Supplementary Material: AerOSeg: Harnessing SAM for Open-Vocabulary Segmentation in Remote Sensing Images

Saikat Dutta[1,2,3] Akhil Vasim[2] Siddhant Gole[2]
Hamid Rezatofighi[3] Biplab Banerjee[2]
[1]IITB-Monash Research Academy [2]IIT Bombay [3]Monash University

1. Ablation Study: Effect of Different VL backbones

In addition to CLIP ViT-B, we have experimented with CLIP ViT-L as our VL backbone. Furthermore, we have explored two Remote Sensing-specific CLIP models: GeoRSCLIP (ViT-L) [3] and SkyCLIP (ViT-L) [2]. Table-1 presents the quantitative results on the iSAID dataset for different VL backbones. Among them, CLIP ViT-L achieves the best performance. Notably, despite being trained on Remote Sensing data, GeoRSCLIP and SkyCLIP do not outperform CLIP ViT-L. This can likely be attributed to the lower image resolution used during their training.

VL backbone	Architecture	s-mIoU	u-mIoU	h-mIoU
CLIP	ViT-B	75.48	51.46	61.20
CLIP	ViT-L	79.85	66.80	72.74
GeoRSCLIP	ViT-L	75.47	50.32	60.38
SkyCLIP	ViT-L	66.75	57.79	61.95

Table 1. Quantitative comparison between different VL backbones on iSAID dataset.

2. Visualization of Refined Correlation feature maps

We visualize refined correlation maps for different classes, comparing the use of CLIP features versus SAM features as guidance in the Spatial Refinement block. Specifically, Fig. 1 presents refined correlation features of different object categories for the Baseline and Config-A models (as discussed in Sec. 4.5 of main text). The results show that when SAM refines the correlation features, the refined features achieve better localization of the object of interest.

3. Additional comparison with state-of-the-art

3.1. Comparison with SegEarth-OV

We have compared our model, AerOSeg with state-of-the-art training-free method SegEarth-OV [1]. Since for training-free OVS, seen-unseen class split is irrelevant, we report mIoU scores over all classes. From Table 2, we can see that our model performs significantly better than SegEarth-OV on all three datasets, highlighting the importance of domain-specific training. Fig. 2 shows qualitative comparison between SegEarth-OV and our model.

Methods	iSAID	DLRSD	OEM
SegEarth-OV	17.87	18.92	29.72
AerOSeg (Ours)	**65.87**	**51.62**	**49.71**

Table 2. Quantitative comparison between training-free method SegEarth-OV and our proposed model.

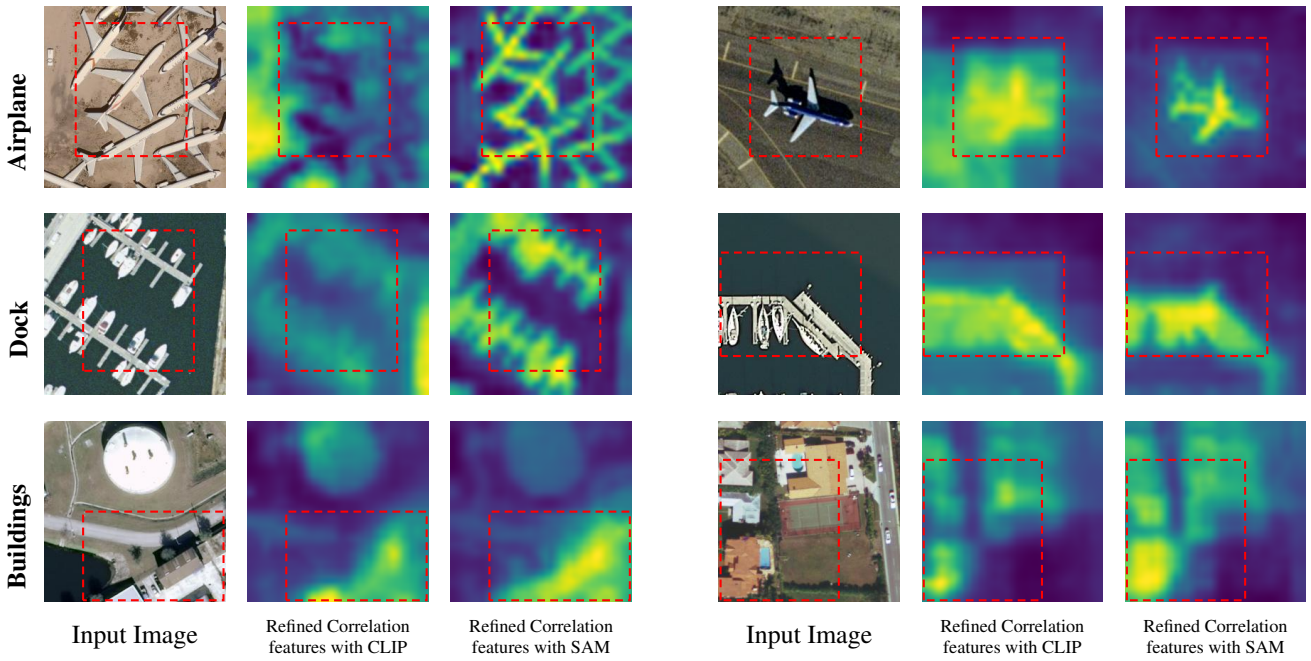
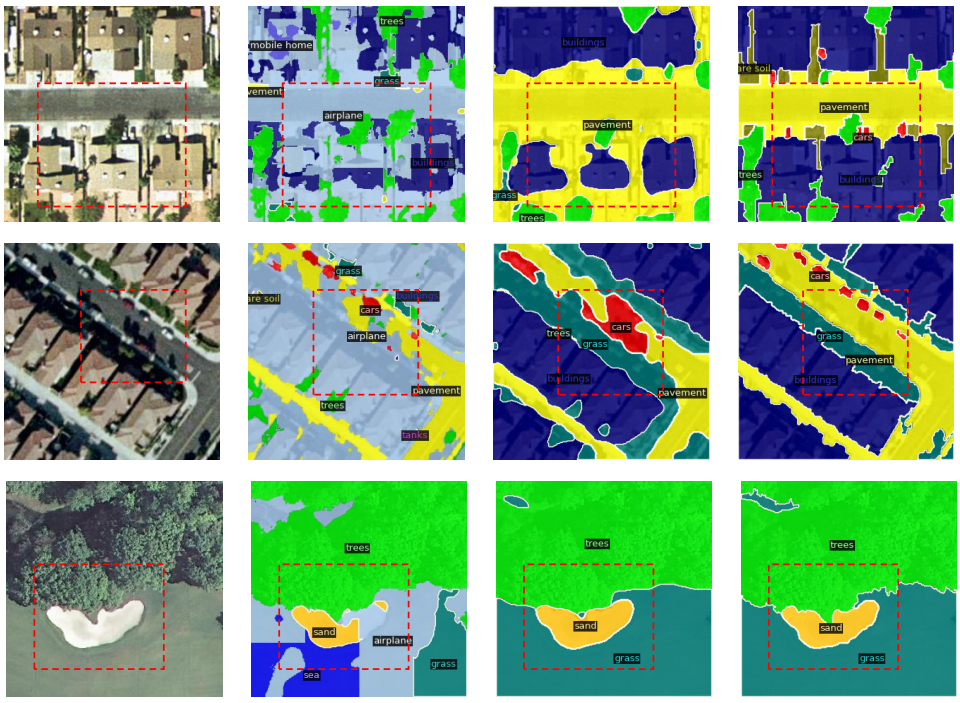


Figure 1. Refined Correlation feature visualization with CLIP and SAM guidance for different classes. SAM-refined correlation features can better discern between foreground and background for respective categories.

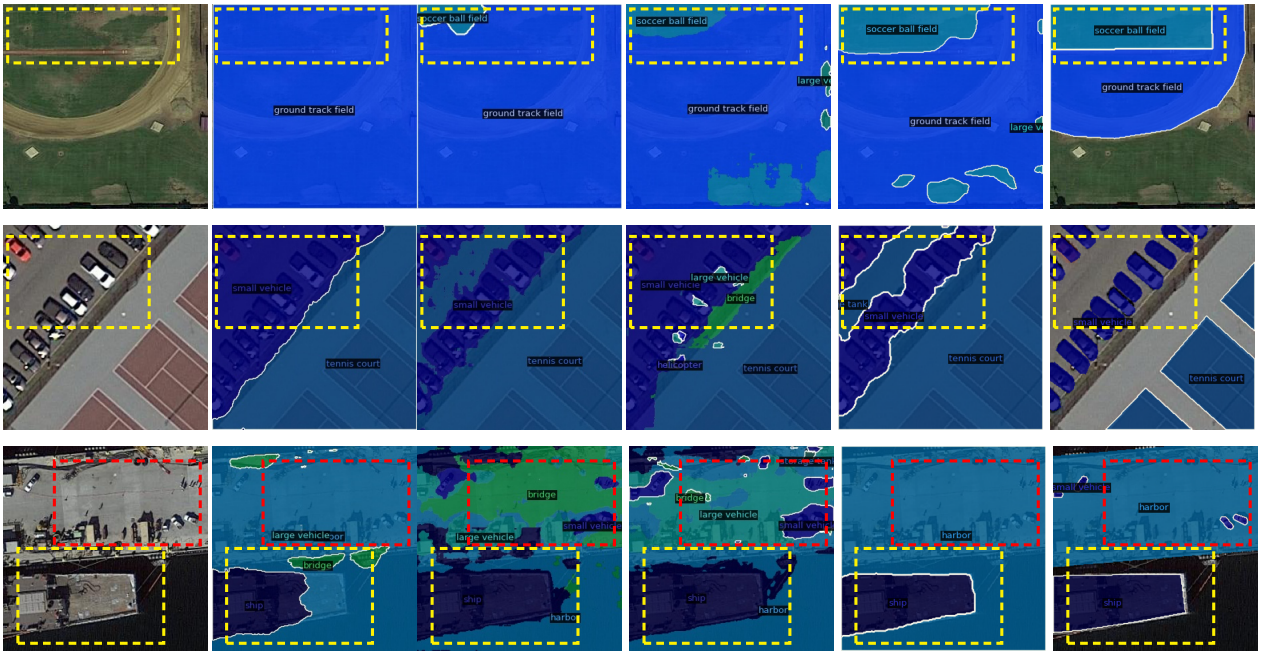


(a) Input Image (b) SegEarth-OV (c) AerOSeg (Ours) (d) Ground Truth

Figure 2. Qualitative comparison between SegEarth-OV and our method. Dashed bounding boxes highlight regions where our model achieves more precise segmentation.

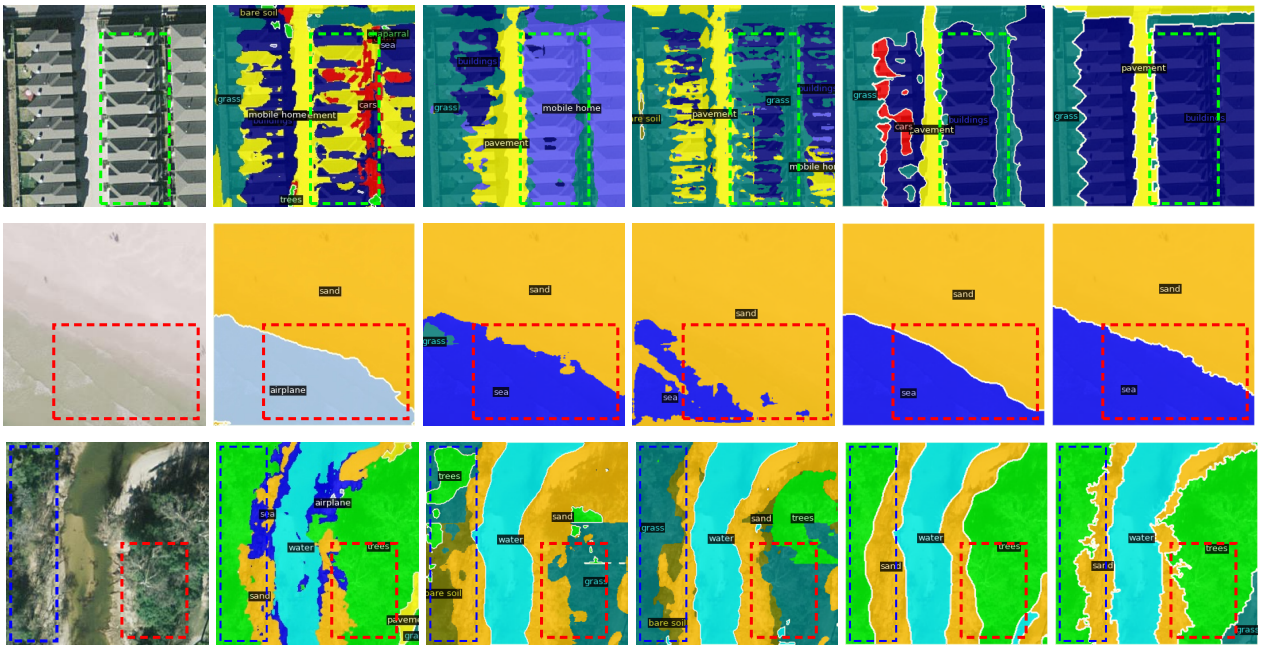
3.2. More qualitative results

In this section, we have shown additional qualitative comparisons with various state-of-the-art models on iSAID, DLRSD and OpenEarthMap datasets in Fig. 3, 4 and 5, respectively.



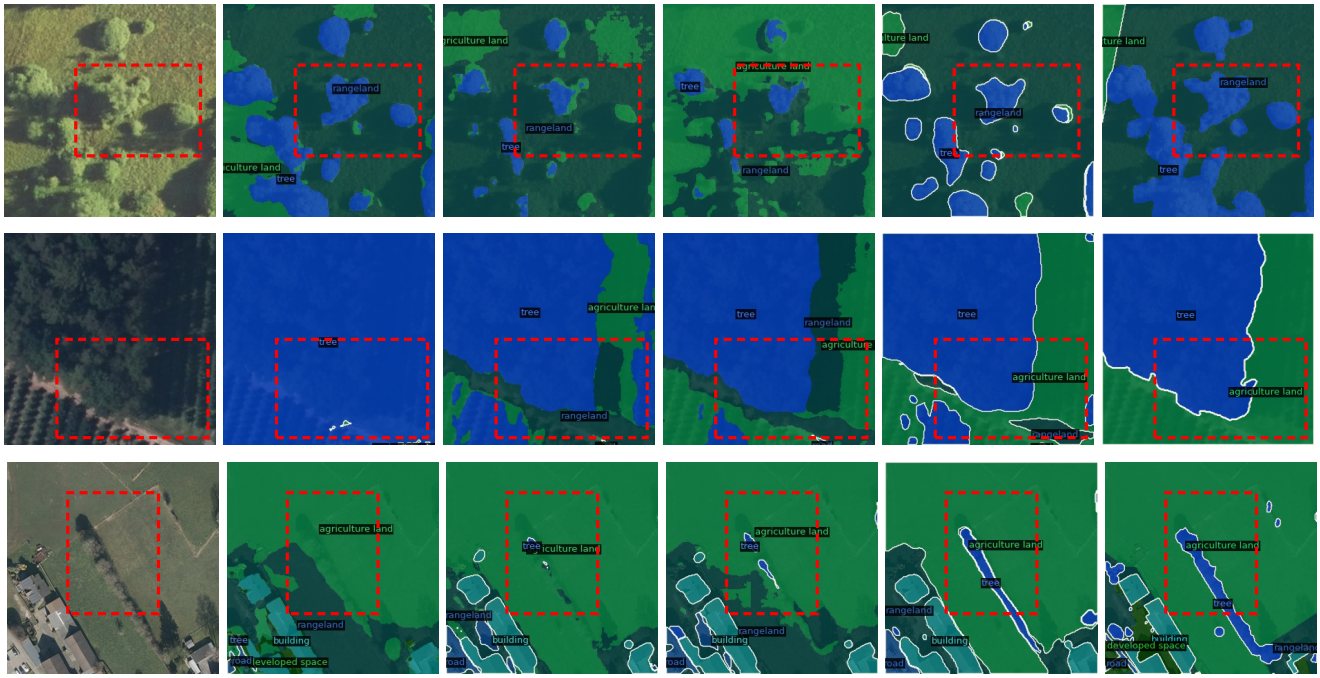
(a) Input Image (b) SED (c) CAT-Seg (d) OVRS (e) AerOSeg (Ours) (f) Ground Truth

Figure 3. Qualitative comparison with state-of-the-art methods on iSAID dataset. Dashed bounding boxes highlight regions where our model achieves more precise segmentation.



(a) Input Image (b) SED (c) CAT-Seg (d) OVRS (e) AerOSeg (Ours) (f) Ground Truth

Figure 4. Qualitative comparison with state-of-the-art methods on DLRSD dataset. Dashed bounding boxes highlight regions where our model achieves more precise segmentation.



(a) Input Image (b) SED (c) CAT-Seg (d) OVRS (e) AerOSeg (Ours) (f) Ground Truth

Figure 5. Qualitative comparison with state-of-the-art methods on OpenEarthMap dataset. Dashed bounding boxes highlight regions where our model achieves more precise segmentation.

References

[1] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. *arXiv preprint arXiv:2410.01768*, 2024. 1

[2] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 1

[3] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1