ReAttnCLIP: Training-Free Open-Vocabulary Remote Sensing Image Segmentation via Re-defined Attention in CLIP

Supplementary Material

7. Datasets

OpenEarthMap consists of 2.2 million segments from 5,000 aerial and satellite images, spanning 97 regions in 44 countries across six continents, and provides manually annotated land cover labels for eight categories at a ground sampling distance (GSD) of 0.25–0.5 meters.

LoveDA dataset contains 5,987 high-resolution images (0.3 m) from three different cities, with 166,768 semantically annotated objects, and encompasses two distinct domains: urban and rural.

iSAID dataset contains 655,451 object instances across 15 categories; these are annotated in 2,806 high-resolution images.

Potsdam dataset, provided by the International Society for Photogrammetry and Remote Sensing (ISPRS), covers an extensive area of Potsdam, Germany. It consists primarily of high-resolution aerial imagery, complemented by detailed ground truth data including annotations for categories such as buildings, roads, and vegetation.

Vaihingen dataset is a benchmark dataset released by the International Society for Photogrammetry and Remote Sensing (ISPRS) for evaluating semantic segmentation and object detection algorithms. It comprises 33 image tiles of varying sizes, which were extracted from a larger top-level orthophoto.

UAVidimg dataset serves as a complementary high-resolution UAV semantic segmentation benchmark, which introduces new challenges such as large-scale variations, moving object recognition, and temporal consistency. This dataset comprises 30 video sequences of 4K high-resolution imagery captured from oblique viewpoints.

UDD5 was captured by a professional-grade UAV (DJI Phantom 4) flying at altitudes between 60 and 100 meters.

VDD comprises 400 RGB images with a resolution of 4000 × 3000 pixels. All images were captured at altitudes ranging from 50 to 120 meters.

$\mathrm{WHU}^{SAT.II}$

Massachusetts encompasses a wide variety of urban, suburban, and rural landscapes over an area exceeding 2,600 km², with a spatial resolution of 1 meter. The ground truth labels are generated by rasterizing road centerlines from the OpenStreetMap project, employing a line thickness of 7 pixels.

8. More Implementation Details

8.1. FeatUp

FeatUp is a multi-level guided upsampling module designed to progressively reconstruct high-resolution feature maps by leveraging hierarchical features from different encoder stages. Instead of relying solely on the final-layer features, FeatUp aggregates intermediate representations in a coarse-to-fine manner, enabling better preservation of spatial details and boundary structures.

Let $\mathcal{F} = \{F_1, F_2, \ldots, F_L\}$ denote the set of feature maps extracted from multiple levels of the encoder, where each $F_l \in \mathbb{R}^{C_l \times H_l \times W_l}$. These features are first projected to a unified channel dimension C using 1×1 convolutions:

$$\hat{F}_l = \mathrm{Conv}_{1\times1}(F_l), \quad \hat{F}_l \in \mathbb{R}^{C \times H_l \times W_l}$$

The reconstruction is performed in a top-down fashion. Starting from the coarsest level, the upsampled features are combined with the corresponding intermediate features through residual refinement:

$$U_l = \mathrm{Upsample}(U_{l+1}) + \phi(\hat{F}_l), \quad l = L-1, \ldots, 1$$

Here, $\mathrm{Upsample}(\cdot)$ denotes bilinear interpolation, and $\phi(\cdot)$ is a lightweight refinement block (e.g., a residual or convolutional unit). The final output feature map U_1 integrates both semantic and spatial cues, and is typically used for dense prediction tasks such as semantic segmentation.

This design allows FeatUp to effectively reconstruct fine-grained structures, offering improved boundary localization and detail preservation compared to conventional single-level upsampling methods.

8.2. SimFeatUp

To address the mismatch between training and inference stages in training-free open-vocabulary segmentation, a simplified variant of FeatUp, termed *SimFeatUp*, is introduced. The key idea is to decouple the upsampling process from the final output tokens of the vision-language model, and instead, utilize earlier-stage features with reduced semantic entanglement and improved generalization.

Let $X \in \mathbb{R}^{(1+hw)\times d}$ denote the token sequence before the final transformer block of CLIP's image encoder. SimFeatUp extracts the patch tokens $X[1 : hw+1]$, then applies a linear projection to reduce their dimensionality:

$$\mathcal{O}' = \mathrm{Proj}(X[1 : hw + 1])$$

Table 9. The prompt class name of the evaluation datasets. {} indicates multiple prompt vocabularies for one class.

Dataset	Class Name
OpenEarthMap	background, {bareland, barren}, grass, pavement, road, {tree, forest}, {water, river}, cropland, {building, roof, house}
LoveDA	background, {building, roof, house}, road, water, barren, forest, agricultural
iSAID	background, ship, store tank, baseball diamond, tennis court, basketball court, ground track field, bridge, large vehicle, small vehicle, helicopter, swimming pool, roundabout, soccer ball field, plane, harbor
Potsdam, Vaihingen	{road, parking lot}, building, low vegetation, tree, car, {clutter, background}
UAVid	background, building, road, car, tree, vegetation, human
UDD5	vegetation, building, road, vehicle, background
VDD	background, facade, road, vegetation, vehicle, roof, water
$\text{WHU}^{\text{Sat.II}}$	background, building
Massachusetts	background, road

These projected features $\mathcal{O}'$ are subsequently upsampled using a modified joint bilateral upsampling (JBU) operator, which defines two learnable kernels: a spatial kernel k_{spatial} that emphasizes proximity in image coordinates, and a range kernel k_{range} that encodes local similarity in the guidance feature space. Formally:

$$k_{\text{spatial}}(p, q) = \exp\left(-\frac{\|p - q\|^2}{2\tau_{\text{spatial}}^2}\right)$$

$$k_{\text{range}}(p, q) =$$

$$\text{softmax}_{(a,b)\in\Omega}\left(\frac{1}{\tau_{\text{range}}^2} \cdot \text{MLP}(G[i, j]) \cdot \text{MLP}(G[a, b])\right)$$

Here, G denotes the guidance feature extracted from a high-resolution RGB image, Ω is a local window centered at position (i, j), and $\tau_{\text{spatial}}, \tau_{\text{range}}$ are learnable temperature parameters.

Unlike the original FeatUp, which operates on semantically rich but inference-sensitive features, SimFeatUp leverages features with less cross-token entanglement. Furthermore, to accommodate the scale diversity in remote sensing imagery—ranging from fine-grained textures to large-scale structures—the kernel window size is enlarged (e.g., 11×11) to provide a broader context.

This design ensures that SimFeatUp offers an effective and lightweight upsampling strategy, enhancing segmentation fidelity while avoiding overfitting to training-specific feature behaviors.

9. Effectiveness of the Rotation Operation

To evaluate the effectiveness of the rotation operation, we visualize its impact in Figure 7. As illustrated, incorporating rotation leads to a noticeable improvement in the preservation of structural details. Specifically, the rotation operation enhances the model's ability to capture spatial variations, thereby recovering fine-grained information that is often lost in standard feature extraction pipelines. This suggests that rotation introduces beneficial diversity to the receptive fields, enabling the network to better model complex geometries and subtle object boundaries.

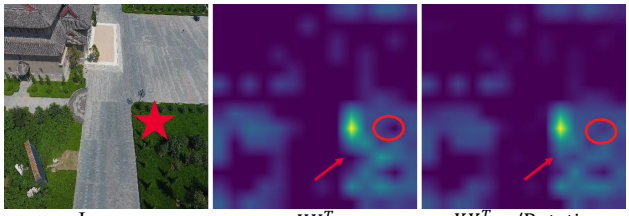

Figure 7. The effect of applying rotation is visualized for better understanding.

10. Effect of Attention Layer Selection on Global Bias Removal

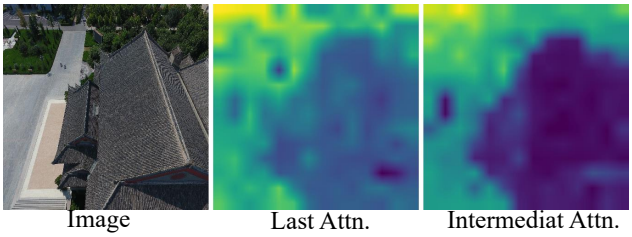

Figure 8. Visualization of the Impact of Attention Maps from Different Layers on Global Bias

To investigate the influence of different attention layers in removing global bias, we visualize the effect of using attention maps from various Transformer layers, as shown in Figure 8. The visualization demonstrates that attention maps from intermediate layers yield more localized and semantically diverse feature representations, effectively suppressing the global bias introduced by the *[CLS]* token.

Table 10. Different backbones.

Backbone	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVidimg	UDD5	VDD	Average
RemoteCLIP	19.8	22.4	19.5	14.5	23.4	23.7	33.3	31.3	23.5
RSCLIP	28.4	31.0	21.5	35.7	23.4	34.0	35.7	37.7	30.9
Ours	**41.1**	**37.0**	**23.2**	**48.7**	**29.9**	**44.0**	**53.7**	**49.7**	**40.9**

Table 11. More results.

Methods	DLRSD	WHDLD	Average
SegEarth-OV	18.9	20.8	19.9
Ours	**20.5**	**22.5**	**21.5**

In contrast, attention maps from the final layer tend to be overly dominated by high-level semantics, leading to less discriminative spatial features. This comparison highlights the importance of layer selection when leveraging attention for bias reduction.

11. Different Backbones.

We replace CLIP with RSCLIP and RemoteCLIP, with results reported in Table 10. As can be observed from the table, models fine-tuned on remote sensing data still underperform the CLIP model.As these models are fine-tuned on a limited number of remote sensing datasets, they are more susceptible to overfitting, which may limit their generalization performance on downstream tasks.

12. Additional Results.

To further validate the generalization ability of our method, we additionally evaluate it on two remote sensing datasets, **DLRSD** and **WHDLD**. As shown in Table 11, our approach consistently outperforms the baseline SegEarth-OV on both datasets, achieving improvements of 1.6 and 1.7 on DLRSD and WHDLD, respectively. These results demonstrate that our method maintains strong performance across different datasets and further confirms its robustness.