

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

Saikat Dutta^{1,2,3*} Akhil Vasim² Siddhant Gole²
Hamid RezaTofighi³ Biplab Banerjee²

¹IITB-Monash Research Academy ²IIT Bombay ³Monash University

Abstract

Image segmentation beyond predefined categories is a key challenge in remote sensing, where novel and unseen classes often emerge during inference. Open-vocabulary image Segmentation addresses these generalization issues in traditional supervised segmentation models while reducing reliance on extensive per-pixel annotations, which are both expensive and labor-intensive to obtain. Most Open-Vocabulary Segmentation (OVS) methods are designed for natural images but struggle with remote sensing data due to scale variations, orientation changes, and complex scene compositions. This necessitates the development of OVS approaches specifically tailored for remote sensing. In this context, we propose **AerOSeg**, a novel OVS approach for remote sensing data. First, we compute robust image-text correlation features using multiple rotated versions of the input image and domain-specific prompts. These features are then refined through spatial and class refinement blocks. Inspired by the success of the Segment Anything Model (SAM) in diverse domains, we leverage SAM features to guide the spatial refinement of correlation features. Additionally, we introduce a semantic back-projection module and loss to ensure the seamless propagation of SAM’s semantic information throughout the segmentation pipeline. Finally, we enhance the refined correlation features using a multi-scale attention-aware decoder to produce the final segmentation map. We validate our SAM-guided Open-Vocabulary Remote Sensing Segmentation model on three benchmark remote sensing datasets: *iSAID*, *DLRS*, and *OpenEarthMap*. Our model outperforms state-of-the-art open-vocabulary segmentation methods, achieving an average improvement of 2.54 h-mIoU.

1. Introduction

Open-vocabulary image segmentation aims to segment objects belonging to an unbounded set of categories, thereby allowing the model to handle novel or previously unseen

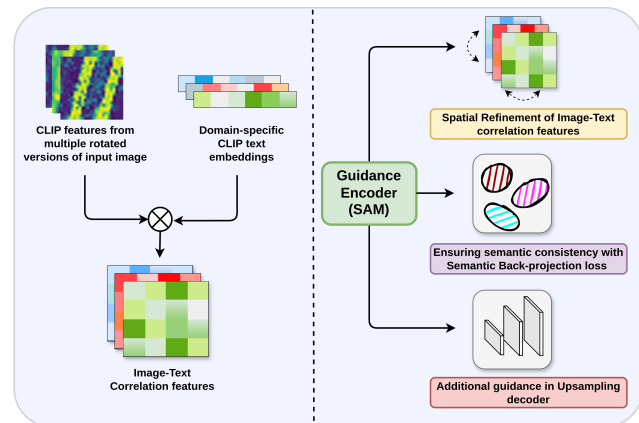


Figure 1. Key aspects of **AerOSeg**: (a) Correlation features computation from CLIP features across rotated inputs and domain-specific text embeddings. (b) Guidance Encoder’s multifaceted contribution.

classes at inference. Unlike traditional semantic segmentation — where the training and testing categories are strictly identical — open-vocabulary segmentation endeavors to learn transferable semantic representations that can facilitate the segmentation of any class not observed during training. In the context of *remote sensing* imagery, this capability becomes especially critical: supervised segmentation approaches typically require large-scale datasets with pixel-level annotations that are notoriously expensive and time-consuming to obtain, primarily due to high spatial resolution, complex scenes, and varying geospatial contexts. Moreover, models trained solely on these well-annotated classes tend to overfit, limiting their ability to generalize to new or rare categories that often arise in real-world remote sensing applications (e.g., emerging structures, seasonal objects, or natural disasters).

Most recent open-vocabulary segmentation methods leverage vision-language models such as CLIP [20] or ALIGN [10], which learn powerful joint embeddings of visual and textual data. Despite their success with natural images, these models often struggle with satellite imagery,

*C-MInDS, IIT Bombay — 23d2031@iitb.ac.in

where objects can appear at drastically different scales, orientations, and resolutions. Remote sensing images are fundamentally distinct from typical natural images due to large scene coverage, small object sizes, and complex backgrounds, necessitating specialized strategies that address these domain-specific challenges.

Motivated by these requirements, researchers have recently proposed open-vocabulary solutions tailored to remote sensing. A pioneering approach, Open-Vocabulary Remote Sensing Segmentation (OVRs) [3], introduces a rotational invariance paradigm by augmenting the visual encoder input with multiple rotated versions of the image, thereby capturing orientation-invariant semantic features. Subsequent modules refine these features spatially and by category and finally leverage a multi-scale decoder for high-resolution segmentation. While effective, OVRs heavily relies on CLIP for open-vocabulary segmentation, which is suboptimal. CLIP, while excelling in open-vocabulary classification, is designed to align global visual and textual semantics, limiting its effectiveness for per-pixel tasks like segmentation.

In this work, we hypothesize that **complementary guidance features** can help preserve semantic richness throughout the refinement pipeline. Specifically, we propose exploiting the **Segment Anything Model (SAM)** as a feature guidance encoder, in conjunction with CLIP, to strengthen the semantic flow in open-vocabulary remote sensing segmentation. SAM, trained on large-scale data for promptable segmentation, produces semantically rich features and performs well across diverse domains and datasets. Given the smaller size of remote sensing datasets, leveraging SAM’s features provides valuable additional guidance for improved segmentation. Our framework, **AerOSeg**, first computes image–text correlation maps from multiple rotated versions of the input image and specialized textual prompts, ensuring orientation-invariant features. Instead of generic prompts, we use domain-specific Remote Sensing prompts, enhancing the model’s robustness. A swin transformer-based spatial refinement block, guided by SAM features, then refines these correlation maps, followed by a class refinement block that further sharpens category-specific responses. To address the loss of semantic fidelity after repeated refinements, we introduce a back-projection module that reproduces SAM features and enforces feature alignment via a *semantic back-projection loss*, ensuring that semantic content is preserved. Finally, the refined correlation maps are upsampled via an Attention-aware Upsampling decoder to produce the final segmentation predictions. The salient aspects of AerOSeg is presented in Fig. 1.

In summary, our main contributions are as follows:

- **SAM as Guidance:** We propose a novel synergy between SAM and CLIP for remote sensing segmentation, utilizing SAM as a feature guidance encoder to enrich semantic

representations and mitigate loss of generalizability.

- **Back-Projection for Semantic Fidelity:** We integrate a back projection module and a corresponding feature reconstruction loss to sustain crucial semantic information derived from SAM, preventing over-refinement that impairs generalization to unseen classes.
- **Comprehensive Experiments:** We conduct extensive evaluations on three benchmark remote sensing segmentation datasets, demonstrating that our approach significantly outperforms state-of-the-art open-vocabulary segmentation methods on both seen and unseen categories.

2. Related Works

A. Semantic segmentation in remote sensing: Image segmentation partitions an image into regions corresponding to specific classes or objects through pixel-wise grouping. In remote sensing, the high resolution and complexity of images present significant challenges. Modern deep learning methods, particularly Convolutional Neural Networks (CNNs), have advanced this field by learning intricate features from large-scale images [11, 19, 43]. Building on CNNs, subsequent work has addressed class imbalance with weighted uncertainty labeling [1], enabled multi-scale feature learning via pyramid attention pooling [46], and enhanced spatial detail capture through global context integration [39]. Despite these advances, CNNs are inherently limited to local feature extraction.

Vision Transformers (ViTs) have emerged as a powerful alternative by leveraging attention mechanisms to model long-range dependencies [26]. This has led to their increasing adoption in remote sensing segmentation. For example, Xu et al. [40] introduced an Efficient Transformer backbone, adapted from the Swin Transformer [18], which integrates an MLP head and auxiliary edge fusion to reduce computational costs and improve edge segmentation. However, due to the heavy computational burden of full attention mechanisms, most recent approaches combine ViTs and CNNs. Hybrid models, such as CVMH-Unet [4] and CCTNet [29], effectively merge global and local feature extraction while maintaining efficiency. Encoder-decoder hybrids like UNetFormer [30] and frameworks employing Swin Transformer encoders with CNN decoders [44] further illustrate the trend toward integrating multi-scale feature representations in remote sensing segmentation. Despite achieving decent performance on benchmark datasets, this class of segmentation models can not be extended to segment an arbitrary number of classes during deployment.

B. Open-Vocabulary segmentation: Open-vocabulary segmentation is a challenging task that seeks to segment objects from an open set of categories defined by textual labels or descriptions. Unlike conventional segmentation, the label sets during training and testing can differ significantly.

Early methods focused on aligning visual embeddings

with pre-defined word embeddings [2, 35]. More recent approaches leverage large-scale visual-language models to align visual and semantic feature spaces. For instance, Li et al. [14] integrate CLIP’s [20] text encoder with a DPT-based image encoder [21], employing a contrastive loss to align pixel embeddings with corresponding textual embeddings. Similarly, Ghiasi et al. [9] introduce a class-agnostic segmentation module based on region-to-image cross-attention [28] that generates segmentation masks used for visual-semantic alignment. In the same vein, Zegformer [8] decouples the task into Maskformer-based class-agnostic segmentation [6] and a zero-shot classification of segments using CLIP embeddings, while Xu et al. [37] exploit internal feature representations from text-to-image diffusion models to predict segmentation masks, classifying segments via CLIP text embeddings.

Additional contributions include the Side Adapter Network [38], which transforms images into visual tokens, appends query tokens, and integrates CLIP features within transformer layers. These are then processed by MLP layers to produce attention biases and mask proposals. Yu et al. [42] demonstrate that a frozen convolutional CLIP backbone can robustly perform open-vocabulary classification and mask generation, even at higher image resolutions. Cho et al. [7] refine a cost-volume based on cosine similarity between CLIP image and text embeddings to generate segmentation maps. Shan et al. [24] perform element-wise addition of frozen CLIP and SAM features, which are then processed by a transformer-based decoder to generate segmentation maps through embedding balancing. Wang et al. [31] introduce a data pipeline for curating extensive segment-text pairs alongside a universal segment embedding model to classify segments into diverse text-defined categories.

Although most research in open-vocabulary segmentation has targeted natural images, only a few works extend to other domains such as remote sensing. For example, Cao et al. [3] compute feature correlations between rotated image embeddings and text embeddings and utilize an attention-aware upsampling decoder to generate segmentation outputs. Similarly, Li et al. [15] employ a feature upsampler to enhance low-resolution features for training-free open-vocabulary segmentation in remote sensing imagery.

C. Segment Anything Model: The Segment Anything Model (SAM) [13] is a zero-shot segmentation framework trained on 1.1 billion masks and 11 million images. Given an image and a visual prompt—such as a bounding box, point annotations, or an initial mask—SAM employs an image encoder and a prompt encoder to generate corresponding embeddings. These embeddings are fused in a lightweight mask decoder to predict segmentation masks, ensuring robust outputs even when prompts are ambiguous.

SAM has demonstrated impressive interactive segmen-

tation performance across diverse domains, including medical imaging, agriculture, food, and remote sensing. Recent work has sought to enhance and extend SAM’s capabilities by developing domain-specific variants [27, 32], incorporating additional prompt modalities [48], improving computational efficiency, and even adapting the model for promptable video object segmentation [22]. Inspired by the generalization ability of SAM in multiple domains, we utilize features extracted by the SAM encoder to refine CLIP image-text correlation features in this work.

3. Proposed Methodology

Problem Definition: Given an input Remote Sensing Image I , the aim is to classify each pixel from a set of categories defined by textual labels or description. Different from traditional set-up, class-set during training \mathcal{C}_{train} can be different from class-set during inference, \mathcal{C}_{test} in Open-vocabulary setup.

In this work, we propose a novel segmentation framework, AerOSeg, composed of six interconnected modules: (i) Vision-Language Backbone, (ii) Guidance Encoder, (iii) Correlation Feature Computation, (iv) Correlation Feature Refinement, (v) Semantic Back-Projection Module, and (vi) Attention-Aware Upsampling Decoder. Figure 2 provides an overview of the entire framework. In the following sections, we describe each component in detail and explain how they collectively contribute to our segmentation pipeline.

3.1. Describing the Model Components

Vision-Language (VL) Backbone: The VL backbone is a critical component of our framework, as it encapsulates the alignment between visual and linguistic feature spaces. In our approach, we employ CLIP ViT-B as the VL backbone. Given an input image I , we first rotate it by a set of pre-defined angles and then pass both the original and rotated images through the CLIP vision encoder \mathcal{F}_v to extract dense image embeddings. These embeddings are subsequently rotated back to the original orientation. Specifically, for each rotation angle $\theta \in \Theta = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, the dense embedding is computed as

$$F_v^\theta = \text{rotate}(\mathcal{F}_v(I^\theta), -\theta) \in \mathbb{R}^{(H \times W) \times d}. \quad (1)$$

This multi-angle strategy enhances robustness to angular variations [3]. To generate textual embeddings, we create multiple text prompts for each candidate class c using remote-sensing-specific prompt templates, as suggested by Li et al. [16]. In contrast to generic prompts (e.g., “A photo of a [CLS] in the scene” [3, 7]), we employ the following templates:

- A satellite image of a [CLS]
- A land use image of a [CLS]
- A remote sensing image of a [CLS]

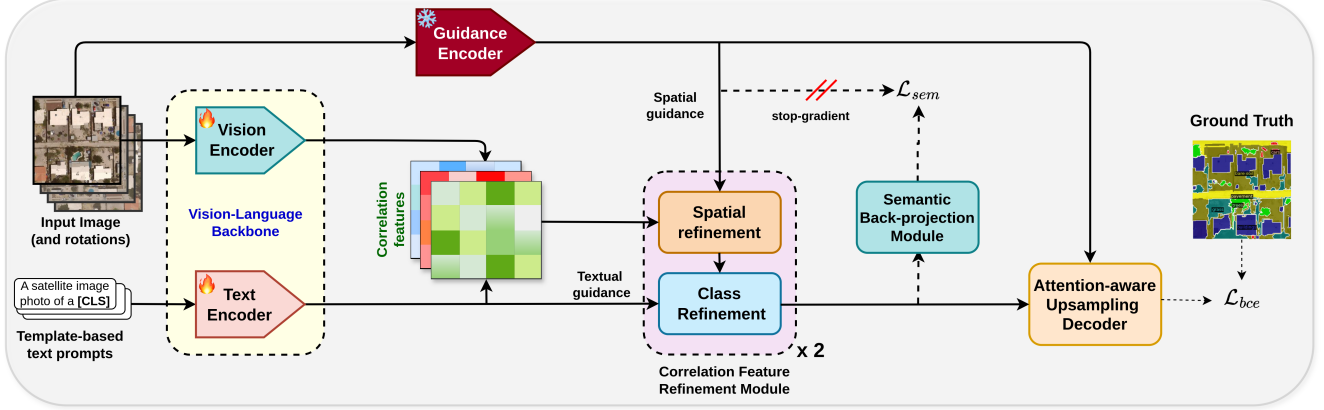


Figure 2. Overview of our proposed framework, **AerOSeg**. The input image (along with its rotated versions) and domain-specific text prompts are first processed by the Vision-Language backbone. The extracted image and text features are then used to generate Correlation features, which are refined through Correlation Feature Refinement blocks. The refined Correlation features are subsequently fed into Attention-aware Upsampling Decoder to yield final segmentation map. Additionally, features from the Guidance encoder are leveraged in both the Correlation Feature Refinement and Attention-aware Upsampling Decoder.

→ An aerial image of a [CLS]

These prompts are processed by the CLIP text encoder \mathcal{F}_l to yield the textual embeddings:

$$F_l = \mathcal{F}_l(T) \in \mathbb{R}^{N_C \times P \times d}, \quad (2)$$

where $P = 4$ denotes the number of prompts per class.

Guidance Encoder: Complementary to the VL backbone, the guidance encoder is designed to provide additional features that improve segmentation performance. Inspired by the success of SAM in diverse domains, we adopt its image encoder—a Vision Transformer (ViT)—as our guidance encoder. To capture a rich hierarchy of representations, we extract features from multiple stages of the encoder. In particular, we use the outputs from the 8th, 16th, and final layers of the SAM-L encoder, denoted as F_g^1 , F_g^2 , and F_g^3 , respectively.

Correlation Feature Computation: With both image embeddings $F_v^\theta(x)$ and text embeddings $F_l(n, i)$ available (where x denotes a 2D spatial position, θ the rotation angle, n the class index, and i the prompt index), we compute the image-text correlation map $C^\theta \in \mathbb{R}^{(H \times W) \times N_C}$ using cosine similarity [23]. For a given spatial location x , class n , and prompt i , the correlation is defined as

$$C^\theta(x, n, i) = \frac{F_v^\theta(x) \cdot F_l(n, i)}{\|F_v^\theta(x)\| \|F_l(n, i)\|}. \quad (3)$$

Correlation maps computed across different rotation angles and prompts are concatenated and processed by a convolutional layer to generate the initial correlation feature $\phi \in \mathbb{R}^{(H \times W) \times N_C \times d_\phi}$, where d_ϕ denotes the feature dimension:

$$\phi(x, n) = \text{conv} \left(\text{concat} \left(\left\{ \left\{ C^\theta(x, n, i) \right\}_{i=1}^P \right\}_{\theta \in \Theta} \right) \right). \quad (4)$$

Correlation Feature Refinement Block: Because CLIP is trained with a global contrastive objective, the dense features it produces can be noisy. To obtain reliable segmentation maps, the initial correlation features ϕ must be refined. Following [7], our refinement module consists of two sequential submodules: (a) a Spatial Refinement Block and (b) a Class Refinement Block.

(a) *Spatial Refinement Block:* To improve the spatial structure of the correlation features, we apply a Swin Transformer [18] module. Refinement is performed independently for each class using two successive Swin Transformer blocks: the first applies window-based multi-head self-attention (W-MSA) over local windows, and the second uses shifted window-based self-attention (SW-MSA). Additionally, to further enhance spatial refinement, we integrate guidance from dense visual features. While prior work [3, 7] utilizes CLIP visual embeddings for this task, it has several drawbacks. First, since correlation features are originally derived from interactions between CLIP image and text embeddings, using CLIP features for refinement is sub-optimal. Second, dense features from the CLIP image encoder tend to be noisy due to its global contrastive learning objective. In contrast, SAM produces more semantic and less noisy feature maps (see Fig. 3). Therefore, we leverage SAM-derived features in the spatial refinement block. The refined feature ϕ' is computed as:

$$\phi'(\cdot, n) = \mathcal{T}^{\text{SP}} \left(\phi(\cdot, n), \mathcal{P}_v(F_g^3) \right), \quad (5)$$

where \mathcal{T}^{SP} denotes the Spatial Refinement Block and \mathcal{P}_v

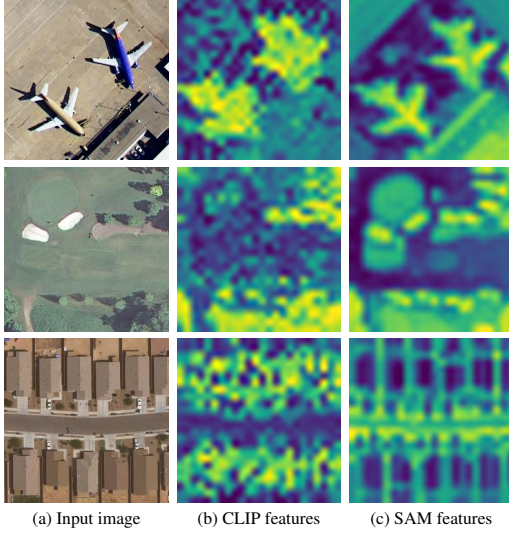


Figure 3. **Feature visualization for CLIP and SAM features.** SAM produces feature maps with richer semantic information and reduced noise compared to CLIP.

is a projection layer. The correlation features and visual embeddings are concatenated to form the query and key, while only the correlation features are used as the value feature in this block. Please refer to Sec. 2 of Supplementary Material for visualizations of the effect of using SAM features in the Spatial Refinement Block.

(b) *Class Refinement Block:* After spatial refinement, we further process the features to incorporate the text modality and explicitly capture inter-class relationships. This step is crucial for addressing the challenges of open-vocabulary segmentation, such as handling a variable number of categories and ensuring invariance to their order. To meet these requirements, we employ a linear transformer layer without positional embeddings [12]. In this block, the text embeddings F_l serve as guidance features, and the refined output is given by

$$\phi''(x, :) = \mathcal{T}^{\text{cls}}\left(\phi'(x, :), \mathcal{P}_l(\bar{F}_l)\right), \quad (6)$$

where \mathcal{T}^{cls} is the Class Refinement Block, \mathcal{P}_l a projection layer, and \bar{F}_l represents the text embeddings averaged over all prompts. In our work, two Correlation Feature Refinement blocks are used.

Semantic Back-Projection Module: To ensure that the refined correlation features remain semantically consistent with SAM features, we introduce a lightweight back-projection module. This module reconstructs the SAM features from the class-wise refined correlation features ϕ'' . Specifically, the correlation features are concatenated along the channel dimension and processed through three linear layers with GELU activation, yielding the reconstructed feature $\psi \in \mathbb{R}^{(H \times W) \times d}$ as follows:

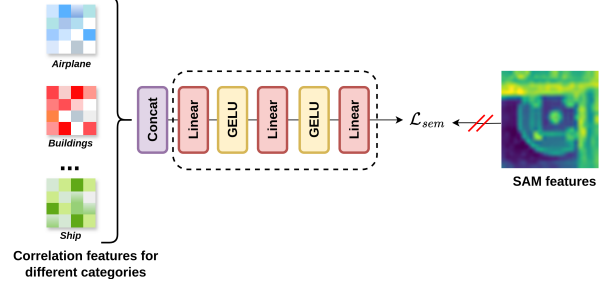


Figure 4. **Semantic Back-projection module and loss.** Class-specific correlation maps are concatenated and passed to an MLP block. Semantic back-projection loss \mathcal{L}_{sem} is computed between reconstructed features and SAM features.

$$\psi = \mathcal{T}^{sem}\left(\text{concat}\left(\left\{\phi''(:, n)\right\}_{n=1}^{N_c}\right)\right). \quad (7)$$

where, \mathcal{T}^{sem} is Semantic back-projection module and N_c is the number of classes. The reconstructed feature is subsequently used to compute a Semantic Back-Projection Loss (see Section 3.2). Fig. 4 shows a schematic of Semantic back-projection module. Our design is inspired by the Recovery Decoder in [45], yet differs in that we reconstruct SAM features using a lightweight MLP, rather than reconstructing CLIP features via multi-head cross-attention.

Attention-aware Upsampling Decoder: Finally, as the refined correlation features ϕ'' are at 1/16th of the input image resolution, an attention-aware upsampling decoder is employed to restore full resolution. Initially, ϕ'' is upsampled by a factor of 2 using a transposed convolution, yielding ϕ'_{2x} . Spatial and channel attention features [3] are then computed from the class-averaged upsampled features $\bar{\phi}'_{2x}$ as follows:

$$A^{sp} = \text{conv}\left(\text{avgpool}_{sp}\left(\bar{\phi}'_{2x}\right)\right), \quad (8)$$

$$A^{ch} = \text{conv}\left(\text{avgpool}_{ch}\left(\bar{\phi}'_{2x}\right)\right), \quad (9)$$

where avgpool_{sp} and avgpool_{ch} denote average pooling across the spatial and channel dimensions, respectively. The intermediate guidance features F'_g from SAM are then transformed via

$$F'_g = A^{sp} \odot \text{up}(F_g^2, 2) + A^{ch} \odot \text{up}(F_g^2, 2) + \text{up}(F_g^2, 2), \quad (10)$$

where \odot represents the Hadamard product and $\text{up}(\cdot, 2)$ denotes nearest neighbor upsampling by a factor of 2. The upsampled correlation feature ϕ_{2x} is obtained by concatenating ϕ'_{2x} with the transformed guidance feature F'_g and applying a convolution:

$$\phi_{2x} = \text{conv}\left([\phi'_{2x}, F'_g]\right). \quad (11)$$

This upsampling process is repeated to generate ϕ_{4x} , which is then refined via a convolution layer and bilinear upsampling to yield the final segmentation map \hat{y} .

Dataset	Classes
iSAID	ship, storage tank, baseball diamond, basketball court, ground track field, large vehicle, swimming pool, roundabout, plane,
	tennis court, bridge, small vehicle, helicopter, soccer ball field, harbor
DLRSD	chaparral, court, dock, field, grass, mobile home, sand, ship, tanks, water,
	airplane, bare soil, buildings, cars, pavement, sea, trees
OEM	bareland, rangeland, road, building,
	developed space, tree, water, agriculture land

Table 1. Class details for different datasets used in our work. Classes in Green are seen during training, whereas classes in Magenta are only encountered in inference.

3.2. Loss Functions

To train our network, we utilize a combination of loss functions that jointly optimize segmentation accuracy and feature semantic consistency.

Binary Cross-Entropy Loss: We use the binary cross-entropy loss to align the predicted segmentation maps with the ground truth. The loss is defined as

$$\mathcal{L}_{bce} = \frac{1}{HW} \sum_{c \in \mathcal{C}} \sum_{i,j} \left[-y_{ij}^c \log(\hat{y}_{ij}^c) - (1 - y_{ij}^c) \log(1 - \hat{y}_{ij}^c) \right], \quad (12)$$

where \hat{y}_{ij}^c and y_{ij}^c denote the predicted probability and ground truth for class c at pixel (i, j) , respectively.

Semantic Back-Projection Loss: To ensure that the reconstructed feature ψ is semantically consistent with the SAM features from the guidance encoder, we introduce the semantic back-projection loss:

$$\mathcal{L}_{sem} = \|\psi - sg(F_g^3)\|_2^2. \quad (13)$$

where $sg(\cdot)$ is stop-gradient operator.

Overall Loss: The final loss function is a linear combination of the binary cross-entropy loss and the semantic back-projection loss:

$$\mathcal{L} = \mathcal{L}_{bce} + \mathcal{L}_{sem}. \quad (14)$$

4. Experiments

4.1. Dataset details

iSAID dataset is originally an instance segmentation dataset developed from large-scale object detection dataset DOTA [33], which consists of aerial images collected from multiple sensors. The dataset consists of varying high-resolution images and their annotations for 15 object categories. We utilize the processed iSAID dataset [3, 41] which consists of 18,076 training and 6,363 validation images of 256×256 resolution.

DLRSD - Dense Labeling Remote Sensing Dataset is an extension of the multi-label Remote Sensing Image Retrieval (RSIR) archive [5]. The images from this archive were semantically divided and assigned predefined pixel labels to enable various downstream tasks beyond image retrieval.

The processed DLRSD dataset [25] consists of 7002 aerial images with a spatial resolution of 256×256 , along with annotations for 17 object categories. The training set contains 5601 images, while the validation set includes 1401 images.

OpenEarthMap [34] (OEM) is a global high-resolution land cover mapping dataset comprising 5,000 aerial and satellite images with annotations for eight categories. It is constructed using images from existing benchmark datasets, covering 97 regions across 44 countries, with a ground sampling distance of 0.25–0.5m. For our experiments, we excluded the xBD subset due to partial availability. Our processed dataset contains 31,154 training images and 5,195 validation images, each of size 256×256 .

For each dataset, we split the classes into seen and unseen categories. Only the seen classes are available during training, while all classes are used during inference. Details of the seen-unseen class split are provided in Table 1.

4.2. Implementation details

We develop our models using PyTorch and Detectron2 framework. We have used initial learning rate of 2×10^{-6} for VL backbone. Following [7], we only fine-tune query and value projection matrices of image and text encoders in the VL backbone keeping other parameters frozen. Initial learning rate is set to 2×10^{-4} for every other module except guidance encoder. We have kept parameters of guidance encoder frozen throughout training. We have used AdamW optimizer with a batch size of 4 to train our models. Our model is trained for 10K, 5K and 15K iterations for iSAID, DLRSD and OEM datasets respectively. We use one NVIDIA A100 80GB GPU to run our experiments.

4.3. Evaluation metrics

We compute Intersection-over-Union (IoU) metric for all the classes. We report mean IoU score over seen classes and unseen classes separately, denoted as **s-mIoU** and **u-mIoU** respectively as used commonly in Open-Vocabulary Segmentation literature [8, 47]. Additionally, we report the harmonic mean of s-mIoU and u-mIoU, denoted as **h-mIoU**, which serves as the key metric for evaluating generalized zero-shot performance.

Method	Venue	iSAID			DLRSD			OEM			Average		
		s-mIoU	u-mIoU	h-mIoU	s-mIoU	u-mIoU	h-mIoU	s-mIoU	u-mIoU	h-mIoU	s-mIoU	u-mIoU	h-mIoU
SAN	CVPR'23	62.28	35.77	45.44	50.18	12.62	20.17	45.36	15.19	22.76	52.61	21.19	29.46
SCAN	CVPR'24	44.31	35.55	39.45	25.86	20.10	22.62	31.85	16.53	21.77	34.01	24.06	27.95
SED	CVPR'24	74.37	36.00	48.51	77.73	27.00	40.08	57.46	46.95	51.68	69.85	36.65	46.76
CAT-Seg	CVPR'24	72.57	42.64	53.70	52.79	24.39	33.36	49.33	39.51	43.87	58.23	35.51	43.64
OVRs	arXiv'24	75.85	49.27	59.73	61.00	31.86	41.86	51.20	47.23	49.13	62.68	42.79	50.24
AerOSeg (Ours)	-	75.48	51.46	61.20	60.37	39.12	47.48	51.27	48.16	49.66	62.37	46.25	52.78

Table 2. **Quantitative comparison with state-of-the-art methods.** Values in red and blue indicate the best and second-best results, respectively. The results show that while other methods tend to overfit seen classes, our approach generalizes well to both seen and unseen classes, achieving the highest average h-mIoU score.

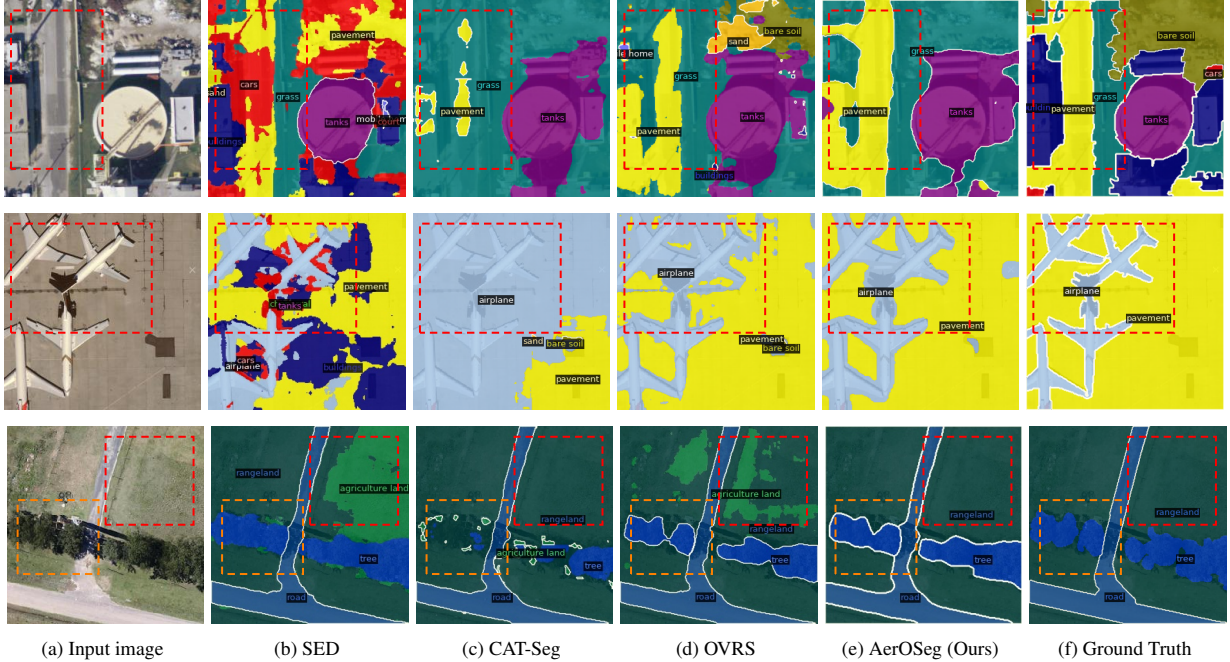


Figure 5. **Qualitative comparison with state-of-the-art methods.** Dashed bounding boxes highlight regions where our model achieves more precise segmentation with sharper boundaries.

4.4. Comparison with state-of-the-art

We compare our approach with several state-of-the-art Open-Vocabulary Segmentation methods: SAN [38], SCAN [17], SED [36], CAT-Seg [7], and OVRs [3]. Among these, OVRs is designed for remote sensing, while the others are originally developed for natural images. These models are trained using their publicly available source codes with default configurations. We use their base versions, where SED employs a CLIP ConvNeXT-B architecture, while the others utilize a CLIP ViT-B/16 as the Vision-Language backbone. Table 2 presents a quantitative comparison of these methods. AerOSeg achieves higher h-mIoU scores than other state-of-the-art methods on the iSAID and DLRSD datasets and ranks second on the OEM dataset. On average, AerOSeg surpasses the next best model, OVRs, by a margin of 2.54 h-mIoU, highlighting its strong generalization capability.

In Fig. 5, we qualitatively compare segmentation pre-

dictions across models. Our model achieves more precise segmentation for “Pavement” in the first image and sharper boundaries for “Airplane” in the second. In the third row, it accurately segments “Tree” and “Rangeland”.

For comparison with training-free method SegEarth-OV [15] and more qualitative results, please refer to Sec. 3 of Supplementary material.

4.5. Ablation study

Main components ablation: We experimented with multiple configurations to evaluate the effectiveness of different components in our framework:

- **(I) Baseline:** Our model without the Guidance Encoder, trained only with \mathcal{L}_{bce} . A single generic prompt, "A photo of a [CLS] in a scene", is used instead of the Remote-Sensing prompt ensemble. Intermediate CLIP features serve as guidance for spatial refinement and the decoder.
- **(II) Config-A:** Our model with the Guidance Encoder and

Configuration	Guidance Encoder	Semantic Back-projection loss	Remote-Sensing Prompts	iSAID	DLRSD	OEM	Average
Baseline	✗	-	✗	57.88	28.69	46.86	44.47
Config-A	✓	✗	✗	58.10	39.93	50.46	49.49
Config-B	✓	✓	✗	59.32	42.34	47.16	49.60
Config-C (Ours)	✓	✓	✓	61.20	47.48	49.66	52.78

Table 3. Ablation study for different components of our framework.

a single generic prompt, trained with \mathcal{L}_{bce} .

- **(III) Config-B:** Our model with the Guidance Encoder and a single generic prompt, trained with \mathcal{L}_{bce} and \mathcal{L}_{sem} .
- **(IV) Config-C:** Our full model.

Table 3 presents the h-mIoU scores for these configurations. Adding the Guidance Encoder in Config-A improves the average h-mIoU by 5.02 over the Baseline. Training with Semantic Back-Projection loss further enhances h-mIoU by 1.22 and 2.41 for iSAID and DLRSD, respectively. Finally, incorporating the Remote Sensing prompt ensemble yields additional gains of 1.88, 5.14, and 2.5 for the iSAID, DLRSD, and OEM datasets, respectively, compared to Config-B.

Guidance encoder training: We experimented with training the Guidance Encoder jointly with the rest of the network, fine-tuning it with an initial learning rate of 2×10^{-6} . A quantitative comparison of h-mIoU scores obtained for the fine-tuned and frozen Guidance Encoder is shown in Table 4. While fine-tuning improves iSAID performance by 0.36 h-mIoU, the frozen Guidance Encoder yields significantly better results for DLRSD and OEM, with gains of 3.07 and 2.63 h-mIoU, respectively. We hypothesize that, given the relatively small dataset size, the frozen Guidance Encoder performs better on average.

Guidance Encoder Training Strategy	iSAID	DLRSD	OEM	Average
Fine-tune 🔥	61.56	44.41	47.03	51.00
Frozen (Ours) ❄️	61.20	47.48	49.66	52.78

Table 4. Quantitative comparison between fine-tuned and frozen Guidance Encoder.

Ablation on Semantic Back-Projection Loss: We experimented with different reconstruction targets for our proposed semantic back-projection module: (i) RGB, (ii) CLIP features, and (iii) SAM features. Table 5 compares the h-mIoU scores on three datasets for different reconstruction targets. Using RGB as the reconstruction target leads to significantly lower performance due to overfitting on known classes. Reconstructing CLIP features improves performance for both seen and unseen categories compared to RGB. However, the best results are achieved when SAM features are used as the reconstruction target.

Effect of Different Guidance Encoders: We investigated various SAM image encoders as our guidance encoder, specifically experimenting with SAM, SAM2 [22], and SAM2.1 [22]. While SAM is originally trained for

Reconstruction target	iSAID	DLRSD	OEM	Average
RGB	14.93	40.03	38.21	31.06
CLIP features	53.27	42.10	48.55	47.97
SAM features (Ours)	61.20	47.48	49.66	52.78

Table 5. Ablation study on different reconstruction targets for the Semantic Back-Projection module.

promptable image segmentation, SAM2 and SAM2.1 are designed for promptable video segmentation. For SAM, we used the ViT-L and ViT-H variants, whereas for SAM2 and SAM2.1, we employed the Hiera-Base+ variant. Table-6 presents the quantitative results on the iSAID dataset using different SAM variants as the guidance encoder. Among these, SAM2.1 achieved the best overall performance.

Guidance Encoder	Architecture	s-mIoU	u-mIoU	h-mIoU
SAM	ViT-L	75.48	51.46	61.20
SAM	ViT-H	74.90	52.81	61.94
SAM2	Hiera-Base+	77.63	53.32	63.22
SAM2.1	Hiera-Base+	77.06	53.93	63.45

Table 6. Quantitative comparison between different Guidance encoders on iSAID dataset.

We have also explored the effect of different VL backbones, please refer to Sec. 1 of Supplementary Material for more details.

5. Conclusion

In this work, we introduce AerOSeg, a novel open-vocabulary segmentation framework for remote sensing that combines CLIP-based vision-language alignment with SAM-guided feature refinement. Our approach mitigates the semantic knowledge gap in CLIP by leveraging SAM as a Guidance Encoder. Additionally, Semantic Back-Projection loss ensures better alignment between image-text correlation features and SAM features, preserving semantic consistency. Extensive experiments on benchmark remote sensing datasets demonstrate that our model outperforms existing methods, excelling in generalization to both seen and novel classes. This work paves the way for advancing Open-Vocabulary Semantic Segmentation in geospatial applications.

Acknowledgements. We would like to thank C-MInDS, IIT Bombay and IITB-Monash Research Academy for financial support.

References

- [1] Patrik Olā Bressan, José Marcato Junior, José Augusto Correa Martins, Maximilian Jaderson de Melo, Diogo Nunes Gonçalves, Daniel Matte Freitas, Ana Paula Marques Ramos, Michelle Taís Garcia Furuya, Lucas Prado Osco, Jonathan de Andrade Silva, Zhipeng Luo, Raymundo Cordero Garcia, Lingfei Ma, Jonathan Li, and Wesley Nunes Gonçalves. Semantic segmentation with labeling uncertainty and class imbalance applied to vegetation mapping. *International Journal of Applied Earth Observation and Geoinformation*, 108:102690, 2022. 2
- [2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3
- [3] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024. 2, 3, 4, 5, 6, 7
- [4] Yice Cao, Chenchen Liu, Zhenhua Wu, Wenxin Yao, Liu Xiong, Jie Chen, and Zhixiang Huang. Remote sensing image segmentation using vision mamba and multi-scale multi-frequency feature fusion, 2024. 2
- [5] Bindita Chaudhuri, Begüm Demir, Subhasis Chaudhuri, and Lorenzo Bruzzone. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2): 1144–1158, 2017. 6
- [6] 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. 3
- [7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryoung Kim. Catseg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 3, 4, 6, 7
- [8] 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. 3, 6
- [9] 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. 3
- [10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1
- [11] Michael Kampffmeyer, Arnt-Børre Salberg, and Robert Jenssen. Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. 2016. 2
- [12] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 5
- [13] 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. 3
- [14] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 3
- [15] 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. 3, 7
- [16] Xiang Li, Congcong Wen, Yuan Hu, and Nan Zhou. Rs-clip: Zero shot remote sensing scene classification via contrastive vision-language supervision. *International Journal of Applied Earth Observation and Geoinformation*, 124:103497, 2023. 3
- [17] 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. 7
- [18] 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, 4
- [19] José Augusto Correa Martins, Keiller Nogueira, Lucas Prado Osco, Felipe David Georges Gomes, Danielle Elis Garcia Furuya, Wesley Nunes Gonçalves, Diego André Sant’Ana, Ana Paula Marques Ramos, Veraldo Liesenberg, Jeferson Alex dos Santos, Paulo Tarso Sanches de Oliveira, and José Marcato Junior. Semantic segmentation of tree-canopy in urban environment with pixel-wise deep learning. *Remote Sensing*, 13(16), 2021. 2
- [20] 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, 3
- [21] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 3
- [22] 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. 3, 8
- [23] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching.

- In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017. 4
- [24] Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024. 3
- [25] Zhenfeng Shao, Ke Yang, and Weixun Zhou. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing*, 10(6), 2018. 6
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 2
- [27] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [28] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 3
- [29] Hong Wang, Xianzhong Chen, Tianxiang Zhang, Zhiyong Xu, and Jiangyun Li. Cctnet: Coupled cnn and transformer network for crop segmentation of remote sensing images. *Remote Sensing*, 14(9), 2022. 2
- [30] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M. Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. 2
- [31] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2024. 3
- [32] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu, and Yueming Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.12620*, 2023. 3
- [33] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beaulongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 6
- [34] 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 (WACV)*, 2023. 6
- [35] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 3
- [36] 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*, pages 3426–3436, 2024. 7
- [37] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 3
- [38] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xi-ang 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. 3, 7
- [39] Zhiyong Xu, Weicun Zhang, Tianxiang Zhang, and Jiangyun Li. Hrcnet: High-resolution context extraction network for semantic segmentation of remote sensing images. *Remote Sensing*, 13(1), 2021. 2
- [40] Zhiyong Xu, Weicun Zhang, Tianxiang Zhang, Zhifang Yang, and Jiangyun Li. Efficient transformer for remote sensing image segmentation. *Remote Sensing*, 13(18), 2021. 2
- [41] Xiwen Yao, Qinglong Cao, Xiaoxu Feng, Gong Cheng, and Junwei Han. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2021. 6
- [42] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [43] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169: 114417, 2021. 2
- [44] Cheng Zhang, Wanshou Jiang, Yuan Zhang, Wei Wang, Qing Zhao, and Chenjie Wang. Transformer and cnn hybrid deep neural network for semantic segmentation of very-high-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–20, 2022. 2
- [45] Yi Zhang, Meng-Hao Guo, Miao Wang, and Shi-Min Hu. Exploring regional clues in clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3270–3280, 2024. 5
- [46] Qi Zhao, Jiahui Liu, Yuewen Li, and Hong Zhang. Semantic segmentation with attention mechanism for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, PP:1–13, 2021. 2
- [47] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 6
- [48] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae

Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3