

# Breaking Rectangular Shackles: Cross-View Object Segmentation for Fine-Grained Object Geo-Localization

Qingwang Zhang, Yingying Zhu✉

College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

zhangqingwang2022@email.szu.edu.cn, zhuyy@szu.edu.cn

## Abstract

This paper addresses the limitations of existing cross-view object geo-localization schemes, which rely on rectangular proposals to localize irregular objects in satellite imagery. These “rectangular shackles” inherently struggle to precisely define objects with complex geometries, leading to incomplete coverage or erroneous localization. We propose a novel scheme, cross-view object segmentation (CVOS), which achieves fine-grained geo-localization by predicting pixel-level segmentation masks of query objects. CVOS enables accurate extraction of object shapes, sizes, and areas—critical for applications like urban planning and agricultural monitoring. We introduce the CVOGL-Seg dataset specifically to support and evaluate the new CVOS scheme. To tackle CVOS challenges, we propose Transformer Object Geo-localization (TROGeo), a two-stage framework. First, the Heterogeneous Task Training Stage (HTTS) employs a single transformer encoder with a Cross-View Object Perception Module (CVOPM) and is trained by learning a heterogeneous task. Second, the SAM Prompt Stage (SPS) utilizes SAM’s zero-shot segmentation capability, guided by HTTS outputs, to generate precise masks. Extensive experiments on both CVOGL and CVOGL-Seg datasets demonstrate that our approach achieves state-of-the-art performance, effectively breaking the rectangular shackles and unlocking new possibilities for fine-grained object geo-localization. Our project page: <https://zqwlearning.github.io/CVOS>.

## 1. Introduction

Consider Figure 1 (a), given a reference (satellite-view) image with geographic information, how can we obtain fine-grained geographic information of the query object in the image if an image taken from the ground or drone view in the area is provided? The above process is the core issue of the cross-view object geo-localization, which can be applied to object localization [42], smart city manage-

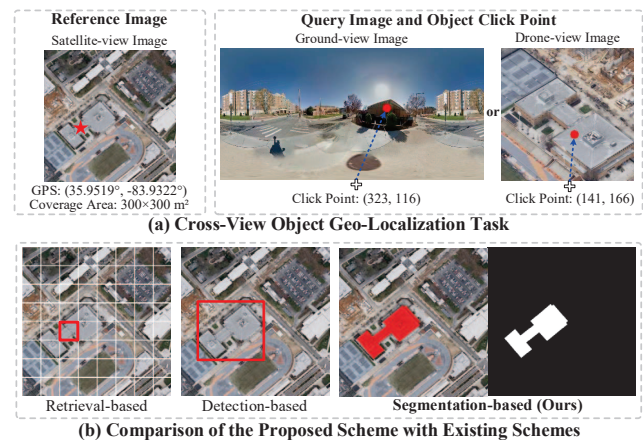


Figure 1. Illustration of the cross-view object geo-localization task and comparison of different schemes. (a) The reference image is typically a satellite image with geographic information (e.g., GPS coordinate and coverage area), while the query image is captured from a ground or drone view, with a click point (red dot) indicating the query object. Cross-view object geo-localization aims to find the geographic location of a specific object in the query image by matching it to the reference image with attached geographic information. (b) The **retrieval-based** scheme divides the satellite image into uniformly-sized patches to construct a reference database and attempts to retrieve the patch that is most semantically similar to the query object as a location proposal. The **detection-based** scheme finds the best bounding box containing the query object in satellite images as a location proposal. Our **segmentation-based** scheme attempts to obtain a segmentation mask of the query object as a location proposal. Clearly, the segmentation mask proposal expresses the location of the irregular object more accurately.

ment [3, 17], and disaster monitoring [14, 45], etc. An intuitive scheme is to try to imitate cross-view image geo-localization [2, 6, 11, 18, 23, 26, 28, 35, 43, 46–49], i.e., retrieval-based scheme. First, the satellite image is divided into uniform-sized patches, and then the patch most semantically similar to the query object is retrieved as the object’s geographic location. However, due to the different shapes, sizes, and locations of the objects, it is diffi-

cult for the pre-sampled uniform-sized rectangles to perfectly cover irregular objects. Subsequent work [37] proposes a detection-based scheme that uses a bounding box (dynamic-sized rectangle) to define the geographic location of the object, which alleviates the problem of pre-sampling fixed-size rectangles in the retrieval-based scheme to a certain extent and achieves significant improvements.

However, retrieval-based and detection-based schemes always have **rectangular shackles**, *i.e.*, *attempting to define the geographic locations of irregular objects with simple geometries (rectangles)*. These inherent limitations result in incomplete object coverage or incorrect inclusion of background and other objects as location proposals, as shown in Figure 1 (b). To address the above challenges, we first propose a segmentation-based scheme for cross-view object geo-localization, termed **cross-view object segmentation** (CVOS). It directly obtains pixel-level segmentation masks for query objects and achieves fine-grained object geo-localization. Combined with geographic information from the reference image, CVOS can extract object properties (*e.g.*, area and shape), which are crucial for urban planning and agricultural monitoring applications but difficult to obtain with existing schemes. Figure 1 (b) compares different cross-view object geo-localization schemes.

Although CVOS offers significant application potential, it faces substantial challenges due to dramatic visual appearance variations of objects across different viewpoints, compounded by the fact that satellite images cover extensive areas, contain noisy backgrounds, and frequently include similar objects. Furthermore, the requirement for high-resolution mask prediction of the object further amplifies the complexity of cross-view object segmentation.

To address the above issues, we propose a novel Transformer Object Geo-localization (TROGeo) framework, which consists of a Heterogeneous Task Training Stage (HTTS) and a SAM Prompt Stage (SPS). The HTTS is based on a single Transformer encoder, combined with our proposed Cross-View Object Perception Module (CVOPM), to perceive the semantic consistency of the object itself and the surrounding environment, helping the model to better learn discriminative features. The heterogeneous task predicts both the bounding box and segmentation mask of the object to help with more precise supervision of the object and achieve better performance. In SPS, SAM [21] with strong zero-shot prompt segmentation capability is used to generate fine segmentation masks for cross-view objects based on the bounding boxes (and center points) proposed in HTTS. Extensive experimental results on the existing CVOGL and our newly created CVOGL-Seg datasets show that our method has significant advantages.

Our contributions can be summarized as follows:

- We propose a cross-view object segmentation (CVOS) scheme to reformulate cross-view object geo-localization,

which uses segmentation masks to define irregular cross-view objects for fine-grained object geo-localization. To support this scheme, we create the CVOGL-Seg dataset to provide high-resolution mask annotations for objects.

- We propose a Transformer Object Geo-localization (TROGeo) framework to implement CVOS, which consists of a Heterogeneous Task Training Stage (HTTS) and a SAM Prompt Stage (SPS). In HTTS, we adopt a single Transformer encoder and introduce a CVOPM to train the model through a heterogeneous task. In SPS, we achieve fine-grained object geo-localization by prompting the SAM to obtain segmentation masks.
- Extensive experiments reveal the attractive potential of CVOS for fine-grained object geo-localization, and our TROGeo achieves state-of-the-art performance.

## 2. Related Work

**Cross-View Image Geo-Localization.** Image-based localization is often formulated as an image retrieval problem and solved by metric learning techniques [18, 25, 34, 40]. Recent studies have made significant progress on the commonly used benchmarks CVUSA [44], CVACT [26], and VIGOR [48], demonstrating excellent performance [11, 35, 43, 46, 49]. These methods can achieve coarse-grained image-level localization, but are not well suited for precise geo-localization of objects. In this paper, we study how to achieve fine-grained localization information of the objects.

**Cross-View Object Geo-Localization.** The task of cross-view object geo-localization is to determine the geographic location information of a given object in a query image from a satellite image with geographic information. Referring to the traditional cross-view image retrieval idea [18, 25, 40], cross-view object geo-localization can be regarded as an image retrieval task. In some works [22, 37], this task is framed as cross-view object detection. However, due to the inherent limitations of the rectangles, it is difficult to use both schemes to localize irregularly shaped objects accurately. To break rectangular shackles, we propose a novel cross-view object segmentation scheme for reframing it.

**Transformer** [39] is a model based on the attention mechanism that initially achieved remarkable success in natural language processing tasks [32, 38]. Due to its powerful global modeling capabilities, Transformer has gradually been introduced into the field of computer vision for tasks such as image classification [12, 16], object detection [7], and image segmentation [36]. Swin Transformer [27] is an improved Vision Transformer [12] designed specifically for computer vision tasks. Its hierarchical architecture and shifted windows allow it to have linear computational complexity while performing well on more visual tasks. We use the Swin Transformer as a feature extraction network.

**Segment Anything Model (SAM)** [21] is one of the most classic and powerful foundation models for computer vi-

sion. The SAM demonstrates excellent zero-shot generalization capabilities when applied to downstream segmentation tasks [10, 19, 24], *e.g.*, satellite image segmentation [5]. In this study, we leverage SAM’s powerful prompt segmentation capabilities to achieve masks of cross-view objects.

### 3. Methodology

In this section, we first introduce the task we address in this paper and discuss different cross-view object geo-localization schemes. Next, we dive into the details of the proposed framework. Finally, we introduce the optimization objective adopted by our method.

#### 3.1. Preliminary

**Problem Statement.** The cross-view object geo-localization task aims to find the specific query object in a query image  $I_q$  (a ground / drone image) from a given reference image  $I_r$  (satellite image) with geographic information (a GPS coordinate and coverage area). The query object is specified by a click point  $p$  in the query image. The *retrieval-based scheme* divides  $I_r$  into uniformly-sized image patches and retrieves the patch that is most semantically similar to the query object. The *detection-based scheme* attempts to find the best bounding box in the satellite image that contains the query object. Both of them try to use rectangles to represent the localizations of the irregular objects.

**Cross-View Object Segmentation.** We propose a segmentation-based scheme to reframe the cross-view object geo-localization task, *i.e.*, based on the query image  $I_q$ , the query object click point  $p$ , and the reference image  $I_r$ , a mask  $I_{\text{mask}}$  of the object is segmented on  $I_r$ , termed “**cross-view object segmentation**” (CVOS) and be formulated as:  $(I_q, p, I_r) \mapsto I_{\text{mask}}$ . With the  $I_{\text{mask}}$  and the GPS coordinate and coverage area attached to the  $I_r$ , CVOS can achieve fine-grained geographic information of the object. This solution breaks the shackles of previous schemes (retrieval and detection) that use rectangles to define object localizations and can obtain detailed geographic information, such as the shape and area of the object.

**Theoretical Upper Bounds for Different Schemes.** We compared the theoretical upper bounds of different schemes for handling pixel-level fine-grained object geo-localization on the CVOGL-Seg dataset, and the results are shown in Table 1. The optimal object masks obtained by different schemes are shown in Figure 2. The retrieval- and detection-based schemes directly generate rectangular masks based on the ground truth masks, and the segmentation-based scheme can theoretically obtain the ground truth masks. We find that retrieval and detection (rectangle) masks often fail to provide complete or accurate object location proposals, while segmentation masks precisely define object locations. Furthermore, segmentation-based schemes have the highest theoretical upper bound

on performance. These results reveal the inherent limitations of the rectangular shackle schemes and highlight the promising potential of the CVOS scheme for achieving fine-grained object localization information. For more discussion and analysis in the *Supplementary Materials*.

Scheme	Drone $\rightarrow$ Satellite & Ground $\rightarrow$ Satellite							
	Validation				Test			
	mIoU $\uparrow$ (%)	mDice $\uparrow$ (%)	AAE $\downarrow$ ( $m^2$ )	ME $\downarrow$ ( $m$ )	mIoU $\uparrow$ (%)	mDice $\uparrow$ (%)	AAE $\downarrow$ ( $m^2$ )	ME $\downarrow$ ( $m$ )
Retrieval-base	32.25	45.89	3736.58	11.98	31.32	44.84	3858.20	12.95
Detection-base	58.16	71.22	2196.57	3.29	57.76	70.92	1945.30	3.08
<b>Segmentation-base</b>	<b>100.00</b>	<b>100.00</b>	<b>0.00</b>	<b>0.00</b>	<b>100.00</b>	<b>100.00</b>	<b>0.00</b>	<b>0.00</b>

Table 1. Theoretical upper bounds of different schemes.



Figure 2. Comparison of optimal masks theoretically obtainable by different schemes. The mask is overlaid on the original image.

#### 3.2. TROGeo

**Overview.** Figure 3 shows details inside the Transformer Object Geo-localization (TROGeo) framework, which consists of two key stages: a Heterogeneous Task Training Stage (HTTS) and a SAM Prompt Stage (SPS). In HTTS, a single Transformer encoder is adopted, and a Cross-View Object Perception Module (CVOPM) is introduced to improve model performance by training a heterogeneous task. In SPS, we achieve fine-grained object geo-localization by prompting SAM to obtain segmentation masks.

##### 3.2.1. Heterogeneous Task Training Stage

First, the coordinate of the query object click point  $p$  is encoded as the point embedding matrix  $I_p$  (the values decrease from the coordinate  $p$  to the surroundings) [37]. It is channel-wise concatenated with the query image  $I_q$  through the Position Embedding (dual convolutional layer with a ReLU layer [30]) and then fed to the query image Transformer encoder  $E_q$  to generate the query image feature map  $F_q$ . The reference image  $I_r$  is directly input to the reference image Transformer encoder  $E_r$  to generate the reference image feature map  $F_r$ .  $E_q$  and  $E_r$  share trainable parameters, *i.e.*, a single encoder. The  $F_q$  and  $F_r$  are then sent to the CVOPM to generate the perceptual feature map  $F_p$ . Finally,  $F_p$  is required to learn a heterogeneous task, *i.e.*, simultaneously predicting the bounding box (bbox) and low-resolution ( $64 \times 64$ ) mask of the object to supervise the model. *The bbox defines the object under rectangular shackles, which allows fair comparison with previous methods. In addition, it can also serve as a SAM prompt.*

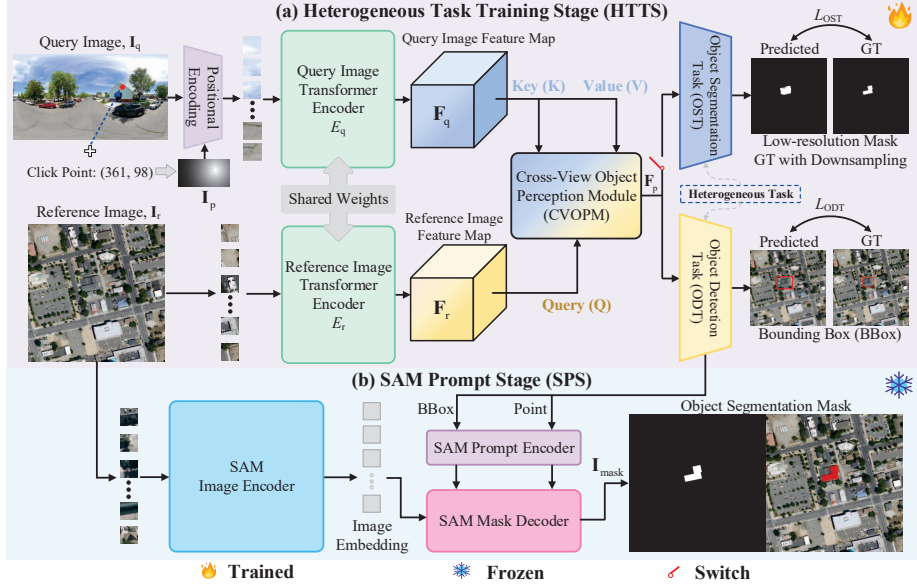


Figure 3. Overview of the proposed TROGeo framework, consisting of (a) Heterogeneous Task Training Stage (HTTS) and (b) SAM Prompt Stage (SPS). HTTS employs a single Transformer encoder and integrates a Cross-View Object Perception Module (CVOPM) to learn cross-view correspondences through a heterogeneous task (Section 3.2.1). SPS uses the bounding box (and its center point) generated by HTTS as external prompts to guide SAM in producing the final object segmentation masks (Section 3.2.2).

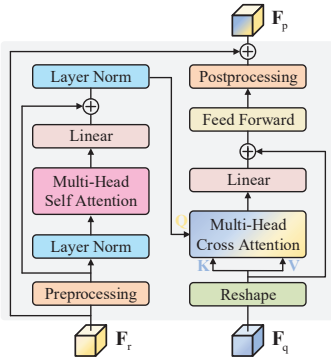


Figure 4. Cross-View Object Perception Module (CVOPM).

**Cross-View Object Perception Module.** Human perception provides a key insight: when searching for cross-view object localization cues, observers consciously integrate the object and its semantic context (*e.g.*, road layout, neighboring buildings). This cognitive process suggests that consistent object matches inherently imply semantic coherence of the surrounding area—a key signal that current methods ignore. Existing methods [11, 37] compress image feature maps into a single query vector, which overemphasizes global representations at the expense of fine-grained details. To bridge this gap, inspired by [33], we propose a Cross-View Object Perception Module (CVOPM), which enables the model to automatically focus on critical local and global localization cues. Figure 4 shows the structure of CVOPM.

First, the reference image feature map  $\mathbf{F}_r$  is preprocessed to obtain a new  $\mathbf{F}_r$ ,  $\mathbf{F}_r = \text{Reshape}(\text{Conv2d}(\text{GroupNorm}(\mathbf{F}_r)))$ ,  $\mathbf{F}_r \in \mathbb{R}^{(H_r \cdot W_r) \times D}$ , where  $\text{Reshape}(\cdot)$  denotes dimensional reshaping,  $\text{Conv2d}(\cdot)$  denotes 2D convolution,  $\text{GroupNorm}(\cdot)$  denotes group normalization [41],  $H_r$  and  $W_r$  denote raw  $\mathbf{F}_r$  height and width, and  $D$  denotes the dimension. The query image feature map  $\mathbf{F}_q$  is reshaped to obtain a new  $\mathbf{F}_q$ ,  $\mathbf{F}_q \in \mathbb{R}^{(H_q \cdot W_q) \times D}$ ,  $H_q$  and  $W_q$  denote raw  $\mathbf{F}_q$  height and width. Then, the preprocessed  $\mathbf{F}_r$  is passed through layer normalization (LN) [4], Multi-Head Self-Attention (MHSA) and linear projection [39], and then connected with the preprocessed  $\mathbf{F}_r$  through residual connection [15], and finally passed through LN as the query  $\mathbf{Q}$  of the Multi-Head Cross-Attention (MHCA), while the reshaped  $\mathbf{F}_q$  is used as the key  $\mathbf{K}$  and value  $\mathbf{V}$ . MHCA is a key component of CVOPM and can be represented as:

$$\mathbf{Q} = \mathbf{F}_r \mathbf{W}_q, \mathbf{K} = \mathbf{F}_q \mathbf{W}_k, \mathbf{V} = \mathbf{F}_q \mathbf{W}_v \quad (1)$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V} \quad (2)$$

where  $\mathbf{W}_q$ ,  $\mathbf{W}_k$  and  $\mathbf{W}_v$  are the projection matrices.

Then,  $\mathbf{A}$  goes through linear projection, residual connection and Feed Forward network [39], and finally goes through post-processing (dimensional reshaping and 2D convolution layer) and residual connection to output the perceptual feature representation  $\mathbf{F}_p$ . The  $\mathbf{F}_p$  contains not

only the visual details of the image but also integrates contextual auxiliary localization cues from the local region of the query object and its surrounding global region, which enables the model to learn discriminative features while obtaining more sensitive object spatial location awareness.

**Heterogeneous Task.** We introduce a Heterogeneous Task (HT) that uses the same input to complete two tasks: Object Detection Task (ODT) and Object Segmentation Task (OST). OST is an optional option controlled by a switch, which has the following advantages: (i) When the segmentation mask is available, it can strengthen the object supervision and thus improve the model performance; (ii) For low-performance devices (which cannot load SAM), the output of OST can be used as a viable alternative; (iii) OST is banned, which facilitates a fair comparison with existing methods. For ODT, we use a fully convolutional network, which includes a stride-2 deconvolution layer (4×4 convolutional kernel) with half the number of channels [31], a ReLU activation layer, and a 1×1 convolutional output layer with 45 output channels. For OST, the rest of the ODT is retained, and only the output channels are changed to 1.

When ground truth segmentation masks are unavailable, OST can be initiated by utilizing the bounding boxes to prompt SAM to generate approximate segmentation masks. *This substitution is still effective*, as shown in Table 2.

### 3.2.2. SAM Prompt Stage

Thanks to the powerful zero-shot prompt segmentation capability of the SAM foundation model [21], it supports many downstream tasks [10, 19]. The SAM comprises three components: the SAM Image Encoder, the SAM Prompt Encoder, and the SAM Mask Decoder. The reference image  $I_r$  is input to the SAM Image Encoder to obtain the image embedding. The SAM Prompt Encoder receives the bbox and its center point generated in the HTTS as external prompts and encodes them as prompt embeddings. Subsequently, the SAM Mask Decoder combines the image embedding and prompt embeddings to predict the segmentation mask  $I_{\text{mask}}$  for the query object on  $I_r$ . The  $I_{\text{mask}}$  implements cross-view object segmentation, which contains fine-grained localization information of the query object.

### 3.3. Optimization Objective

The optimization objective for the HTTS is denoted as:

$$\mathcal{L} = \mathcal{L}_{\text{ODT}} + \alpha \mathcal{L}_{\text{OST}} \quad (3)$$

where  $\mathcal{L}_{\text{ODT}}$  and  $\mathcal{L}_{\text{OST}}$  represent the ODT and OST optimization objectives, respectively;  $\alpha$  is a hyperparameter that controls the importance of the two tasks.

**ODT optimization objective.** We use the same anchor-based detection method as in previous work [37] to define the ODT optimization objective, which is expressed as:  $\mathcal{L}_{\text{ODT}} = \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{conf}}$ . Where  $\mathcal{L}_{\text{reg}}$  denotes the regression loss, which is used to minimize the distance between

the predicted bounding box and the ground truth, and  $\mathcal{L}_{\text{conf}}$  denotes the confidence loss, which is used to assess the confidence of the existence of the object, and is implemented by the binary cross-entropy loss [8].

**OST optimization objective.** The binary cross-entropy loss  $\mathcal{L}_{\text{BCE}}$  and the Dice loss [29]  $\mathcal{L}_{\text{Dice}}$  are used to supervise the low-resolution mask predicted:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^N \left[ \frac{1}{H_l W_l} \sum_{j=1}^{H_l} \sum_{k=1}^{W_l} (\beta \cdot y_{ijk} \cdot \log(\sigma(x_{ijk})) + (1 - y_{ijk}) \cdot \log(1 - \sigma(x_{ijk}))) \right] \quad (4)$$

$$\mathcal{L}_{\text{OST}} = (1 - \lambda) \cdot \mathcal{L}_{\text{BCE}} + \lambda \cdot \mathcal{L}_{\text{Dice}} \quad (5)$$

where  $N$  is the number of samples;  $H_l$  and  $W_l$  are the height and width of the low-resolution mask;  $y_{ijk}$  and  $x_{ijk}$  are the ground truth (downsampling) low-resolution (64×64) mask and model-predicted values for the  $i$ -th sample at position  $(j, k)$ , respectively;  $\beta$  is the positive category weight, which is important for the category imbalance problem;  $\lambda$  is the hyperparameter controlling the weight of  $\mathcal{L}_{\text{BCE}}$  and  $\mathcal{L}_{\text{Dice}}$ .

## 4. Experiment

### 4.1. Dataset and Evaluation Metrics

**CVOGL** [37] is a large-scale cross-view object geo-localization dataset, which uses bounding box (bbox) annotations for query objects in reference images. It includes 5,836 high-resolution satellite images (1024×1024), 5,279 drone images (256×256), and 5,279 ground images (256×512), enabling two tasks: the “Drone → Satellite” task, using drone images as query images, and the “Ground → Satellite” task, using ground images as query images. In both tasks, satellite images serve as reference images. The CVOGL dataset is divided into two subsets corresponding to the distinct tasks. Each subset is further split into training (4,343), validation (923), and test (973) sets.

**Evaluation metrics.** On the CVOGL dataset, we use the  $\text{Acc}@K$  ( $K \in \{50\%, 25\%\}$ ) metric for evaluation [37].  $\text{Acc}@K$  represents the proportion of queries in the validation/test set whose IoU ratio between the predicted bbox and the ground truth bbox is greater than  $K$ .

**CVOGL-Seg.** To support the cross-view object segmentation (CVOS) scheme, we develop the CVOGL-Seg dataset, based on the CVOGL dataset [37]. It provides high-resolution (1024×1024) segmentation masks for each query object within satellite images, totaling 6,079 masks. The supported tasks and dataset partitions are identical to those of the CVOGL dataset. CVOGL-Seg leverages the bbox annotations provided in CVOGL, which are linked to OpenStreetMap [1] to generate mask annotations for objects. These annotations are subsequently subjected to manual inspection to ensure their quality and accuracy. The dataset includes objects of various categories and sizes, such as

Method	Drone → Satellite				Ground → Satellite			
	Validation		Test		Validation		Test	
	Acc@50%	Acc@25%	Acc@50%	Acc@25%	Acc@50%	Acc@25%	Acc@50%	Acc@25%
CVM-Net [18]	3.47%	20.04%	3.29%	20.14%	0.87%	5.09%	0.51%	4.73%
RK-Net [23]	3.03%	19.94%	2.67%	19.22%	0.98%	8.67%	0.82%	7.40%
L2LTR [43]	5.96%	38.68%	6.27%	38.95%	1.84%	12.24%	2.16%	10.69%
TransGeo [49]	5.42%	34.78%	6.37%	35.05%	3.25%	21.67%	2.88%	21.17%
SAFA [35]	6.39%	36.19%	6.58%	37.41%	3.25%	20.59%	3.08%	22.20%
Sample4Geo [11]	16.25%	52.87%	18.40%	52.83%	4.33%	24.27%	5.24%	25.80%
DetGeo [37]	55.15%	59.81%	57.66%	61.97%	43.99%	46.70%	42.24%	45.43%
VAGeo [22]	59.59%	64.25%	61.87%	66.19%	44.42%	47.56%	45.22%	48.21%
TROGeo (w/o OST)	65.87%	72.38%	68.65%	74.31%	44.20%	48.86%	45.53%	49.44%
<b>TROGeo (w OST)</b>	<b>66.63%</b>	<b>73.35%</b>	<b>68.96%</b>	<b>76.16%</b>	<b>46.59%</b>	<b>51.46%</b>	<b>46.56%</b>	<b>51.08%</b>

Table 2. Comparison with previous works on the CVOGL dataset. “w/o OST” and “w OST” refer to whether the Object Segmentation Task (OST) is learned in the Heterogeneous Task Training Stage (HTTS) of TROGeo. **Bold** denotes the best results.

buildings, baseball fields, and roundabouts, with areas ranging from under  $50 m^2$  to over  $20,000 m^2$ , making precise localization in cross-view images a challenging task. Additionally, the satellite images span three different area sizes ( $300 \times 300 m^2$ ,  $512 \times 512 m^2$ , and  $1024 \times 1024 m^2$ ), adding to the dataset’s complexity, utility, and challenge.

**Evaluation metrics.** On the CVOGL-Seg dataset, we use the mean Intersection over Union (**mIoU**), mean Dice coefficient (**mDice**), Absolute Area Error (**AAE**), and Meter Error (**ME**) to evaluate the effectiveness of the methods. The mIoU and mDice are commonly used metrics in semantic segmentation tasks [10, 13], while AAE and ME are commonly used to evaluate the area error of the predicted area from the actual area and the meter-level error of the predicted position from the actual position, respectively.

## 4.2. Implementation Details

In the HTTS, the Swin-S [27] is used as the Transformer encoder of the query/reference image, and the off-the-shelf weights pre-trained on ImageNet-1K [9] are used for initialization. In the CVOPM, the embedding dimension is set to 768, the number of heads for MHSA and MHCA is 12, and the dimension of each head is 64 [33]. We use the Adam optimizer [20] with an initial learning rate of 0.0001, decaying by half every 10 epochs, a batch size of 8, and training 25 epochs. In the SPS, we use the SAM Huge [21] model and freeze all trainable parameters.

## 4.3. Comparison with State-of-the-art Methods

**Rectangular Shackles: Defining Geographic Locations of Objects Using Rectangular Regions.** The CVOGL dataset only provides bounding box (bbox) annotations, supporting the evaluation of cross-view object geo-localization with rectangular regions as outputs. For the retrieval-based scheme, we select some of the classic and outstanding methods for comparison in the cross-view image geo-localization, such as CVM-Net [18], RK-Net [23], L2LTR [43], TransGeo [49], SAFA [35], and Sample4Geo [11]. Sample4Geo is the best retrieval-based

method implemented by us; other results are from [37]. In the training phase, the retrieval-based scheme uses the best matching patch with the query image as the positive sample and the rest as the negative sample. In the inference phase, for each query image, the most similar patch is retrieved from all the partitioned patches of its corresponding reference image. The detection-based methods are DetGeo [37] and VAGeo [22]. DetGeo, for the first time, uses bboxes in a detection setting to define the query object location and performs significantly better than retrieval-based methods.

To make a fair comparison, our TROGeo does not include the SAM Prompt Stage (SPS) by default when running on the CVOGL dataset and only uses the bbox outputs in the HTTS for evaluation. Our method’s comparison with state-of-the-art methods on the CVOGL dataset is shown in Table 2. For our method, we report two configurations: TROGeo (w/o OST) and TROGeo (w OST). For the case with only bbox annotations, the OST in TROGeo is supervised by the *pseudo segmentation masks* generated after prompting SAM [21] with the bbox annotations. The results demonstrate that TROGeo achieves superior performance, particularly excelling on the Drone → Satellite task, where it outperforms the second-best method by a large margin. The above experiments demonstrate the superior advantage of our method in handling cross-view object geo-localization under rectangular shackles.

**Breaking Rectangular Shackles: Defining Geographic Locations of Objects Using Segmentation Masks.** The CVOGL-Seg dataset provides segmentation mask annotations of the objects to support the cross-view object segmentation. We select the two best-performing methods on the CVOGL dataset, Sample4Geo [11] (retrieval-based) and DetGeo [37] (detection-based) as baselines. These methods only take rectangular regions as outputs, and we first report the metrics calculated with the ground truth masks after converting the rectangular regions to rectangular masks. Then, we apply the SPS to these methods so that they can produce segmentation masks, and then report the relevant metrics, as shown in Table 3. Figure 5 shows the visualiza-

Method	Drone → Satellite								Ground → Satellite							
	Validation				Test				Validation				Test			
	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ ( $m^2$ )	ME ↓ (m)	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ ( $m^2$ )	ME ↓ (m)	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ ( $m^2$ )	ME ↓ (m)	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ ( $m^2$ )	ME ↓ (m)
Sample4Geo [11]	16.83	25.41	3736.58	62.96	16.62	25.29	3858.20	57.49	5.30	8.56	3736.58	127.92	5.92	9.56	3858.20	128.81
Sample4Geo + SPS	19.21	25.37	2270.92	62.51	18.67	24.83	2178.15	57.47	5.03	6.94	2403.32	128.58	6.26	8.57	2347.28	130.22
DetGeo [37]	30.81	39.40	2885.87	80.69	30.50	39.12	2517.70	85.35	25.77	32.65	2510.07	100.89	25.40	32.27	3005.94	103.71
DetGeo + SPS	42.68	48.24	1096.89	80.19	42.27	47.98	1149.01	85.06	34.76	39.18	1229.54	100.46	34.04	38.69	1721.25	103.39
<b>TROGeo</b>	<b>55.54</b>	<b>62.80</b>	<b>762.15</b>	<b>49.22</b>	<b>56.59</b>	<b>64.24</b>	<b>927.00</b>	<b>50.45</b>	<b>38.79</b>	<b>43.76</b>	<b>1169.61</b>	<b>90.59</b>	<b>38.12</b>	<b>43.34</b>	<b>1126.56</b>	<b>95.02</b>

Table 3. Comparison with previous works on the CVOGL-Seg dataset. “+ SPS” denotes that our SAM Prompt Stage (SPS) is added to obtain the segmentation mask using the original rectangular region output as a prompt.

tion results. Experimental results show that the performance of our method has a significant advantage over previous methods, verifying the superiority of TROGeo in handling CVOS to accomplish fine-grained localization of irregular objects. Meanwhile, the use of SPS effectively improves the performance of existing methods, demonstrating its superiority. After adding SPS, there is a significant performance gap between other methods and our method, which shows that our method provides more accurate prompts for SAM.

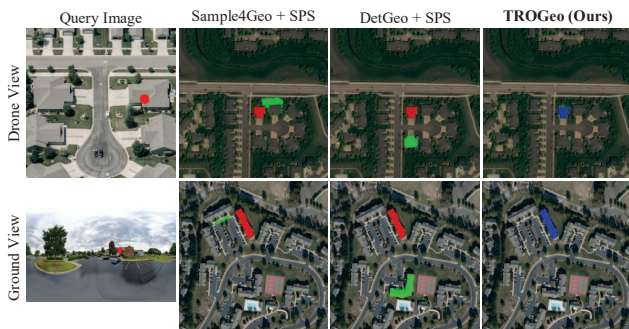


Figure 5. Visual comparison with state-of-the-art methods. Click points are indicated by red dots in the query images. Red, green and blue regions represent ground truth, prediction and overlapping regions, respectively. Best viewed on screen with zoom-in.

#### 4.4. Ablation Studies

To evaluate each component of TROGeo, we performed ablation studies on the CVOGL-Seg and CVOGL datasets.

**Importance of SAM Prompt Stage (SPS).** The SPS is essential for effective cross-view object segmentation. To examine its impact, we removed it, predicting only the bounding boxes and converting them into rectangular masks, referred to as “w/o SPS”. The qualitative results are shown in the third and fourth columns of Figure 7. The quantitative results are shown in Table 4. Simple rectangular masks can hardly accurately depict irregular objects and use the background or other objects as location proposals. In addition, SPS can be effectively combined with existing methods to enhance their performance, as demonstrated by the improved results marked as “+ SPS” in Table 3.

Setting	Drone → Satellite			
	mIoU (%) ↑	mDice (%) ↑	AAE ( $m^2$ ) ↓	ME (m) ↓
<b>All Components (Ours)</b>	<b>55.54</b>	<b>62.80</b>	<b>762.15</b>	<b>49.22</b>
w/o SPS	39.62	50.90	2294.23	50.00
w/o CVOPM	35.45	40.05	1318.28	101.37
w/o OST	54.09	61.06	886.85	56.65
w/o Shared W.	37.91	42.60	1218.00	97.07

Setting	Ground → Satellite			
	mIoU (%) ↑	mDice (%) ↑	AAE ( $m^2$ ) ↓	ME (m) ↓
<b>All Components (Ours)</b>	<b>38.79</b>	<b>43.76</b>	<b>1169.61</b>	<b>90.59</b>
w/o SPS	27.95	35.84	2599.18	91.06
w/o CVOPM	35.62	40.12	1256.62	105.65
w/o OST	36.79	41.45	1245.60	96.33
w/o Shared W.	36.60	41.26	1264.66	100.68

Table 4. Ablation study of different components of TROGeo on the CVOGL-Seg validation set. Segmentation masks serve as the output results (breaking rectangular shackles).

**Impact of different SAM prompts.** In order to investigate the impact of different SAM prompts, we provide different external prompts for the SPS: point, mask, bbox, and bbox+point, where the point is the center of the bbox, and the prompts are all generated by the HTTS. The ablation study on the CVOGL-Seg dataset is shown in Figure 6. The bbox+point achieves the best performance since it is more precise compared to the ambiguity of the point and bbox. SAM usually requires the mask to come from a previous iteration of itself, and the mask we produce in the HTTS is a coarse, low-resolution mask, so it is not the best choice.

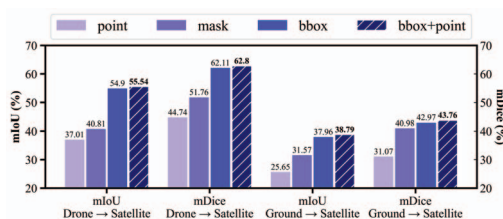


Figure 6. Ablation study of different SAM prompts.

**Effectiveness of Cross-View Object Perception Module (CVOPM).** We use the query image feature map  $F_q$  and the reference image feature map  $F_r$  to replace CVOPM directly by element-wise summation, and the experimental results are shown in Tables 4 and 5. Removing the CVOPM significantly affects the performance of the model on the CVOGL-Seg and CVOGL datasets. This highlights the benefits of

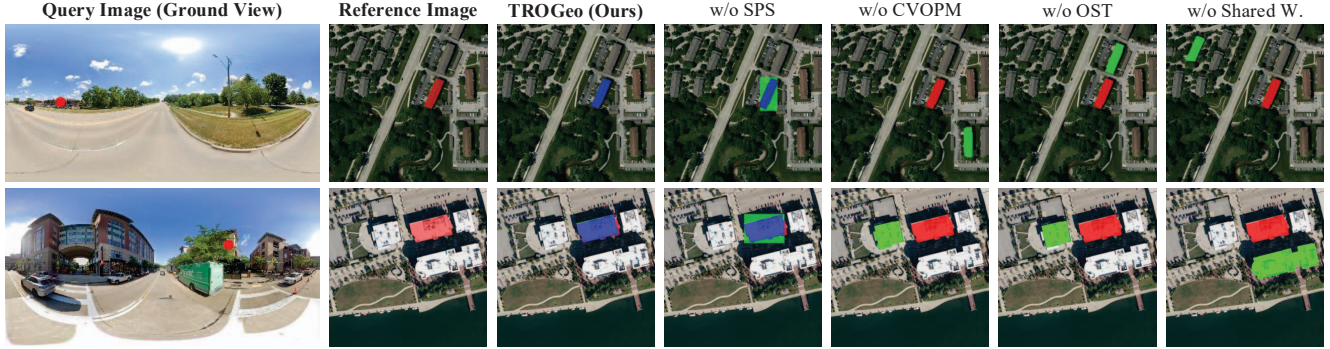


Figure 7. Visual comparison on the CVOGL-Seg (Ground → Satellite) test set. Description is the same as Figure 5.

using CVOPM to automatically perceive object contextual information and discover critical localization cues for cross-view object geo-localization.

**Effectiveness of Object Segmentation Task (OST).** OST is optional in the TROGeo framework. When OST is disabled, the performance of the model on both CVOGL-Seg and CVOGL datasets will decrease, as shown in the Tables 4 and 5. This is because OST provides more detailed supervision on the object region, rather than treating the entire bbox as the object, which causes semantic ambiguity.

**Effectiveness of Shared Weights (Shared W.).** For our method, we ablate both unshared (w/o Shared W.) and shared weights configurations for the query image and the reference image Transformer encoders. The experimental results are shown in Tables 4 and 5. Shared weights show better performance, which we believe is because shared weights reduce architectural complexity while retaining discriminability to improve generalization and performance.

Setting	Drone → Satellite		Ground → Satellite	
	Acc@50%↑	Acc@25%↑	Acc@50%↑	Acc@25%↑
<b>All Components (Ours)</b>	<b>66.63%</b>	<b>73.35%</b>	<b>46.59%</b>	<b>51.46%</b>
w/o CVOPM	43.88%	47.67%	43.99%	47.13%
w/o OST	65.87%	72.37%	44.20%	48.86%
w/o Shared W.	45.29%	50.27%	44.53%	48.65%

Table 5. Ablation study of different components of TROGeo on the CVOGL validation set. Rectangles serve as the output results.

**Hyperparameter Selection.** Our heterogeneous task optimization objective contains three hyperparameters:  $\alpha$ ,  $\beta$ , and  $\lambda$ . The value of  $\beta$  is set to 15, which is obtained by calculating the ratio of foreground to background in the training set masks. The value of  $\lambda$  is usually set to 0.8 [10]. The hyperparameter  $\alpha$  is used to control the relative importance of the two tasks, *i.e.* Object Detection Task (ODT) and Object Segmentation Task (OST). Table 6 shows the experimental results when  $\alpha$  takes different values. The results show that the value of  $\alpha$  has a significant impact on the results, and when  $\alpha = 2$ , the result is optimal. It can be seen that in the heterogeneous task, it is crucial to reason-

ably allocate the weights of ODT and OST.

$\alpha$	CVOGL		CVOGL-Seg			
	Acc@50%↑	Acc@25%↑	mIoU ↑	mDice ↑	AAE ↓	ME ↓
1	65.76%	73.46%	54.80%	61.86%	765.93 $m^2$	54.44 $m$
2	<b>68.47%</b>	<b>74.87%</b>	<b>55.54%</b>	<b>62.80%</b>	<b>762.15 <math>m^2</math></b>	<b>49.22 <math>m</math></b>
4	65.44%	71.51%	53.37%	60.17%	793.81 $m^2$	60.60 $m$
8	66.96%	73.69%	54.75%	61.80%	834.02 $m^2$	57.30 $m$

Table 6. Ablation study of different values of  $\alpha$  on the CVOGL and CVOGL-Seg validation sets (Drone → Satellite).

**Qualitative Analysis.** We visualized some localization results for qualitative comparison with w/o SPS, w/o CVOPM, w/o OST, and w/o Shared W. as shown in Figure 7. As illustrated in the figure, rectangular masks struggle to perfectly cover irregularly shaped objects (w/o SPS). Meanwhile, removing CVOPM, Shared W., or OST leads to varying degrees of localization errors (completely wrong results or partially correct results). These visualizations highlight the necessity of employing cross-view object segmentation to frame the cross-view object geo-localization task, facilitating fine-grained object localization and demonstrating the effectiveness of each component of TROGeo.

## 5. Conclusion

In this paper, we propose a novel Cross-View Object Segmentation (CVOS) scheme to break the “rectangular shackles” of existing (retrieval-based and detection-based) schemes by predicting pixel-level segmentation masks of query objects for fine-grained geo-localization. We create a new CVOGL-Seg dataset to support and evaluate this scheme. To address the challenge of CVOS, we propose a Transformer Object Geo-localization (TROGeo) framework, which consists of two stages. Extensive experiments show that our method achieves state-of-the-art performance compared to existing models. This study demonstrates the necessity of segmentation masks to represent the geographic location of objects, opening up new possibilities for fine-grained cross-view object geo-localization.

## Acknowledgement

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

## References

- [1] <https://www.openstreetmap.org>. 5
- [2] Woo-Jin Ahn, So-Yeon Park, Dong-Sung Pae, Hyun-Duck Choi, and Myo-Taeg Lim. Bridging viewpoints in cross-view geo-localization with siamese vision transformer. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [3] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, and Jameela Al-Jaroodi. Applications of big data to smart cities. *Journal of Internet Services and Applications*, 6(1):1–15, 2015. 1
- [4] Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4
- [5] L Sant'anna Bins, LM Garcia Fonseca, Guaraci José Erthal, and F Mitsuo Ii. Satellite imagery segmentation: a region growing approach. *Simpósio Brasileiro de Sensoriamento Remoto*, 8(1996):677–680, 1996. 3
- [6] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8391–8400, 2019. 1
- [7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2
- [8] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinfeld. A tutorial on the cross-entropy method. *Annals of operations research*, 134:19–67, 2005. 5
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [10] Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9622–9631, 2024. 3, 5, 6, 8
- [11] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 1, 2, 4, 6, 7
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2
- [13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 6
- [14] Ronny Hänsch, Jacob Arndt, Dalton Lunga, Matthew Gibb, Tyler Pedelose, Arnold Boedihardjo, Desiree Petrie, and Todd M Bacastow. Spacenet 8-the detection of flooded roads and buildings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1472–1480, 2022. 1
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4
- [16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2
- [17] Danfeng Hong, Jingliang Hu, Jing Yao, Jocelyn Chanussot, and Xiao Xiang Zhu. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 178:68–80, 2021. 1
- [18] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 1, 2, 6
- [19] Duojuan Huang, Xinyu Xiong, Jie Ma, Jichang Li, Zequn Jie, Lin Ma, and Guanbin Li. Alignsam: Aligning segment anything model to open context via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3205–3215, 2024. 3, 5
- [20] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 5, 6
- [22] Zhongyang Li, Xin Yuan, Wei Liu, and Xin Xu. Vageo: View-specific attention for cross-view object geo-localization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 2, 6
- [23] 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. 1, 6
- [24] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 3

- [25] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2
- [26] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 1, 2
- [27] 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*, pages 10012–10022, 2021. 2, 6
- [28] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations. *arXiv preprint arXiv:2403.13965*, 2024. 1
- [29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5
- [30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 3
- [31] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. 5
- [32] R Paulus. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017. 2
- [33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 4, 6
- [34] Yujiao Shi and Hongdong Li. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17010–17020, 2022. 2
- [35] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 6
- [36] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 2
- [37] Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geolocalization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 2, 3, 4, 5, 6, 7
- [38] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 2
- [39] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2, 4
- [40] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 2
- [41] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4
- [42] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024. 1
- [43] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34:29009–29020, 2021. 1, 2, 6
- [44] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017. 2
- [45] Chenhui Zhang and Sherrie Wang. Good at captioning, bad at counting: Benchmarking gpt-4v on earth observation data. *arXiv preprint arXiv:2401.17600*, 2024. 1
- [46] Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7251–7259, 2024. 1, 2
- [47] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [48] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2
- [49] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 1, 2, 6