

ConInfer: Context-Aware Inference for Training-Free Open-Vocabulary Remote Sensing Segmentation

Wenyang Chen¹, Zhanxuan Hu¹, Yaping Zhang¹, Hailong Ning², Yonghang Tai^{1*}

¹Yunnan Normal University

²Xi'an University of Posts & Telecommunications

{dog.yang000, zhanxuanhu}@gmail.com, zhangyp@ynnu.edu.cn
ninghailong@xupt.edu.cn, taiyonghang@126.com

Abstract

Training-free open-vocabulary remote sensing segmentation (OVRSS), empowered by vision-language models, has emerged as a promising paradigm for achieving category-agnostic semantic understanding in remote sensing imagery. Existing approaches mainly focus on enhancing feature representations or mitigating modality discrepancies to improve patch-level prediction accuracy. However, such independent prediction schemes are fundamentally misaligned with the intrinsic characteristics of remote sensing data. In real-world applications, remote sensing scenes are typically large-scale and exhibit strong spatial as well as semantic correlations, making isolated patch-wise predictions insufficient for accurate segmentation. To address this limitation, we propose **ConInfer**, a context-aware inference framework for OVRSS that performs joint prediction across multiple spatial units while explicitly modeling their inter-unit semantic dependencies. By incorporating global contextual cues, our method significantly enhances segmentation consistency, robustness, and generalization in complex remote sensing environments. Extensive experiments on multiple benchmark datasets demonstrate that our approach consistently surpasses state-of-the-art per-pixel VLM-based baselines such as SegEarth-OV, achieving average improvements of 2.80% and 6.13% on open-vocabulary semantic segmentation and object extraction tasks, respectively. The implementation code is available at [here](#).

1. Introduction

Semantic segmentation plays a crucial role in remote sensing by enabling detailed scene understanding and supporting diverse downstream applications such as land cover mapping [8], urban planning [22], and environmental monitoring [31]. Traditional remote sensing segmentation meth-

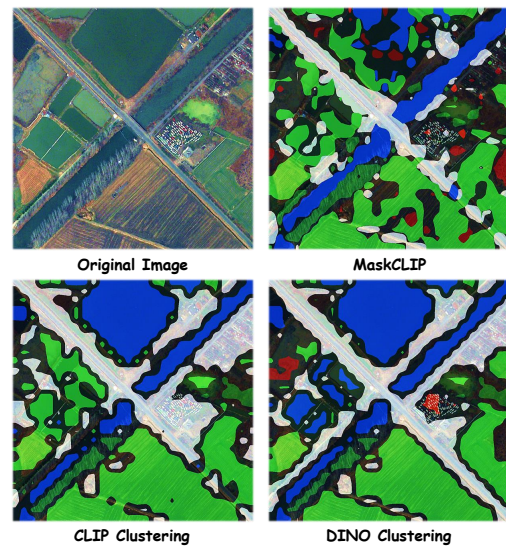


Figure 1. Compared with the direct predictions of MaskCLIP, the segmentation maps generated by clustering patch features from CLIP or DINOv3 exhibit superior spatial consistency.

ods [20, 37, 51], however, are typically developed under *closed-set assumptions*, where the model is restricted to a fixed set of predefined categories. Such a setting fails to reflect dynamic real-world scenarios, where novel or unseen objects frequently appear in large-scale and heterogeneous environments. This limitation motivates the exploration of *Open-Vocabulary Remote Sensing Segmentation* [18, 19, 49], which aims to achieve category-agnostic semantic understanding and generalize to arbitrary concepts beyond the training label space.

Empowered by the zero-shot capabilities of vision-language models (VLMs), train-free open-vocabulary segmentation (OVS) has achieved remarkable success on natural images, and most existing train-free OVRSS approaches directly build upon these techniques to process remote sensing data. Two major research directions have emerged: (1)

*Corresponding author

enhancing patch-level representations [18, 52] to strengthen the discriminative power of visual tokens for dense prediction, and (2) mitigating the modality gap between visual and textual spaces [3, 38] to improve cross-modal consistency. While these methods have achieved promising results, they largely inherit the paradigm of *independent prediction*, where each patch is treated as an isolated unit.

However, remote sensing imagery differs fundamentally from natural images. Remote sensing scenes are captured at large spatial scales, encompassing heterogeneous objects (e.g. buildings, vegetation, water bodies) with complex spatial organization and strong semantic correlations [30]. Moreover, visually similar regions (e.g. roads and runways) may convey different semantics depending on their spatial context, making purely independent predictions prone to fragmentation or misclassification. Consequently, directly transferring OVS techniques from natural images fails to fully exploit the contextual dependencies intrinsic to remote sensing imagery, leading to suboptimal segmentation consistency in large-scale scenes, as shown in Fig. 1.

To overcome these limitations, we argue that contextual information should be explicitly incorporated during inference to capture spatial dependencies and improve segmentation consistency. As illustrated in Fig. 1, compared to the results obtained from a text classifier, clustering the patch features extracted by DINOv3 [32] produces segmentation maps with superior spatial coherence. However, DINOv3 features alone lack semantic alignment with the language space, making them unsuitable for direct open-vocabulary prediction. To this end, we propose a *Context-aware Inference framework (ConInfer)* for OVRSS that jointly exploits the strengths of DINOv3 and VLMs. Specifically, our approach leverages the contextual cues encoded by DINOv3 to refine the VLM’s predictions, enforcing spatial and semantic consistency across patches (See Fig. 2). This design effectively bridges the gap between semantic understanding and spatial reasoning. Notably, our framework is entirely *training-free* and orthogonal to existing methods, allowing it to serve as a *plug-and-play module* for enhancing various VLM-based segmentation models. In comparison with Segearth-OV [18], our approach build upon a simple baseline [15] outperforms it by 2.80% and 6.13% mIOU on open-vocabulary semantic segmentation, and object extraction tasks, respectively. The main contributions of this work can be summarized as follows:

- **A new perspective.** To the best of our knowledge, this is the *first work* to explicitly incorporate contextual information of remote sensing imagery during the inference stage, offering a new perspective on open-vocabulary remote sensing segmentation.
- **A simple yet effective framework.** We propose a *simple yet effective context-aware inference framework* that establishes a paradigm for integrating the contextual rea-

soning capability of vision foundation models with the semantic alignment ability of vision-language models.

- **Promising results.** Despite its simplicity and training-free nature, our approach achieves consistently superior performance across multiple benchmarks, showing clear potential for advancing OVRSS in real-world scenarios.

2. Related Work

2.1. Open-Vocabulary Segmentation

OVS aims to enable pixel-level recognition for arbitrary categories beyond the closed-set vocabulary, typically by leveraging the zero-shot capabilities of VLMs [50]. Existing studies can be broadly categorized into *training-based* [36, 42, 44–46] and *training-free* [15, 16, 47, 52, 56] approaches. The former often fine-tune the visual or textual components of VLMs with additional supervision, while the latter perform segmentation directly using additional pre-trained models, such as DINO [27, 32] and SAM [14, 29], without any further training. Compared with training-based approaches, training-free methods avoid the additional cost of annotation and large-scale computation, making them increasingly attractive to the research community.

In the natural image domain, training-free OVS has achieved remarkable progress by adapting image-level representations from VLMs for dense, pixel-level prediction. Early studies focused on enhancing the spatial localization of visual features, for instance, by refining the attention maps or activations from VLMs to improve object boundary delineation [10, 15, 34, 48, 56]. Subsequent works sought to enrich representation quality and reduce the modality gap through multi-scale feature fusion [1, 4] or prototype optimization [2, 13, 47]. Inspired by these advances, recent efforts have explored extending OVS frameworks to remote sensing imagery [11, 17, 18, 49]. Typical strategies include domain-specific feature calibration [17, 18], prototype alignment [53, 54], or pseudo-label-based supervision to mitigate the domain gap between natural and remote sensing images [38]. While such methods show promising results, they mostly inherit the paradigm of *independent prediction*, neglecting the strong spatial and semantic correlations [30] of remote sensing data. Consequently, *independent prediction* fails to fully exploit the contextual dependencies intrinsic to remote sensing imagery.

2.2. Context Modeling in Segmentation

Context modeling forms a cornerstone in semantic segmentation, aiming to capture spatial and semantic dependencies among pixels or regions to resolve ambiguities arising from local predictions and ensure global consistency of segmentation results. Existing methods focus on aggregating contextual information via multi-scale feature fusion [6] and feature aggregation techniques [11]. In the remote sensing

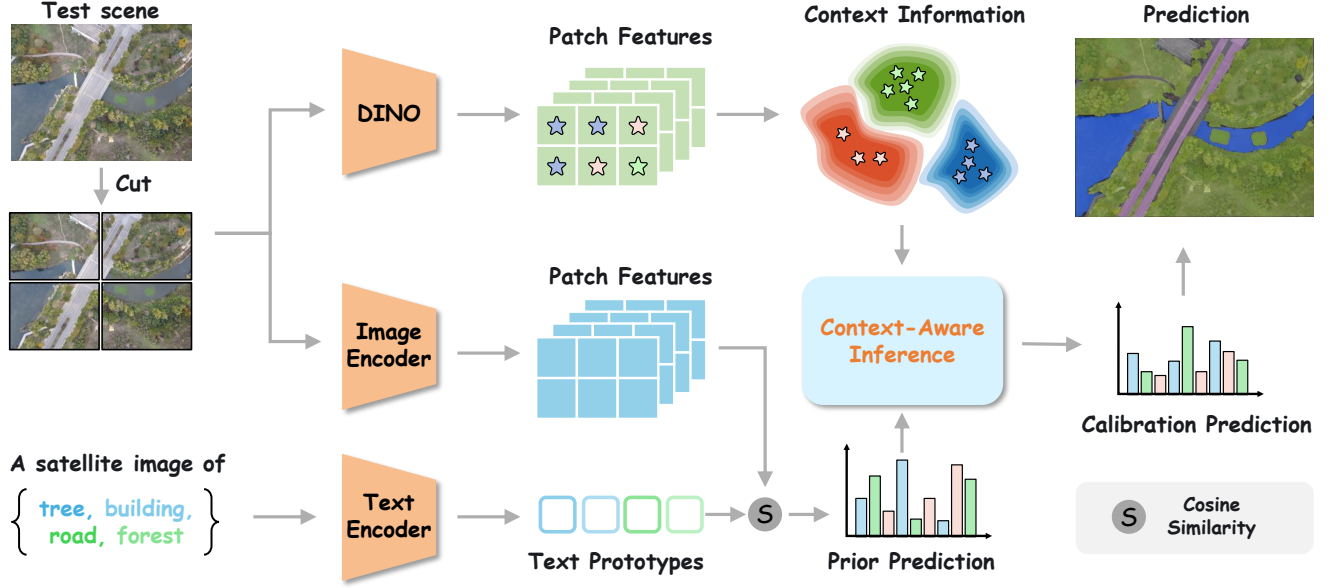


Figure 2. Overview of the proposed Context-Aware Inference (*ConInfer*) framework. *ConInfer* performs joint, training-free inference by integrating semantic priors from CLIP with structural regularization from DINO. This formulation enables globally consistent open-vocabulary segmentation without any retraining.

domain, due to the large spatial scale, dense distribution of objects, and strong spatial correlations, the proper exploitation of contextual cues is essential for suppressing fragmented predictions and improving the coherence of overall scene structure [30].

However, a common characteristic of current methods is that they integrate context modeling into the training stage by embedding attention modules, graph convolutional layers, or multi-scale fusion units within the network architecture [21, 40, 55]. This paradigm suffers from two main limitations: first, the context modules are tightly coupled with backbone networks, reducing flexibility and generality; second, they cannot operate independently during inference, thus failing to serve as plug-and-play components for arbitrary pre-trained VLMs. In training-free OVRSS, such structural dependencies constrain the generalization of pre-trained VLMs and weaken their adaptability to novel categories and large-scale scenes.

3. Method

Fig. 2 illustrates the overall pipeline of our proposed *ConInfer* framework. Given a large-scale remote sensing scene, we first partition it into fixed-size, non-overlapping tiles. Each tile is processed independently to extract two types of patch-level features: (1) high-level contextual features from a vision foundation model (e.g. DINOv3) that capture structural patterns within the tile, and (2) semantic-aligned visual features from *any* VLM-based OVRSS model that produce open-vocabulary recognition scores via cosine sim-

ilarity with text prototypes. The DINO-derived features implicitly encode *contextual information* across patches, enabling spatial grouping of visually coherent regions that often correspond to semantically related areas in remote sensing imagery. Meanwhile, the VLM branch yields a *prior prediction* distribution for each patch from text prototypes, but operates independently and is agnostic to scene context. *ConInfer* bridges these two complementary sources through a *training-free context-aware inference* module, yielding a *calibrated prediction* that integrates semantic alignment from the language domain with contextual reasoning from the visual domain. In the following sections, we present the design and formulation of each component within *ConInfer*.

3.1. Prior Prediction from VLM-based OVRSS

We follow the standard OVRSS paradigm to generate the *prior prediction*. Specifically, given an input tile $\mathcal{I}$, a vision encoder $f(\cdot)$ extracts patch-level features:

$$\mathbf{V} = f(\mathcal{I}) \in \mathbb{R}^{n \times d}, \quad (1)$$

where n denotes the number of patches, and d denotes the feature dimension. For a candidate category set $\mathcal{C}$, each category name $c \in \mathcal{C}$ is converted into a prompt template $T(c)$ and encoded using a text encoder $g(\cdot)$ to obtain text prototype vectors:

$$\mathbf{T} = \{g(T(c)) \mid c \in \mathcal{C}\} \in \mathbb{R}^{|\mathcal{C}| \times d}. \quad (2)$$

Prediction is then achieved by computing patch-to-text similarity via cosine similarity:

$$\mathbf{s}_{ij} = \cos(\mathbf{V}_i, \mathbf{T}_j). \quad (3)$$

After softmax normalization, we obtain the per-patch category probability distribution:

$$\mathbf{p}_{ij} = \frac{\exp(\mathbf{s}_{ij}/\tau)}{\sum_{k=1}^{|\mathcal{C}|} \exp(\mathbf{s}_{ik}/\tau)}. \quad (4)$$

The resulting distribution $\mathbf{p}_{ij}$ forms the *prior prediction* for each patch. Importantly, the VLM-based OVRSS model used here is not restricted to a specific architecture or implementation: any existing OVRSS model can serve as this branch. *ConInfer* only consumes prediction scores and does not modify or fine-tune any model parameters.

3.2. Context Information Extraction from VFMs

To capture contextual dependencies in remote sensing imagery, we leverage a vision foundation model (VFM), specifically DINOv3, to extract high-level visual features from all patches across all tiles of the scene. Unlike the VLM-based OVRSS branch, which is aligned with textual semantics, the VFM is trained purely on visual data and therefore encodes structural layout, spatial organization, and appearance regularities present within the scene.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ denote the set of DINO-derived patch features, where each $\mathbf{x}_i \in \mathbb{R}^d$ is one patch feature in the entire scene. To model the contextual structure, we fit a Gaussian Mixture Model (GMM) with $|\mathcal{C}|$ mixture components:

$$p(\mathbf{x}) = \sum_{k=1}^{|\mathcal{C}|} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad (5)$$

where π_k , $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixture weight, mean, and covariance of each Gaussian component, respectively. The GMM parameters can be estimated using Expectation-Maximization (EM) in a training-free manner.

For each patch feature $\mathbf{x}_i$, we compute its posterior assignment over mixture components:

$$\mathbf{q}_i = [q_{i1}, \dots, q_{i|\mathcal{C}|}], \quad q_{ik} = p(z = k \mid \mathbf{x}_i), \quad (6)$$

which reveals the context information. Patches belonging to semantically relevant or spatially coherent regions tend to exhibit similar posterior distributions $\mathbf{q}_i$. As such, the GMM posterior encapsulates structural context cues that reflect *how patches are spatially and visually related across the scene*. However, despite capturing strong contextual regularities, the GMM posterior does not reveal semantic meaning because it is not aligned with the textual label space. This complementary property naturally motivates the next step.

3.3. Context-Aware Inference

Given the semantic prior distributions $\{\mathbf{p}_i\}_{i=1}^N$ (from the OVRSS branch) and the contextual posterior distributions $\{\mathbf{q}_i\}_{i=1}^N$ (from the VFM branch), our goal is to obtain calibrated probability distributions $\{\mathbf{z}_i\}_{i=1}^N$. The calibrated distributions should jointly reflect both semantic alignment (from VLM) and contextual coherence (from VFM).

Joint Latent Distribution Optimization. For each patch, we formulate $\mathbf{z}_i$ as a latent categorical distribution that is encouraged to stay close to both sources of information. We achieve this by minimizing the sum of two KL divergences:

$$\min_{\{\mathbf{z}_i\}_{i=1}^N} \sum_{i=1}^N \left[\text{KL}(\mathbf{z}_i \parallel \mathbf{p}_i) + \text{KL}(\mathbf{z}_i \parallel \mathbf{q}_i) \right], \quad (7)$$

where N denotes the number of patches. Eq. (7) encourages $\mathbf{z}_i$ to lie in the agreement region between the semantic prediction $\mathbf{p}_i$ and the contextual structure encoded by $\mathbf{q}_i$.

Directly optimizing Eq. (7) is of limited value because $\mathbf{q}_i$ is obtained from a context model learned *without* semantic information; thus $\mathbf{q}_i$ is not aligned with $\mathbf{p}_i$ in the semantic space. A straightforward remedy is to initialize the GMM with the semantic prior $\mathbf{p}_i$ so that the learned $\mathbf{q}_i$ carries initial semantic alignment; one can then apply Eq. (7) to obtain a consensus prediction [39]. However, such an isolated procedure prevents the GMM parameter estimation from benefiting from the evolving $\mathbf{z}_i$. We therefore propose *joint learning* of the latent consensus distributions and the GMM parameters, leading to the following objective:

$$\min_{\{\mathbf{z}_i\}_{i=1}^N, \boldsymbol{\mu}, \boldsymbol{\Sigma}} \sum_{i=1}^N \left[\text{KL}(\mathbf{z}_i \parallel \mathbf{p}_i) + \text{KL}(\mathbf{z}_i \parallel \mathbf{q}_i) \right], \quad (8)$$

where $\mathbf{q}_i$ is the GMM posterior computed from the DINO-derived features and depends on the GMM parameters (means $\boldsymbol{\mu}$ and covariances $\boldsymbol{\Sigma}$).¹ This joint objective enables $\mathbf{z}_i$ and $\mathbf{q}_i$ to be mutually refined: $\mathbf{z}_i$ becomes increasingly context-aware while $\mathbf{q}_i$ gradually absorbs semantic alignment from the VLM prior $\mathbf{p}_i$. Ablation results on joint learning are shown in Tab. 4.

3.4. Optimization

We optimize Eq. (8) using an alternating manner.

Initialization. We first initialize the GMM parameters using the semantic prior distribution $\mathbf{p}_i$. Let $\mathbf{q}_{ik}^{(0)} \propto \mathbf{p}_{ik}$ denote the initial responsibilities, based on which initial mix-

¹The mixture weights π are held fixed to a uniform distribution ($\pi_k = 1/|\mathcal{C}|$) and are therefore not optimization variables.

ture parameters are computed as:

$$\mu_k^{(0)} = \frac{\sum_i q_{ik}^{(0)} \mathbf{x}_i}{\sum_i q_{ik}^{(0)}}, \quad (9)$$

$$\Sigma_k^{(0)} = \frac{\sum_i q_{ik}^{(0)} (\mathbf{x}_i - \mu_k^{(0)}) (\mathbf{x}_i - \mu_k^{(0)})^\top}{\sum_i q_{ik}^{(0)}}. \quad (10)$$

Alternating Updates. Given the initialization, we iteratively update $\mathbf{q}$, $\mathbf{z}$, and the GMM parameters. For notational simplicity, we omit the superscripts hereafter.

(1) Update $\mathbf{q}$ (E-step). With the current mixture parameters $\{\pi_k, \mu_k, \Sigma_k\}$, the GMM posterior is obtained by:

$$q_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)}{\sum_{\ell=1}^C \pi_\ell \mathcal{N}(\mathbf{x}_i | \mu_\ell, \Sigma_\ell)}. \quad (11)$$

(2) Update $\mathbf{z}$. Given the semantic prior $\mathbf{p}_i$ and the contextual posterior $\mathbf{q}_i$, the closed-form solution of $\mathbf{z}$ minimizing the KL-based agreement objective is:

$$\mathbf{z}_{ik} \propto \mathbf{p}_{ik} \odot \mathbf{q}_{ik}, \quad \sum_{k=1}^C \mathbf{z}_{ik} = \mathbf{1}. \quad (12)$$

(3) Update GMM (M-step). We treat the updated $\mathbf{z}_i$ as soft responsibilities and re-estimate mixture parameters:

$$\mu_k = \frac{\sum_i \mathbf{z}_{ik} \mathbf{x}_i}{\sum_i \mathbf{z}_{ik}}, \quad (13)$$

$$\Sigma_k = \frac{\sum_i \mathbf{z}_{ik} (\mathbf{x}_i - \mu_k) (\mathbf{x}_i - \mu_k)^\top}{\sum_i \mathbf{z}_{ik}}. \quad (14)$$

In practice, these three steps are executed in an alternating manner until the algorithm converges, which usually takes fewer than 10 iterations (See Fig. 5).

3.5. Convergence Analysis

Our alternating optimization process shows stable convergence behavior in practice. In each iteration, updating $\mathbf{z}$ decreases the objective by moving it toward the agreement region of $\mathbf{p}$ and $\mathbf{q}$, while updating $\mathbf{q}$ (via GMM re-estimation) makes the contextual posterior better match the consensus distribution $\mathbf{z}$. As a result, the two steps monotonically reduce the objective value and progressively align $\mathbf{z}$ and $\mathbf{q}$. We empirically observe that the optimization converges within a small number of iterations (See Fig. 5).

3.6. Final Prediction

At inference time, we directly use the calibration prediction $\mathbf{Z}$ as the semantic probability for each patch. Since prediction is performed at the patch level, the final segmentation mask for the large-scale remote sensing scene is obtained by mapping each patch back to its original tile position and performing tile-wise upsampling to reconstruct full-resolution dense segmentation.

Algorithm 1: Optimization

Require: VFM features $\mathbf{X} \in \mathbb{R}^{N \times d}$, VLMs prior prediction $\mathbf{P} \in \mathbb{R}^{N \times C}$, max iterations L

Ensure : Consensus distribution $\mathbf{Z}$ and GMM parameters (μ, Σ)

- 1 Initialize GMM parameters (μ^0, Σ^0) using $\mathbf{P}$ via Eqs. (9) and (10);
 - 2 **for** $l = 1$ **to** L **do**
 - 3 Compute contextual posteriors $\mathbf{Q}$ via GMM E-step using Eq. (11);
 - 4 Update consensus distribution $\mathbf{Z}$ via Eq. (12);
 - 5 Re-estimate (μ, Σ) via GMM M-step using $\mathbf{Z}$ according to Eqs. (13) and (14);
 - 6 **end**
 - 7 **return** $\mathbf{Z}, (\mu, \Sigma)$
-

4. Experiments

4.1. Dataset

We follow the evaluation protocol of SegEarth-OV [18] and conduct experiments on 17 open-vocabulary remote sensing segmentation datasets, covering both multi-class semantic segmentation and single-class extraction tasks. For multi-class semantic segmentation, we evaluate on eight datasets: OpenEarthMap [43], LoveDA [35], iSAID [41], Potsdam, Vaihingen², UAVid [24], UDD5 [7], and VDD [5]. These datasets encompass diverse land-cover categories and varied spatial structures, with the first five mainly containing high-resolution satellite imagery and the latter three focusing on unmanned aerial vehicle (UAV) urban scene observations. For single-class extraction, we concentrate on three key tasks: building extraction, using four datasets (WHU^{Aerial} [12], WHU^{Sat.II} [12], Inria [25], and xBD^{pre} [9]); road extraction, also with four datasets (CHN6-CUG [57], DeepGlobe³, Massachusetts [26], and SpaceNet [33]); and flood detection, with the WBS-SI5 dataset⁴. These latter datasets emphasize fine-grained structural targets (*e.g.*, buildings and roads) or dynamic hydrological regions. See Appendix for dataset details.

4.2. Setup

Implementation. Our implementation is built upon the MMSegmentation toolkit⁵. For the vision backbones, we load CLIP [28] using the official ViT-B/16 pretrained weights from OpenAI. For DINOv3, we adopt the ViT-H/16 distilled model for the VDD and UAVid datasets, while using the ViT-L/16 distilled checkpoint pretrained

²<https://www.isprs.org/education/benchmarks/UrbanSemLab>

³<http://deepglobe.org>

⁴<https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images>

⁵<https://github.com/open-mmlab/mms Segmentation>

Table 1. Quantitative comparison of open-vocabulary semantic segmentation performance (mIoU, %) across eight remote sensing datasets. These datasets comprise high-resolution satellite imagery (first five) and UAV-based urban scenes (latter three). **Bold** and underlined values indicate the best and second-best performance, respectively.

Methods	Satellite Images					UAV Images			Average
	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAvid	UDD5	VDD	
CLIP [28]	9.16	9.59	6.53	11.62	7.84	8.94	7.91	11.27	9.11
MaskCLIP [56]	22.30	25.25	14.32	31.41	22.78	30.09	32.74	29.12	26.13
SCLIP [34]	23.85	24.60	15.08	31.84	21.96	29.69	33.25	34.26	26.82
GEM [4]	29.96	26.95	<u>20.01</u>	38.24	21.59	33.49	36.02	38.46	30.59
ClearCLIP [15]	29.94	28.06	16.62	40.20	20.55	34.35	38.07	40.24	31.00
SegEarth-OV [18]	37.81	35.79	18.14	47.94	24.96	42.81	49.64	46.73	37.98
MaskCLIP*(baseline)	32.46	31.58	18.17	44.49	19.94	37.20	42.17	42.54	33.57
ConInfer(Ours)	41.95 _{↑9.49}	39.33 _{↑7.75}	20.08 _{↑1.91}	49.99 _{↑5.50}	31.37 _{↑11.43}	46.40 _{↑9.20}	46.86 _{↑4.69}	50.29 _{↑7.75}	40.78 _{↑7.21}

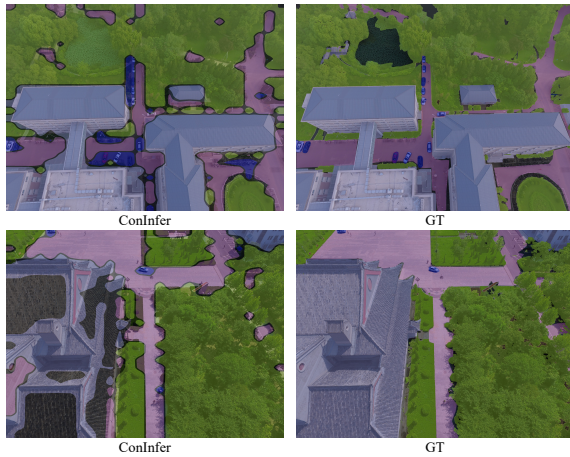


Figure 3. Examples from the UDD5 dataset illustrating cases where ConInfer encounters difficulties, including weak boundaries on small objects (vehicles) and category confusion (roof tiles).

on SAT-493M [32] for all other datasets. For the text encoder, we use the standard ImageNet prompt template (e.g. “a photo of a class name”). In single-class extraction tasks, we further enhance the background prompt by composing it with multiple fine-grained negative concepts (e.g. forest, vegetation, bare land, river, paved road); detailed prompts are provided in the Appendix. Input images are resized to 448×448 pixels, and inference is performed without sliding windows. All experiments are conducted on a single NVIDIA RTX 4090 GPU. During inference, after processing every 50 images, we apply the iterative correction procedure with the number of iterations set to 10.

Evaluation Protocol. We evaluate the performance on semantic segmentation and single-class extraction using mean Intersection over Union (mIoU) and Intersection over Union (IoU) of the foreground class, respectively.

Baseline. As in natural image tasks, the MaskCLIP [56] framework is also applicable to remote sensing scenarios. Based on this framework, we remove the feed-forward network (FFN) and residual connections in the final Trans-

Table 2. Proposed method is a plug-and-play module that can be seamlessly integrated into existing OVS methods.

Methods	OpenEarthMap	WHU ^{Aerial}	WBS-SI
MaskCLIP [56]	22.30	32.00	40.30
+SegEarth-OV	25.94	38.18	47.75
+ConInfer	29.28 _{↑6.98}	54.98 _{↑22.98}	52.43 _{↑12.13}
SCLIP [34]	23.85	37.89	42.91
+SegEarth-OV	27.98	44.61	48.51
+ConInfer	33.52 _{↑9.67}	59.37 _{↑21.48}	49.38 _{↑6.47}
ClearCLIP [15]	29.94	43.59	46.55
+SegEarth-OV	36.12	49.89	55.80
+ConInfer	41.36 _{↑11.42}	56.38 _{↑12.79}	56.08 _{↑9.53}

former layer, following [15]. In addition, we replace the final self-attention module with a multi-level fusion modulated attention mechanism [23, 48], yielding our baseline model, MaskCLIP*. Further architectural details are provided in Appendix.

4.3. Comparison with the State-of-the-Art Methods

To evaluate the effectiveness of the proposed *ConInfer* framework, we compare it against 6 state-of-the-art training-free OVS methods. This includes five representative approaches originally developed for natural image domains, namely Vanilla CLIP [28], MaskCLIP [56], SCLIP [34], GEM [4], and ClearCLIP [15], as well as SegEarth-OV [18], the current state-of-the-art in OVRSS. Notably, SegEarth-OV introduces a universal feature up-sampler during inference to refine dense predictions and better capture the fine-grained characteristics of remote sensing imagery.

Semantic Segmentation. As shown in Tab. 1, the re-implemented MaskCLIP* baseline already brings notable gains over the original MaskCLIP, indicating that refining attention maps remains an effective means of enhancing training-free OVRSS performance. Building upon this stronger baseline, our *ConInfer* further achieves consistent

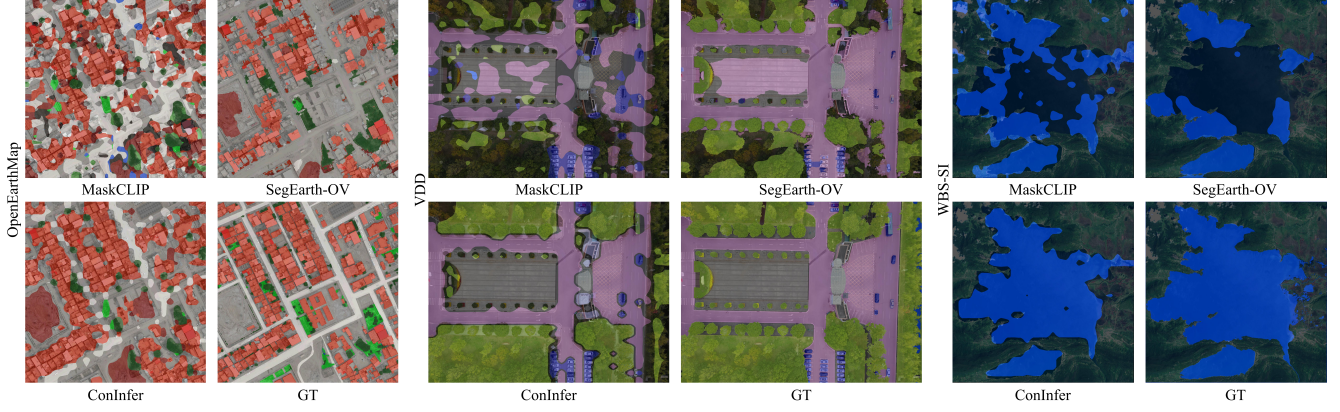


Figure 4. Qualitative comparison of different OVSS methods on the OpenEarthMap (land cover mapping), VDD (UAV aerial scene), and WBS-SI (water body segmentation) datasets. GT denotes the ground truth.

Table 3. Quantitative comparison on the open-vocabulary single-class remote sensing extraction task, measured by *Foreground* IoU. **Bold** and underlined values indicate the best and second-best performance, respectively.

Method	Building Extraction				Road Extraction				Flood Detection	
	WHU ^{Aerial}	WHU ^{Sat.11}	Inria	xBD ^{pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI	Average
CLIP [28]	6.62	1.45	15.12	8.88	3.74	1.13	3.50	3.20	8.98	5.85
MaskCLIP [56]	32.00	11.77	33.15	26.92	16.93	11.53	8.74	17.39	40.30	22.08
SCLIP [34]	37.89	7.82	33.92	29.42	13.86	7.28	5.56	8.13	42.91	20.75
GEM [4]	41.48	15.98	40.74	33.48	22.40	13.43	7.51	17.92	50.64	27.06
ClearCLIP [15]	43.59	13.72	37.08	29.83	25.20	12.86	7.18	17.02	46.55	25.89
SegEarth-OV [18]	52.90	20.99	45.41	35.95	35.13	18.07	10.43	20.85	57.38	33.01
MaskCLIP*(baseline)	45.76	15.17	37.26	31.40	29.12	14.68	8.42	19.83	52.57	28.25
ConInfer(Ours)	58.54 _{↑12.78}	39.35 _{↑24.18}	55.65 _{↑18.39}	41.34 _{↑9.94}	40.00 _{↑10.88}	19.85 _{↑5.17}	12.16 _{↑3.74}	24.02 _{↑4.19}	61.38 _{↑8.81}	39.14 _{↑10.89}

improvements across all eight datasets, reaching an average mIoU of 40.78%. Compared to the strongest competitor, SegEarth-OV, *ConInfer* attains a higher mean performance (+2.80%) without relying on any pretrained upsampler, demonstrating that modeling cross-region semantic dependency at inference is a more direct and generalizable solution than introducing additional parametric modules. We note that *ConInfer* shows slightly lower performance on the UDD5 dataset. Under low-altitude UAV viewpoints, the visual distribution differs substantially from satellite imagery, leading to two main issues when compared with the ground truth: (1) weak boundary sensitivity for small objects, such as vehicles, and (2) misclassification of roof tiles as background rather than belonging to the building class, as illustrated in Fig. 3. Despite these issues, *ConInfer* still produces coherent and semantically consistent segmentation results.

Single-Class Extraction. As presented in Tab. 3, *ConInfer* delivers notably larger gains in single-class extraction than in multi-class segmentation, increasing the baseline average IoU from 28.25% to 39.14% and surpassing SegEarth-OV by 6.13% on average. The improvement is particularly significant on WHU^{Sat.11}, Inria, CHN6-CUG, and WBS-SI, indicating that context-aware inference is highly effective when the target class exhibits strong structural regularity. Buildings and roads typically form spatially continuous, ho-

mogeneous patterns, and the contextual priors encoded by DINOv3 allow *ConInfer* to reinforce such structured dependencies, leading to more coherent and complete extraction regions. In contrast, independent prediction schemes often produce fragmented or discontinuous results. This trend is further corroborated by the visual comparisons in Fig. 4, where *ConInfer* yields segmentation masks with clearer boundaries and improved global connectivity.

Qualitative Results. Fig. 4 presents a qualitative comparison of MaskCLIP, SegEarth-OV, and *ConInfer*. Key observations include: (1) MaskCLIP often produces fragmented regions and redundant classes (e.g. bare ground misidentified as water) and suffers from notable misclassification (e.g. rooftops labeled as background). (2) SegEarth-OV provides finer details, but the lack of contextual reasoning leads to spatial discontinuities (e.g. background appearing within buildings or missing regions in water bodies). (3) *ConInfer* delivers more coherent and semantically consistent segmentation by introducing context-aware inference. More visualizations can be found in Appendix.

4.4. Ablation Studies

ConInfer is Plug-and-Play. By design, *ConInfer* solely utilizes prediction scores and does not alter or fine-tune any model parameters, rendering it an ideal plug-and-play mod-

Table 4. Ablation study of *ConInfer*. We analyze the impact of alternative context information extraction (via CLIP features) and the removal of joint latent distribution optimization (applying prior knowledge post-EM).

Methods	Multi-class segmentation				Single-class extraction			
	OpenEarthMap	LoveDA	iSAID	Vaihingen	WHU ^{Sat.11}	Inria	xBD ^{pre}	WBS-SI
Baseline	32.46	31.58	18.17	19.94	15.17	37.26	31.40	52.57
Context _{CLIP}	35.12 $\uparrow 2.66$	33.21 $\uparrow 1.63$	19.05 $\uparrow 0.88$	25.10 $\uparrow 5.16$	17.38 $\uparrow 2.21$	38.27 $\uparrow 1.01$	34.98 $\uparrow 3.58$	54.48 $\uparrow 1.91$
No joint-learning	36.23 $\uparrow 3.77$	36.27 $\uparrow 4.69$	18.86 $\uparrow 0.69$	32.80 $\uparrow 12.86$	28.56 $\uparrow 13.39$	54.47 $\uparrow 17.21$	36.30 $\uparrow 4.90$	56.78 $\uparrow 4.21$
ConInfer	41.95 $\uparrow 9.49$	39.33 $\uparrow 7.75$	20.08 $\uparrow 1.91$	31.37 $\uparrow 11.43$	39.35 $\uparrow 24.18$	55.65 $\uparrow 18.39$	41.34 $\uparrow 9.94$	61.38 $\uparrow 8.81$

ule for existing OVRSS frameworks. The results presented in Tab. 2 validate its exceptional efficacy in this capacity. When integrated, *ConInfer* delivers significant and consistent performance gains to all baseline models (MaskCLIP, SCLIP, and ClearCLIP) across three distinct datasets. Notably, it boosts MaskCLIP’s performance by an impressive 22.98% on the WHU^{Aerial} dataset. Furthermore, our approach consistently surpasses the performance of SegEarth-OV [18] across all experimental settings. These findings firmly establish *ConInfer* as a highly effective, versatile, and superior solution for enhancing segmentation frameworks with minimal integration effort.

ConInfer Benefits from Better VFMs. We investigate the impact of feature choice for context clustering by replacing DINOv3 features with visual features from CLIP. The results repoted in Tab. 4 indicate consistent improvements across most datasets, with the exception of UDD5 and Potsdam, demonstrating the robustness of our overall design. However, the performance gains are marginal. We attribute this to CLIP’s limited capability in capturing fine-grained details, which suggests that employing more powerful VFMs could lead to further enhancements.

ConInfer Benefits from Joint-learning. To better understand the effect of the proposed joint optimization in Eq. (8), we compare it with a *Decoupled-learning* variant. The decoupled baseline follows a two-stage pipeline: (1) the GMM is initialized using the VLM prior $\mathbf{p}_i$, and its parameters are then optimized independently on DINO features to obtain $\mathbf{q}_i$; (2) the final prediction is obtained by fusing the fixed GMM posterior $\mathbf{q}_i$ with the VLM prior $\mathbf{p}_i$ through Eq. (7). As shown in Tab. 4, *Decoupled-learning* can still yield noticeable improvements. However, its performance consistently lags behind *Joint-learning*. The reason is that *Joint-learning* establishes a bi-directional refinement loop: as $\mathbf{z}_i$ becomes more semantically meaningful, it reshapes the GMM’s posterior structure, which in turn provides more reliable contextual cues for updating $\mathbf{z}_i$. This iterative interplay enables the contextual model and semantic priors to reinforce each other.

4.5. Further Analysis

The convergence behavior of our iterative method is illustrated in Fig. 5. We observe only a modest gap between the

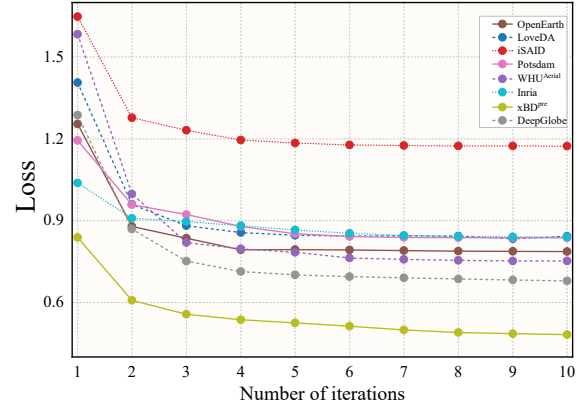


Figure 5. Convergence analysis was conducted on 8 datasets. During the iterative process, the *loss* rapidly converged and subsequently remained stable.

initial and final loss values, confirming the effectiveness of our prior prediction-based initialization strategy. The loss decreases rapidly within the first few iterations and stabilizes thereafter, reaching a smooth plateau without noticeable oscillations. This demonstrates that the optimization procedure converges efficiently in practice.

5. Conclusion

In this work, we presented *ConInfer*, a training-free context-aware inference framework for open-vocabulary remote sensing segmentation. By explicitly modeling cross-patch semantic dependencies and jointly reasoning over multiple spatial units, *ConInfer* substantially improves segmentation consistency, and consistently surpasses existing training-free and state-of-the-art VLM-based baselines across diverse remote sensing datasets. Our results demonstrate that context-awareness is a critical missing component in training-free OVRSS and provides a new inference-level paradigm beyond independent patch prediction. Despite these promising results, *ConInfer* currently operates at the patch level, which limits its capability in fine-grained boundary delineation. In future work, we will explore pixel-level contextual modeling to further enrich structural detail, enhance boundary precision, and strengthen generalization across heterogeneous remote sensing domains.

6. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62365017, the Yunnan Provincial Major Science and Technology Special Project under Grants 202503AP140044 and 202502AS080003, the Key Project of Yunnan Provincial Basic Research Fund under Grant 202401AS070038, and the Yunnan Provincial Science and Technology Talent and Platform Development Project under Grant 202605AS350005.

References

- [1] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation, 2024. arXiv preprint arXiv:2411.15869. 2
- [2] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *CVPR*, pages 3689–3698, 2024. 2
- [3] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *ICCV*, pages 22025–22035, 2025. 2
- [4] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, pages 3828–3837, 2024. 2, 6, 7
- [5] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025. 5
- [6] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary high-resolution remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025. 2
- [7] 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*, pages 347–359, 2018. 5
- [8] Runmin Dong, Lichao Mou, Mengxuan Chen, Weijia Li, Xin-Yi Tong, Shuai Yuan, Lixian Zhang, Juepeng Zheng, Xiaoxiang Zhu, and Haohuan Fu. Large-scale land cover mapping with fine-grained classes via class-aware semi-supervised semantic segmentation. In *ICCV*, pages 16783–16793, 2023. 1
- [9] 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. arXiv preprint arXiv:1911.09296. 5
- [10] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, pages 5061–5071, 2025. 2
- [11] Shiqi Huang, Shuting He, Huaiyuan Qin, and Bihan Wen. Score: Scene context matters in open-vocabulary remote sensing instance segmentation. In *ICCV*, pages 12559–12569, 2025. 2
- [12] 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. 5
- [13] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *ECCV*, pages 299–317, 2024. 2
- [14] 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 *ICCV*, pages 4015–4026, 2023. 2
- [15] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, pages 143–160, 2024. 2, 6, 7
- [16] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclick: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, pages 70–88, 2024. 2
- [17] 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. 2
- [18] 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 *CVPR*, pages 10545–10556, 2025. 1, 2, 5, 6, 7, 8
- [19] Ke Li, Di Wang, Ting Wang, Fuyu Dong, Yiming Zhang, Luyao Zhang, Xiangyu Wang, Shaofeng Li, and Quan Wang. Rsvg-zeroov: Exploring a training-free framework for zero-shot open-vocabulary visual grounding in remote sensing images, 2025. arXiv preprint arXiv:2509.18711. 1
- [20] Zhengpeng Li, Jun Hu, Kunyang Wu, Jiawei Miao, and Jiansheng Wu. Comprehensive attribute difference attention network for remote sensing image semantic understanding. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [21] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 3
- [22] Yuanyuan Liu, Dingyuan Chen, Ailong Ma, Yanfei Zhong, Fang Fang, and Kai Xu. Multiscale u-shaped cnn building instance extraction framework with edge constraint for high-spatial-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):6106–6120, 2020. 1
- [23] 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 *ICCV*, pages 10012–10022, 2021. 6

- [24] 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. 5
- [25] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International geoscience and remote sensing symposium*, pages 3226–3229, 2017. 5
- [26] Volodymyr Mnih. *Machine learning for aerial image labeling*. PhD thesis, University of Toronto, 2013. 5
- [27] 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, 2023. arXiv preprint arXiv:2304.07193. 2
- [28] 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 *ICML*, pages 8748–8763, 2021. 5, 6, 7
- [29] 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, 2024. arXiv preprint arXiv:2408.00714. 2
- [30] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning, 2024. arXiv preprint arXiv:2402.01444. 2, 3
- [31] Abubakar Sani-Mohammed, Wei Yao, and Marco Heurich. Instance segmentation of standing dead trees in dense forest from aerial imagery using deep learning. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 6:100024, 2022. 1
- [32] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. arXiv preprint arXiv:2508.10104. 2, 6
- [33] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series, 2018. arXiv preprint arXiv:1807.01232. 5
- [34] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332, 2024. 2, 6, 7
- [35] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, 2021. arXiv preprint arXiv:2110.08733. 5
- [36] Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *CVPR*, pages 14824–14834, 2025. 2
- [37] 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. 1
- [38] Wenzhen Wang, Aoran Xiao, Wei He, Hongyuan Zhu, and Liang Xiao. Text-to-image activation for open-vocabulary semantic segmentation in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 2
- [39] Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 4
- [40] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *AAAI*, pages 5805–5813, 2024. 3
- [41] 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 *CVPRW*, pages 28–37, 2019. 5
- [42] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *ECCV*, pages 320–337, 2024. 2
- [43] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *WACV*, pages 6254–6264, 2023. 5
- [44] 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. 2
- [45] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022.
- [46] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: side adapter network for open-vocabulary semantic segmentation. *IEEE TPAMI*, 45(12):15546–15561, 2023. 2
- [47] Xiwei Xuan, Ziquan Deng, and Kwan-Liu Ma. Reme: A data-centric framework for training-free open-vocabulary segmentation. In *ICCV*, pages 20954–20965, 2025. 2
- [48] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *CVPR*, pages 29968–29978, 2025. 2, 6
- [49] Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *AAAI*, pages 9436–9444, 2025. 1, 2
- [50] Maxime Zanella, Benoît Gérin, and Ismail Ayed. Boosting vision-language models with transduction. *NeurIPS*, 37: 62223–62256, 2024. 2

- [51] Qiaolin Zeng, Jingxiang Zhou, Jinhua Tao, Liangfu Chen, Xuerui Niu, and Yumeng Zhang. Multiscale global context network for semantic segmentation of high-resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024. [1](#)
- [52] Dengke Zhang, Fagui Liu, and Quan Tang. Corclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *ICCV*, pages 24677–24687, 2025. [2](#)
- [53] Shijie Zhang, Bin Zhang, Yuntao Wu, Huabing Zhou, Junjun Jiang, and Jiayi Ma. Segclip: Multimodal visual-language and prompt learning for high-resolution remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. [2](#)
- [54] Xiaokang Zhang, Chufeng Zhou, Jianzhong Huang, and Lefei Zhang. Tpv-seg: Textually enhanced prompt tuning of vision-language models for open-vocabulary remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. [2](#)
- [55] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model, 2023. arXiv preprint [arXiv:2306.11300](#). [3](#)
- [56] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712, 2022. [2](#), [6](#), [7](#)
- [57] 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. [5](#)