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

Qingwang Zhang, Yingying Zhu✉
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
zhangqingwang2022@email.szu.edu.cn, zhuyy@szu.edu.cn

Overview

In this supplementary material, we provide the following items for a better understanding of our main paper.

1. CVOGL-Seg dataset. § 1
2. Theoretical upper bounds. § 2
3. Qualitative analysis. § 3
4. Computational costs. § 4
5. Implementation details. § 5

1. CVOGL-Seg Dataset

CVOGL-Seg, introduced in this paper, represents the first dataset specifically developed for cross-view object segmentation. While the main paper provides a general overview, this section offers a more detailed introduction.

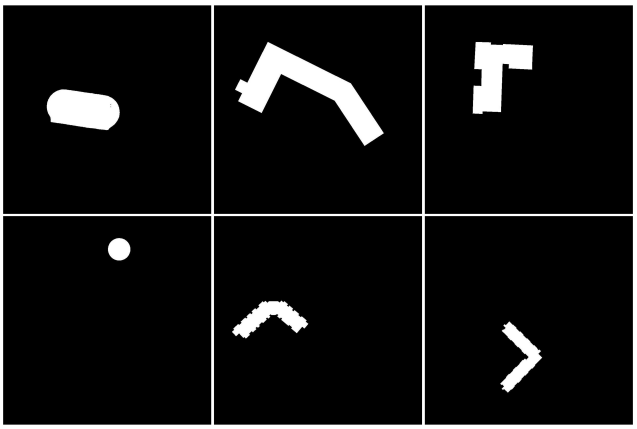


Figure 1. Examples of object mask annotations in the CVOGL-Seg dataset (white pixels represent the objects/foregrounds, and black pixels represent the backgrounds).

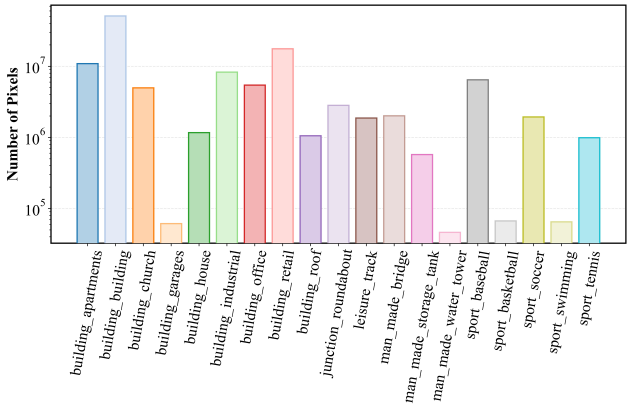
Mask Annotation. The CVOGL-Seg dataset creates segmentation mask annotations for objects using bounding box (bbox) annotations from the CVOGL dataset [8]. The pipeline involves converting the bbox coordinates to GPS formats, linking them to OpenStreetMap [1] to find relevant elements in the specified region, and then creating masks for the corresponding elements. These masks are saved as binary images (1024×1024), where white pixels represent the foreground and black pixels represent the background. Figure 1 illustrates examples of object mask annotations, highlighting the diversity of object shapes.

Figure 2. Comparison of simultaneously visualizing the *bounding boxes* and *mask annotations* onto satellite images.

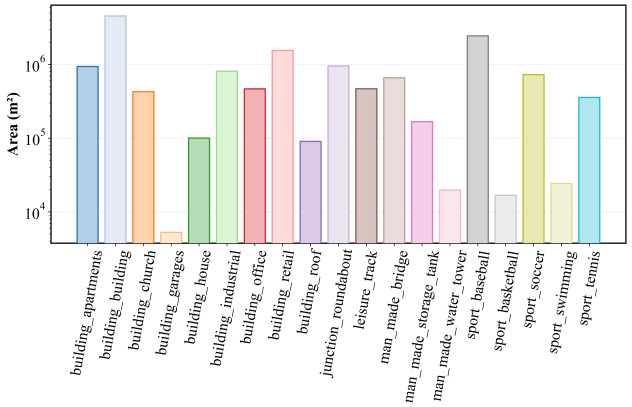
Annotation Quality. To ensure the high quality of the annotations, we manually review all masks by comparing them against their corresponding bounding boxes. Any masks with noticeable errors are re-annotated using the SAM annotation tool [5]. Figure 2 shows the comparison of simultaneously visualizing the bounding boxes and mask annotations onto satellite images. From these comparisons, we can observe that the object mask annotations generated by our pipeline closely correspond to the existing bounding box (bbox) annotations. Additionally, the mask annotations can perfectly cover the object without considering other regions (non-masked regions within the bboxes) as objects as well, leading to imprecise semantic annotations.

Dataset Statistics. Figures 3a and 3b illustrate the distri-

bution of the number of annotated pixels and the area occupied by different categories in the CVOGL-Seg dataset. These results reflect the compositional characteristics of the dataset and provide an important reference for future expansion of the dataset and model training. The CVOGL-Seg dataset is partitioned in the same way as the CVOGL [8] dataset, which also supports the Drone → Satellite task and the Ground → Satellite task. Both tasks aim to locate objects based on cross-view images. The primary distinction lies in the view of the query images: in the Drone → Satellite task, they are captured from drone view, whereas in the Ground → Satellite task, they are taken from ground view. Table 1 shows the statistics of the CVOGL-Seg dataset.



(a) Finely annotated pixels per category.



(b) Finely annotated areas per category.

Figure 3. Finely annotated pixels and areas per category.

Task	Split	#Instance	#Query Image	#Reference Image	#Mask Pixel
Drone → Satellite & Ground → Satellite	Traing	4,343	3,674	4,063	83,471,648
	Validation	923	793	863	17,718,563
	Test	973	812	910	16,695,911

Table 1. Statistics of the CVOGL-Seg dataset.

2. Theoretical Upper Bounds

In the main paper, we directly present the theoretical upper bounds for different schemes in achieving pixel-level precise object localization. In this section, we will provide a detailed explanation of how these upper bounds are derived.

Retrieval-based scheme. In cross-view image geo-localization [2, 3, 7, 10], the reference images (satellite images) in the reference database are always square-shaped, representing special rectangles with equal width and height. Thus, when employing a retrieval-based method to frame cross-view object geo-localization, we use squares to partition the satellite images. Specifically, for a given satellite image, sliding windows with varying sizes (w) and strides (s) are used for sampling. A common configuration involves setting $w = s$, as illustrated in Figure 4, which provides examples of different sampling parameters used to construct the reference image database.

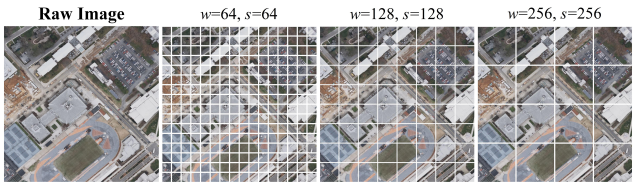


Figure 4. Comparison when w and s are set to different values.

Setting	Drone → Satellite & Ground → Satellite							
	Validation				Test			
	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ (m^2)	ME ↓ (m)	mIoU ↑ (%)	mDice ↑ (%)	AAE ↓ (m^2)	ME ↓ (m)
$w = 64, s = 64$	22.86	35.29	**2177.36**	11.96	23.16	35.51	**2020.65**	12.07
$w = 128, s = 128$	23.80	36.36	3736.58	22.74	23.89	36.38	3858.20	23.19
$w = 256, s = 256$	13.68	22.44	15017.48	45.01	13.15	21.64	16469.80	47.19
$w = 64, s = 32$	27.54	40.63	**2177.36**	**7.18**	28.13	41.14	**2020.65**	**7.15**
$w = 128, s = 64$	**32.25**	**45.89**	3736.58	11.98	**31.32**	**44.84**	3858.20	12.95
$w = 256, s = 128$	18.60	28.92	15017.48	22.79	18.04	28.07	16469.80	24.51

Table 2. Theoretical upper bounds of the retrieval-based scheme for different settings of w and s on the CVOGL-Seg dataset.

We use an overlapping sampling strategy [9], *i.e.*, the setting $s = \frac{w}{2}$, to improve the construction of the reference image database. We select the patch with the highest overlap and closest centroid to the mask annotation in the CVOGL-Seg dataset as the retrieval result from the reference image database (consisting of pre-sampled satellite image patches). Then, we convert the square region of the selected patch in the satellite image into the retrieval mask used to compute the theoretical upper bound. The theoretical upper bounds of the retrieval-based scheme under different w and s configurations on the CVOGL-Seg dataset are shown in Table 2. Experimental results indicate that the configuration $w = 128$ and $s = 64$ provides the most balanced performance. Accordingly, we report the corresponding results in the main paper.

It is important to emphasize that while more sophisticated sampling strategies may further enhance the performance of the retrieval-based scheme, the rectangular shackles remain inherent limitations of this scheme.

Detection-based scheme. The bounding boxes provided by the CVOGL [8] dataset are the best results that can be obtained by the detection-based scheme, as this is the optimization objective of the detection scheme. We convert the bounding box annotations in the CVOGL dataset to detection masks and then compute the theoretical upper bound with the mask annotations in the CVOGL-Seg dataset, as shown in Table **1** in the main paper.

Since the detection-based scheme can theoretically identify arbitrary rectangles, the theoretical upper bound of the retrieval-based scheme (retrieving only from pre-sampled rectangles) cannot surpass that of the detection-based scheme. However, the detection-based scheme is still limited to rectangular shackles.

Segmentation-based scheme. In the segmentation setting, pixel-wise mask outputs can represent any irregularly shaped objects. Therefore, the segmentation-based scheme can theoretically achieve perfect pixel-level localization of any object, leading to optimal performance as shown in Table **1** of the main paper.

Figure 5 illustrates the optimal masks that can theoretically be obtained for different schemes. We treat the irregular object segmentation mask as homogeneous and use its centroid as the object's position coordinates (the mean of the horizontal and vertical coordinates). Rectangular masks are a special case of irregular masks.

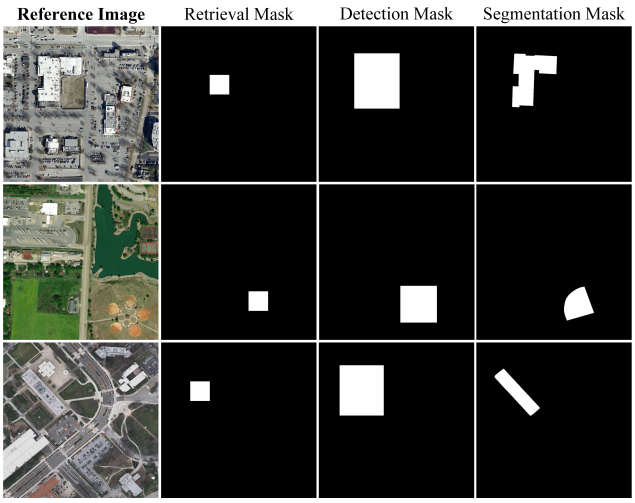


Figure 5. Comparison of the optimal masks that can be theoretically obtained in different schemes. Retrieval masks are transformed from patches (uniformly sized squares). Detection masks are transformed from bounding boxes (flexibly sized rectangles). Segmentation masks can take arbitrary, irregular shapes.

3. Qualitative Analysis

In the main paper, we show some localization visualization results on the CVOGL-Seg (Drone $\rightarrow$ Satellite) test set. In Figures 7 and 8, we provide more localization visualization results on the CVOGL-Seg test set for the Ground $\rightarrow$ Satellite and Drone $\rightarrow$ Satellite tasks. These visualizations highlight the necessity of employing cross-view object segmentation to frame the cross-view object geo-localization task, facilitating pixel-level, fine-grained object localization and demonstrating the effectiveness of each component of Transformer Object Geo-localization (TROGeo).

Failure Cases and Challenging Task. Figure 6 illustrates several failure cases. TROGeo incorrectly identifies objects with appearances highly similar to the query object and located nearby, resulting in localization failures. These outcomes are understandable, as even remote sensing experts may struggle to accurately distinguish these objects. Moreover, it highlights that cross-view object geo-localization (cross-view object segmentation) is an exceptionally challenging task, warranting further exploration by researchers.

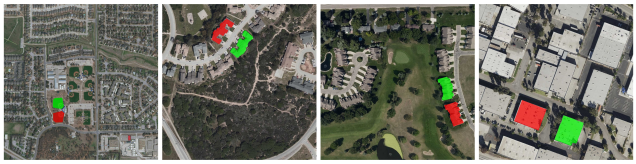


Figure 6. Localization failure cases. *Red* and *green* regions represent *ground truth* and *prediction* regions, respectively. Best viewed on screen with zoom-in.

4. Computational Costs

In this section, we discuss the computational costs of representative methods for different schemes. The experimental results are shown in Table 3, and the metrics computations are calculated by processing one drone image (256×256) and one satellite image (1024×1024) from the Drone $\rightarrow$ Satellite task on a single NVIDIA V100 GPU. Sample4Geo [2] is the best retrieval-based method ($w = 128$, $s = 64$), but there is a significant gap between its Frames Per Second (FPS) at training and testing. This is because the method selects only the best patch (128×128) for training, while all 225 patches ($225 = ((1024 - 64)/64)^2$) need to be computed to retrieve the best patch during testing. DetGeo [8] is the best detection-based method and outputs bounding boxes in both the training and testing so that the relevant metrics remain consistent. TROGeo is the segmentation-based method with the least trainable parameters in the training stage. In the testing stage, the introduction of SAM [5] (freezing all parameters) to obtain highly accurate object masks increases the parameters and reduces the FPS. Nevertheless, TROGeo performs much better than

the other methods in both detection and segmentation settings and is able to provide pixel-level object localization information, which is difficult to achieve by the other methods. TROGeo completes a single instance of pixel-level cross-view object geo-localization in approximately 0.6 s.

Method	Parameters (M) ↓	FPS ↑	Output	Acc@50% (%) ↑	mIoU (%) ↑
Sample4Geo 🔥 [2]	87.57	**31.05**	patch	18.40	16.62
Sample4Geo ❄ [2]		0.30			
DetGeo 🔥 [8]	73.80	26.86	bbox	57.66	30.50
DetGeo ❄ [8]					
TROGeo 🔥	**71.28**	14.80	bbox	**70.09**	40.83
TROGeo ❄	707.66	1.59	mask		**56.59**

Table 3. Comparison of computational costs of different methods. All methods are tested on a single NVIDIA V100 GPU. One drone image and one satellite image are processed at a time. 🔥 denotes the training and ❄ denotes the testing.

5. Implementation Details

Our method is implemented based on PyTorch [6]. The drone images, ground images, and satellite images are input into our model with resolutions of 256×256, 256×512, and 1024×1024, respectively, following [8]. We use Adam [4] optimizer to train our model with a learning rate of 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Our model is trained for 25 epochs on two NVIDIA V100 GPUs with a batch size of 8. During training, we apply random flipping and rotation to satellite images, while horizontal flipping is applied to query images (data augmentation).

References

[1] https://www.openstreetmap.org. 1

[2] 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. 2, 3, 4

[3] 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. 2

[4] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[5] 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. 1, 3

[6] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 4

[7] 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. 2

[8] Yuxi Sun, Yunming Ye, Jian Kang, Ruben Fernandez-Beltran, Shanshan Feng, Xutao Li, Chuyao Luo, Puzhao Zhang, and Antonio Plaza. Cross-view object geo-localization in a local region with satellite imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 2023. 1, 2, 3, 4

[9] 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

[10] 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. 2

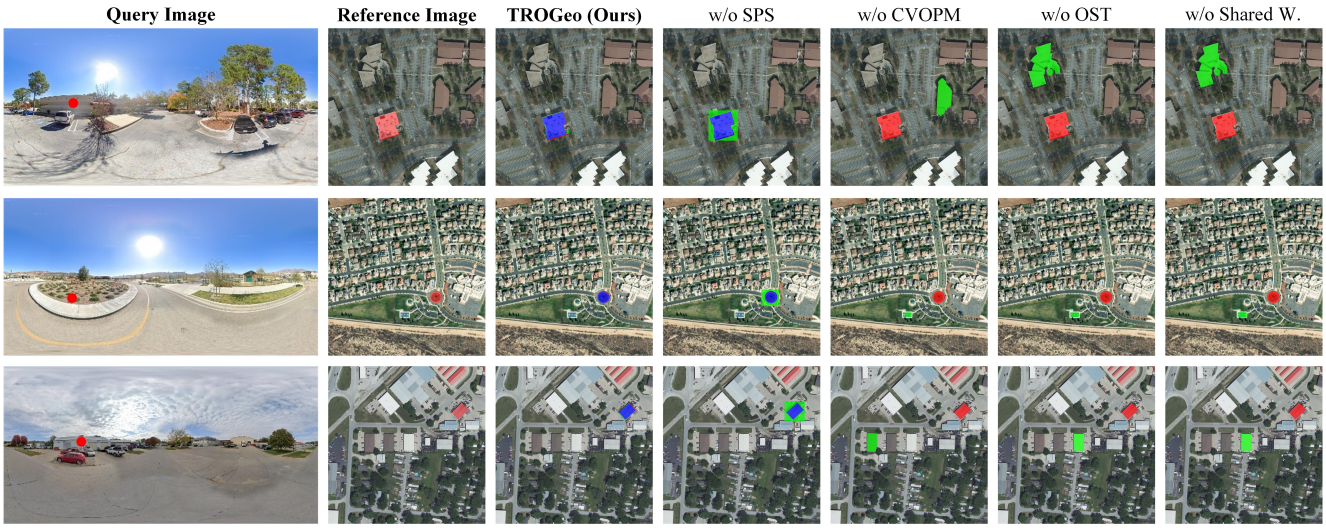


Figure 7. Visual comparison on the CVOGL-Seg (Ground → Satellite) test set. Click points are indicated by *red* dots in the query images. *Red*, *green* and *blue* regions represent *ground truth*, *prediction* and *overlapping* regions, respectively. Best viewed on screen with zoom-in.

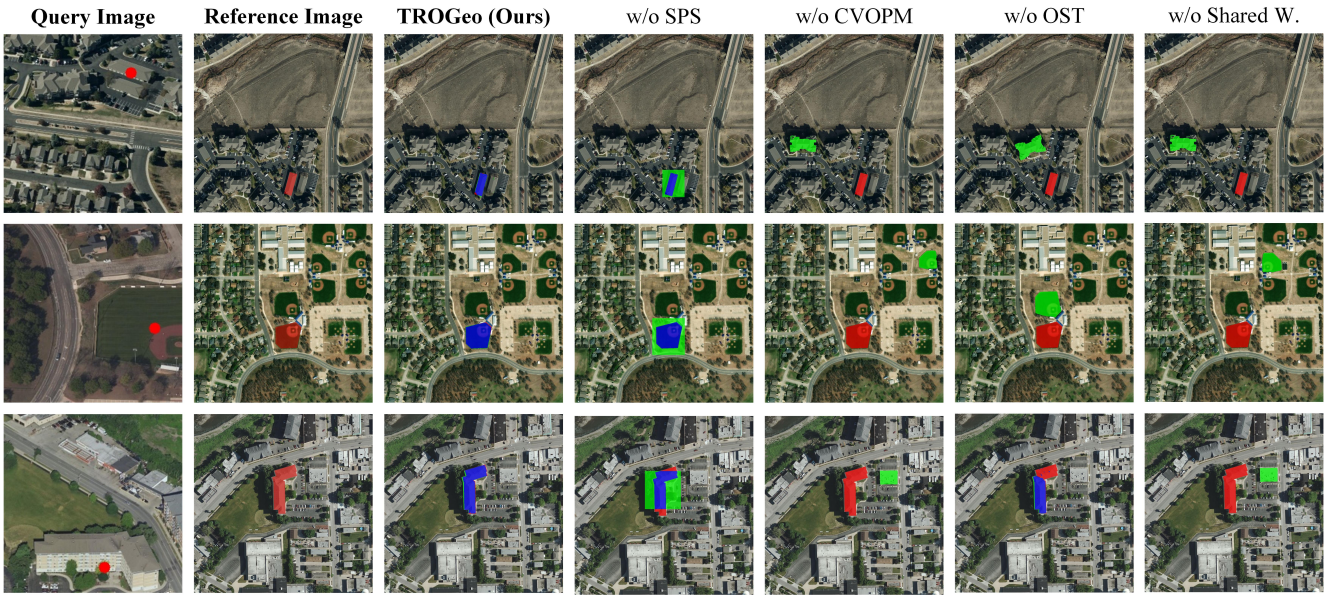


Figure 8. Visual comparison on the CVOGL-Seg (Drone → Satellite) test set. Click points are indicated by *red* dots in the query images. *Red*, *green* and *blue* regions represent *ground truth*, *prediction* and *overlapping* regions, respectively. Best viewed on screen with zoom-in.