Open-Vocabulary Semantic Segmentation in Remote Sensing via Hierarchical Attention Masking and Model Composition

Mohammadreza Heidarianbaei
heidarianbaei@ipi.uni-hannover.de
Mareike Dorozynski
dorozynski@ipi.uni-hannover.de
Hubert Kanyamahanga
kanyamahanga@ipi.uni-hannover.de
Max Mehltretter
mehltretter@ipi.uni-hannover.de
Franz Rottensteiner
rottensteiner@ipi.uni-hannover.de

Institute of Photogrammetry and GeoInformation
Leibniz University Hannover
Germany

Abstract

In this paper, we propose ReSeg-CLIP, a new training-free Open-Vocabulary Semantic Segmentation method for remote sensing data. To compensate the problems of vision language models such as CLIP in semantic segmentation caused by inappropriate interactions within the self-attention layers, we introduce a hierarchical scheme utilizing masks generated by SAM to constrain the interactions at multiple scales. We also present a model composition approach that averages the parameters of multiple RS-specific CLIP variants, taking advantage of a new weighting scheme that evaluates representational quality using varying text prompts. Our method achieves state-of-the-art results across three RS benchmarks without additional training.
https://github.com/aemrhb/ReSeg-CLIP.

1 Introduction

Semantic segmentation is the task of assigning a class label to each pixel in an image, e.g., representing land cover in the context of remote sensing (RS). Despite recent advancements [28, 29], existing methods face two fundamental challenges: they typically require large sets of training data to perform well, and models trained on a specific dataset often do not generalize well to other domains. Recently, vision language models (VLMs) such as CLIP [30] and ALIGN [18] have emerged as promising tools to overcome these limitations. Trained via contrastive learning to align images and text in a shared embedding space, these models exhibit strong zero-shot performance in image classification. This has motivated adaptations for Open-Vocabulary Semantic Segmentation (OVSS), resulting in models that can recognize categories beyond those seen during training. Following the adaptation of CLIP for OVSS,

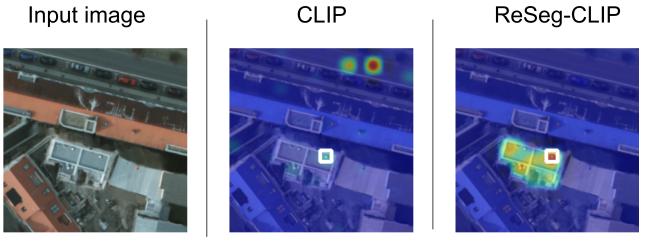

Figure 1: Example of distorted inter-patch attention for a selected patch (white squares). Left to right: input image, attention map obtained by the original CLIP vision encoder [30], and the attention map obtained by our method. Blue corresponds to low attention, red to high attention in relation to the selected patch. CLIP often assigns high attention to arbitrary patches without any relevance for the selected patch. Applying our method results in high attention concentrated on patches associated with the same object as the selected patch.

early research focused on fine-tuning to enhance pixel-level predictions [13, 43]. However, fine-tuning is often limited to smaller and less diverse datasets, typically leading to a reduction in the model's zero-shot capacity [43]. In response, training-free approaches have been proposed, mainly focusing on natural images [1, 14, 24, 32, 35, 45]. Training-based adaptations of VLMs for OVSS in RS are proposed in [2, 5, 12]; Li et al. [23], though not training CLIP itself, introduces an upsampling module that requires training. Observing this lack of completely training-free OVSS solutions in RS, we propose **ReSeg-CLIP,** a new such method for high-resolution RS imagery which is based on two main contributions:

- A **hierarchical masking strategy** to refine attention computations by imposing constraints based on hierarchical segmentation results obtained by the Segment Anything Model (SAM) [19], addressing problems of existing methods to obtain accurate pixel-wise predictions. Introducing these masks at different vision encoder stages, context is considered at different scales while mitigating the impact of unrelated patches.

- A **combination of multiple domain-adapted CLIP variants** for OVSS, to improve the generalization capabilities of existing models. For that purpose, we propose the **Prompt Variant Separation Margin**, a new metric quantifying each model's semantic representational quality by exploiting synthetic text prompts, and use it to compute model-specific weights for the averaging of the model parameters.

Refinement is needed because VLMs like CLIP align text with global image features (via the `[CLS]` token in ViTs [11]), causing attention weights to overlook semantically related regions [24]. For instance, Figure 1 shows some patches ("outlier patches" [32]; white squares) that attract disproportionately high attention from the rest of the image when using CLIP; focusing attention on such irrelevant regions causes problems for dense prediction. Consequently, several studies [1, 24, 32, 35] have tried to refine attention scores such that semantically related patches attend more strongly to one another. Although these methods boost inter-patch correlations, they still suffer from patches interacting with unrelated regions [45]. Zhang et al. [45] constrain attention to regions defined by SAM masks but only at a single scale. To address varying object sizes, we extend their approach with a hierarchical masking strategy that enables the model to capture information across multiple levels.

Our second contribution mentioned above is relevant because CLIP, pretrained on natural images, often underperforms on RS data due to a significant domain gap. Prior work [25, 36, 50] tried to solve this problem by fine-tuning CLIP on RS data. However, our own experiments based on GeoRSCLIP [50] and RemoteCLIP [25] show that these models still struggle to generalize across classes unseen during training, an essential requirement for OVSS. Inspired by model composition techniques such as [2, 8, 20, 37, 47], we thus propose to combine several CLIP models, each fine-tuned on a different RS dataset, by averaging the model parameters and applying the combined model for inference. We introduce a new metric called Prompt Variant Separation Margin (PVSM), and we use this metric for computing the weights to be used for averaging. PVSM measures the representational quality of the individual models based on the variability of the text embeddings generated for different text prompts related to the same class.

2 Related work

Vision Language Models seek to learn joint representations from both visual and textual data, thereby enabling cross-modal understanding and reasoning. CLIP [30] and ALIGN [18] accomplish this by contrastive learning. Early work on applying VLMs to RS data fine-tuned CLIP using RS5M, a RS-specific dataset [50]. Subsequent works tried to obtain better models by fine-tuning them on more curated RS datasets [25, 36]. Nevertheless, VLMs continue to perform poorly on semantic segmentation. In RS, this challenge is compounded by the limited scale and diversity of datasets, restricting zero-shot performance [25].

Open-vocabulary semantic segmentation aims to assign a class label to every pixel in an image, specifying the set of classes by textual descriptions at test time. Existing VLM-based OVSS methods fall into three categories: (1) Training-based adaptations [10, 22, 27, 40], which often generalize poorly due to limited datasets. (2) Two-stage approaches, which first generate mask proposals and subsequently apply VLMs [9, 41], are limited by pretraining on full images. (3) Training-free methods that modify the computation of self-attention. Li et al. [24] reveal inconsistencies in the relations between semantic regions formed by self-attention layers, with variants such as GEM [1], ClearCLIP [21] and SCliP [35], aiming to enhance the attention mechanism. Instead of modifying the attention mechanism, potentially introducing discrepancies between training and inference, CorrCLIP [45] uses SAM to restrict the spatial extent of patch interactions (though only at a single scale). All these works are primarily developed for natural images and do not account for the unique characteristics of RS data.

Open-vocabulary semantic segmentation for RS is addressed in [44, 49], via contrastive training for pixel-text alignment and the combination of text embeddings from the CLIP text encoder with features from an image encoder, respectively. However, such methods do not leverage existing VLMs trained on large text-image datasets. This is achieved in [42], where CLIP is combined with a specialist RS image branch in a dual-stream image encoder. Cao et al. [3] use CLIP to generate orientation-adaptive similarity maps, followed by some refinement layers. Dutta et al. [12] incorporate visual features from SAM to enhance semantic representations. However, all of these methods require training to achieve a good performance. To the best of our knowledge, SegEarth-OV [23] is the only approach for OVSS in RS which is claimed to be training-free; however, while its CLIP-based predictions need no training, its upsampling module still does. In contrast, our method is entirely training-free.

Model merging aims to combine independently trained models of identical architecture,

often employing a linear interpolation between the model parameters [17, 31, 37, 48]. Ilharco et al. [16] define task vectors as the differences between fine-tuned and pre-trained weights to capture task-specific directions and to enable model control via simple arithmetic operations; Zhang et al. [46] extend this by learning the weights for computing a linear combination. Recent studies propose compositional approaches for low-rank adaptation [15, 38], as well as sample-wise interpolation strategies [6, 26]. Typically introducing learnable parameters, these methods are not training-free; also, they neither address OVSS nor RS. In contrast, we compose RS-adapted model merging for OVSS in a training-free manner.

Discussion: Most related to our work are [23] and [45]. While in [45], SAM masks are used to limit the patch interactions at a single scale, we use SAM masks at multiple scales to consider different context regions. Li et al. [23] also address OVSS in RS, but focus on improving the spatial resolution of a prediction obtained by a CLIP variant, requiring training an upsampling module. We improve the CLIP-based predictions without any training.

3 ReSeg-CLIP

The goal of our proposed OVSS method, ReSeg-CLIP, is to map an RS input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ to a dense label map $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W}$ (see Fig. 2) in a training-free manner, i.e., we rely on pre-trained CLIP models and do not introduce training at any other stage. H and W denote the image height and width, respectively, and the number of channels is 3, because CLIP is pre-trained on RGB images. Besides proposing, to the best of our knowledge, the first entirely training-free OVSS method for RS, our main contributions concern (1) a hierarchical guidance of attention in the image encoder, aiming to achieve interactions between semantically related patches (Sec. 3.1), and (2) a new model merging approach, aiming to enhance the model's generalization capabilities (Sec. 3.2).

Before presenting $\mathbf{X}$ to the image encoder, a ViT composed of L blocks, it is partitioned into $N = \frac{H\cdot W}{P^2}$ disjoint patches $\{\mathbf{x}_i\}_{i=1}^{N}$, $\mathbf{x}_i \in \mathbb{R}^{P\times P\times 3}$ that are flattened and projected to a sequence $\mathbf{Z} = [\mathrm{CLS}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N] \in \mathbb{R}^{(N+1)\times D}$ of patch embeddings $\mathbf{z}_i \in \mathbb{R}^D$ (D: embedding dimension) and a class token CLS. $\mathbf{Z}$ is processed by the modified vision encoder (see Sec. 3.1 for details), resulting in image embeddings $\mathbf{f}_i \in \mathbb{R}^D$, where the set of all $\mathbf{f}_i$ is denoted by $\mathbf{F} \in \mathbb{R}^{(N+1)\times D}$. The text encoder, a standard Transformer [34], receives class-specific base prompts. Here, a base prompt is defined to be a text string (e.g., *"an aerial image of [c] in the city"*) for all semantic classes $c \in \{1, \ldots, C\}$ (C: total number of classes). This results in text embeddings $\mathbf{t}_c \in \mathbb{R}^D$, where the set of all $\mathbf{t}_c$ is denoted by $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_C] \in \mathbb{R}^{C\times D}$. In contrast to the standard CLIP, relying on the CLS token, the patch level predictions are obtained by computing the cosine similarity between image embeddings $\mathbf{F}$ and text embeddings $\mathbf{T}$ via

$$\mathrm{Sim}_{i,c} = \frac{\langle \mathbf{f}_i \mathbf{t}_c \rangle}{\|\mathbf{f}_i\|_2 \, \|\mathbf{t}_c\|_2} \in \mathbb{R}, \tag{1}$$

resulting in the similarity map $\mathbf{Sim}_{\mathrm{low}} \in \mathbb{R}^{H/P\times W/P\times C}$. $\mathbf{Sim}_{\mathrm{low}}$ is bilinearly upsampled to the original image resolution, resulting in $\mathbf{Sim} \in \mathbb{R}^{H\times W\times C}$, and the pixel predictions are obtained by $\hat{Y}_{(x,y)} = \arg\max_c \mathrm{Sim}_{(x,y),c}$, where $\mathrm{Sim}_{(x,y),c} \in \mathbf{Sim}$ is the score for class c at pixel (x,y).

3.1 Refining the attention map of CLIP

The main goal of the proposed refinement of CLIP's attention maps for OVSS is to enhance feature interactions among patches from relevant regions and suppress interference with ir-

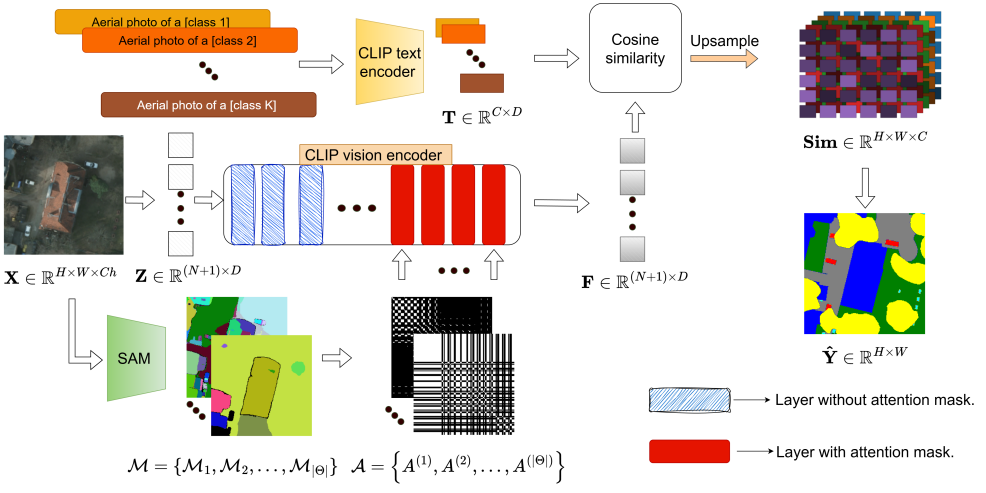

Figure 2: ReSeg-CLIP consists of CLIP-based vision and text encoders and SAM. The input image $\mathbf{X}$ is processed by both the vision encoder and SAM, resulting in features $\mathbf{F}$. SAM produces hierarchical masks $\mathcal{M}$, converted into attention masks $\mathcal{A}$ for the final vision encoder layers (red blocks). Text prompts for each class are encoded into embeddings $\mathbf{T}$, which are compared with the features $\mathbf{F}$ via cosine similarity. The results are upsampled to score map $\mathbf{Sim}$, and the segmentation $\hat{\mathbf{Y}}$ assigns to each pixel the class with highest similarity.

relevant ones. To do so, we use class-agnostic masks provided by SAM to generate attention masks that constrain feature aggregation in the CLIP vision encoder. For generalization across scenes, containing both long-range and fine-grained patterns, we propose a hierarchical masking strategy to enable multi-scale feature aggregation, by varying the SAM mask generator and producing coarse masks at earlier stages in the encoder to encourage broad attention, and fine-grained masks at later stages to emphasize detailed semantic structures.

We restrict the attention of the last $r = 1, 2, \ldots, |\Theta|$ layers (cf. Fig. 2), by utilizing SAM masks with different hyperparameter configurations $\theta_r \in \Theta$ (Θ: set of all considered configurations; cf. App. A.1). The $L_u = L - |\Theta|$ initial layers of the vision encoder are not modified. $|\Theta|$ is a hyperparameter, enabling control over the depth at which attention constraints are introduced. Each mask generator segments an input image $\mathbf{X}$ into Q_r regions, encoded by a set of binary masks $\mathcal{M}_{\theta_r} = \{M_{1_r}, M_{2_r}, \ldots, M_{Q_r}\}$, where each mask $M_{q_r} \in \{0,1\}^{H\times W}$ indicates a distinct region. These masks are combined to form a label image $\mathbf{S}_r \in \{0, 1, \ldots, Q_r\}^{H\times W}$, where each pixel (x, y) is labeled with its corresponding region index RI_r: $S_r(x, y) = q_r$ if $M_{q_r}(x, y) = 1$, and $S_r(x, y) = 0$ if the pixel does not belong to any segment (i.e., background).

The dominant RI of each patch $\mathbf{x}_i \subset \mathbf{X}$ is obtained via majority voting. Let $\mathbf{RI_r} \in \mathbb{R}^{H/P\times W/P}$ denote the patch-level RI, $a, b \in \{0, 1, \ldots, N\}$ the indices over the input sequence $\mathbf{Z}$, where index 0 corresponds to the class token and indices 1 to N to patch tokens. The mask $\mathbf{A}^{(r)} \in \{0,1\}^{(N+1)\times(N+1)}$, applied to the attention mechanism of the vision encoder at layer $l = L_u + r$, prevents high attention between patches not being in the same regions:

$$\mathbf{A}^{(r)}_{(a,b)} = \begin{cases} 1 & \text{if } (a = b = 0) \vee \left[(a > 0) \wedge (b > 0) \wedge \left(\mathbf{RI}_{r_a} = \mathbf{RI}_{r_b}\right)\right] \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where the class token is restricted to attend only to itself (case $a = b = 0$). The process is re-

peated independently for each set of hyperparameters $\theta_r \in \Theta$, each yielding a segmentation-aware attention mask. The final set of attention masks for the last $|\Theta|$ layers with restricted attention is represented as $\mathcal{A} = \{\mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(|\Theta|)}\}$. To integrate the masks into the vision encoder, we apply a large negative bias to masked logits. Specifically, for layer l we set $\overline{QK}^{(l)}_{(a,b)} = \frac{Q^{(l)}_a \cdot K^{(l)}_b}{\sqrt{D}}$ if the attention mask $\mathbf{A}^{(r)}_{(a,b)} = 1$, and $\overline{QK}^{(l)}_{(a,b)} = -\infty$ otherwise, where $Q^{(l)}_a$ and $K^{(l)}_b$ are the query and key vectors of tokens a and b, and D is the embedding dimension. This ensures that tokens only attend to others within the same SAM-derived region.

3.2 Model composition based on PVSM

To improve generalization while remaining training-free, we merge several CLIP variants by averaging their parameters with weights derived from a new metric, the *Prompt Variant Separation Margin (PVSM)*. Given a set of models $\{\text{Mod}_o\}_{o=1}^{O}$ with parameters ϕ_o, we obtain fused parameters $\phi_f = \sum_{o=1}^{O} w_o \phi_o$, where the weights w_o are based on a new metric that measures the similarity of text embeddings obtained for augmented text prompts for the same class. Images are not considered to compute this metric (and, thus, the weights) because encoding augmented images across a large dataset would be computationally too expensive.

For each class c, we define a base prompt pr_c. To introduce lexical variation, we define a set $\text{Syn}_c = \{\text{s}_{c,n_s}\}_{n_s=1}^{N_s}$ of N_s class-specific synonyms syn_{c,n_s} and sets of prefixes $\Pi = \{\pi_{n_\pi}\}_{n_\pi=1}^{N_\pi}$ and suffixes $\Sigma = \{\sigma_{n_\sigma}\}_{n_\sigma=1}^{N_\sigma}$, where each prefix π_{n_π} is a natural language phrase that precedes the synonym, and each suffix σ_{n_σ} is a phrase that follows the synonym. For every class c, we randomly generate K (K is a hyperparameter) natural language variants $v_{c,z}$, $z = 1, ..., K$ by combining a random prefix $\pi_{z_\pi} \in \Pi$, a synonym $s_{c,z_s} \in \text{Syn}_c$, and a random suffix $\sigma_{z_\sigma} \in \Sigma$ as $v_{c,z} = \pi_{z_\pi} + \text{" of "} + \text{syn}_{c,z_s} + \text{" "} + \sigma_{z_\sigma}$, $z, z_\pi, z_s, z_\sigma = 1, ..., K$. The resulting variant set of K prompts $v_{c,z}$ for class c is denoted by $\mathcal{V}_c = \{v_{c,1}, \ldots, v_{c,K}\}$. Given a pre-trained CLIP model Mod_o, each prompt variant $v_{c,z} \in \mathcal{V}_c$ is tokenized and encoded as $\mathbf{t}^o_{c,z} = \frac{\text{Mod}_o(\texttt{tokenize}(v_{c,z}))}{\|\text{Mod}_o(\texttt{tokenize}(v_{c,z}))\|_2}$, where `tokenize` refers to the CLIP preprocessing of text input for the model's encoder. Let $\mathbf{T}^o_c = \{\mathbf{t}^o_{c,1}, \ldots, \mathbf{t}^o_{c,K}\}$ be the embeddings for class c and model Mod_o. We compute the intra-class similarity $\mu^{(c),o}_{\text{intra}}$ as the average cosine similarity across all unordered pairs within $\mathbf{T}^o_c$, i.e., $\mu^{(c),o}_{\text{intra}} = \frac{2}{K(K-1)} \sum_{1 \le m' < n' \le K} \langle \mathbf{t}^o_{c,m'}, \mathbf{t}^o_{c,n'} \rangle$, and the inter-class similarity $\mu^{(c),o}_{\text{inter}}$ as the average similarity between embeddings of class c and those of all other classes $c' \neq c$, i.e., $\mu^{(c),o}_{\text{inter}} = \frac{1}{K^2(C-1)} \sum_{c' \neq c} \sum_{i'=1}^{K} \sum_{j'=1}^{K} \langle \mathbf{t}^o_{c,i'}, \mathbf{t}^o_{c',j'} \rangle$. The separation margin for class c, defined as $\delta^{(c),o} = \mu^{(c),o}_{\text{intra}} - \mu^{(c),o}_{\text{inter}}$, reflects how tightly grouped the class embeddings are and how distinct they are from other classes, indicating how well the model Mod_o has learned the underlying class concepts. We define the margin for model Mod_o as:

$$\text{PVSM}_o = \frac{1}{C} \sum_{c=1}^{C} \delta^{(c),o} \tag{3}$$

and use it to define a weight w_o of Mod_o as the normalized separation margin, thus $w_o = \text{PVSM}_o / \sum_{u'=1}^{O} \text{PVSM}_{u'}$. The parameters ϕ_f of the fused model Mod_f, are obtained via a linear combination of the individual model parameters ϕ_o: $\phi_\text{f} = \sum_{o=1}^{O} w_o \cdot \phi_o$. This formulation enables the weighted interpolation of models in parameter space, with the weights indicating how well a model is able to produce meaningful text embeddings from varying prompts.

4 Experimental setup

Datasets: We evaluate our method on the validation set of three high-resolution RS benchmark datasets: Potsdam [33] consists of orthophotos with a ground sampling distance (GSD) of 5 cm and 6 classes, UDD5 [4] consists of low-altitude oblique UAV images (4 + 1 classes) and OpenEarthMap [39] provides 0.25–0.5 m satellite/aerial images (8 + 1 classes) from various regions on earth. Compared to the datasets used for training CLIP, GeoRSCLIP and RemoteCLIP (we use their model parameters in our experiments), vehicles and roads are far finer in Potsdam, UDD5 diverges from the straight-down and medium-resolution satellite views and OpenEarthMap covers additional land-cover categories and varied sensors.

General Setup: All experiments employ the CLIP-L/14 backbone. We parametrize our model as a weighted ensemble of RemoteCLIP [25] and GeoRSCLIP [30], selecting them because they supply pretrained weights that are compatible with our used architecture. RemoteCLIP was pretrained on 828,725 image–caption pairs automatically generated from 17 public datasets spanning satellite and UAV platforms with GSDs from 5 cm to 1 m [25]. GeoRSCLIP was pretrained on RS5M, a dataset of 5 million RS image–text pairs, comprising roughly 3 million web-filtered aerial images and 2 million captioned satellite and aerial scenes from BigEarthNet, FMoW, and MillionAID [30]. The weights for Remote-CLIP and GeoRSCLIP are set to 0.37 and 0.63, determined according to Section 3.2. Input images are divided into tiles of 224×224 pixels. As these tiles may not fully capture the spatial context of all objects, we use a sliding window approach with a stride of 50 pixels and average the probabilities of a pixel across all overlapping tiles. Our vision encoder has $L = 24$ layers in total and we apply attention masking to the final $|\Theta_r| = 6$ layers. The SAM hyperparameters Θ_r controlling the generation of masks at varying levels of granularity are detailed in Appendix A.1. All hyperparameters were determined using 5% of the Potsdam training set. Details on the text prompts we have used are given in Appendices A.2-A.4.

Evaluation strategy: We evaluate our method using the mean Intersection over Union (mIoU), and compare it against established training-based OVSS frameworks for RS, i.e., SegEarth-OV [23] and the method proposed in [2]. As, to the best of our knowledge, there are no training-free OVSS RS approaches, we also compare our approach to training-free methods designed for general-purpose segmentation. These baselines are initialized with the original CLIP weights, following prior work [23]. Additionally, we compare our method against a naive CLIP-based baseline by computing cosine similarity between the image and patch tokens. We conduct ablation studies utilizing the pretrained CLIP, RemoteCLIP and GeoRSCLIP models, assessing the zero-shot semantic segmentation performance of each model individually and exploring different variants of model composition (cf. Section 3.2).

5 Results and discussion

Method performance and comparison: Comparing the results of our method and those of other RS OVSS frameworks (see Tab. 1), our method achieves an 8 percentage points (pp) higher mIOU on the Potsdam dataset compared to [2], and 7.4 pp to 8.8 pp lower mIoU compared to SegEarth-OV. This performance gap can be attributed to the use of FeatureUp. This effect is also evident in Figure 3, where the label map generated by SegEarth-OV appears more consistent and homogeneous. In contrast, our method achieves more precise spatial localization and clearer class distinction in adjacent regions (red circles), and also demonstrates

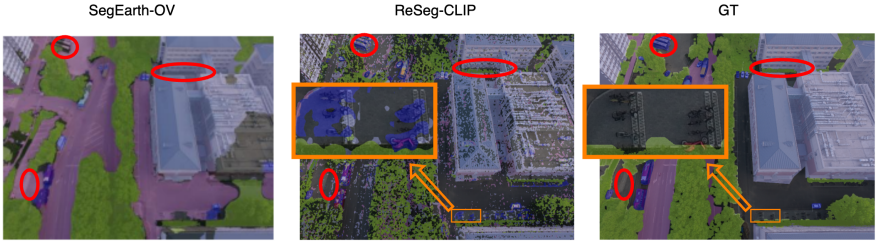

Figure 3: Results on the UDD5 dataset. SegEarth-OV yields more homogeneous masks, our method offers better class distinction in adjacent regions (red circles). Despite some interpolation-induced noise, our model effectively detects mislabeled areas (orange square).

Method	Potsdam	UDD5	OEM	Training
Cao et al. [2]	30.3	-	-	✓
Ye et al. [42]	45.7	-	-	✓
SegEarth-OV	47.1	50.6	40.3	✠
ReSeg-CLIP	38.3	43.2	32.4	✗

Table 1: Comparison of mIoU [%] across OVSS RS methods. Last column: ✓: full-network training; ✠: training of the upsampling module; ×: no training.

robustness in scenarios with mislabeled areas (orange square). While effective, FeatureUp requires to be trained, which conflicts with our training-free design and hinders a fully fair comparison. However, as FeatureUp is model-agnostic, it could be optionally integrated into our method.

As listed in Table 2, compared to a naive CLIP-based model parametrized based on [30] and to a variant of ReSeg-CLIP without using any SAM-based attention masks in the vision encoder (called C+P), our method achieves significantly higher performances across all datasets. These gains can be attributed to our proposed modifications; the refinement of attention maps and the focus on semantically relevant regions. Additionally, our approach outperforms other training-free methods, including MaskCLIP and SCLIP, across all three benchmarks. When compared to GEM, our model performs better on Potsdam and UDD5, with gains of 1.8 pp and 2.0 pp, respectively, though it lags on OEM by 1.5 pp. Relative to ClearCLIP, our method shows superior performance on UDD5 and OEM by 1.4 pp but falls short by 2.6 pp on Potsdam. This variation suggests that OVSS models, originally developed for natural image domains, can still generalize reasonably well to RS tasks, though their performance tends to be inconsistent across datasets. In contrast, our method demonstrates greater consistency, achieving best or second-best performance on all benchmarks. This underscores the effectiveness of our SAM-based hierarchical attention mechanism compared to modifications of the attention module without semantic guidance for inter-patch attention.

Analyzing the per-class IoU values given in Table 3, it can be seen that our method achieves good results of about 60% for the classes *Building* and *Vegetation*. On the other hand, the IoU values for *Vehicle* and *Background* are particularly low, which is also the case for all other methods listed. This indicates that segmenting smaller objects and obtaining a meaningful representation of such a heterogeneous class as background poses particular challenges to training-free methods in general and requires further investigations.

Dataset	CLIP	MaskCLIP	SCLIP	GEM	ClearCLIP	C+P	Ours
Potsdam	14.5	31.7	36.6	36.5	40.9	18.8	38.3
UDD5	9.5	32.4	38.7	41.2	41.8	15.0	43.2
OEM	12.0	25.1	29.3	33.9	31.0	15.3	32.4

Table 2: Comparison of mIoU [%] (best in red, second best in blue) across training-free general-purpose methods. C+P refers to ReSeg-CLIP without using any SAM-based attention masks in the vision encoder, i.e., $|\Theta| = 0$.

Parametrization	Weighting	0	1	2	3	4	5	mIoU
CLIP [30]	-	25.9	44.4	34.6	41.5	6.1	3.4	24.5
GeoRSCLIP [50]	-	34.8	54.4	50.1	51.0	14.7	5.6	33.0
RemoteCLIP [25]	-	36.5	61.4	39.1	48.1	2.8	2.5	30.4
[50] + [25]	PVSM	41.7	60.2	53.3	59.3	11.3	3.7	38.3
[50] + [25]	equal	39.1	56.5	49.9	55.6	10.6	3.5	35.9
[50] + [25] + [30]	PVSM	29.0	48.9	34.9	44.1	8.1	2.8	28.9

Table 3: Per-class IoU [%] and mean IoU (mIoU) [%] (best in red, second best in blue) on the Potsdam dataset. All variants with $|\Theta| = 6$ layers with SAM-based attention masks in the vision encoder (cf. Sec. 3.1). Classes: 0: Artificial Surface, 1: Building, 2: Natural Surface, 3: Vegetation, 4: Vehicle, 5: Background.

Ablation studies: We conducted two ablation studies. The results in Table 3 show that initializing our model with the original CLIP model weights performs poorly (24.5% mIOU), while using the weights of RemoteCLIP and GeoRSCLIP achieves a 5.9 pp and 8.5 pp higher mIOU, respectively. As CLIP was not exposed to RS data during training, this is an expected outcome. Among all pairwise combinations, merging RemoteCLIP and GeoRSCLIP yields the best results, highlighting the effectiveness of fusing complementary information for more generalizable representations. Employing the proposed PVSM strategy increases the mIoU by 2.4 pp compare to equal weighting, demonstrating the semantic expressiveness of our metric. In contrast, combining all three models performs worse than the best pairwise combination. We hypothesize that this is due to parameter oversmoothing, which may suppress critical neural activations and dilute the individual strengths of the fine-tuned models. Our second ablation study investigates the impact of varying the number of final layers in the vision encoder that restrict attention based on SAM-derived masks. The results show that increasing this number up to 6 progressively improves the mIoU up to 38.3% (cf. Tab. 4). This highlights the effectiveness of hierarchical feature aggregation guided by SAM masks. However, when the number of masked layers is increased beyond 6, the performance drops. This indicates that preserving global context in early layers while applying localized attention in later ones is optimal for the segmentation performance.

6 Conclusion

In this work, we introduce ReSeg-CLIP, a fully training-free method for OVSS of RS images. Our method tackles two challenges of VLMs: disrupted patch-level attention in dense prediction tasks and poor generalization across domains. To address these challenges, we

|■|	0	1	3	6	12	18
mIoU [%]	18.8	25.5	33.7	38.3	32.1	19.4

Table 4: Effect of number of layers in ReSeg-CLIP's vision encoder with SAM-based attention masks $|\Theta|$ on Potsdam dataset (best in red, second best in blue).

propose a hierarchical attention masking strategy that applies multi-scale SAM-generated masks at different vision encoder depths and a weight-space model composition technique using weights derived from our novel data-driven PVSM metric. Extensive experiments across three RS benchmarks demonstrate improvements in accuracy and robustness due to our hierarchical masking strategy and model composition technique, respectively. Combining both contributions, our method achieves promising results particularly for buildings and vegetation, outperforms existing training-free approaches and achieves competitive results with partially trained ones. Future work following the training-free paradigm may explore incorporating image-aware model fusion criteria, optimizing hierarchical masking for efficiency, and improving the alignment of masks with the true semantic boundaries.

References

[1] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024.

[2] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024.

[3] 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.

[4] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018.

[5] Yuxing Chen and Lorenzo Bruzzone. Toward open-world semantic segmentation of remote sensing images. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5045–5048. IEEE, 2023.

[6] Feng Cheng, Ziyang Wang, Yi-Lin Sung, Yan-Bo Lin, Mohit Bansal, and Gedas Bertasius. Dam: Dynamic adapter merging for continual video qa learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6805–6817. IEEE, 2025.

[7] Alexandra Chronopoulou, Matthew E Peters, Alexander Fraser, and Jesse Dodge. Adaptersoup: Weight averaging to improve generalization of pretrained language models. *arXiv preprint arXiv:2302.07027*, 2023.

[8] Nico Daheim, Thomas Möllenhoff, Edoardo Ponti, Iryna Gurevych, and Mohammad Emtiyaz Khan. Model merging by uncertainty-based gradient matching. In *The Twelfth International Conference on Learning Representations*, 2024.

[9] 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.

[10] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10995–11005, 2023.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Saikat Dutta, Akhil Vasim, Siddhant Gole, Hamid Rezatofighi, and Biplab Banerjee. Aeroseg: Harnessing sam for open-vocabulary segmentation in remote sensing images. *arXiv preprint arXiv:2504.09203*, 2025.

[13] 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.

[14] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5061–5071. IEEE, 2025.

[15] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic loRA composition. In *First Conference on Language Modeling*, 2024.

[16] Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023.

[17] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[18] 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.

[19] 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.

[20] Jedrzej Kozal, Jan Wasilewski, Bartosz Krawczyk, and Michał Woźniak. Continual learning with weight interpolation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4187–4195, 2024.

[21] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024.

[22] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *ArXiv*, abs/2201.03546, 2022.

[23] 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.

[24] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, page 111409, 2025.

[25] 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*, 2024.

[26] Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems*, 37:78905–78935, 2024.

[27] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023.

[28] Jinna Lv, Qi Shen, Mingzheng Lv, Yiran Li, Lei Shi, and Peiying Zhang. Deep learning-based semantic segmentation of remote sensing images: a review. *Frontiers in Ecology and Evolution*, 11:1201125, 2023.

[29] Sergiu Nedevschi et al. Semantic segmentation of remote sensing images with transformer-based u-net and guided focal-axial attention. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[30] 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.

[31] David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

[32] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156, 2024.

[33] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. Bigearthnet: A large-scale benchmark archive for remote sensing image understanding. In *IGARSS 2019-2019 IEEE international geoscience and remote sensing symposium*, pages 5901–5904. IEEE, 2019.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[35] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332, 2024.

[36] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5805–5813, 2024.

[37] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

[38] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of loRA experts. In *The Twelfth International Conference on Learning Representations*, 2024.

[39] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023.

[40] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18134–18144, 2022.

[41] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753, 2022.

[42] Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9436–9444, 2025.

[43] 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*, 36:32215–32234, 2023.

[44] Valérie Zermatten, Javiera Castillo-Navarro, Diego Marcos, and Devis Tuia. Learning transferable land cover semantics for open vocabulary interactions with remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 220:621–636, 2025.

[45] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing correlations in clip with off-the-shelf foundation models for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.10086*, 2024.

[46] Frederic Z Zhang, Paul Albert, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. Knowledge composition using task vectors with learned anisotropic scaling. *Advances in Neural Information Processing Systems*, 37:67319–67354, 2024.

[47] Jinghan Zhang, Junteng Liu, Junxian He, et al. Composing parameter-efficient modules with arithmetic operation. *Advances in Neural Information Processing Systems*, 36: 12589–12610, 2023.

[48] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019.

[49] 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.

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