SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensing Images

Kaiyu Li[1], Ruixun Liu[2], Xiangyong Cao[2,4†], Xueru Bai[6], Feng Zhou[6], Deyu Meng[3,4,5], Zhi Wang[1]
[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, 710049, China
[2]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China
[3]School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China
[4]Ministry of Education Key Laboratory of Intelligent Networks and Network Security,
Xi'an Jiaotong University, Xi'an, 710049, China [5]Pengcheng Laboratory [6]Xidian University

Figure 1. Visualization and performance of SegEarth-OV on open-vocabulary semantic segmentation of remote sensing images. We evaluate our SegEarth-OV on 17 remote sensing datasets (including semantic segmentation, building extraction, road extraction, and flood detection tasks), and our SegEarth-OV consistently generates high-quality segmentation masks.

Abstract

Current remote sensing semantic segmentation methods are mostly built on the close-set assumption, meaning that the model can only recognize pre-defined categories that exist in the training set. However, in practical Earth observation, there are countless new categories, and manual annotation is impractical. To address this challenge, we first attempt to introduce training-free[1] open-vocabulary semantic segmentation (OVSS) into the remote sensing context. However, due to the sensitivity of remote sensing images to low-resolution features, distorted target shapes and ill-fitting boundaries are exhibited in the prediction mask. To tackle these issues, we propose a simple and universal upsampler, i.e. SimFeatUp, to restore lost spatial information of deep features. Specifically, SimFeatUp only needs to learn from a few unlabeled images, and can upsample arbitrary remote sensing image features. Furthermore, based on the observation of the abnormal response of patch tokens to the `[CLS]` *token in CLIP, we propose to execute a simple subtraction operation to alleviate the global bias in patch tokens. Extensive experiments are conducted on 17 remote sensing datasets of 4 tasks, including semantic segmentation, building extraction, road detection, and flood detection. Our method achieves an average of 5.8%, 8.2%, 4.0%, and 15.3% improvement over state-of-the-art methods on the 4 tasks. Code is available at* `https://likyoo.github.io/SegEarth-OV`.

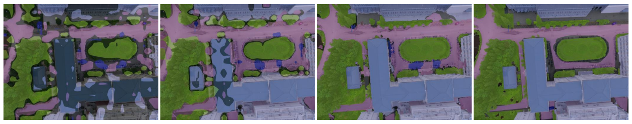
1. Introduction

Remote sensing images have changed the way humans observe and understand the Earth. It enables us to monitor land cover/use types, respond effectively to natural disasters (e.g., fires, earthquakes, floods), gain insight into food and water resources, etc. Among the 17 Sustainable Development Goals (SDGs)[2] issued by the United Nations, remote sensing images can provide important data support for several goals including "Zero Hunger", "Clean Water

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

[1]Strictly "annotation-free". We adopt the term "training-free" to emphasize the ability of trained SimFeatUp to generalize across datasets.

[2]https://sdgs.un.org



(a) MaskCLIP (b) ClearCLIP (c) SegEarth-OV (d) GT

Figure 2. Limitations of state-of-the-art OVSS methods in remote sensing images. The two predictions on the left present distorted target shapes and ill-fitting boundaries (best viewed digitally with zoom, especially for the object edges).

and Sanitation", "Industry, Innovation and Infrastructure", "Climate Action", "Life on Land", etc. [60]. Notably, remote sensing data can be considered as a distinct modality in computer vision. It involves more diverse spatial resolutions (from centimeters to kilometers), temporal dimensions (from hours to decades), and object perspectives (overhead and oriented) than natural images. Therefore, solutions designed for other data modalities (e.g., natural images) may be sub-optimal for remote sensing data [53].

Recently, raw remote sensing images have been available from various sources (e.g., Landsat, Sentinel), but obtaining large-scale annotated categories is still a challenge due to countless potential categories and expensive manual costs. In addition, on the broad surface of the earth, "stuff" (e.g., grassland, forest, etc.) occupies much more area than "things" (e.g. buildings, ships, etc.) [74]. Therefore, for remote sensing images, pixel-level perception, i.e., segmentation, is applied more frequently than instance-level perception, and the demand for pixel-level annotation exacerbates the difficulty of obtaining large-scale labels. OpenStreetMap [21] is a popular solution that aims to create a freely usable, editable, and shareable map of the world. However, the completeness of the annotations in OpenStreetMap is affected by regional income levels, resulting in limited data availability [22]. The rise of vision language model (VLM) brings us new inspirations with its capabilities of open-vocabulary semantic segmentation (OVSS). However, through some exploratory experiments, we find that the solution designed for natural images is sub-optimal on remote sensing images. A notable phenomenon is the presence of distorted target shapes and ill-fitting boundaries in the prediction mask, as shown in Fig. 2 (a)(b).

Empirically, these issues can be largely attributed to the excessively low resolution of the features [19, 76]. In the current CLIP-based OVSS paradigm, the feature maps from CLIP [51] are downsampled to $1/16th$ of the original image (ViT-B/16). Hence, in this paper, we propose a simple and general feature upsampler, i.e. SimFeatUp, which is trained with the goal of reconstructing content-invariant high-resolution (HR) features on a few unlabeled images and can upsample arbitrary remote sensing image features after training. Owning to this property of SimFeatUp, it can be used as a universal external unit for training-free OVSS framework. Further, CLIP is trained at the image level, it uses the `[CLS]` token as a representation of the entire image, and attaches global properties to the local token [48, 52, 62]. However, this global property biases local features against patch-level inference in OVSS. To effectively reduce global bias, we propose a simple subtraction operation of local patch features and global features. Extensive quantitative and qualitative experiments demonstrate the superior segmentation quality of our method over other state-of-the-art methods.

In summary, our contributions are as follows.

- We propose SimFeatUp, a general feature upsampler for training-free OVSS, which robustly upsamples low-resolution (LR) features and maintains semantic consistency with image content.
- We propose an extremely simple and straightforward way to alleviate the global bias problem of CLIP, i.e., executing subtraction operations of local and global tokens.
- Our proposed SegEarth-OV model can achieve state-of-the-art performance on 17 remote sensing datasets of 4 tasks, including semantic segmentation, building extraction, road extraction, and flood detection.

2. Related Work

Vision-Language Model. Recently, foundation models, especially visual language models, have energized the field of computer vision. One phenomenal advance is contrastive language-vision pretraining, i.e., CLIP [51], which elegantly bridges the gap between images and natural language. By training with massive data in a multimodal embedding space, CLIP gains strong transfer capabilities, achieving leaps in zero-shot learning and making OV learning possible [67]. Subsequently, related research has gradually emerged, from the data [10, 58, 70, 73], training [16, 35, 73] or model [33, 34] side. However, CLIP focuses only on global `[CLS]` tokens, and even though patch-level tokens can be generated, they are inevitably contaminated by global bias [48, 52, 62], which is detrimental to dense prediction. In addition, several remote sensing VLMs emerge, they adapt general VLMs to remote sensing contexts [37, 49, 65, 75] or mine the characteristics of remote sensing data [23, 50].

Supervised semantic segmentation. Semantic segmentation aims to discriminate images at the pixel level. The prediction head (aka decoder), as an essential component of segmentation models, is able to upsample LR feature maps into HR predictions. Typical prediction heads contain upsampling operators (e.g., bilinear interpolation, JBU [29]) and HR encoder features (as guidance), e.g., UNet [55], UperNet [69], Semantic FPN [27], MaskFormer [9], etc. Some works [38, 41, 78] focus on dynamic, learnable upsampling operators that make this process content-aware. FeatUp [17] constructs a model-agnostic upsampling scheme that uses multi-view consistency loss with deep analogies to NeRFs [45]. **However, it only explores the condition with labels.** Inspired by FeatUp and built on top of it, the SimFeatUp proposed in this work is able to significantly improve OVSS without any labels.

Open-Vocabulary Semantic Segmentation. As VLMs have shown remarkable zero-shot inference in image classification [51], which naturally extends to semantic segmentation. They empower the segmentation pipeline to recognize seen and unseen categories, and users can segment almost any category in an image using prompt vocabulary [67, 79]. We divide current CLIP-based OVSS methods into two groups: training-required and training-free. The former allows models to be trained on some base classes in a supervised or weakly supervised manner. Typically, some works [18, 42, 48, 52] try to train a localization-aware CLIP which can naturally make dense predictions, while others [11, 15, 32, 39, 71, 72] select a subset of the CLIP's pre-trained parameters and/or introduce a limited number of trainable parameters into the frozen CLIP, i.e., fine-tuning the CLIP to adapt to dense prediction on base classes. Still, training-free OVSS methods emphasize tapping into CLIP's inherent localization capabilities with limited surgery of features or structures. MaskCLIP [77] pioneers the removal of query and key projections at the attention pooling layer of CLIP's image encoder. Following it, subsequent studies [3, 30, 36, 62] adequately explore self-self attention (i.e., *q*-*q*, *k*-*k* or *v*-*v* self-attention), and these modifications somewhat mitigate noisy activations and spatial invariant perception of CLIP. Another stream [2, 25, 56, 59] is the two-stage method, which first generates category-agnostic mask proposals and then classifies the masks. Besides, some other foundation models (e.g. SAM [28], Stable Diffusion [54]) can be introduced to enhance the localization ability of CLIP, and these explorations also make sense [2, 31, 64].

Different from previous methods, we focus on the inherent characteristics of remote sensing images rather than the general attributes of natural images. The only contemporaneous work is [6], but it is training-required, like [11, 72]. Our SimFeatUp component, although it needs to be trained on a few images-only data beforehand, this process is independent of the semantic segmentation process and the trained weights can be used for almost any remote sensing data (like the foundation model in other works [2, 31]), so our method can still be seen as a training-free method.

3. Preliminaries

3.1. CLIP

In ViT-based CLIP, the image encoder consists of a series of Transformer blocks. Let $X = [x_{cls}, x_1, ..., x_{h\times w}]^\mathsf{T} \in \mathbb{R}^{(hw+1,d)}$ denotes the input of the last block, where h and w denote the height and width of the feature map, d denotes the dimension of tokens, and x_{cls} is a learnable global token and the others are local tokens from different image patches. The forward process of the last Transformer block can be formulated as follows:

$$\begin{aligned} \boldsymbol{q} = \mathrm{Emb}_q(X), \boldsymbol{k} = \mathrm{Emb}_k(X), \boldsymbol{v} = \mathrm{Emb}_v(X), \\ \boldsymbol{y} = X + \mathrm{SA}\left(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}\right), \\ \boldsymbol{z} = \boldsymbol{y} + \mathrm{FFN}(\mathrm{LN}(\boldsymbol{y})), \end{aligned} \tag{1}$$

where $\boldsymbol{q}$, $\boldsymbol{k}$, $\boldsymbol{v}$ denote Query, Key, and Value, respectively. Emb consists of a layer normalization (LN) layer and a linear layer, and FFN denotes the feed-forward network. SA denotes a standard self-attention module, i.e., $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}) = \mathrm{softmax}(\frac{\boldsymbol{q}\cdot\boldsymbol{k}^\mathsf{T}}{\sqrt{d}})\cdot\boldsymbol{v}$. Then, a projection layer maps $\boldsymbol{z}$ to a multi-modal embedding space:

$$\mathcal{O} = \mathrm{Proj}(\boldsymbol{z}), \tag{2}$$

where $\mathcal{O} = [o_{cls}, o_1, ..., o_{h\times w}]^\mathsf{T} \in \mathbb{R}^{(hw+1,c)}$ denotes the output of the image encoder, c denotes token dimension after the projection layer, and $c < d$. During CLIP training, o_{cls} is used for image-level learning; while during OVSS inference, $\mathcal{O}[1 : hw+1]$ is used for patch-level prediction.

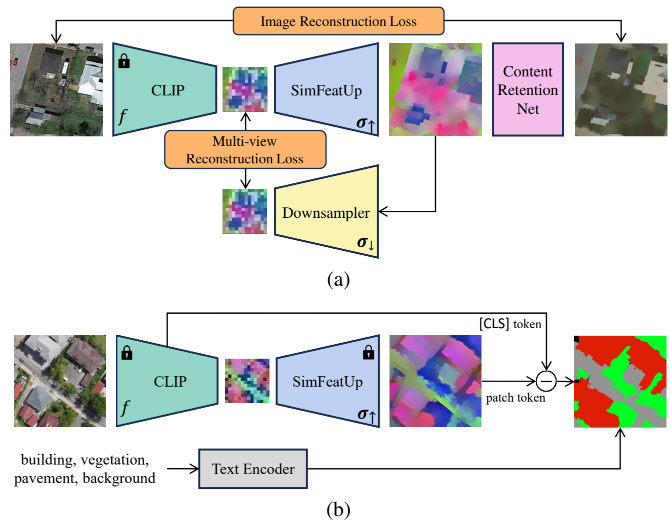


Figure 3. Illustration of the proposed method. (a) is the training process of SimFeatUp. CLIP is frozen and only SimFeatUp is useful in reasoning. (b) is the reasoning process of SegEarth-OV. The LR feature maps from CLIP are upsampled by SimFeatUp and then the `[CLS]` token is subtracted to alleviate global bias. For better presentation, the color renderings follow [17].

3.2. FeatUp

FeatUp [17] aims to train a model-agnostic upsampler. It executes an upsampling operation on LR features $\mathcal{O}[1 : hw+1]$ from a frozen backbone network via a learnable upsampler $\sigma_\uparrow$, and then reconstructs the LR features using a learnable downsampler $\sigma_\downarrow$. Its critical insights can

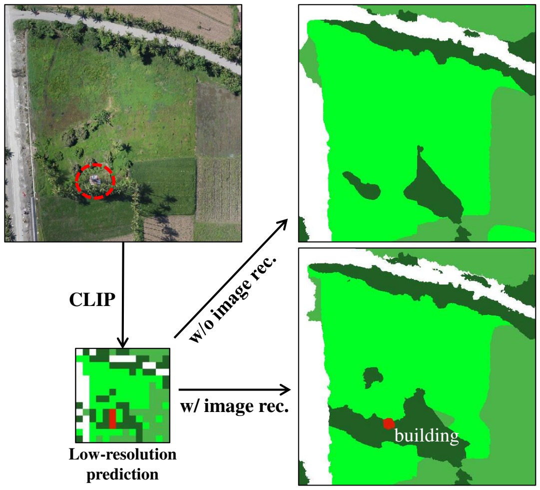


Figure 4. Comparison of with and without image reconstruction loss (Eq. (4)). the LR prediction is obtained directly using the output of CLIP (without bilinear interpolation). Color: building, tree, cropland, grass.

be briefly summarized by the following loss function:

$$\mathcal{L}_{rec} = \|\mathcal{O}[1:hw+1] - \sigma_{\downarrow}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2. \quad (3)$$

FeatUp instantiates $\sigma_{\uparrow}$ as stacked parameterized JBU operators [29]. The upsampled HR feature element is estimated by weighting the neighboring elements of the LR feature. For weight generation, JBU considers two factors, the similarity and distance between neighboring elements and the center element in the guidance feature, corresponding to kernel k_{range} and $k_{spatial}$. For brevity, we omit the multi-view consistency constraint in [17].

4. Method

In the following, we introduce SegEarth-OV by first describing SimFeatUp's training, design and explaining why it is suitable for OVSS. Second, we discuss the impact of global token on dense prediction and present our method for alleviating global bias.

4.1. SimFeatUp

FeatUp provides us with an excellent training paradigm for general upsamplers. However, it lacks some considerations for the training-free setting, resulting in sub-optimal for the OVSS task, especially in remote sensing contexts.

Image content retention. As described in Sec. 3.2, the goal of FeatUp is to minimize the original LR features and the LR features after the up-down-sampling (i.e. $\sigma_{\downarrow}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))$). Since both $\sigma_{\uparrow}$ and $\sigma_{\downarrow}$ are learnable, with such a weak constraint, the up-down-sampling process becomes a black box, and there is no guarantee that the intermediate HR features are complete and consistent with the original image in content. A direct example is shown in Fig. 4, where a small building in the original image is present in the LR prediction but disappears in the HR prediction (top right). To solve this issue, we introduce an additional image reconstruction loss to constrain the HR features:

$$\mathcal{L}_{img} = \|I - \mathrm{CRN}(\sigma_{\uparrow}(\mathcal{O}[1:hw+1]))\|_2^2, \quad (4)$$

where I denotes the input image, CRN denotes a content retention net. CRN is a very lightweight network that receives HR features as input and reconstructs the original image. Specifically, CRN consists of two 2D convolutional layers with normalization and a *Tanh* activation layer, where the *Tanh* layer is designed to constrain the output to [-1, 1], cf. VAEs [26]. Finally, the loss for training SimFeatUp consists of $\mathcal{L}_{rec}$ and $\mathcal{L}_{img}$ with a weight γ, i.e.,

$$\mathcal{L} = \mathcal{L}_{rec} + \gamma\mathcal{L}_{img}. \quad (5)$$

Which feature to upsample? FeatUp takes the final output of CLIP, i.e., $\mathcal{O}[1:hw+1]$ in Eq. (2), as input to the upsampler. This can work well in training-based settings, e.g., linear probe [1]. However, in training-free OVSS, as described in Sec. 2, vanilla self-attention leads to inferior performance. Therefore, the current OVSS method modulates it to self-self attention, and this law also works in remote sensing images. Under this premise, the SA in Eq. (1) would be replaced by other modules, and direct upsampling of $\mathcal{O}[1:hw+1]$ would lead to the mismatch between training and inference. Motivated by this, we propose to upsample CLIP features at an earlier layer. Specifically, we select the input of the last Transformer block of the CLIP's image encoder, i.e., $X[1:hw+1]$ in Eq. (1). Furthermore, the high dimension of tokens in X leads to a high-cost upsampler. Therefore, we retain the projection layer. Ultimately, the features $\mathcal{O}'$ which need upsampling are formulated as:

$$\mathcal{O}' = \mathrm{Proj}(X[1:hw+1]). \quad (6)$$

Larger upsampling kernel. We follow the upsampling operator in FeatUp, i.e., the parameterized JBU. As mentioned in Sec. 3.2, the upsampling kernels k_{range} and $k_{spatial}$ of the JBU are computed from the elements within a window in the guidance feature. The generation of k_{range} and $k_{spatial}$ can be formulated as follows:

$$k_{spatial}(p,q) = \exp\left(\frac{-\|p-q\|_2^2}{2\tau_{spatial}^2}\right), \quad (7)$$

$$k_{range}(p,q) = \mathrm{softmax}_{(a,b)\in\Omega}\left(\frac{1}{\tau_{range}^2} MLP(G[i,j]) \cdot MLP(G[a,b])\right), \quad (8)$$

where (p,q) denotes the position in the kernel. Ω denotes a window centered at (i,j) in the guidance feature G, which

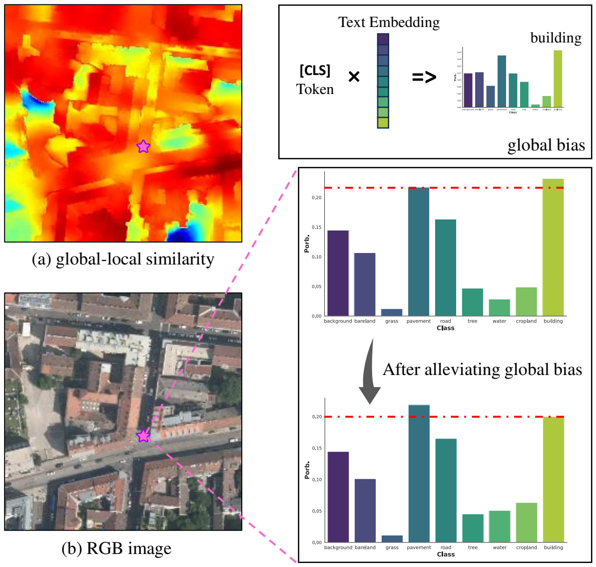


Figure 5. Comparison of before and after alleviating the global bias. (a) is the similarity map of patch tokens and cls tokens, some "non-building" regions also present high response, (b) is the original RGB image. Note that the right-hand histograms stretch the raw values for better presentation.

is extracted from HR RGB image. $\tau_{spatial}$ and τ_{range} are learnable factors. In remote sensing images, unlike natural images, the size of the target presents a logarithmic scale spanning from the meter scale (e.g., trees, gardens) to the kilometer scale (e.g., forests, rangelands) [53]. Therefore, we set larger upsampling kernels to obtain a wider receptive field. Here, we expand the window size to 11×11, compared to 7×7 in FeatUp. A possible concern is that a larger receptive field may introduce more irrelevant context, but with $k_{spatial}$, more distant points consistently contribute lower weights, which makes it more reasonable to use larger upsampling kernels.

Simplify. On the structural side, we simplify the components in FeatUp. In FeatUp, the parameterized JBU modules are stacked 4 times for $16\times$ upsampling, and the parameters of each JBU module are independent. Although we fed HR features into the CRN to ensure the integrity of its content, the behavior of each JBU module is indeterminable. Therefore, in SimFeatUp, we change "JBU_Stack" to "JBU_One", i.e., only one parameterized JBU is used for upsampling. If $16\times$ upsampling is required, then it only needs to repeat the execution 4 times. Further, "JBU_One" significantly reduces the number of trainable parameters in the upsampler and provides the possibility of upsampling arbitrary multiples.

4.2. Alleviating global bias

As described in Sec. 3.1, in the training phase of CLIP, the `[CLS]` token, which contains the global information of the whole image, is optimized with the text embedding in the multi-modal space via contrastive learning. However, in the inference phase of OVSS, `[CLS]` token is generally discarded and only patch tokens are used for similarity computation with the prompt vocabulary. This means that there is a gap between training and inference. Indeed, previous work [48, 52, 62] also demonstrates that: each local visual token in CLIP focuses on a wide range of positions, and the attention maps typically share similar patterns. This suggests that the global attribute is attached to the patch tokens in CLIP. This property is generally not a concern in classification task, but it significantly impairs performance in dense prediction.

The visualization in Fig. 5 demonstrates the above elaboration. We extract the `[CLS]` token using CLIP for the RGB image in Fig. 5(b), and compute its similarity with the candidate text embeddings. The image is recognized as the building, which is reasonable because the building covers the maximum range in the image. Then, we calculate the similarity of the `[CLS]` token with patch tokens as shown in Fig. 5(a). The highly responsive regions are not only the regions with buildings, some roads and pavements are also activated, which indicates that the global bias contaminates the local patch tokens. Motivated by this observation, we propose to "subtract" some global bias from the patch token. This solution is very straightforward and simple, it can be formulated as follows.

$$\hat{\mathcal{O}} = \mathcal{O}[1 : hw + 1] - \lambda\mathcal{O}[0], \tag{9}$$

where λ denotes an intensity factor. $\mathcal{O}[0]$ is broadcast to the same dimension as $\mathcal{O}[1 : hw + 1]$.

5. Experiments

5.1. Dataset

In remote sensing application contexts, not only multi-class semantic segmentation but also extraction of certain land cover types (e.g., buildings, roads, water bodies) is required, e.g., Google's Open Buildings project[3]. Therefore, we select 17 typical datasets covering common semantic segmentation, building extraction, road extraction, and water body segmentation (flood detection) tasks.

Semantic segmentation. We evaluate SegEarth-OV on 8 remote sensing semantic segmentation datasets including OpenEarthMap [68], LoveDA [63], iSAID [66], Potsdam, Vaihingen[4], UAVid [43], UDD5 [8] and VDD [5]. Among them, the first 5 datasets consist of mainly satellite images and the last 3 consist of UAV images. They contain custom foreground classes and a background class. Detailed descriptions of these datasets can be found in Appendix 7.1.

[3] https://sites.research.google/gr/open-buildings

[4] https://www.isprs.org/education/benchmarks/UrbanSemLab

Table 1. Open-vocabulary semantic segmentation quantitative comparison on remote sensing datasets. Evaluation metric: mIoU. **Best** and **second best** performances are highlighted.

Methods		OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid^{img}	UDD5	VDD	Average
CLIP [51]	ICML'21	12.0	12.4	7.5	14.5	10.3	10.9	9.5	14.2	11.4
MaskCLIP [77]	ECCV'22	25.1	27.8	14.5	31.7	24.7	28.6	32.4	32.9	27.2
SCLIP [62]	arXiv'23	29.3	30.4	16.1	36.6	**28.4**	31.4	38.7	37.9	31.1
GEM [3]	CVPR'24	33.9	31.6	17.7	36.5	24.7	33.4	41.2	**39.5**	32.3
ClearCLIP [30]	ECCV'24	**31.0**	**32.4**	**18.2**	**40.9**	27.3	**36.2**	**41.8**	39.3	**33.4**
SegEarth-OV	Ours	**40.3**	**36.9**	**21.7**	**47.1**	**29.1**	**42.5**	**50.6**	**45.3**	**39.2**

Single-class extraction. We select 4 building extraction datasets (i.e., WHU^{Aerial} [24], WHU$^{Sat.\text{II}}$ [24], Inria [44], and xBD [20]), 4 road extraction datasets (i.e., CHN6-CUG [80], DeepGlobe [14], Massachusetts [46], and SpaceNet [61]), and 1 flood detection dataset (i.e., WBS-SI[5]) for the evaluation of single-class extraction. These datasets contain 1 foreground class (building, road or flood) and 1 background class. Detailed descriptions are in Appendix 7.2-7.4.
Training dataset for SimFeatUp. SimFeatUp requires only image data for training, moreover, to avoid unfair comparisons, we use a public remote sensing image classification dataset, Million-AID [40], which is mainly collected from Google Earth. We randomly selected only 16k of these images to train SimFeatUp.

5.2. Setup

Implementation. Our implementations are based on MMSegmentation [12] toolkit. If not specified, we use the original pretrained weights of CLIP (ViT-B/16) provided by OpenAI. For the text part, we use the `OpenAI ImageNet template` as input for the text encoder, e.g., "a photo of a {class name}". In addition, since the definition of certain classes may vary in some datasets, we use slight class rename tricks for all methods. For example, we rename "clutter" to "background" and "building" to {"building", "house", "roof"}, and the sub-class with the highest probability in {} will be the probability of that class. Detailed prompt class names for all datasets are listed in Appendix Tab. 7. For the image part, we resize input images with a long side of 448 and perform slide inference with a 224 $\times$ 224 window and 112 stride. For SimFeatUp training, we randomly crop 224 $\times$ 224 image patches on the original image. We use two 4090 GPUs to train 1 epoch with batch size set to 8. We retain the multi-view consistency constraint in FeatUp, and random flipping, translation and zoom are applied. For the hyper-parameters mentioned, the value of γ is set to 0.1 and λ is set to 0.3 for all datasets.
Evaluation. We evaluate the semantic segmentation using the mean intersection over union (mIoU) metric. For single-class extraction, the IoU of the foreground class is used.

[5]https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images

Baseline. We take some lessons from natural image OVSS, which are also suitable for remote sensing scenes: we remove the FFN and residual connection of the last Transformer block, insights from [36] and [30]. In addition, the last self-attention is replaced by our modulated attention, i.e., the summation of q-q, k-k and v-v as the weights of v:

$$\text{M-SA}(\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}) = \sum_{\boldsymbol{i}\in\{\boldsymbol{q},\boldsymbol{k},\boldsymbol{v}\}} \text{softmax}(\frac{\boldsymbol{i}\cdot\boldsymbol{i}^{\top}}{\sqrt{d}})\cdot\boldsymbol{v}. \tag{10}$$

5.3. Comparison to State-of-the-art

Since the proposed SegEarth-OV is a training-free method and there is no previous OVSS method designed for remote sensing images, we select 5 state-of-the-art training-free OVSS models of natural images for comparison, including vanilla CLIP [51], MaskCLIP [77], SCLIP [62], GEM [3] and ClearCLIP [30].
Semantic segmentation. As listed in Tab. 1, SegEarth-OV achieves the best performance on all 8 semantic segmentation datasets. SegEarth-OV achieves more than 40% mIoU on 5 datasets and more than 50% on the UDD5 dataset, which implies that the OVSS method is feasible in remote sensing scenarios. Compared to the previous method, SegEarth-OV achieves a performance gain of more than 5% on 5 datasets and an average gain of 5.8% on 8 datasets. On the iSAID dataset, the mIoU of SegEarth-OV is only 21.7%, which is due to the fine-grained category delineation in this dataset, which covers 16 categories (see Appendix Tab. 7).
Single-class extraction. In the building extraction task, the increase delivered by SegEarth-OV is more significant, as listed in Tab. 2. Considering that the "building" class occupies a small area (see Appendix Fig. 1), we evaluate the setup for larger scale images, i.e., resizing the long side of the input image to 896 $\times$ 896. This setup significantly improves the IoU of Inria and xBD, which on the other hand supports our view that spatial detail preservation is essential for remote sensing OVSS. In the road extraction task, although SegEarth-OV achieves the best IoU, overall, the performance of all methods on the 4 datasets is unsatisfactory, with a best IoU of only 35.4%. There may be two reasons for this phenomenon: (1) The special shape of the road makes it difficult to be extracted in a training-free OVSS manner; (2) The labels of some data are generated based on

Figure 6. Qualitative comparison between different training-free OVSS methods on OpenEarthMap [68], UDD5 [8] and WHU^{Aerial} [24] datasets (best viewed digitally with zoom, especially for the edges of the object).

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

Method	Building Extraction				Road Extraction				Flood Detection
	WHU^{Aerial}	WHU$^{Sat.II}$	Inria	xBD^{pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
448 × 448:									
CLIP [51]	17.7	3.5	19.6	16.0	7.7	3.9	4.9	7.1	18.6
MaskCLIP [77]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8
SCLIP [62]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1
GEM [3]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5
ClearCLIP [30]	36.6	**20.8**	39.0	30.1	25.5	5.7	6.4	16.3	44.9
SegEarth-OV	**49.2**	**28.4**	**44.6**	**37.0**	**35.4**	**17.8**	**11.5**	**23.8**	**60.2**
896 × 896:									
SegEarth-OV	**49.9**	-	**48.9**	**43.1**	32.8	**20.1**	**17.2**	**29.1**	**57.9**

OpenStreetMap vector shapes with fixed widths attached, which are inherently imprecise. Again, the extraction of roads can generally benefit from larger size images. For the flood detection task, where "water" class features can be easily recognized, the IoU of SegEarth-OV is improved by 15.3% over the previous best method, up to 60.2%. Due to the small size of the original images in the WBS-SI dataset, resizing to a larger size does not result in a positive gain.

Qualitative results. We present qualitative results for MaskCLIP, ClearCLIP, and SegEarth-OV in Fig. 6. Some observations are summarized as follows: (1) There are some incorrect category predictions in MaskCLIP, e.g., water on the road and bareland on the cropland. (2) ClearCLIP can generate correct category predictions, but lacks precise localization capability, with distorted target shapes and ill-fitting boundaries of the prediction mask. (3) SegEarth-OV is capable of generating more fine-grained masks that fit the target edges and maintain correct category discrimination. More visualizations can be found in Appendix Fig. 8-10.

5.4. Ablation Study and Analysis

Plug and play. Two key insights of this work, SimFeatUp and global bias alleviation, which can be attached to other OVSS methods as plug-and-play modules. As listed in Tab. 4, a revealing observation is that on both the OpenEarthMap and WHU^{Aerial} datasets, as the base capability of the model improves (from MaskCLIP to ClearCLIP), the increases delivered by our method also improve (**↑3.3**, **↑5.1**, **↑8.1** on OpenEarthMap, **↑5.6**, **↑6.1**, **↑14.5** on WHU^{Aerial}). This suggests that our method has the potential to improve localization and discrimination for stronger models.

Ablation study. To assess each of the proposed components, we perform a detailed ablation analysis, as listed in Tab. 5. FeatUp (CLIP) denotes the original FeatUp upsampler, which provides a 1.5% improvement on OpenEarthMap but decreases the performance on WHU$^{Sat.II}$ and WBS-SI (more comparisons between FeatUp and the proposed method are shown in the bottom-right of Fig. 1). FeatUp (MaskCLIP) denotes using $\boldsymbol{v}$ of self-attention as the upsampled feature, which somewhat mitigates the possible negative effects of FeatUp (CLIP). In SimFeatUp, the input feature X of the last block is used to upsample, which presents a significant improvement in all 3 datasets. A substantial improvement is also delivered after replacing the training material for the upsampler from natural images to remote sensing images. "JBU_One" reduces the parameters by nearly $4\times$ while delivering a slight IoU gain (only <0.3M parameters). The introduction of CRN with image reconstruction loss brings 1.7%, 0.4%, and 1.6% improvement on 3 datasets, respectively. Note that the CRN only

Table 3. Quantitative comparison of vanilla CLIP and remote sensing CLIPs (ViT-B/32). Evaluation metric: mIoU.

Models		OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid^{img}	UDD5	VDD	Average
CLIP [51]	ICML'21	25.7	27.2	16.2	**40.0**	**25.1**	**31.6**	**39.7**	**39.1**	30.6
RemoteCLIP [37]	TGRS'23	18.2	**37.8**	**18.9**	21.9	22.9	16.1	27.1	28.1	23.9
GeoRSCLIP [75]	TGRS'24	**35.0**	30.8	**23.6**	38.0	22.3	**34.0**	**39.1**	**40.5**	**32.9**
SkyCLIP [65]	AAAI'24	**28.6**	**33.0**	15.3	**41.7**	**24.1**	**31.6**	38.2	35.8	**31.0**

Table 4. The proposed method is a plug-and-play module. "+ ours" indicates using SimFeatUp (Sec. 4.1) and alleviating the global bias (Sec. 4.2).

Methods	OpenEarthMap	WHU^{Aerial}	WBS-SI
MaskCLIP	25.1	29.8	39.8
+ ours	28.4↑3.3	35.4↑5.6	48.8↑9.0
SCLIP	29.3	33.4	32.1
+ ours	34.4↑5.1	39.5↑6.1	53.4↑21.3
ClearCLIP	31.0	36.6	44.9
+ ours	39.1↑8.1	51.1↑14.5	60.4↑15.5

Table 5. Detailed ablation results for each component. "X"↑ indicates upsampling earlier stage features, i.e. Eq. (6). "+ RS Data" indicates using Million-AID [40] to train the upsampler, before using the images in COCO-Stuff [4].

	OpenEarthMap	WHU$^{Sat.II}$	WBS-SI
Baseline	32.4	22.7	46.9
FeatUp (CLIP)	33.9	20.2	39.6
FeatUp (MaskCLIP)	33.8	25.2	45.9
"X"↑	34.6	26.0	54.2
+ RS Data	36.0 ↑1.4	26.2	56.4
+ JBU_One	36.3 ↑0.3	26.0	57.1
+ Rec. Image	37.6 ↑1.3	26.4	58.7
+ Alleviate Global Bias	39.3 ↑1.7	27.9	59.5
+ Large Kernel	40.3 ↑1.0	28.4	60.2

participates in SimFeatUp's training and is discarded during inference. Global bias alleviation shows significant improvement for all 3 datasets, with an average 1.3% improvement. Finally, expanding the upsampling kernel to 11×11 also exhibits consistent improvement across all datasets.

Results on natural images. We evaluate SimFeatUp as an external unit on 3 natural image datasets: PASCAL Context59 [47], COCOStuff [4] and Cityscapes [13]. As listed in Tab. 6, after upsampling the visual features of MaskCLIP, SCLIP, and ClearCLIP using SimFeatUp, their mIoUs are improved by 5.7%, 1.2%, and 1.2%, respectively. This reveals the potential of our method to inspire general vision.

Remote sensing CLIPs for OVSS. We evaluate the performance of remote sensing CLIPs on OVSS, including RemoteCLIP [37], GeoRSCLIP [75], and SkyCLIP [65], which are trained on 0.8M, 5M, and 2.6M remote sensing data, respectively, without changing the model structure of

Table 6. OVSS quantitative comparison on natural image datasets. The basic results are cited from [30].

Methods	Context59 [47]	Stuff [4]	Cityscapes [13]	Average
TCL [7]	30.3	19.6	23.1	24.3
Reco [57]	22.3	14.8	21.1	19.4
MaskCLIP	26.4	16.4	12.6	18.5
+ SimFeatUp	28.7	18.0	25.8	24.2 ↑5.7
SCLIP	33.0	21.1	29.1	27.7
+ SimFeatUp	34.1	22.0	30.5	28.9 ↑1.2
ClearCLIP	35.9	23.9	30.0	29.9
+ SimFeatUp	37.5	25.1	30.7	31.1 ↑1.2

CLIP. Since these works do not provide the ViT-B/16, we uniformly use ViT-B/32. Hence, we repeat the JBU operation 5 times in SimFeatUp. For fair comparison, we train the respective SimFeatUp for each model. As listed in Tab. 3, RemoteCLIP performs suboptimally to vanilla CLIP, which indicates that a small amount of domain data diminishes the model's transfer capability. GeoRSCLIP achieves the best performance against SkyCLIP, which suggests that domain VLMs can benefit from more diverse domain-specific data. Moreover, the OVSS task effectively reflects the model's discrimination and localization capabilities, and can serve as an evaluation metric for remote sensing VLMs.

6. Conclusion

In this paper, we present SegEarth-OV, a training-free OVSS method for remote sensing images. The design of SegEarth-OV was motivated by the observation that OVSS methods currently used for natural images do not perform well on remote sensing images. The two key insights of SegEarth-OV, i.e., SimFeatUp and global bias alleviation, exhibit consistent improvements on 17 remote sensing datasets spanning the tasks of semantic segmentation, building extraction, road extraction, and flood detection, well beyond the previous state-of-the-art methods. More importantly, as the first exploration of training-free OVSS method in remote sensing scenarios, this work demonstrates that the OVSS solution is feasible in earth perception tasks even if the VLMs are pre-trained on natural images. We expect that this work will inspire more OVSS methods and more capable remote sensing VLMs, and open up new possibilities for the Earth vision community.

Acknowledgment. This work is supported by the National Key R&D Program of China (2022YFA1004100), National Natural Science Foundation of China (No. 62272375, No. 62425113, No. 62192781), Major Key Project of PCL (PCL2024A06), and Tianyuan Fund for Mathematics of the National Natural Science Foundation of China (No. 12426105).

References

[1] Guillaume Alain. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. 4

[2] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 3

[3] 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. 3, 6, 7

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 8

[5] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 5

[6] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024. 3

[7] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 8

[8] 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, 7

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2

[11] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 3

[12] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 8

[14] Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. Deepglobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 172–181, 2018. 6

[15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 3

[16] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[17] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 4

[18] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 3

[19] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems*, 35:1140–1156, 2022. 2

[20] 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. 6

[21] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008. 2

[22] Benjamin Herfort, Sven Lautenbach, João Porto de Albuquerque, Jennings Anderson, and Alexander Zipf. A spatio-temporal analysis investigating completeness and inequalities of global urban building data in openstreetmap. *Nature Communications*, 14(1):3985, 2023. 2

[23] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[24] 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. 6, 7

[25] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *European Conference on Computer Vision and Pattern Recognition (ECCV)*, 2024. 3

[26] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[27] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 2

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[29] Johannes Kopf, Michael F Cohen, Dani Lischinski, and Matt Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics (ToG)*, 26(3):96–es, 2007. 2, 4

[30] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 3, 6, 7, 8

[31] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2025. 3

[32] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 3

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[35] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 2

[36] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 3, 6

[37] 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. 2, 8

[38] Wenze Liu, Hao Lu, Hongtao Fu, and Zhiguo Cao. Learning to upsample by learning to sample. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6027–6037, 2023. 2

[39] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024. 3

[40] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:4205–4230, 2021. 6, 8

[41] Hao Lu, Wenze Liu, Zixuan Ye, Hongtao Fu, Yuliang Liu, and Zhiguo Cao. Sapa: Similarity-aware point affiliation for feature upsampling. *Advances in Neural Information Processing Systems*, 35:20889–20901, 2022. 2

[42] 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. 3

[43] 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

[44] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 6

[45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 3

[46] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 6

[47] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 8

[48] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 2, 3, 5

[49] Chao Pang, Jiang Wu, Jiayu Li, Yi Liu, Jiaxing Sun, Weijia Li, Xingxing Weng, Shuai Wang, Litong Feng, Gui-Song Xia, et al. H2rsvlm: Towards helpful and honest remote sensing large vision language model. *arXiv preprint arXiv:2403.20213*, 2024. 2

[50] Li Pang, Datao Tang, Shuang Xu, Deyu Meng, and Xiangyong Cao. Hsigene: A foundation model for hyperspectral image generation. *arXiv preprint arXiv:2409.12470*, 2024. 2

[51] 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. 2, 3, 6, 7, 8

[52] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5571–5584, 2023. 2, 3, 5

[53] Esther Rolf, Konstantin Klemmer, Caleb Robinson, and Hannah Kerner. Mission critical–satellite data is a distinct modality in machine learning. *arXiv preprint arXiv:2402.01444*, 2024. 2, 5

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[56] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024. 3

[57] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35:33754–33767, 2022. 8

[58] Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2505–2515, 2024. 2

[59] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 3

[60] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27695–27705, 2024. 2

[61] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 6

[62] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 2, 3, 5, 6, 7

[63] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates, Inc., 2021. 5

[64] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 3

[65] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 2, 8

[66] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 5

[67] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3

[68] 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. 5, 7

[69] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 2

[70] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[71] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[72] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 3

[73] Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Alip: Adaptive language-image pre-training with synthetic caption. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2922–2931, 2023. 2

[74] D. Zanaga, R. Van De Kerchove, D. Daems, W. De Keersmaecker, C. Brockmann, G. Kirches, J. Wevers, O. Cartus, M. Santoro, S. Fritz, M. Lesiv, M. Herold, NE Tsendbazar, P. Xu, F. Ramoino, and O. Arino. Esa worldcover 10 m 2021 v200, 2022. 2

[75] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300*, 2023. 2, 8

[76] Zhuo Zheng, Yanfei Zhong, Junjue Wang, Ailong Ma, and Liangpei Zhang. Farseg++: Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[77] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3, 6, 7

[78] Minghao Zhou, Hong Wang, Yefeng Zheng, and Deyu Meng. A refreshed similarity-based upsampler for direct high-ratio feature upsampling. *arXiv preprint arXiv:2407.02283*, 2024. 2

[79] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[80] 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. 6