

SegEarth-OV3: Exploring SAM 3 for Open-Vocabulary Semantic Segmentation in Remote Sensing Images

Kaiyu Li^{1*}, Shengqi Zhang^{1*}, Yupeng Deng², Zhi Wang¹, Deyu Meng¹, Xiangyong Cao^{1†}

¹Xi'an Jiaotong University ² Chinese Academy of Sciences

Abstract

Most existing methods for training-free Open-Vocabulary Semantic Segmentation (OVSS) are based on CLIP. While these approaches have made progress, they often face challenges in precise localization or require complex pipelines to combine separate modules, especially in remote sensing scenarios where numerous dense and small targets are present. Recently, Segment Anything Model 3 (SAM 3) was proposed, unifying segmentation and recognition in a promptable framework. In this paper, we present a preliminary exploration of applying SAM 3 to the remote sensing OVSS task without any training. First, we implement a mask fusion strategy that combines the outputs from SAM 3's semantic segmentation head and the Transformer decoder (instance head). This allows us to leverage the strengths of both heads for better land coverage. Second, we utilize the presence score from the presence head to filter out categories that do not exist in the scene, reducing false positives caused by the vast vocabulary sizes and patch-level processing in geospatial scenes. We evaluate our method on extensive remote sensing datasets. Experiments show that this simple adaptation achieves promising performance, demonstrating the potential of SAM 3 for remote sensing OVSS. Our code is released at <https://github.com/earth-insights/SegEarth-OV-3>.

1. Introduction

Semantic segmentation is a fundamental task in remote sensing analysis, enabling dense, pixel-level understanding of Earth observation scenes. Traditionally, segmentation models were restricted to a closed set of predefined categories, limiting their applicability in dynamic, open-world scenarios where visual concepts are virtually infinite. To overcome this limitation, Open-Vocabulary Semantic Segmentation (OVSS) has emerged as a critical research di-

rection [40]. By leveraging the rich semantic knowledge embedded in pre-trained Vision-Language Models (VLMs), remote sensing OVSS aims to segment and recognize image regions based on arbitrary text descriptions, effectively generalizing to categories unseen during training. This capability is essential for diverse applications, *e.g.*, urban planning [37] and disaster monitoring [57], where the model must handle a vast vocabulary [38–40].

Currently, the dominant paradigm for training-free OVSS relies heavily on VLMs, particularly CLIP [48]. Early works, such as MaskCLIP [72] and SCLIP [55], attempt to extract dense features directly from the CLIP image encoder. However, CLIP is pre-trained for image-level classification, and adapting its patch-level features for pixel-level localization often results in coarse boundaries. To address this, subsequent research has focused on integrating auxiliary Visual Foundation Models (VFM). For instance, ProxyCLIP [34] and CorrCLIP [70] utilize structural guidance from DINO [47] and SAM [32] to refine CLIP's attention maps, while SegEarth-OV [39, 40] builds an upsampler to reconstruct high-resolution features. Although these methods improve boundary quality, they rely on complex pipeline and feature alignment to bridge the gap between different representations. Moreover, they lack simultaneous semantic and instance segmentation capabilities, limiting their utility in complex geospatial analysis.

Recently, the Segment Anything Model 3 (SAM 3) [10] was introduced. Unlike CLIP-based paradigm, SAM 3 is a unified model that supports promptable concept segmentation. It builds upon the DETR [9] and MaskFormer [16, 17] architectures, employing a query-based Transformer design. Crucially, SAM 3 utilizes a decoupled architecture in which a presence head is specifically designed to predict the probability that the prompted concept exists in the image. Meanwhile, a Transformer decoder and a semantic segmentation head generate precise masks for discrete instances and continuous semantic regions, respectively. Although SAM 3 demonstrates impressive zero-shot capabilities on some natural image semantic segmentation benchmarks, remote sensing images present distinct challenges,

[†]Corresponding author (caoxiangyong@mail.xjtu.edu.cn)

This report is only a preliminary version; more details and further exploration will be included in future updates.

e.g., the intricate coexistence of dense small objects and vast amorphous backgrounds. Therefore, a tailored exploration to adapt SAM 3 for geospatial scenarios remains valuable.

In this paper, we present a preliminary exploration of adapting SAM 3 for the remote sensing OVSS task without additional training. We investigate whether SAM 3’s unified architecture can offer a stronger, yet simpler baseline than complex CLIP-ensemble methods for Earth observation. Our proposed method, namely SegEarth-OV3, consists of two simple strategies tailored to SAM 3’s design:

- **(1) Dual-Head Mask Fusion:** Remote sensing images exhibit a distinct duality: amorphous “stuff” (*e.g.*, road, bareland) requiring pixel-wise semantic continuity, and countable “things” (*e.g.*, buildings, vehicles) demanding instance-level boundary precision. We identify that SAM 3’s decoupled architecture naturally aligns with this duality. We propose to assign the semantic head to maintain land-cover completeness and the Transformer decoder to capture fine-grained instance details, thereby unifying these complementary strengths to ensure robust segmentation across diverse geospatial targets.
- **(2) Presence-Guided Filtering:** In remote sensing OVSS, a complete vocabulary list might include global land cover types, but a single image patch covers a minute geographical area (*e.g.*, hundreds of meters). This results in a high category sparsity, where the vast majority of queried concepts are physically absent in the local view. We leverage the presence score to address this global-local discrepancy, explicitly suppressing irrelevant categories to eliminate false positives caused by the vast vocabulary against limited visual content [14, 53].

We evaluate our approach on 17 remote sensing semantic segmentation benchmarks and some general segmentation benchmarks. Our results demonstrate the strong capability of SAM 3 for remote sensing OVSS, which is further enhanced by our proposed improvements.

2. Related Work

2.1. Training-based OVSS

Training-based OVSS methods fine-tune pre-trained VLMs on annotated datasets, typically adopting either a mask classification or a dense feature adaptation paradigm. Mask classification methods, including OpenSeg [24], OVSeg [41], ZegFormer [21], and MasQCLIP [66], *etc.*, leverage generated class-agnostic masks for subsequent CLIP classification. In contrast, dense feature adaptation methods, ranging from early pixel-alignment works like LSeg [35] to advanced adapter-based models like SAN [65], SED [62] and CAT-Seg [18], refine CLIP’s feature maps directly for dense prediction via side networks or cost aggregation. Other methods such as SegCLIP [43] and GroupViT [64] explore weakly-supervised grouping from

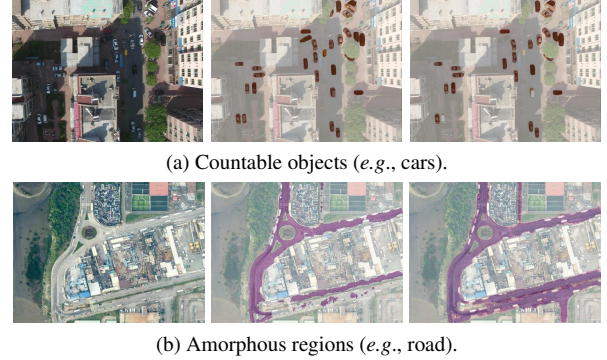


Figure 1. The Transformer decoder of SAM 3 excels at delineating countable objects but produces fragmented masks for amorphous regions, while the semantic head preserves continuity for amorphous regions but lacks boundary precision for small targets. (Left: Remote sensing image. Middle: Prediction of Transformer Decoder. Right: Prediction of semantic segmentation head.)

image-text pairs. Despite achieving strong performance, these methods require computationally expensive training on large-scale datasets, limiting their flexibility compared to training-free alternatives.

2.2. Training-free OVSS

To avoid training costs, training-free methods directly adapt pre-trained VLMs for dense prediction. Pure CLIP-based methods such as SCLIP [55] and ClearCLIP [33] modify the internal mechanisms of CLIP, *e.g.*, removing pooling layers or refining self-attention maps, to extract dense features [5, 26, 29, 72]. However, due to CLIP’s image-level pre-training objective, these methods often suffer from coarse localization. To address this, the VFM-assisted methods [3, 31, 34, 51, 70] integrate auxiliary models like DINO [11, 47] or SAM [32, 49] to provide structural guidance. By utilizing attention maps or object proposals from these foundation models, they achieve clearer boundaries. However, these methods operate as disjointed pipelines, relying on separate, heavy models for segmentation and recognition, which leads to significant system complexity.

2.3. Remote Sensing OVSS

Extending OVSS to remote sensing images faces unique challenges, such as extreme scale variations and arbitrary orientations, which often degrade the performance of methods designed for natural images. To mitigate the coarse localization of CLIP, recent methods like SegEarth-OV [40] and GSNet [69] employ feature upsampling modules or dual-stream architectures to incorporate domain-specific priors, while SkySense-O [74] advances this by pre-training vision-centric VLMs on large-scale remote sensing data. Addressing the specific geometric complexities of aerial views, Cao *et al.* [8] introduce rotation-aggregative modules to handle orientation diversity. Furthermore, AerOSeg [22]

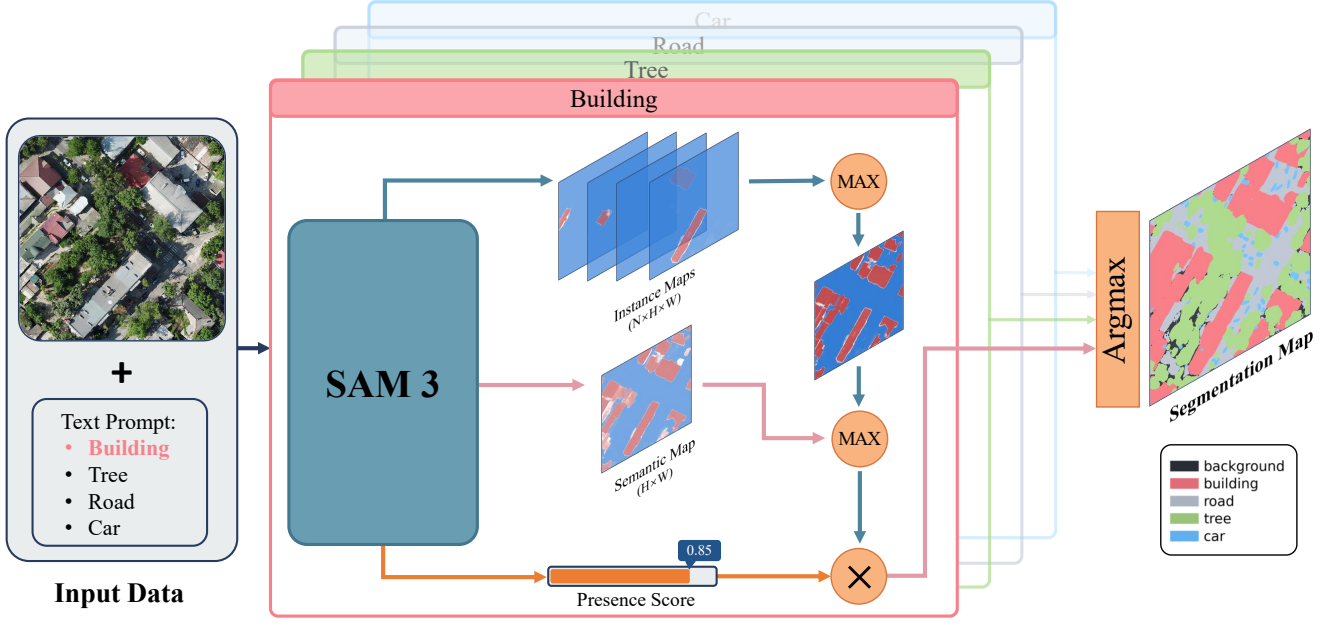


Figure 2. The overall inference pipeline of SegEarth-OV3. Given an input image and a list of text prompts, we leverage SAM 3’s decoupled outputs. The pipeline involves: (1) instance aggregation to consolidate sparse object queries; (2) dual-head mask fusion to combine the fine-grained instance details with the global coverage of the semantic head; and (3) presence-guided filtering (using the presence score) to suppress false positives from absent categories. MAX denotes the element-wise maximum operation, and \times denotes multiplication.

and SCORE [27] leverage SAM features for structural guidance and regional context, while RemoteSAM [68] and InstructSAM [71] utilize SAM-based pipelines to unify segmentation with broader interpretation tasks. However, these methods typically rely on complex multi-stage pipelines or require expensive domain-specific training. In contrast, we explore SAM 3 as a unified, training-free framework to simplify remote sensing OVSS.

3. Methods

3.1. Preliminaries

We adopt SAM 3 as our foundational architecture. Unlike standard semantic segmentors that map an image $I \in \mathbb{R}^{H \times W \times 3}$ directly to a label map $L \in \mathbb{R}^{H \times W}$, SAM 3 operates as a prompt-conditioned predictor. Given an image I and a specific text prompt t (e.g., a category name “building”), the model predicts the probability that a pixel or region belongs to the concept defined by t . Architecturally, SAM 3 consists of a vision encoder and a text encoder that extract image embeddings and text embeddings, respectively. These are processed by a Fusion Encoder, which outputs prompt-conditioned image features F_{cond} . Based on F_{cond} , the model utilizes three decoupled heads to generate predictions:

- **Presence Head:** Predicts a scalar score $S_{pres} \in [0, 1]$, indicating the global probability that the concept t exists

in the image.

- **Semantic Segmentation Head:** A dense prediction module (FCN-style) that maps F_{cond} to a semantic probability map $P_{sem} \in [0, 1]^{H \times W}$.
- **Transformer Decoder (Instance Head):** A query-based module that outputs a set of N instance predictions $\{(P_{inst}^{(k)}, s_{conf}^{(k)})\}_{k=1}^N$. Here, $P_{inst}^{(k)} \in [0, 1]^{H \times W}$ is the probability map for the k -th object query, and $s_{conf}^{(k)} \in [0, 1]$ is its associated confidence score.

For OVSS, a naive baseline is to rely solely on the instance predictions from the Transformer decoder [10], aggregating them to form a segmentation mask. However, in remote sensing, this baseline often overlooks amorphous regions and is prone to false positives when querying a large vocabulary. Motivated by this, we introduce some strategies to address these limitations.

3.2. Inference Pipeline

We propose a training-free inference strategy designed to tackle the specific challenges of OVSS in remote sensing scenarios, as shown in Figure 2. Our pipeline processes each category in the vocabulary \mathcal{V} sequentially and aggregates the results into a final segmentation map.

Instance Aggregation. Remote sensing scenes often contain dense clusters of small, identical objects (e.g., vehicles in a parking lot, ships in a harbor). Standard semantic segmentation methods often leads to boundary adhesion in

dense clusters. To address this, we first leverage the Transformer decoder, which treats objects as discrete queries, effectively isolating individual instances even in crowded scenes. It generates a set of N discrete instance predictions $\{(P_{inst}^{(k)}, s_{conf}^{(k)})\}_{k=1}^N$. We aggregate these sparse predictions into a single category-level map P_{inst_agg} by taking the maximum weighted probability at each pixel:

$$P_{inst_agg}(h, w) = \max_{k=1}^N \left(P_{inst}^{(k)}(h, w) \cdot s_{conf}^{(k)} \right). \quad (1)$$

This aggregation effectively consolidates individual object instances into a unified semantic layer, preserving the fine-grained localization capabilities of the instance head even in crowded scenes.

Dual-Head Mask Fusion. While the instance head excels at delineating countable objects (“things”), it may produce fragmented predictions for large-scale continuous and amorphous regions (“stuff”) like road or bareland, which are prevalent in remote sensing images, as shown in Figure 1. Conversely, the semantic head provides dense, global coverage but often blurs the boundaries of small targets or misses them entirely. To reconcile these complementary strengths, we fuse the aggregated instance map P_{inst_agg} with the dense probability map P_{sem} from the semantic head using a max-fusion strategy:

$$P_{fused}(h, w) = \max(P_{sem}(h, w), P_{inst_agg}(h, w)). \quad (2)$$

This fusion ensures robust segmentation performance across diverse categories, capturing both the distinct boundaries of small instances and the completeness of large-scale amorphous regions.

Presence-Guided Filtering. A critical challenge in OVSS arises from querying a massive global vocabulary against a localized image patch. Although the vocabulary \mathcal{V} might enumerate a comprehensive list of global land-cover types (e.g., various biomes, infrastructure), a single inference patch restricts the view to a small geographical extent (e.g., a few hundred meters). This results in a high category sparsity, where the valid targets in a given image are only a small subset of \mathcal{V} . The model becomes prone to hallucinating absent categories due to textural ambiguities common in geospatial data, e.g., confusing “barren land” with “sports field”. To mitigate this, we utilize the global presence score S_{pres} to explicitly suppress irrelevant categories. We apply a soft gating operation, $P_{final}^{(c)} = P_{fused}^{(c)} \cdot S_{pres}^{(c)}$, which reduces the weight of the probability maps of categories predicted to be absent. Finally, we assign each pixel to the category with the highest probability:

$$M(h, w) = \arg \max_{c \in \mathcal{V}} P_{final}^{(c)}(h, w). \quad (3)$$

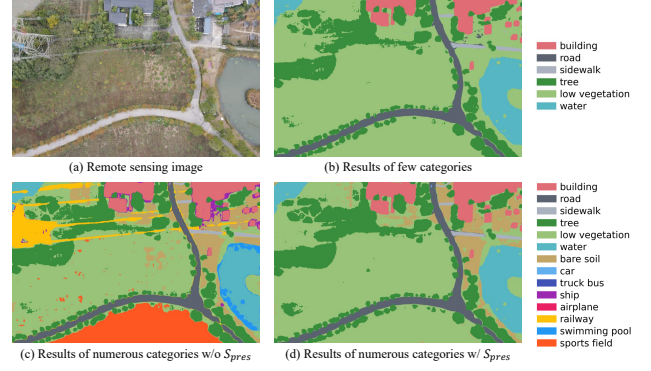


Figure 3. Impact of vocabulary size and our filtering strategy. Querying a vast vocabulary introduces severe noise due to distracting candidates (b to c). Our presence-guided filtering leverages presence scores to suppress absent categories, effectively eliminating interference and restoring segmentation quality.

To handle ambiguous regions, pixels with a maximum probability below a threshold τ are assigned to the “background” category (if it exists). Note that for categories with multiple prompts, we select the one with the highest probabilities to ensure robustness.

4. Experiments

4.1. Setup

Datasets. Following SegEarth-OV [40], we evaluate our method on 17 remote sensing datasets, covering diverse scenes, resolutions, and tasks.

- **Semantic Segmentation:** We use eight benchmarks (OpenEarthMap [61], LoveDA [56], iSAID [58], Potsdam, Vaihingen¹, UAVid [44], UDD5 [15] and VDD [7]) to assess multi-class segmentation performance across satellite, aerial, and UAV platforms.
- We further test on nine binary extraction datasets focusing on critical geospatial objects: building extraction (WHU^{Aerial} [28], WHU^{Sat.II} [28], Inria [45], and xBD [25]), road extraction (CHN6-CUG [73], DeepGlobe [20], Massachusetts [46], and SpaceNet [54]), and flood detection (WBS-SI²).

Additionally, we report results on general scene datasets (Pascal VOC20 [23], COCO Stuff [6], and Cityscapes [19]) to demonstrate universality. The evaluation metric is mIoU for multi-class segmentation and IoU of the foreground class for binary extraction tasks.

Implementation Details. We use the official SAM 3 model equipped with the Perception Encoder-Large+ (PE-L+) [4] backbone. Input images are resized to 1008×1008. Text prompts are derived directly from category names (e.g.,

¹<https://www.isprs.org/education/benchmarks/UrbanSemLab>

²<https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images>

Table 1. Open-vocabulary semantic segmentation quantitative comparison on remote sensing datasets. Evaluation metric: mIoU. **Best** and **second best** performances are highlighted. SCAN, SAN, SED, Cat-Seg, OVRS, GSNet, RSKT-Seg are tuned on dataset (7,002 images with 17 categories). SkySense-O is tuned on Sky-SA dataset (35,000 images with 1,763 categories). “Oracle” is achieved by a fully supervised SegFormer-b0 [63] model using full training data.

Methods	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid ^{img}	UDD5	VDD	Avg
<i>Training on remote sensing segmentation data</i>									
SCAN _{CVPR2024} [42]	-	23.2	44.3	27.5	15.2	20.3	34.1	29.2	-
SAN _{CVPR2023} [65]	-	25.3	49.6	37.3	39.2	23.5	37.2	35.8	-
SED _{CVPR2024} [62]	-	24.6	51.2	29.4	39.0	21.3	35.7	32.5	-
Cat-Seg _{CVPR2024} [18]	-	28.6	53.3	35.8	42.3	25.7	40.2	39.1	-
OVRS _{TGRS2025} [8]	-	31.5	52.7	36.4	43.5	24.1	40.8	37.2	-
GSNet _{AAAI2025} [69]	-	32.5	53.7	37.9	44.1	24.2	40.9	37.3	-
RSKT-Seg _{AAAI2026} [36]	-	33.2	54.3	38.4	42.7	25.7	42.1	39.7	-
SkySense-O _{CVPR2025} [74]	40.8	38.3	43.9	54.1	51.6	-	-	-	-
<i>Training-free</i>									
CLIP _{ICML2021} [48]	12.0	12.4	7.5	15.6	10.8	10.9	9.5	14.2	11.4
MaskCLIP _{ECCV2022} [72]	25.1	27.8	14.5	33.9	29.9	28.6	32.4	32.9	27.2
SCLIP _{ECCV2024} [55]	29.3	30.4	16.1	39.6	35.9	31.4	38.7	37.9	31.1
GEM _{CVPR2024} [5]	33.9	31.6	17.7	39.1	36.4	33.4	41.2	39.5	32.3
ClearCLIP _{ECCV2024} [33]	31.0	32.4	18.2	42.0	36.2	36.2	41.8	39.3	33.4
SegEarth-OV _{CVPR2025} [40]	40.3	36.9	21.7	48.5	40.0	42.5	50.6	45.3	39.2
ProxyCLIP _{ECCV2024} [34]	38.9	34.3	21.8	49.0	47.5	35.8	40.8	47.8	39.5
CorrCLIP _{ICCV2025} [70]	32.9	36.9	25.5	51.9	47.0	38.3	46.1	47.3	40.7
SegEarth-OV3	42.9	47.4	27.6	57.8	60.8	54.7	71.7	64.5	53.4
Oracle	64.4	50.0	36.2	74.3	61.2	59.7	56.5	62.9	58.2

Table 2. Open-vocabulary building / road / flood extraction quantitative comparison on remote sensing datasets. Evaluation metric: IoU of the foreground class, i.e. building, road or flood. **Best** and **second best** performances are highlighted.

Method	Building Extraction				Road Extraction				Flood Detection
	WHU ^{Aerial}	WHU ^{Sat.II}	Inria	xBD ^{Pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
CLIP [48]	17.7	3.5	19.6	16.0	7.7	3.9	4.9	7.1	18.6
MaskCLIP [72]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8
SCLIP [55]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1
GEM [5]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5
ClearCLIP [33]	36.6	20.8	39.0	30.1	25.5	5.7	6.4	16.3	44.9
SegEarth-OV [40]	49.2	28.4	44.6	37.0	35.4	17.8	11.5	23.8	60.2
SegEarth-OV3	86.9	44.2	72.4	64.3	49.6	39.3	27.7	35.6	75.6

“building”, “road”), and some categories are augmented with synonyms³. The background threshold τ and the initial confidence threshold for the Transformer decoder are manually tuned for each dataset to achieve roughly optimal performance. No test-time augmentation and extra post-processing method [1] is used.

4.2. Main Results

Semantic segmentation. We report the quantitative comparison on eight remote sensing semantic segmentation benchmarks in Table 1. Our method, SegEarth-OV3,

achieves a new state-of-the-art, demonstrating a substantial performance leap over existing methods. Specifically, it achieves an average mIoU of 53.4%, surpassing the best previous training-free method, CorrCLIP (40.7% mIoU), by a remarkable margin of +12.7% mIoU. This performance also consistently outperformed training-based OVSS methods fine-tuned on remote sensing data, e.g., RSKT-Seg (39.7% mIoU) and Cat-Seg (39.1% mIoU). A notable observation is that our zero-shot method outperforms the fully supervised Oracle on some datasets. On UDD5 and VDD datasets, SegEarth-OV3 achieves 71.7% mIoU and 64.5% mIoU respectively, exceeding the Oracle baselines (56.5% and 62.9% mIoU). This surprising result suggests that the rich semantic knowledge and robust segmentation capabil-

³The prompt setting is manually curated based on dataset features and has not been systematically explored in this version.

Table 3. Ablation study on dual-head mask fusion. “Instance Only” denotes using predictions solely from the Transformer decoder, while “Semantic Only” relies exclusively on the semantic segmentation head. **Bold** indicates the best performance.

Method	LoveDA	Uavid	xBD ^{pre}	CHN6-CUG
Instance Only	32.2	50.4	61.4	38.4
Semantic Only	35.4	47.1	44.9	39.5
SegEarth-OV3	47.4	54.7	64.3	49.6

Table 4. Comparison with state-of-the-art OVSS methods on Pascal VOC20, COCO Stuff, and Cityscapes benchmarks. **Bold** indicates the best performance.

Method	Size	VOC20	Stuff	City
Training-based				
TCL [12]		83.2	22.4	24.0
CLIP-DINOiser [60]		80.9	24.6	31.7
CoDe [59]	ViT-B/16	-	23.9	28.9
CAT-Seg [18]		94.6	-	-
Training-free				
CLIP [48]		41.9	4.4	5.0
MaskCLIP [72]		74.9	16.4	12.6
ClearCLIP [33]		80.9	23.9	30.0
SCLIP [55]		80.4	22.4	32.2
ProxyCLIP [34]		80.3	26.5	38.1
LaVG [30]		82.5	23.2	26.2
CLIPtrase [50]	ViT-B/16	81.2	24.1	-
NACLIP [26]		83.0	25.7	38.3
Trident [51]		84.5	28.3	42.9
ResCLIP [67]		86.0	24.7	35.9
SC-CLIP [2]		84.3	26.6	41.0
CLIPer [52]		85.2	27.5	-
CASS [31]		87.8	26.7	39.4
CorrCLIP [70]		88.8	31.6	49.4
FreeDA [3]		87.9	28.8	36.7
CaR [53]		91.4	-	-
ProxyCLIP [34]		83.2	25.6	40.1
ResCLIP [67]	ViT-L/14	85.5	23.4	33.7
SC-CLIP [2]		88.3	26.9	41.3
CLIPer [52]		90.0	28.7	-
CorrCLIP [70]		91.5	34.0	51.1
ProxyCLIP [34]		83.3	26.8	42.0
Trident [51]	ViT-H/14	88.7	28.6	47.6
CorrCLIP [70]		91.8	32.7	49.9
SegEarth-OV3	PE-L+/14	96.8	42.8	69.7

ities inherent in the SAM 3 foundation model can, in certain scenarios, exceed the generalization ability of models trained on domain-specific data. In Figure 4, we visualize the results on a large-scale remote sensing image.

Single-class extraction. We further evaluate the single-class land-cover extraction capability of our method on nine

benchmarks focusing on buildings, roads, and floods (Table 2). In building extraction, our method achieves unprecedented gains, reaching 86.9% IoU on WHU^{Aerial} and 72.4% IoU on Inria, outperforming the previous state-of-the-art SegEarth-OV by massive margins of +37.7% and +27.8% respectively. Similarly, for road extraction, it consistently surpasses previous methods, achieving 49.6% IoU on CHN6-CUG. In the flood detection task, SegEarth-OV3 attains 75.6% IoU, marking a +15.4% improvement. These consistent and dramatic improvements across diverse geospatial targets, emphasize the robustness of SegEarth-OV3 and the generalization capability of SAM 3.

4.3. Ablation Studies

We investigate the contribution of the dual-head fusion strategy by evaluating the performance of each head individually on four representative datasets, as listed in Table 3. Using only Transformer decoder predictions yields strong results on object-centric tasks like building extraction, but underperforms on complex scenes, due to struggles with amorphous regions. Conversely, the “Semantic Only” baseline struggles with precise object boundaries, achieving only 44.9% IoU on xBD. By fusing both heads, SegEarth-OV3 achieves significant improvements. On LoveDA and CHN6-CUG, SegEarth-OV3 reaches 47.4% mIoU and 49.6% IoU, surpassing the best single-head baselines by +12.0% and +10.1%. Even on the instance-heavy xBD dataset, fusion further boosts performance by 2.9%. These results validate our hypothesis that the semantic and instance heads provide complementary information, *i.e.*, global coverage and fine-grained precision, and their combination is essential for remote sensing OVSS.

4.4. Results on General Scene Datasets

We benchmark SegEarth-OV3 on three standard general scene datasets: Pascal VOC20, COCO Stuff, and Cityscapes (Table 4). For Pascal VOC20 and Cityscapes, we refine category names to better align with visual concepts; for instance, the “terrain” class in Cityscapes is expanded to “grass, horizontal vegetation, soil, sand” based on official definitions⁴. SegEarth-OV3 achieves dominance across all benchmarks. On Pascal VOC20, it achieves 96.8% mIoU, surpassing both the best training-free method CorrCLIP (91.8% mIoU) and the training-based CAT-Seg (94.6% mIoU). On COCO Stuff, it reaches 42.8% mIoU, significantly exceeding CorrCLIP by +8.8%. The most significant improvement is observed on Cityscapes, where our method achieves 69.7% mIoU, representing an increase of 18.6% mIoU over the previous best result. These results further emphasize the powerful capabilities of SAM 3 and the effectiveness of SegEarth-OV3.

⁴<https://www.cityscapes-dataset.com/dataset-overview>

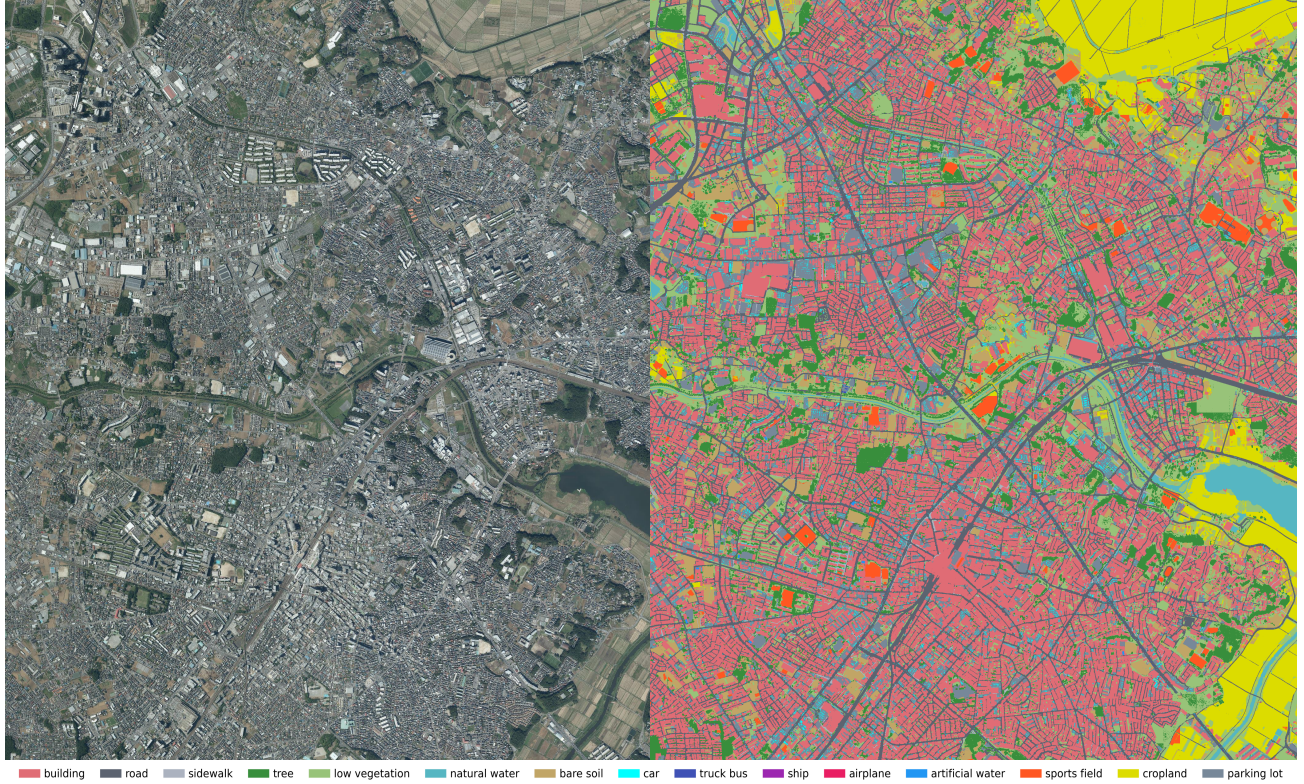


Figure 4. Inference results of SegEarth-OV3 on a remote sensing image exceeding $10k \times 10k$ resolution. The image originates from [13].

5. Conclusion

This paper explores the potential of SAM 3 for remote sensing OVSS. Building upon the powerful zero-shot capabilities of the SAM 3 foundation model, we introduce an inference strategy to adapt its prompt-based architecture for remote sensing OVSS. By fusing semantic and instance outputs and leveraging presence-guided filtering, we effectively address the unique challenges of remote sensing. Our results on 20 diverse datasets show that SegEarth-OV3 not only sets a new training-free state-of-the-art but also outperforms supervised baselines in specific cases. These results highlight the significant potential of SAM 3 for specific domain tasks and demonstrate that, through appropriate adaptation strategies, training-free paradigms can serve as a powerful alternative to traditional supervised learning.

References

- [1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4253–4262, 2020. 5
- [2] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 6
- [3] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 2, 6
- [4] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 4
- [5] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 2, 5
- [6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 4
- [7] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 4
- [8] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024. 2, 5
- [9] 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. 1
- [10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryal, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 1, 3
- [11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2
- [12] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 6
- [13] Hongruixuan Chen, Cuiling Lan, Jian Song, Clifford Broni-Bediako, Junshi Xia, and Naoto Yokoya. Land-cover change detection using paired openstreetmap data and optical high-resolution imagery via object-guided transformer. *arXiv preprint arXiv:2310.02674*, 2023. 7
- [14] Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23124–23134, 2025. 2
- [15] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018. 4
- [16] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 1
- [17] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1
- [18] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2, 5, 6
- [19] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4
- [20] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 4
- [21] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 2
- [22] Saikat Dutta, Akhil Vasim, Siddhant Gole, Hamid Rezaatofighi, and Biplab Banerjee. Aeroseg: Harnessing sam for open-vocabulary segmentation in remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2254–2264, 2025. 2
- [23] 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(2):303–338, 2010. 4
- [24] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2
- [25] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. 4
- [26] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5061–5071. IEEE, 2025. 2, 6
- [27] Shiqi Huang, Shuting He, Huaiyuan Qin, and Bihan Wen. Score: Scene context matters in open-vocabulary remote sensing instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12559–12569, 2025. 3
- [28] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 4
- [29] Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20291–20300, 2025. 2
- [30] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision and Pattern Recognition (ECCV)*, 2024. 6
- [31] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15033–15042, 2025. 2, 6
- [32] 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, 2

- [33] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 2, 5, 6
- [34] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2025. 1, 2, 5, 6
- [35] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2
- [36] Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring efficient open-vocabulary segmentation in the remote sensing. *arXiv preprint arXiv:2509.12040*, 2025. 5
- [37] Ke Li, Fuyu Dong, Di Wang, Shaofeng Li, Quan Wang, Xinbo Gao, and Tat-Seng Chua. Show me what and where has changed? question answering and grounding for remote sensing change detection. *arXiv preprint arXiv:2410.23828*, 2024. 1
- [38] Kaiyu Li, Xiangyong Cao, Yupeng Deng, Chao Pang, Zepeng Xin, Deyu Meng, and Zhi Wang. Dynamicearth: How far are we from open-vocabulary change detection? *arXiv preprint arXiv:2501.12931*, 2025. 1
- [39] Kaiyu Li, Xiangyong Cao, Ruixun Liu, Shihong Wang, Zixuan Jiang, Zhi Wang, and Deyu Meng. Annotation-free open-vocabulary segmentation for remote-sensing images. *arXiv preprint arXiv:2508.18067*, 2025. 1
- [40] 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10545–10556, 2025. 1, 2, 4, 5
- [41] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 2
- [42] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 5
- [43] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 2
- [44] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. 4
- [45] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 4
- [46] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 4
- [47] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2
- [48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5, 6
- [49] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2
- [50] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024. 6
- [51] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23487–23497, 2025. 2, 6
- [52] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23199–23209, 2025. 6
- [53] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 2, 6
- [54] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 4
- [55] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 1, 2, 5, 6
- [56] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates, Inc., 2021. 4
- [57] Junjue Wang, Weihao Xuan, Heli Qi, Zhihao Liu, Kunyi Liu, Yuhang Wu, Hongruixuan Chen, Jian Song, Junshi Xia, Zhuo Zheng, et al. Disasterm3: A remote sensing vision-language dataset for disaster damage assessment and response. *arXiv preprint arXiv:2505.21089*, 2025. 1
- [58] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling

- Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 4
- [59] Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26794–26803, 2024. 6
- [60] Monika Wysoczkańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 320–337. Springer, 2024. 6
- [61] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 4
- [62] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 5
- [63] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 5
- [64] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022. 2
- [65] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2, 5
- [66] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masq-clip for open-vocabulary universal image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 887–898, 2023. 2
- [67] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29968–29978, 2025. 6
- [68] Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3027–3036, 2025. 3
- [69] Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9436–9444, 2025. 2, 5
- [70] Dengke Zhang, Fagui Liu, and Quan Tang. Corclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24677–24687, 2025. 1, 2, 5, 6
- [71] Yijie Zheng, Weijie Wu, Qingyun Li, Xuehui Wang, Xu Zhou, Aiai Ren, Jun Shen, Long Zhao, Guoqing Li, and Xue Yang. Instructsam: A training-free framework for instruction-oriented remote sensing object recognition. *arXiv preprint arXiv:2505.15818*, 2025. 3
- [72] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2, 5, 6
- [73] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li. A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:353–365, 2021. 4
- [74] Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, et al. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14733–14744, 2025. 2, 5