Lu, and Sen Jia. Geographic semantic network for cross-view image geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022b.

A Appendix

A.1 More Experimental Results

A.1.1 Detailed Introduction of Evaluation Metrics.

Definition of Acc@0.25 and Acc@0.50. 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 box b_g as

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

where

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

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

A.1.2 Implementation Details

We conduct all experiments on four NVIDIA GeForce RTX 4090 GPUs, with implementations based on PyTorch Paszke et al. (2019). For training, we adopt the AdamW Loshchilov & Hutter (2017) as the optimizer and set the initial learning rate to 0.0025, the weight decay rate to 0.0001, and the batch size to 16. We train our network for 300 epochs for all the experiments. Since CVOGL is a relatively new task, we follow the pioneering work Sun et al. (2023) and select five CVIGL approaches Hu et al. (2018); Lin et al. (2022); Yang et al. (2021); Shi et al. (2019); Zhu et al. (2022a) and the existing CVOGL approaches Sun et al. (2023); Li et al. (2025); Huang et al. (2025) as our comparison methods. The results of CVIGL comparison methods can be found in previous works Sun et al. (2023); Huang et al. (2025); Li et al. (2025). Additionally, we adopt swin transformer (Swin-t) Liu et al. (2021) as the image encoder for its superior performance on various fields Zhang et al. (2024a); Shi et al. (2024); Chi et al. (2023); Zamir et al. (2022). We set the hyper-parameter m to 6 during training.

A.1.3 More Ablation Study Results

Detailed explanation of performance degradation with excessive steps in our recurrent localization. The recurrent mechanism in ReCOT aims to progressively refine the predicted location. However, as reported in Table 1, excessive recurrent steps may sightly degrade performance. This is because our recurrent localization can be interpreted as predicting and applying residual corrections to the bounding box. After the prediction becomes sufficiently accurate, the remaining residuals are mainly high-frequency noise or background fluctuations. As observed in other iterative refinement works Cao et al. (2022; 2023), continuing to update on such residuals may amplify noise and introduce attention drift, causing the learnable tokens to overfit spurious cues rather than consolidate stable object semantics Cai & Vasconcelos (2021). From an optimization perspective, when the objective is already close to its minimum, further iterations may overshoot or oscillate around the optimum Teed & Deng (2020), leading to a slight decline in IoU.

How to determine m on new datasets? To determine m on a new dataset, one can monitor IoU gain per step on a validation set and stop when the gain becomes negligible or unstable. Tasks with large viewpoint gaps typically require larger m, while nearly aligned views (*e.g.*, Drone→Satellite) benefit from smaller m for a better balance of accuracy and efficiency.

Effect of the parameter α. Table 5 shows how the performance of ReCOT varies with different values of the balancing coefficient α, which controls the weight between the localization loss $\mathcal{L}_{\text{local}}$ and the auxiliary loss $\mathcal{L}_{\text{aux}}$. We observe that $\alpha = 1$ yields the best performance on both Ground→Satellite and Drone→Satellite. A smaller value ($\alpha = 0.1$) weakens the supervision of token alignment and SAM distillation, slightly reducing accuracy. Conversely, a larger value ($\alpha = 10$) overemphasizes auxiliary objectives, causing a notable drop. These results indicate that $\alpha = 1$ achieves the optimal trade-off.

Table 5: Ablation study on the hyper-parameter α in terms of Acc@$0.25(\%)$ ↑ and Acc@$0.50(\%)$ ↑ on the test set of CVOGL-DetGeo dataset.

α	Ground→Satellite		Drone→Satellite	
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
0.1	51.28	48.10	75.13	69.68
1	52.00	48.10	78.21	72.35
10	48.51	45.02	76.05	70.20

Effect of the hyper-parameter n. Table 6 investigates the impact of the token number n on the performance of ReCOT. We observe a clear performance gain when increasing n from 1 to 100. This demonstrates that a sufficient number of tokens provides a richer representation of task-specific intent and better coverage of cross-view semantics, which benefits the recurrent refinement process. However, when n is further increased to 200, performance drops across all metrics. This decline is likely due to over-parameterization and token redundancy, which can introduce noise and hinder effective attention learning Dosovitskiy et al. (2021), as also observed in token-scaling studies Dosovitskiy et al. (2021); Touvron et al. (2021). Therefore, we set n to 100 in this work.

Table 6: Ablation study on the hyper-parameter n in terms of Acc@$0.25(\%)$ ↑ and Acc@$0.50(\%)$ ↑ on the test set of CVOGL-DetGeo dataset.

n	Ground→Satellite		Drone→Satellite	
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
1	47.89	43.37	74.31	68.86
100	52.00	48.10	78.21	72.35
200	50.46	47.28	72.25	67.01

Table 7: Comparisons between DetGeo Sun et al. (2023), DetGeo using Swin-t Liu et al. (2021) (DetGeo*), and our methods in terms of Acc@$0.25(\%)$ ↑ and Acc@$0.50(\%)$ ↑ on the test set of CVOGL-DetGeo dataset.

Method	Ground→Satellite		Drone→Satellite	
	Acc@0.25	Acc@0.50	Acc@0.25	Acc@0.50
DetGeo	45.43	42.24	61.97	57.66
DetGeo*	46.01	41.44	66.60	56.11
VaGeo	48.21	45.22	66.19	61.87
OCGNet	<u>51.49</u>	<u>47.69</u>	<u>68.39</u>	<u>63.93</u>
ReCOT (Ours)	**52.00**	**48.10**	**78.21**	**72.35**

A.1.4 More Comparison Results.

Quantitative Results.

As shown in Table 7, we replace the backbone of DetGeo Sun et al. (2023) with Swin-t Liu et al. (2021) for a more comprehensive and fair comparison. While the upgraded DetGeo* shows slight improvements on the Acc@0.25, it still lags behind our ReCOT by a large margin across all evaluation metrics. Moreover, the backbone replacement does not lead to consistent gains, as DetGeo* fails to outperform other recent CVOGL methods Li et al. (2025); Huang et al. (2025) in Table 7. This demonstrates that the key factor of CVOGL does not lie in the backbone. In contrast, our proposed ReCOT achieves superior performance through a more effective and task-aligned framework, highlighting the importance of architectural innovations for CVOGL.

Qualitative Results.

Fig. 6 provides more visualization results of the cross attention weights beteen the token and reference features during recurrent process. As the recurrent steps progress, object-relevant tokens gradually focus and strengthen around the expected object, enabling the bounding box to converge toward the correct location. In contrast, object-irrelevant tokens weaken and gradually stabilize over iterations. This behavior highlights the ability of ReCOT to dynamically disentangle object-focused information from irrelevant features, which is a key factor driving the success of our recurrent localization strategy.

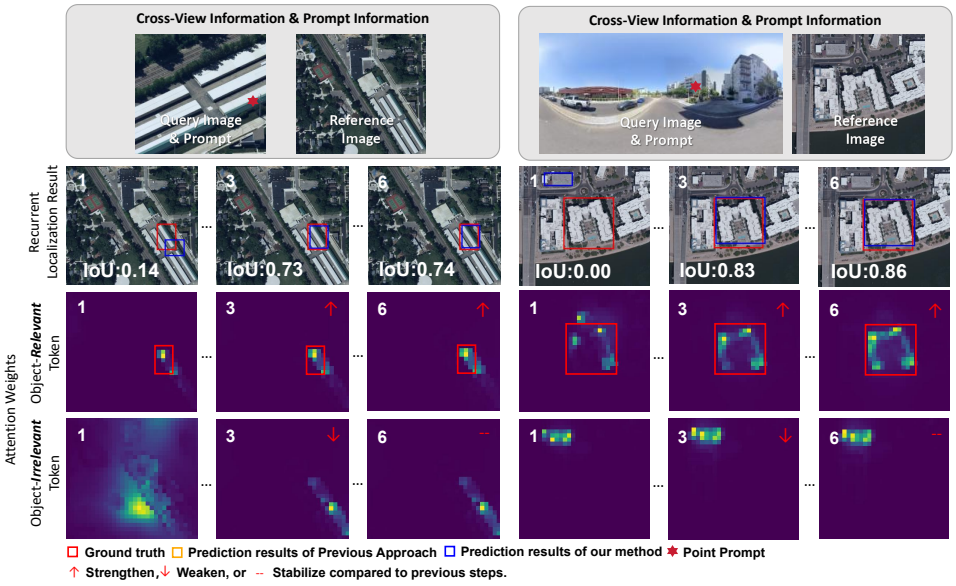

Figure 6: Visualizations of how our ReCOT works. The object-relevant tokens progressively focus and strengthen around the indicated object, while object-irrelevant tokens weaken and stabilize to background patterns.

A.2 Limitations and Future Work

As shown in Fig. 7, some failure cases of ReCOT are caused by ambiguous or imprecise point prompts, which often highlight only a part of the object rather than its entirety. This ambiguity misguides the token-driven recurrent refinement process, leading to suboptimal localization results. In the future, we will incorporate multi-modal prompts ($e.g.$, textual descriptions or bounding boxes) to provide richer and more accurate prompt semantics. Such multi-modal guidance could reduce ambiguity and further improve the robustness and precision of recurrent localization.

A.3 Statement

A.3.1 Ethics Statement

This work addresses cross-view object geo-localization in the recurrent manner. All experiments are conducted on publicly available datasets that do not contain sensitive information. The authors affirm that this work complies with the code of ethics.

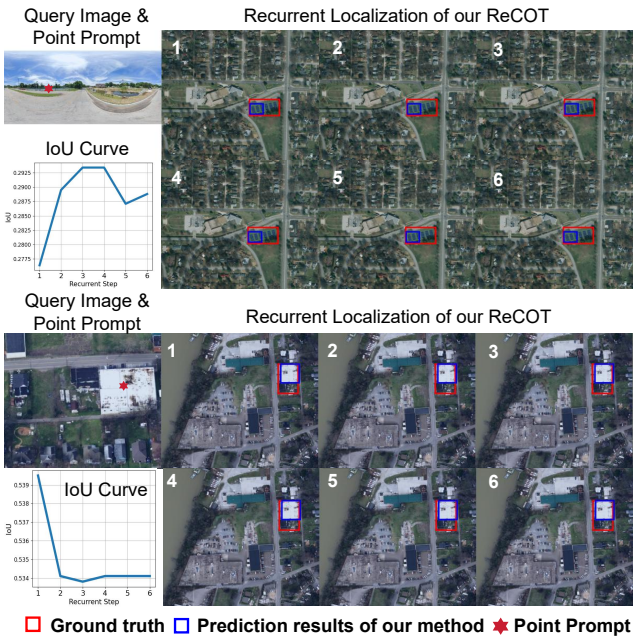

Figure 7: Examples of failure cases. Please refer to the zoomed-in view for better visualization.

A.3.2 Reproducibility Statement

We provide the source code in the supplementary material. Upon acceptance, we will release the source code, repurposed datasets, and model weights to ensure the reproducibility of the experimental results.

A.3.3 The Use of Large Language Model (LLM)

We used the LLM (*e.g.*, ChatGPT) only to aid English writing and polish the presentation of this paper (grammar, clarity, and minor wording suggestions). The research ideas, algorithms, theoretical analysis, experiment design, implementation, and all reported results were entirely developed and validated by the authors without automated assistance. No content was directly copied from model outputs, and all technical claims have been carefully verified by the authors.