POLITECNICO DI TORINO

Master's Degree in Data Science and Engineering

Master's Degree Thesis

SAM Meets FC-CLIP: Advancing Open Vocabulary Segmentation in Satellite Imagery

Supervisors

Prof. Paolo GARZA

Prof. Edoardo ARNAUDO

Candidate

Jacopo LUNGO VASCHETTI

December 2024

Abstract

Our work addresses the challenge of open vocabulary semantic segmentation for very high-resolution satellite imagery. This computer vision task goes beyond traditional semantic segmentation, which assigns predefined category labels to each image pixel. Instead, the open vocabulary approach enables the dynamic identification of any object or region through natural language queries, eliminating the constraints of fixed classification categories. This flexible approach represents a critical advancement in remote sensing applications given the highly diverse scenes captured in satellite observations. We propose two novel solutions that build upon and enhance FC-CLIP, a state-of-the-art open vocabulary model originally designed for natural images. Our first solution, Remote FC-CLIP, integrates a remote sensing-specific CLIP model (Remote CLIP) into the baseline model's architecture, followed by fine-tuning on the OpenEarthMap (OEM) dataset. The second approach, SAM-FC-CLIP, combines a Segment Anything Model for mask extraction with modified classification components from FC-CLIP. This model was trained on a custom-built dataset that combines OEM and iSAID datasets, demonstrating an effective approach to tackle the persistent scarcity of comprehensive training data in the remote sensing domain. Results demonstrate that Remote FC-CLIP achieves superior performance compared to the baselines. While it excels on classes present in the training set, it exhibits reduced generalization to novel categories. In contrast, our SAM-based solution demonstrates remarkable open vocabulary capabilities, surpassing both baseline models and Remote FC-CLIP in identifying previously unseen classes. Despite the challenges posed by the scarcity of comprehensive satellite imagery datasets, these findings represent a step forward within this emerging field while also revealing promising directions for future research.

Table of Contents

List of Tables

List of Figures

Acronyms

CNN

Convolutional Neural Networks

OVS

Open Vocabulary Segmentation

RS

Remote Sensing

VHR

Very High Resolution

SOTA

State-of-the-art

VLM

Vision-Language Models

rle

Run-length encoding

IoU

Intersection over Union

mIoU

mean Intersection over Union

CE

Cross Entropy

Chapter 1

Introduction

In recent years, the accessibility of satellite imagery has increased substantially, driven by initiatives such as Copernicus [1] and the Maxar Open Data Program [2]. Of particular significance are Very High-Resolution (VHR) satellite images, which capture Earth's surface at resolutions of 0.25-0.5 meters per pixel, enabling the identification of fine-grained features ranging from individual vehicles to detailed building structures. However, despite this growing abundance of raw imagery, the limited availability of annotated data remains a significant challenge. This scarcity can be attributed to the substantial time investment and specialized technical expertise required to create high-quality annotations in this domain.

Task - Semantic segmentation has emerged as a fundamental technique for extracting meaningful information from satellite photographs. This computer vision task involves the pixel-wise classification of images into meaningful categories, facilitating the creation of detailed semantic maps that can be useful for a wide range of applications. For example, in urban planning, they help identify and monitor building footprints, road networks, and green spaces, supporting decisions about infrastructure development and city expansion. In environmental monitoring, they facilitate the tracking of deforestation, coastal erosion, and changes in agricultural land use over time. Disaster response teams utilize them to rapidly assess damage after natural disasters, identifying damaged buildings and infrastructure to guide emergency operations. However, traditional semantic segmentation approaches are confined to pre-defined categories, with limited flexibility and applicability across diverse scenarios. This limitation is particularly problematic in the context of satellite imagery, where the variety of objects and features of interest can be vast and unpredictable. Given these constraints, we pursued open vocabulary semantic segmentation (OVS) as our research direction, exploring this emerging paradigm that enables the identification and segmentation of novel, previously unseen categories through natural language queries.

Challenges - The implementation of OVS for VHR satellite imagery presents several challenges. Direct application of models designed for natural images yields poor results due to fundamental differences in domain characteristics. For example, the top-down perspective of satellite imagery eliminates depth information present in conventional photographs, objects appear at dramatically different scales, and satellite scenes typically contain a higher density of objects compared to natural images. As previously mentioned, these domain-specific issues are compounded by the scarcity of comprehensive annotated training data which limits the effectiveness of fine-tuning existing models developed for natural images.

Our Contributions - This thesis presents two novel approaches to address these challenges. Our first solution, Remote FC-CLIP, enhances FC-CLIP, a state-of-the-art open vocabulary model originally designed for natural images, by incorporating Remote CLIP, a vision-language foundation model specifically fine-tuned for remote sensing applications. Our second approach, SAM-FC-CLIP, combines the Segment Anything Model (SAM) with a modified version of Remote FC-CLIP, addressing the limitations in mask extraction for fine-grained objects and demonstrating superior performance in identifying previously unseen classes. The development of this latter approach necessitated the creation of a novel unified dataset, merging OpenEarthMap (OEM) and iSAID to encompass 23 distinct classes. These contributions advance the practical application of OVS for VHR satellite imagery while identifying promising directions for future research.

Thesis Structure - The following is a description of the thesis structure. Chapter 2 provides comprehensive background information and reviews related works. It begins with Section 2.1, which presents a systematic overview of segmentation tasks, progressing from basic semantic segmentation to more advanced approaches like instance and panoptic segmentation, before introducing OVS and standard evaluation methods. Section 2.2 explores frameworks designed for natural images, including a detailed classification of OVS methods, relevant foundation models, like SAM and CLIP, and most recent OVS models. Section 2.3 begins with an analysis of differences between satellite and natural imagery and continues presenting a review of important remote sensing datasets used in this research, including OpenEarthMap, iSAID, LoveDA, and FMARS. Chapter 3 outlines our methodology, beginning with the formal problem statement. It then describes baseline frameworks, FC-CLIP and RemoteCLIP, before introducing our two novel approaches: Remote FC-CLIP and SAM-FC-CLIP. The chapter concludes with a description of our custom ensemble dataset. Chapter 4 presents our experimental work and results. It starts with detailed implementation specifications for both of our solutions. Then, the results section [4.2] provides a comprehensive analysis, beginning with baseline performance metrics, followed by evaluations of Remote FC-CLIP and SAM-FC-CLIP, including both quantitative and qualitative analyses.

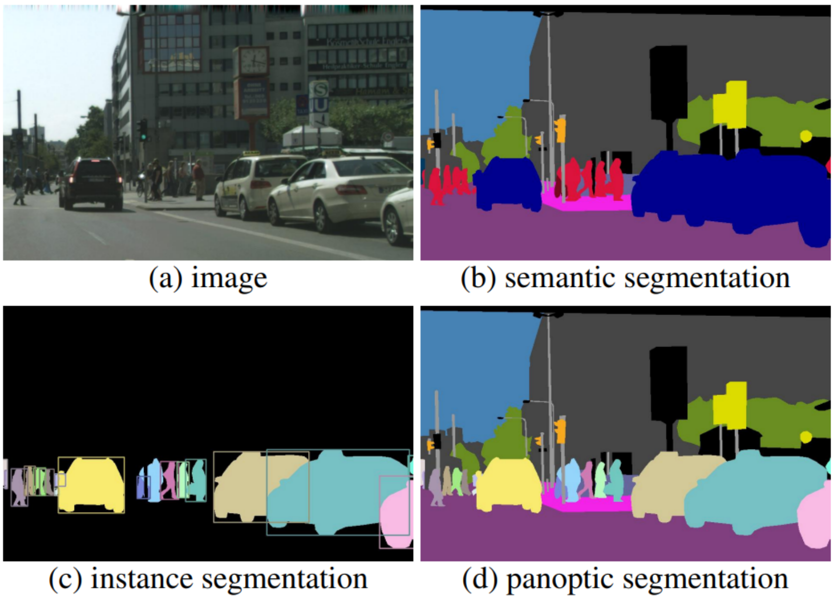
Chapter 5 concludes the thesis by summarizing our findings and contributions while suggesting future works to improve OVS of satellite imagery.

Chapter 2

Background

2.1 Related Works

Image segmentation is a fundamental task in computer vision that involves partitioning an image into multiple segments or regions, each associated with a class. In the context of satellite imagery analysis, segmentation plays a crucial role in extracting meaningful information from complex, high-resolution images. This section provides an overview of four principal segmentation approaches, exemplified in Figure 2.1: semantic, instance, panoptic and open vocabulary segmentation.


(a) image

(b) semantic segmentation

(c) instance segmentation

(d) panoptic segmentation

Figure 2.1: Different types of segmentation (Image from [3]).

2.1.1 Semantic Segmentation

Semantic segmentation aims to classify each pixel in an image into a predefined set of categories. In satellite imagery, these categories might include land cover types such as water, vegetation, urban areas as well as more specific objects like vehicles, individual trees and buildings. The output of semantic segmentation is a label map where each pixel is assigned a class label. Formally, given an input image I of size $h \times w$, semantic segmentation produces a label map L of the same size, where each pixel L_{ij} is assigned a class label from a predefined set $C = \{c_1, c_2, ..., c_K\}$. The goal is to learn a mapping function $f : I \to L$ that accurately classifies each pixel. The advent of deep neural networks, particularly Convolutional Neural Networks (CNN) [4], has revolutionised the field of semantic segmentation. Fully Convolutional Networks (FCN) [5] were the first ent-to-end CNNs for semantic segmentation. FCN replaced fully connected layers with convolutional layers, enabling variable input sizes and efficient dense prediction. Building upon this foundation, U-Net [6] gained prominence, especially in medical image segmentation. This model is based on an encoder-decoder architecture with skip connections. These allow the network to retain detailed spatial information that otherwise would be lost. Further advancements came with the DeepLab series [7, 8], which introduced various innovations. These include atrous convolutions that allow for a larger receptive field and atrous spatial pyramid pooling (ASPP) improving the multi-scale context integration. The success of transformers in the natural language processing field, which became famous thanks to [9], has led to their adoption also in computer vision tasks, including semantic segmentation. This paradigm shift is exemplified by architectures such as SETR (Segmenter Transformer) [10], which applies a pure transformer to semantic segmentation, moving away from traditional FCN-based approaches. SegFormer [11] further refined this approach by combining the strengths of CNNs and transformers in a lightweight, flexible architecture. Mask2Former [12] represents another significant development, offering a more general framework applicable to various segmentation tasks, including semantic segmentation. Due to its relevance to the project, this model will be discussed in more detail in Section 2.2.2. As of 2024, significant progress has been made in interactive segmentation models, with the Segment Anything Model (SAM) [13] and its variants [14, 15, 16]. These models enable interactive segmentation where users can prompt the model with points or bounding boxes to obtain masks of objects of interest. It is important to notice, that while these architectures offer unprecedented flexibility and user control in the segmentation process, they produce agnostic segmentations. That is, masks are not inherently associated with specific semantic classes. This characteristic distinguishes SAM and its variants from traditional semantic segmentation models, which directly assign class labels to each segmented region.

2.1.2 Instance Segmentation

While semantic segmentation groups pixels by class, instance segmentation complicates the task by differentiating between individual objects of the same class while ignoring uncountable objects in the background (e.g. the sky and the road). This is particularly useful in satellite imagery for tasks such as building detection or vehicle counting, where distinguishing between multiple instances of the same object type is crucial. In instance segmentation, for each detected object, a model should output a pixel-wise mask indicating the object's extent and its class label. In the beginning, this task was achieved by models belonging to the R-CNN family (R-CNN [17], Fast R-CNN [18] and Faster R-CNN [19]). These are based on region proposals and CNN-based feature extraction. Compared to previous models, which were quite slow, Faster R-CNN [19], made considerable advancements, with its Region Proposal Network (RPN), resulting in enhanced speed and accuracy. Subsequently, Mask R-CNN extended instance segmentation capabilities by adding a mask prediction branch. It introduced RoIAlign to preserve spatial information, crucial for accurate segmentation. Later models aimed to further improve efficiency and real-time processing. YOLACT (You Only Look At CoefficienTs) [20] introduced a real-time approach by generating prototype masks and per-instance coefficients. Then, SOLO (Segmenting Objects by Locations) [21] brought a paradigm shift. It treated instance segmentation as a direct pixel-to-instance classification problem. More recent models, like DETR [22] and the already cited Mask2Former [12], are based on transformers. They allow the models to capture better long-range dependencies, improving the overall performance.

2.1.3 Panoptic Segmentation

Panoptic segmentation further increases the difficulty of the task by unifying semantic and instance segmentation. It aims to provide a coherent segmentation of an entire scene, distinguishing classes between *things* and *stuff*. The *thing* classes are countable objects with instances, such as cars and people. In comparison, *stuff* classes, represent amorphous uncountable regions (e.g. road, sky and grass). We require the model to perform instance segmentation on *things* classes and semantic segmentation on *stuff* classes, combining the strengths of both tasks. In the context of satellite imagery, panoptic segmentation can be helpful to simultaneously identify large-scale land cover types (e.g. cropland and roads) and individual objects like buildings or vehicles. Mathematically, panoptic segmentation produces a label map where each pixel is assigned a tuple (l, i), where l is the semantic label and i is the instance ID. In the beginning, the panoptic task was tackled by adding a semantic segmentation head to existing models such as Mask R-CNN [23] and Feature Pyramid Network (FPN) [24]. Panoptic FPN [25] is based on this, integrating semantic and instance segmentation more cohesively. Then, UPSNet [26] brought end-to-end training capabilities with a parameter-free panoptic head. Axial-DeepLab [27] employed axial-attention for contextual modelling, while EfficientPS [28] achieved real-time performance using EfficientNet as a backbone. Panoptic-DeepLab [29], adapted its innovations (atrous convolutions and dual-ASPP modules), already employed for multi-scale processing in semantic segmentation, for this specific task. As already written, transformer architectures revolutionized numerous domains in deep learning, including panoptic segmentation tasks. The MaX-DeepLab model [30] was the first truly end-to-end panoptic segmentation model without post-processing requirements. To conclude, the Mask2Former model mentioned above [12] has also been successfully applied in this domain.

2.1.4 Open Vocabulary Segmentation

Open-vocabulary segmentation (OVS) further extends the previous tasks, addressing their limitation of being tied to a predefined set of object categories. Its primary objective is to enable models to segment objects from an unrestricted range of classes, including those not encountered during training. We distinguish between *base* categories observed during training and *novel* ones encountered only at inference time. Users are enabled to recall these classes through natural language descriptions, allowing for greater flexibility and generalisation. In the RS field, open vocabulary segmentation offers significant potential for addressing the diversity and complexity of Earth observation data. It could allow for the identification and segmentation of rare objects and land cover types. The majority of approaches leverage the semantic understanding capabilities of large-scale vision-language models, such as CLIP [31], to bridge the gap between visual features and textual descriptions. By doing so, these models can generalise to novel object categories without requiring explicit training data for these classes. According to the most recent survey on the topic [32], OVS methodologies can be categorized into four types: Region-aware training, Pseudo-labeling, Knowledge distillation and Transfer learning.

Region-aware training techniques establish correspondences between image regions and textual descriptions using weakly supervised learning on image-caption pairs. They strongly rely on caption datasets like COCO Captions [33]. Notable models in this category include: GroupViT [34], which progressively groups image tokens into semantically coherent segments, ViL-Seg [35], incorporating local-to-global correspondence learning and PACL [36], which introduces patch-level contrastive learning.

Pseudo-labelling methods generate pseudo annotations for novel classes, often utilizing pre-trained vision-language models to create these labels. The necessity of knowing the novel category in advantage, at training time, is a salient drawback of this method.

Knowledge distillation approaches transfer the semantic understanding of large

vision-language models to task-specific architectures, enabling them to recognize a broader range of objects. These techniques employ a teacher-student paradigm where the student learns to mimic visual knowledge from the teacher image encoder, which is typically a large vision-language model (e.g. CLIP [31]). An example of this approach is SAM-CLIP [37] which fuses SAM and CLIP image encoders through cosine distillation, using memory replay on a subset of pretraining data.
Transfer learning techniques aim to adapt pre-trained vision-language models (VLMs), particularly their image encoders, directly for downstream detection and segmentation tasks, often employing parameter-efficient fine-tuning methods. With respect to the already presented approaches, this one has demonstrated superior performance, with state-of-the-art models predominantly falling within this category. We enumerate some notable models:

- OpenSeeD [38] presents a unified framework for open-vocabulary detection and segmentation tasks, employing a novel decoupled architecture. It addresses task discrepancies through foreground-background decoding and compensates for data disparities via conditioned mask decoding, enabling effective performance across multiple vision tasks with a single model.
- OVSeg [39] fine-tunes CLIP's image encoder on constructed mask-category pairs to address the domain gap between masked image crops and natural images
- ODISE [40] leverages text-to-image diffusion models for open-vocabulary panoptic segmentation, combining CLIP with diffusion model capabilities
- MaskCLIP [41] modifies CLIP's attention pooling layer to enhance local semantic consistency for dense prediction tasks
- FC-CLIP [42] uses a frozen CNN-based CLIP image encoder for panoptic segmentation tasks
- PosSAM [43] combines the frozen CLIP and SAM image encoders, fusing their output visual features via cross-attention for panoptic segmentation.

Despite the great advancement in solving the task, the OVS field still faces some open challenges including mitigating biases towards base classes, improving the quality of pseudo-labels and addressing the domain gap between pretraining data and downstream tasks.

2.1.5 Evaluation Methods

This section outlines the main quantitative metrics and evaluation protocols adopted to assess the model's performance on the segmentation tasks. We examine both established metrics widely used in semantic and instance segmentation literature, including pixel-wise accuracy, mean intersection over union, and average precision, as well as more recent evaluation methods such as Panoptic Quality. Furthermore, we introduce evaluation techniques specifically designed for OVS.

Semantic Segmentation Evaluation Metrics

Pixel-wise accuracy is one of the most straightforward metrics used in semantic segmentation evaluation. It is defined as the ratio of correctly classified pixels to the total number of pixels in the image:

$$\text{Pixel-wise Accuracy} = \frac{\text{Number of correctly classified pixels}}{\text{Total number of pixels}}$$

It is intuitive and easy to compute, however, it can be misleading in cases of class imbalance, which is common in satellite imagery where certain class types may dominate the scene. For instance, the class representing vehicles typically occupies a substantially smaller fraction of the image compared to the more expansive *grass fields* class. Consequently, a model which achieves high accuracy by correctly classifying the dominant class, but which fails to identify less prevalent but equally important classes (such as vehicles), could be wrongly considered a good model.

Mean Intersection over Union (mIoU) has emerged as the de facto standard for semantic segmentation evaluation due to its robustness to class imbalance. For each class, the IoU is calculated as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

where TP, FP and FN represent true positives, false positives and false negatives, respectively. Figure 2.2 provides a clear visualization of the formula. The mIoU is then computed by averaging the IoU values across all classes:

$$\text{mIoU} = \frac{1}{k}\sum_{i=1}^{k} \text{IoU}_i$$

where k is the number of classes. It provides a balanced measure of segmentation quality across all classes, making it particularly useful for multi-class segmentation tasks.

Instance Segmentation Evaluation Metrics

Average Precision (AP) is the main metric for instance segmentation evaluation. It is derived from object detection literature but adapted for segmentation tasks.

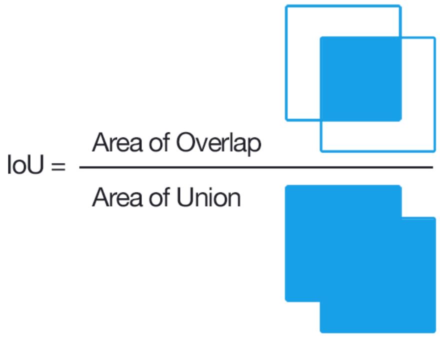


Figure 2.2: Visualization of the IoU metric. The IoU score is calculated by dividing the area where two masks overlap (intersection) by the total area covered by both masks combined (union). The two squares represent the ground truth and predicted segmentation masks, respectively.

Given a IoU threshold t and a confidence threshold c, considering only model's predictions with confidence above c, we define:

$$\text{Precision}_t = \frac{\text{TP}_t}{\text{TP}_t + \text{FP}_t}$$

$$\text{Recall}_t = \frac{\text{TP}_t}{\text{TP}_t + \text{FN}_t}$$

where TP_t, FP_t and FN_t represent true positives, false positives and false negatives at threshold t, respectively. The Precision-Recall (PR) curve, at a fixed IoU, is created by varying the confidence threshold. The AP is then computed as the area under the PR curve:

$$\text{AP} = \int_0^1 p(r)dr$$

where $p(r)$ is the precision at recall r. In practice, this integral is approximated using a finite set of recall values.

Mean Average Precision (mAP) extends the AP concept across multiple classes:

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i$$

where N is the number of classes and AP_i is the Average Precision for class i. To have a more comprehensive evaluation, mAP is often calculated at multiple IoU thresholds and averaged. A common approach consists in computing mAP at IoU thresholds from 0.5 to 0.95 with a step size of 0.05, yielding the metric referred to as mAP@[0.5:0.05:0.95].

Panoptic Segmentation Evaluation Metrics

Panoptic Quality (PQ), defined in [3], is the main metric used to evaluate panoptic segmentation. Its mathematical formulation is the following:

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}|}}_{\text{segmentation quality}} \times \underbrace{\frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}|\mathrm{FP}| + \frac{1}{2}|\mathrm{FN}|}}_{\text{recognition quality}}$$

p and g represent predicted and ground truth segments, respectively. Meanwhile, as stated before, TP, FP and FN represent true positives, false positives and false negatives. The first term, the segmentation quality, measures the quality of predicted masks for matched segments. The second term, the recognition quality, measures how well the model recognizes and identifies different segments.

OVS Evaluation Methods

The evaluation of OVS models presents peculiar challenges due to the inherent disparity in difficulty between segmenting base and novel categories. Typically, researchers employ a *generalized* evaluation protocol examining model performance both on base and novel classes. The used metrics are the same as previously presented, including mIoU for semantic segmentation and mAP for instance segmentation, however distinguishing between base and novel classes. Additionally, some studies utilize a *cross-dataset transfer evaluation* (CDTE) protocol to assess model generalization across different datasets without fine-tuning. This involves training on a source dataset and evaluating on target datasets with potentially overlapping vocabularies. Common benchmark datasets for OVS evaluation include Pascal VOC [44], COCO Stuff [45] and ADE20K [46], each offering diverse semantic categories and scene complexities.

2.2 Natural Image Frameworks

While open vocabulary segmentation represents an active and rapidly evolving area of research for natural images, its application to Very High Resolution (VHR) satellite imagery remains relatively unexplored, with no established and openly available techniques to the best of our knowledge. For this reason, we conducted a comprehensive review of state-of-the-art methods from the standard photograph domain, with the intent of adapting and applying them to remote sensing data. This section begins with a categorization of existing approaches, analyzing their respective strengths and limitations. The analysis demonstrates that transfer learning approaches based on CLIP are the most promising direction for our research, offering superior performance and adaptability to the remote sensing domain. We then provide an in-depth examination of foundation models that form the basis of our proposed solutions: CLIP, Segment Anything Model (SAM), and Mask2Former. Finally, we analyze several state-of-the-art OVS frameworks that serve as our experimental baselines, including OVSeg, MaskCLIP, and MasQCLIP.

2.2.1 OVS Methods Categorization

A link between visual and textual concepts is required when performing an open vocabulary task. Based on how this alignment is obtained, the four methodologies discussed in Section 2.1.4 can be further categorized into two groups: (a) methods building this alignment by training on extensive text-image pair datasets and (b) methods that leverage pre-trained vision-language models (VLMs) We decided to discard the first category from our study since it is not easily adaptable to the RS domain. This approach requires a substantial amount of diverse and labelled data (e.g. COCO [47], Objects365 [48], ADE20K [46]) which are often lacking in the RS domain. Also, this method necessitates extensive computational resources for training. Our investigation primarily focuses on models leveraging VLMs, especially those based on CLIP, given the existence of RemoteCLIP [49], a CLIP model specifically fine-tuned on RS images. Within this latter group, we exclude *Pseudo-labelling* methods since we aim to a truly OVS model, without having to know the novel classes at training time. Additionally, we discarded the *Knowledge distillation* methods since the most performant model in this category lacks open-source code. Consequently, our research concentrates on the *Transfer learning* category, which has emerged as a dominant approach. This methodology presents several advantages: (a) it leverages the powerful visual-language representations learned by foundation models such as CLIP, (b) the majority of recent approaches to solve OVS adopt this framework, and (c) it consistently achieves superior performances compared to alternative solutions.

2.2.2 Related Foundation Models

Segment Anything Model - The Segment Anything Model (SAM) [13] represents a paradigm shift in computer vision being the first foundation model for image segmentation able to generalize across diverse domains. It allows users to specify what to segment using various input types such as simple points and bounding boxes. Its architecture consists of three main components. First, a Vision Transformer serves as the image encoder, processing the input image to create a rich visual representation. Second, a versatile prompt encoder transforms different types of user inputs into a standardized format that the model can process. The third component, a mask decoder, integrates the representations from both encoders to generate the final segmentation masks. This lightweight decoder can rapidly produce multiple potential segmentations, each representing a different prompt interpretation. When processing an image, SAM needs to run it through the image encode only once, regardless of the number of subsequent prompts. This design choice significantly improves computational efficiency in interactive scenarios. The development of SAM was made possible through an innovative data collection approach that bootstrapped the model's own capabilities. Starting with manual annotations assisted by the model, progressing to semi-automatic labelling, and finally achieving fully automatic mask generation, this iterative process created an unprecedented dataset of over one billion masks. This massive scale of high-quality training data enables this model to generalize effectively to diverse domains. SAM allows for both interactive and automatic mask generation. In the latter case, a grid of points is prompted across the entire image. This automatic process is governed by several hyperparameters, which we elaborate upon due to their crucial role in the analyses presented in subsequent sections:

- `points_per_side`: the number of points per side in the prompting grid. The total number of prompted points will be `points_per_side`^{2}.
- `pred_iou_thresh`: a threshold in the range $[0, 1]$ to filter produced masks based on the IoU score predicted by the model itself.
- `stability_score_thresh`: a threshold in the range $[0, 1]$ used to filter produced masks based on a stability score. This score is computed as the IoU between two masks acquired by cutting off mask logits at different values.
- `stability_score_offset`: The amount to shift of the cutoff when calculating the stability score. The larger this value the harsher the generator is in considering a mask as *stable*. In other words, the larger, the lower the stability scores will be.
- `box_nms_thresh`: the maximal suppression's box IoU threshold used to remove duplicate masks.
- `crop_n_layers`: if this value is greater than zero, the automatic mask extraction will be run again on individual crops, in addition to the complete image.
- `min_mask_region_area`: disconnected regions and holes with areas smaller than the indicated value will be removed.

Several OVS approaches in the literature employ SAM in their pipelines. A notable implementation involves utilizing SAM in conjunction with an open-vocabulary object detector. A representative example is Grounded SAM [50], which leverages Grounding DINO [51] for object detection and subsequently prompts SAM with the detected bounding box. Unfortunately, this is not directly applicable to satellite imagery, since objects like roads or rivers can extend along the entire image, potentially resulting in a catastrophic SAM prompt.

CLIP - CLIP [31] is a foundation model developed by OpenAI that aims to associate text and images. Its core innovation stands in the training approach. In this phase, it leverages an enormous dataset of image-text pairs collected from the internet, learning to associate photos with a natural language description of their content. Its architecture comprises two different encoders to handle images and text. They respectively produce a vector representation of the image and the associated textual description, called embedding. By employing a contrastive loss the encoders are trained simultaneously to maximize the similarity between embeddings of matched image-text pairs while minimizing it for unrelated pairs. This training allows CLIP to learn a shared embedding space for images and text, making it useful for tasks combining images and natural language. CLIP has been evaluated on a wide range of tasks, including image classification. Remarkably, it demonstrated competitive performance with state-of-the-art models on many benchmarks, despite not being specifically trained for these tasks. Another significant aspect is its robustness to domain shifts. The model showed improved performance on various out-of-distribution datasets compared to traditional ImageNet-trained models, suggesting that its diverse pre-training data and learning approach contribute to its generalization capabilities. This feature is particularly relevant to our research since we aim to obtain a model able to maintain high performance even on unseen domains. While CLIP demonstrates significant capabilities, its training on low-resolution images can lead to suboptimal performance for fine-grained classification tasks, such as segmentation. Adequate techniques and architectures must be employed to enhance its performance in such tasks.

Mask2Former - Mask2Former [12] is a famous model used for segmentation tasks that build upon MaskFormer [52]. It addresses panoptic, instance, and semantic segmentation tasks without architectural modification. This is enabled through a mask classification approach since the model generates a set of binary masks with corresponding class labels, in contrast to conventional approaches that perform pixel-wise classification independently. At its core, Mask2Former consists of three main components: a backbone feature extractor, a pixel decoder, and a transformer decoder. The key innovation lies in the transformer decoder, which employs a novel masked attention mechanism. Unlike standard cross-attention, it is constrained to areas inside the predicted masks. This innovation leads to faster convergence and improved performance. The pixel decoder is essential to handle small objects effectively. Its pyramid feature maps are fed into consecutive transformer layers, allowing for a multi-scale strategy without excessive computational overhead. Also, Mask2Former introduces several optimization improvements, such as: switching the order of self-attention and masked attention, making query features learnable, and removing dropout. These modifications enhance model performance without increasing resource requirements.

2.2.3 Natural Images OVS Models

OVSeg - Open-Vocabulary Semantic Segmentation (OVSeg) [39] is a recent model developed by Meta, that tackles the problem of open vocabulary segmentation. It consists of two main components: a segmentation architecture based on MaskFormer [52] for generating class-agnostic mask proposals, and a CLIP model for open-vocabulary classification. This paper verifies that simply classifying the masks, cropped from the background, leads to poor results. To overcome this limitation, it proposes adapting CLIP for masked images by fine-tuning it on a diverse dataset of masked image-category pairs. These pairs are collected from existing image-caption datasets, such as COCO Captions, and adapted to the task using a self-labelling strategy. Furthermore, OVSeg introduces mask prompt tuning a technique that replaces *zero tokens* resulting from masked areas with learnable prompt tokens, allowing the model to encode more useful information.

MaskCLIP - The architecture of MaskCLIP [41] consists of a class-agnostic mask proposal network, a visual encoder, referred to as *MaskCLIP Visual Encoder* and a CLIP text encoder. The main novelty is its visual encoder, since it exploits a pretrained CLIP ViT model adding alongside the Relative Mask Attention (RMA) module and introducing a Mask Class Token. Each of these tokens captures the semantics of a single mask proposal by attending the class tokens and all image tokens inside the mask area. Differently from OVSeg, this allows for parallel inference for multiple masks from the same image, being more efficient. The RMA module optimizes the utilization of mask information and refines initial mask predictions.

MasQCLIP - The architecture of MasQCLIP [53] consists of two main stages: a mask generator for mask extraction, and an encoder-only module for mask classification. Building upon the Mask Class Token strategy already used in MaskCLIP, it introduces two innovations: (a) a student-teacher self-training module, and (b) a new fine-tuning strategy. The first allows for a progressive distillation process enabling the model to generate mask proposals beyond the base classes seen at training time. The second, called MasQ-Tuning, fine-tunes CLIP by optimizing only the query parameters within its vision transformer encoder. This approach allows the model to better adapt to the representation of mask regions while preserving the generalization capabilities of the pre-trained CLIP model.

2.3 Remote Sensing Datasets

In recent years, deep learning has revolutionized computer vision tasks on natural images, largely due to the availability of large-scale, diverse datasets such as ImageNet [54], COCO [47], and Pascal VOC [44]. These collections, containing millions of annotated images across hundreds of categories, have been essential in training robust models. However, the remote sensing domain presents a significantly different landscape in terms of annotated data availability and characteristics. While the field has seen growing attention from the research community, the number of comprehensive, large-scale datasets remains limited compared to those available for natural images. This scarcity is particularly notable when considering datasets that provide detailed annotations across a wide range of classes. The unique characteristics of aerial and satellite imagery, including varying spatial resolutions, diverse geographical contexts, and complex object relationships, make the creation of such datasets both resource-intensive and technically challenging. The process requires not only significant manual annotation effort but also domain expertise to accurately label features such as land cover types, urban structures, and natural elements. This section first examines the differences between conventional and aerial photography. It then provides a brief overview of datasets adopted in this work, namely: OpenEarthMap, encompassing semantic segmentation annotation of landcover types, iSAID, which focuses on instance-level annotations, LoveDA, which provides semantic labels across urban and rural domains, and FMARS, a compact but meticulously annotated test dataset. While each makes valuable contributions to the field, they also highlight the ongoing need for more comprehensive resources that can match the scale and diversity seen in natural image datasets. We will analyze their characteristics, limitations, and potential applications, focusing particularly on their utility for the OVS task.

2.3.1 Satellite vs. Natural Images

Satellite images present considerable challenges and important differences compared to natural photographs. Figure 2.3 provides immediate insight into their inherent complexity disparity. Domain variation makes it difficult for models trained on natural images to be performant on satellite ones without adequate modifications. Some of the most important distinguishing characteristics include: (a) scale disparity, (b) resolution gap, (c) scene complexity, (d) complex shapes, (e) orientation invariance, and (f) lack of three-dimensionality. (a) The scale of objects in satellite imagery differs notably from that in natural images. This disparity requires models to identify objects at vastly different scales than those encountered in training data. (b) RS images typically exhibit lower resolution compared to everyday photos. This results in reduced details for individual objects, increasing the difficulty of distinguishing and segmenting them. (c) They often encompass large areas with numerous small objects of interest. This high object density and scene complexity can be challenging for models trained on simpler standard images, which often encompass fewer and bigger objects. (d) Also, aerial images frequently contain elongated, complex shapes that can span the entire photo (e.g. rivers and roads). Usually in natural images objects are more compact. Additionally, (e) in regular photos they often have a fixed or limited orientation, meanwhile, in our domain of interest, they can appear at arbitrary angles. To conclude, (f) satellite imagery presents a flattened, top-down perspective, eliminating most of the three-dimensionality. This can cause classes that are distinct in natural images to appear similar when viewed from above, challenging the model's ability to differentiate between them. An exemplification of some of these challenges is the segmentation of a tree. In natural images, they typically present a log and a crown. Meanwhile, seen from above they appear more similar to bushes as their trunk is hidden under the canopy in the top-down view.

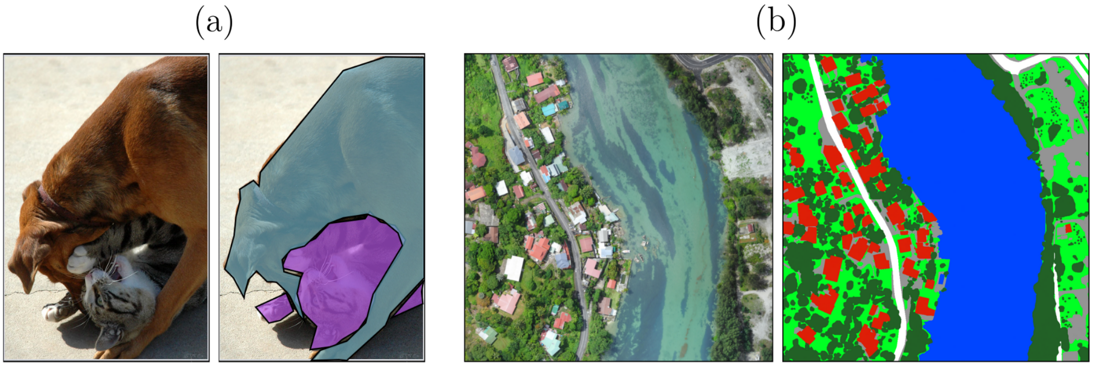


Figure 2.3: Comparison of segmentation complexity between natural and satellite imagery. The figure shows the plain photos and their respective annotations. The image on the left (a), from the COCO [47] dataset, has a much simpler class separation, compared to the aerial photo on the right (b), from the OpenEarthMap [55] dataset.

2.3.2 OpenEarthMap

The OpenEarthMap (OEM) [55] dataset is a benchmark dataset for high-resolution land cover mapping. It comprises 2.2 million annotated segments derived from 5,000 aerial and satellite images. They cover 97 regions across 44 countries on 6 continents. Image resolution ranges from 0.25 to 0.5 meters/pixel. It contains 8 manually annotated classes: *bareland*, *rangeland*, *developed space*, *road*, *tree*, *water*, *agriculture land* and *building*. The labelling was made by a team of several people who conducted quality checks to ensure accuracy. This dataset is ideal for training semantic segmentation models. It is divided into training (3,000 images), validation (500 images) and test (1,500 images) sets. Some sample images are shown in 2.4. Various baseline experiments have been conducted on this dataset using state-of-the-art semantic segmentation models, both CNN-based and Transformer-based.

Notably, models trained with these annotations have shown robust results when applied to RS imagery from a different domain. This feature suggests OEM's potential for training a globally adaptable, open-vocabulary model, capable of generalizing across varied geographical contexts. Despite the high quality of these annotations and our interest in the classes they encompass, the dataset presents some limitations for our specific objective. Specifically, the semantic annotations, while valuable for multiple tasks, do not provide any indication about the specific instance, which would be ideal for our study. For instance, the absence of individual building delineations does not allow us to recognize separate entities, and therefore directly apply these data for training a panoptic segmentation model.

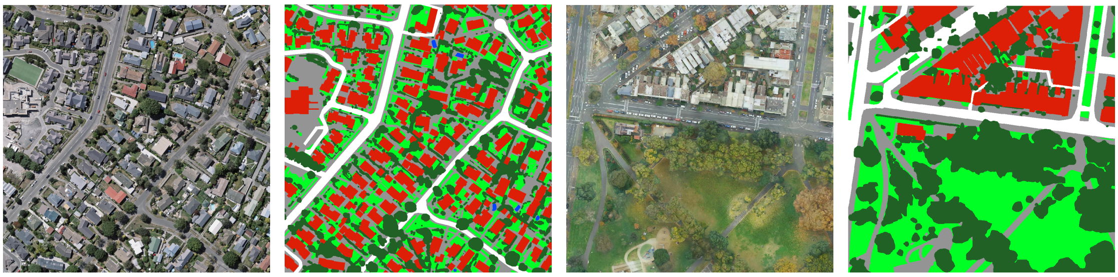
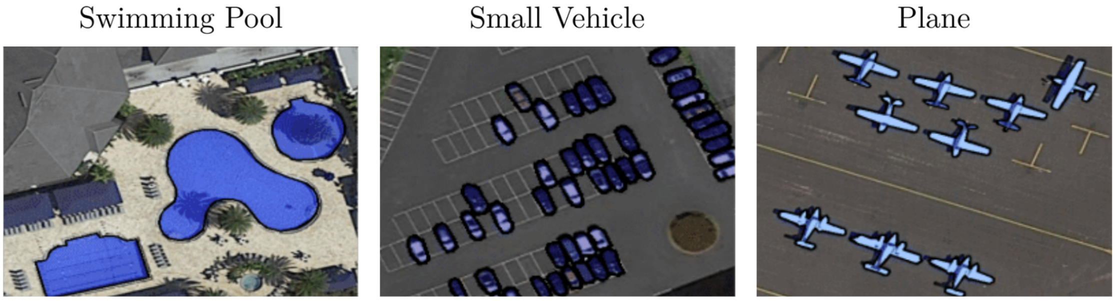


Figure 2.4: Two images and their corresponding ground truth masks from the OEM dataset.

2.3.3 iSAID

Instance Segmentation in Aerial Images Dataset (iSAID) [56] is a large-scale benchmark dataset, specifically designed for the task of instance segmentation in aerial imagery. It significantly surpasses previous RS datasets in terms of both category count and instance count, comprising 2,806 high-resolution images, for a total of 655,451 annotated object instances across 15 categories. These categories include various man-made structures and vehicles commonly observed in aerial views, specifically: *plane*, *ship*, *storage tank*, *baseball diamond*, *tennis court*, *basketball court*, *ground track field*, *harbor*, *bridge*, *large vehicle*, *small vehicle*, *helicopter*, *roundabout*, *swimming pool* and *soccer ball field*. iSAID's images exhibit several challenging characteristics typical of aerial imagery, making it particularly suitable for real-world applications. These include high object density, with up to 8,000 instances per image, significant variations in object scale and orientation, and the presence of tiny objects. The dataset also captures the real-world imbalance in object frequencies, with some categories being much more prevalent than others (see Figure 2.6). Some sample images are shown in Figure 2.5. This work also establishes baseline performance metrics using state-of-the-art instance segmentation methods like Mask R-CNN and PANet. These experiments highlighted the need for specialized solutions to handle the unique characteristics of aerial imagery. In conclusion, this dataset provides high-quality instance-level annotations across diverse semantic categories. Nevertheless, the absence of some important categories, such as *buildings* and *high vegetation*, significantly limits its applicability to our research objectives.



Swimming Pool

Small Vehicle

Plane

Figure 2.5: Three images from the iSAID dataset encompassing different classes.

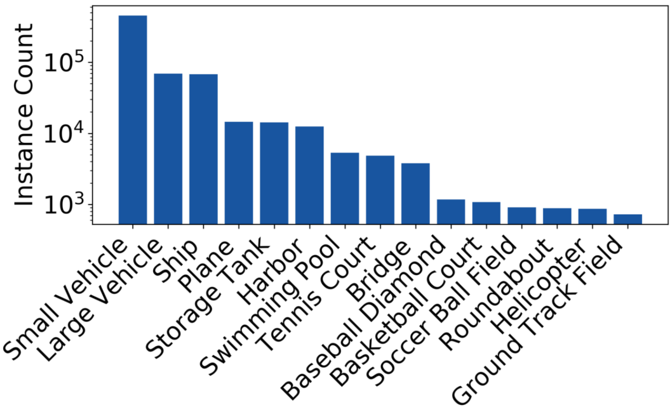


Figure 2.6: Histogram of the number of instances per class (sorted by frequency) of iSAID dataset. (Image from [56]).

2.3.4 LoveDA

The Land-cOVEr Domain Adaptive semantic segmentation (LoveDA) [57] dataset provides semantic labels for remote sensing land-cover mapping, encompassing 5,987 high-resolution images with over 166,000 annotated objects collected from three cities in China. It contains seven common land-cover classes: *buildings*, *roads*, *water*, *forest*, *agriculture*, *barren land* and *background*. It presents inherent challenges due to its diverse coverage of 18 complex urban and rural areas, leading to significant scale variations for objects within the same class across different scenes. Additionally, the dataset exhibits inconsistent class distributions, with urban areas containing more artificial structures like buildings and roads, while rural areas present a higher proportion of natural elements such as forests and agricultural land. Encompassing images from two different domains, LoveDA challenges models to develop a robust, generalized understanding of land cover classes that can effectively perform across both urban and rural environments. This is particularly valuable to develop land cover mapping applications or an OVS systems that need to operate seamlessly across diverse geographical settings, from densely populated cities to expansive rural landscapes. However, being more focused on land-cover classes, this dataset presents coarser annotations, especially on complex classes like buildings and roads. For this reason, when used as a test dataset in a CDTE framework (see Sec. 2.1.5), it could produce misleading results where the fine model predictions are not evaluated correctly because the ground truth is too rough. Also, some images are off-nadir, having a substantial prospective shift with respect to other common RS datasets. Some examples can be seen in Figure 2.7.

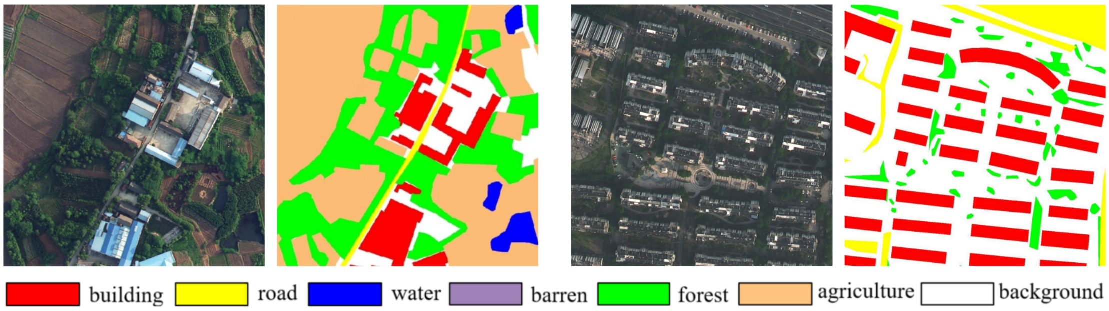


Figure 2.7: LoveDA's sample images and corresponding masks from rural (left) and urban (right) domains.

2.3.5 FMARS

The FMARS [58] dataset includes a compact yet meticulously manually annotated test subset, used to facilitate rigorous evaluation of its semantic segmentation models. This partition comprises 45 high-resolution satellite images sourced from 15 diverse geographical areas. It encompasses four distinct semantic classes: *road*, *tree*, *building* and *background*. This manually curated subset is an excellent way to evaluate RS segmentation models due to its reliable ground truth labels and the inherent geographical diversity of its images. This design not only evaluates segmentation accuracy but also challenges the models' ability to generalize across different domains, providing insights into their robustness against domain shift, a critical factor in real-world remote sensing applications. Clearly, given its small size, this dataset is not suitable to perform any training. However, we can test our models on it, according to the CDTE methodology, to assess their performances on a different domain dataset. Figure 2.8 displays some FMARS's sample images.

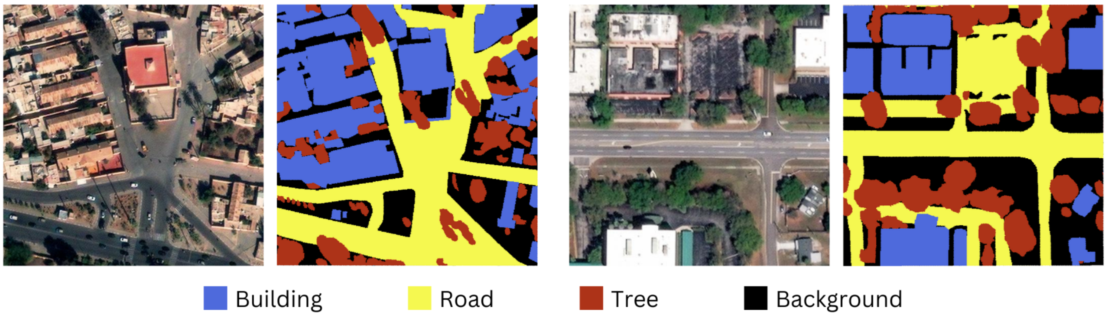


Figure 2.8: Sample images and their corresponding semantic segmentation masks from the FMARS dataset, depicting terrain from Morocco (left) and Florida (right).

Chapter 3

Methodology

3.1 Problem Statement

Remote sensing through satellite imagery has become an invaluable tool for Earth observation, enabling applications ranging from urban planning to environmental monitoring. However, traditional semantic segmentation approaches for satellite imagery are constrained by closed-set assumptions, where models can only identify and segment a predefined set of categories specified during training. This limitation becomes especially evident in RS images, where the diversity of ground features, temporal changes, and regional variations demand more flexible and adaptable capabilities. Our objective is to develop a semantic segmentation system able to identify both base and previously unseen classes across diverse geographical domains by leveraging an open vocabulary approach. It would enable category delineation through arbitrary textual descriptions and overcome the limitations of traditional closed-set paradigms. More formally, we aim to learn a mapping function f that takes as input an RGB image $I \in \mathbb{R}^{H \times W \times 3}$ and a set of user-provided semantic labels $C = c_1, c_2, ..., c_n$, where each c_i is a textual description of a class (e.g., "tree", "building", "car"). By default, this set is augmented with a background class c_0 to contain all the possible pixel assignments $C^+ = C \cup c_0$. This supplementary category should include all elements not explicitly specified in the user's query. The function should produce a semantic segmentation map $S \in 0{,}1, ..., n^{H \times W}$, where each pixel is assigned to one of the provided classes or the background class (denoted as 0). The mapping function f can be expressed as:

$$f : (C, I) \rightarrow S$$

It must be able to handle a variable number of input classes $|C|$, with no fixed upper limit and a minimum requirement of $|C| \geq 1$. Figure 3.1 shows a simple scheme of the pipeline. This formulation presents several challenges: the model must effectively bridge the gap between textual descriptions and visual features, maintain consistent performance across varying numbers of input classes, and properly handle the background class to avoid false positives. Moreover, the system should generalize to novel categories not seen during training while maintaining competitive performance on common ones. The following sections present our approach to addressing these challenges and detail the methodology used to develop a flexible, open-vocabulary semantic segmentation system for remote sensing applications.

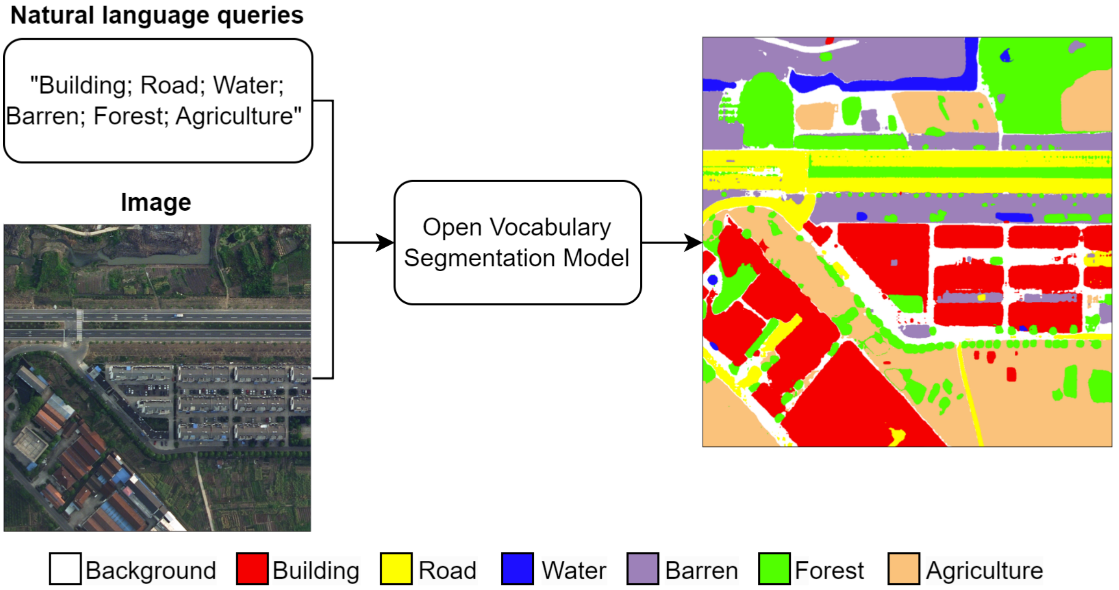


Figure 3.1: Overview of the OVS pipeline. The system takes two inputs: an aerial/satellite image and natural language queries specifying the target classes. The OVS model processes these inputs to produce a semantic segmentation map, where different colors represent distinct categories specified in the query.

3.2 Baseline framework

This section presents and explains in detail the baseline models that form the foundation for our adaptations to the RS field. The focus is posed on two fundamental models: FC-CLIP and RemoteCLIP. This in-depth study offers the necessary context to fully understand our proposed solution presented in later sections.

3.2.1 FC-CLIP

FC-CLIP is an open-vocabulary model, able to perform semantic, instance and panoptic segmentation tasks. As seen in Figure 3.2, it employs a single-stage framework built on a frozen convolutional CLIP backbone, offering a simpler and more efficient alternative to existing two-stage pipelines. The model's architecture comprises three main components: a class-agnostic mask generator, an *in-vocabulary* classifier, and an *out-of-vocabulary* classifier. By leveraging a shared frozen convolutional CLIP backbone for all three components, FC-CLIP achieves near state-of-the-art performance while significantly reducing computational costs and model parameters. A key contribution of this work is the empirical demonstration that convolutional CLIP architectures exhibit enhanced cross-resolution generalization capabilities in comparison to their transformer-based counterparts. This enables the model to process high-resolution images efficiently. The model's design splits into two branches to improve *in-vocabulary* and *out-of-vocabulary* classification.

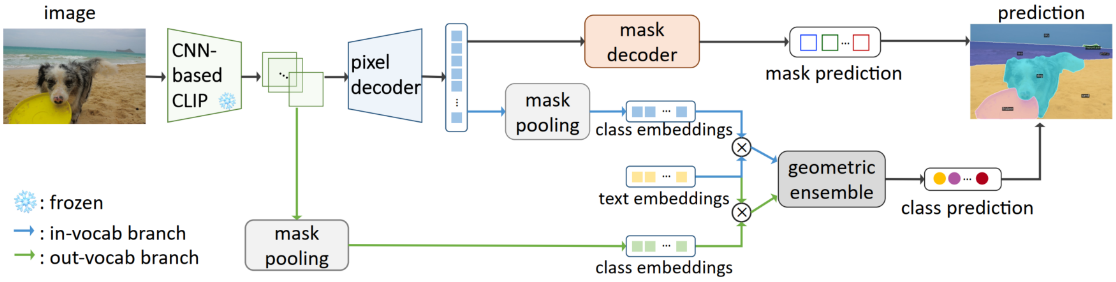


Figure 3.2: FC-CLIP scheme illustrating its main components: the frozen backbone, the pixel decoder, the *in-vocab* branch and the *out-vocab* one, and the mask decoder. Notice that the masks extracted by the mask decoder are fed into the two mask pooling modules, it is not shown for simplicity. (Source: [42]).

Pixel Decoder - At inference time, visual features extracted by the CLIP backbone are fed into a pixel decoder employing multi-scale deformable attention [59], similar to the approach used in Mask2Former [12]. In FC-CLIP, this decoder operates on three different scales of features. It produces *pyramidal* features along with *pixel-wise* features, which is a high-resolution features map that refines and summarizes information extracted at different levels.

Mask Decoder - The features extracted by the pixel decoder are subsequently used by the mask decoder module. This component consists of a series of mask decoders, each of them handling features at different scales. A single decoder comprises: (i) a masked cross-attention layer [60] operating on the *pyramid* feature at the corresponding level, (ii) a self-attention [9] layer, (iii) a feed-forward network layer. The mask decoders generate query vectors, which are matrix-multiplied with the *pixel* features to produce mask logits (one mask for each query vector) and, consequently, the mask prediction. Each mask prediction is matched with a ground truth mask via Hungarian matching [61], enabling the computation of the loss function. Differently from the usual Mask2Former implementation, unmatched masks are not penalized. This modification encourages the network to propose a diverse set of masks, a characteristic that is advantageous in the context of OVS.

The classification is achieved by computing cosine similarity between vectors associated with masks (class embeddings, v) and vectors associated with class names (text embeddings, t). The text embeddings are generated by processing class names through the CLIP textual encoder. On the other hand, the class embeddings are independently predicted by the two distinct branches: the *out-vocab* and *in-vocab* branches.

Out-vocab branch - The *out-vocab* branch employs a straightforward approach. It utilizes the mask predictions from the mask decoder to perform a mask pooling operation on the CLIP backbone features. This process yields class embeddings, $v_{i,out}$, each corresponding to a specific mask. This branch is exclusively used during inference since it is not trainable. The backbone is frozen, and the mask pooling operation is implemented as an average pooling, considering only the spatial positions relevant to the querying mask. This design choice constitutes the strength and purpose of this branch, aiming to mitigate potential biases that might be introduced through the training of the alternative branch of the network.

In-vocab branch - The *in-vocab* branch obtains class embeddings $v_{i,in}$ by applying mask pooling on the *pixel-wise features* of the pixel decoder. These can be thought as high-resolution, refined versions of CLIP features, specifically trained to respond more effectively to *in-vocab* classes when mask pooled. The mask pooling procedure is applied to these features using the same spatial aggregation methodology as employed on the CLIP backbone features.

To obtain class predictions, class embeddings vectors $v_{i,out}$ and $v_{i,out}$, extracted by the two branches, are independently used to compute Equation 3.1

$$\hat{c}_i = \text{softmax}\left(\frac{1}{T}\left[\cos(v_i, t_1), \cos(v_i, t_2), \cdots, \cos(v_i, t_{|C|})\right]\right) \tag{3.1}$$

where C represents the set of classes, T is a learnable temperature parameter and *cos* is cosine similarity. The class predictions from the two branches are then unified via Equation 3.2

$$\hat{c}_i(j) = \begin{cases} (\hat{c}_{i,in}(j))^{(1-\alpha)} \cdot (\hat{c}_{i,out}(j))^{\alpha}, & \text{if } j \in C_{train} \\ (\hat{c}_{i,in}(j))^{(1-\beta)} \cdot (\hat{c}_{i,out}(j))^{\beta}, & \text{otherwise} \end{cases} \tag{3.2}$$

where $\hat{c}_i(j)$ indicates the final score of the class j in the mask i, the addition of the subscript *in* and *out* indicate if the score is computed by the in or *out-vocab* branch and α, $\beta \in [0, 1]$ balance the weights given to the two branches.

3.2.2 RemoteCLIP

RemoteCLIP [49] is a domain-specific adaptation of the CLIP model for RS applications. A significant challenge in adapting foundation models like CLIP to specialized domains is the requirement for large-scale, domain-specific image-text pair datasets. However, as previously mentioned, such resources are scarce in the RS field. To address this limitation, RemoteCLIP employs data scaling techniques. It propose methods such as Box-to-Caption(B2C) and Mask-to-Box (M2B) conversions to leverage and unify heterogeneous annotations from various remote sensing datasets. Using B2C technique, it generates multiple textual descriptions based on detection annotations. Meanwhile, the M2B method converts segmentation datasets to detection datasets to ultimately get a textual description of the starting image. Overall, 17 datasets have been used, comprising: (i) three retrieval datasets (RET-3), including RSICD [62] and RSITMD [63], (ii) ten detection datasets (DET-10), including DOTA [64] and DIOR [65], (iii) four segmentation datasets (SEG-4), including iSAID [56] and LoveDA [56]. This approach allows RemoteCLIP to create a training dataset that is significantly larger and more diverse than any other existing image-text pairs collection in the remote sensing domain. RemoteCLIP is released in three versions, each leveraging different backbone architectures derived from the corresponding OpenAI CLIP models. These include a RemoteCLIP ResNet-50 (38 million parameters), a RemoteCLIP ViT-Base-32 (87 million parameters) and a RemoteCLIP Vit-Large-14 version (304 million parameters), respectively the small, medium and big versions. RemoteCLIP has been extensively evaluated across multiple downstream tasks and datasets, demonstrating its versatility and effectiveness as a vision-language foundation model for remote sensing applications. Some evaluation tasks it underwent include cross-modal retrieval, object counting and zero-shot image classification. The results from these tasks underscore RemoteCLIP's effectiveness as a foundation model for remote sensing applications. Figure 3.3 provides a visual comparison between CLIP and RemoteCLIP's responses to some textual queries. The heat maps, generated by processing 4096 image patches with 1/3 overlap alongside the text query, highlight the superior precision of RemoteCLIP in interpreting remote sensing imagery.

3.3 Remote FC-CLIP

Our custom solution, adapted for the RS domain, leverages FC-CLIP as the foundational architecture. This choice is driven by a series of factors: (i) consistent performance metrics on RS benchmark datasets, (ii) architectural advantages, and (iii) model scalability. FC-CLIP exhibits robust performance on satellite images, particularly with its largest version. It is the most recent architecture, among

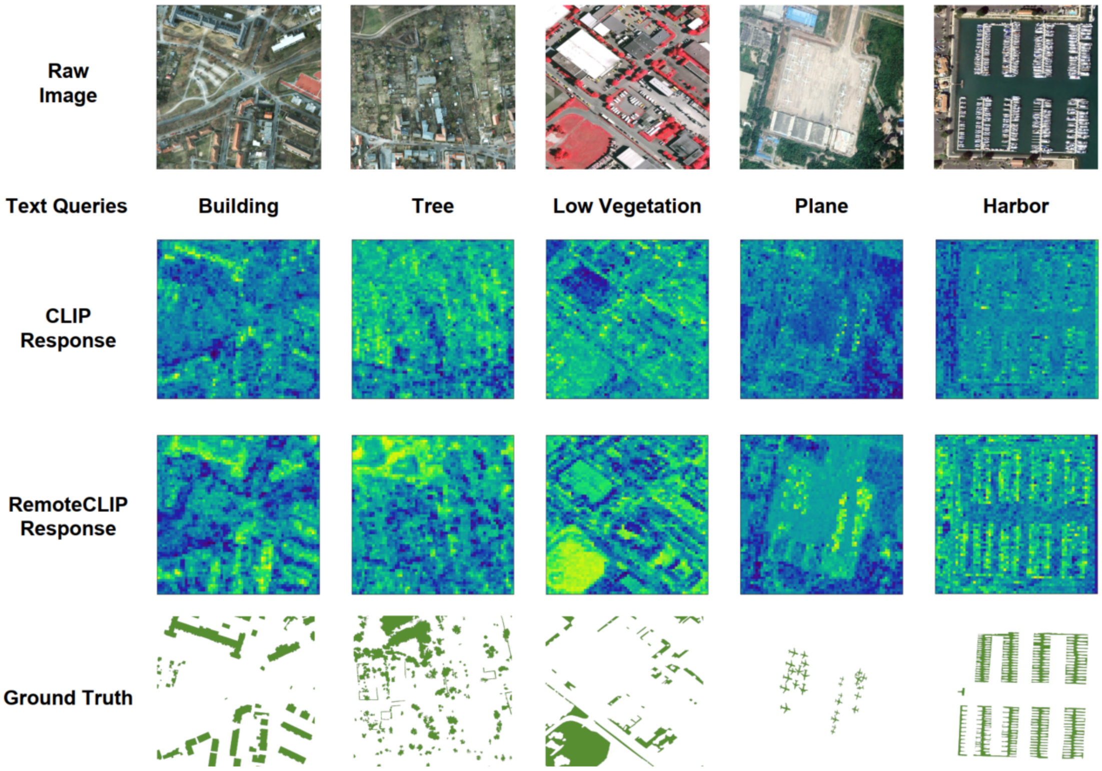


Figure 3.3: Comparative analysis of CLIP vs RemoteCLIP responses to different text queries. Heat maps are generated by feeding the model 4096 patches of the image, with 1/3 overlap, along with the text query. (Source: [49]).

the considered ones, incorporating state-of-the-art techniques. Furthermore, it is available in six variants with different model sizes, allowing us to choose the version that best meets our necessities. To adapt FC-CLIP for RS applications, we replace its CLIP components with their counterparts from RemoteCLIP. Specifically, we integrated the visual backbone and text encoder of a ResNet50-based RemoteCLIP model into the FC-CLIP architecture, replacing the original components (see Figure 3.4). The availability of a ResNet50 variant of the original model facilitated a seamless transition to the remote version. As detailed in Section 3.2.1, the utilization of a convolution-based CLIP architecture, as opposed to a ViT-based one, is crucial to scales better inference on more resolute images. After this architectural change, the decoder's pre-trained weights are no longer compatible with the new encoder configuration since their alignment has been disrupted. For this reason, we perform fine-tuning using the OEM dataset. We initialize our model's decoder with the pre-trained weights from the available FC-CLIP ResNet50 version and conduct a brief training session. This limited fine-tuning approach was chosen to mitigate the risk of overfitting on the relatively small OEM dataset while still allowing for the encoder-decoder alignment. To assess the effectiveness of this procedure, we conducted a series of evaluations on multiple datasets. Results are presented in Section 4.2.2. It is important to note the potential biases that might be introduced by this fine-tuning process. The chosen dataset, being relatively small and comprising only 8 segmentation classes, may skew the model towards these specific categories, potentially compromising its generalization ability and diminishing its open vocabulary capabilities. We anticipate that the mask decoder may exhibit a preference for extracting segmentation masks rather than instance masks, particularly for the classes present in the OEM dataset. While similar biases are also present in the original FC-CLIP weights, they were less pronounced when tested on natural images due to the high diversity in training datasets, such as COCO [47] and ADE20K [46].

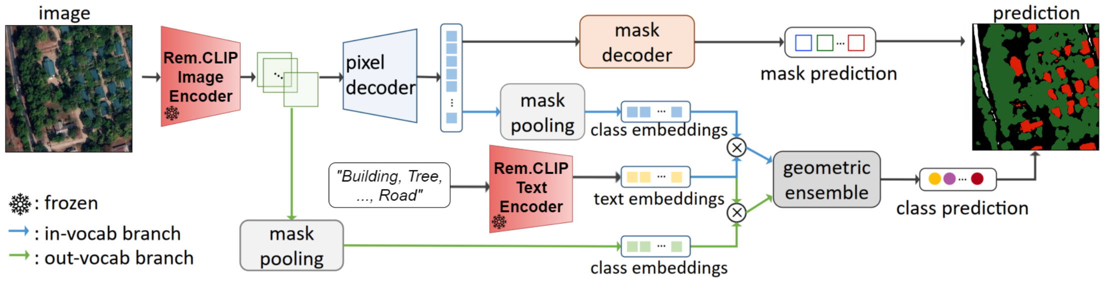


Figure 3.4: Architectural scheme of the Remote FC-CLIP model. Red modules indicate modifications from the base model.

3.4 SAM-FC-CLIP

To mitigate potential biases that could affect Remote FC-CLIP, we explore an alternative solution. The absence of a comprehensive dataset with a large number of distinct object categories presents a significant challenge in obtaining a general agnostic mask extractor for satellite imagery. Training on a single segmentation or instance dataset inherently limits the network's ability to generalize, as it becomes biased towards objects present in those data. To address this limitation, we shift our focus towards leveraging existing mask generator models while concentrating on the classification of the proposed masks. Our solution, outlined in Figure 3.5, integrates a SAM-based model (EfficientViT-SAM [15]) for mask generation with a classification module adapted from Remote FC-CLIP, similar in concept to the R-CNN architecture [17]. As described in Section 2.2.2, SAM offer a versatile approach to mask generation, allowing for both interactive and automatic mask extraction. It supports points or bounding boxes as user prompts, however, we are particularly interested in its automatic mask-generation capabilities. In this modality, it employs a grid of points as prompts. Each point leads up to three masks, which undergo a filtering and a post-processing phase before being returned to the user. Given the quantity and the variety of images SAM has been trained

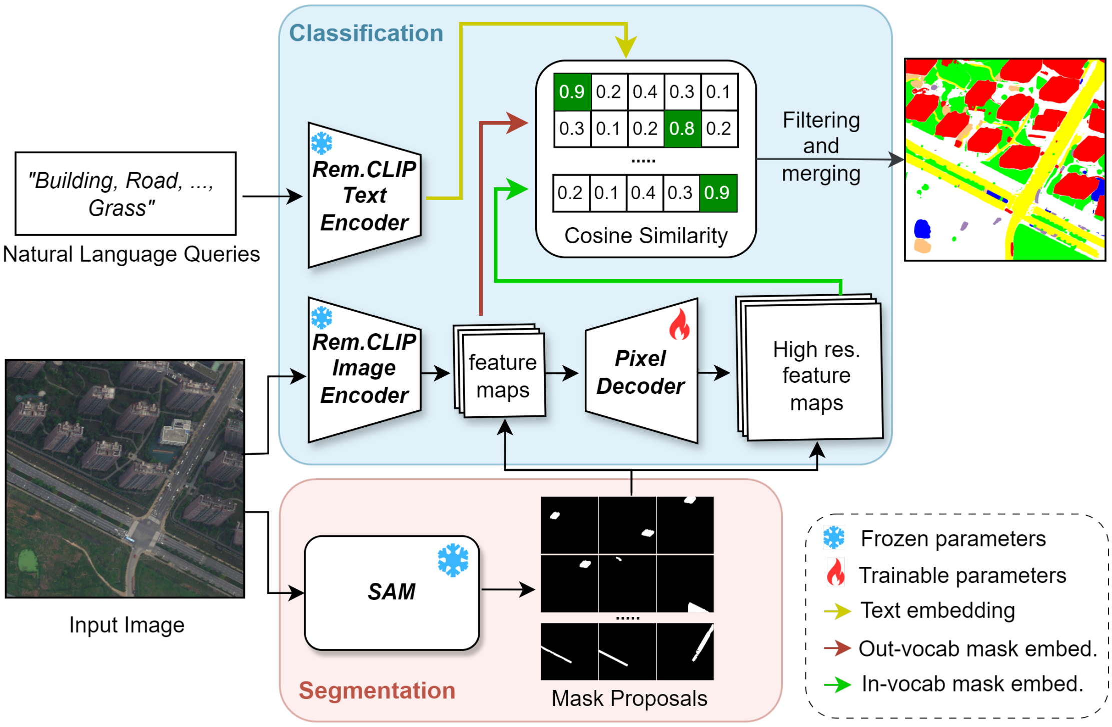


Figure 3.5: SAM-FC-CLIP scheme. The mask proposals generated by SAM are used to query the RemoteCLIP feature map and those enhanced by the pixel decoder. The resulting mask embeddings are compared with the text embeddings to find associations and classify each mask. The CLIP and SAM weights are frozen. The only trainable module is the Pixel Decoder.

on, keeping its weights frozen, the given masks will not be biased as in the previous approach. For this reason, we believe that it could obtain satisfactory results across various RS images. For the classification of the generated masks, we adapt the Remote FC-CLIP architecture removing its mask decoder. The remaining parts are the ResNet50 RemoteCLIP image and text encoder, the pixel decoder, and the in and out-vocab branches. A more comprehensive description of these parts is provided in Section 3.2.1. This classification module is fine-tuned specifically for remote sensing imagery, enabling accurate classification of the masks generated by SAM. The resulting model operates in two stages: SAM automatically generates masks from the input image and the adapted Remote FC-CLIP classifies them based on user queries. After classification, all masks are merged to obtain a segmentation prediction. Areas where no masks have been generated are considered as background. The efficacy of our classification and segmentation is strictly tied to the quality and completeness of the proposed masks. Hypothesizing a perfect mask classifier, we anticipate that suboptimal segmentation could arise from three factors: (i) sparse mask generation, resulting in insufficient masks for the classifier and consequently leaving areas unclassified, (ii) noisy masks including multiple objects of distinct classes, potentially confusing the classifier, and (iii) fragmented object representation, where a single object is partitioned into multiple masks. To mitigate these risks and optimize SAM mask proposals, we carried out an extensive hyperparameter search, more details are provided in Section 4.2.3. The mask classification network is trained using instance segmentation datasets, with a methodology analogous to the baseline model. However, the main difference lies in the utilization of ground truth masks to query the pixel decoder output, as opposed to proposals from the mask decoder. Furthermore, the way it operates allows training on an ensemble of instance datasets. More specifically, the pixel decoder features are optimized only in areas demarcated by the querying ground truth masks, effectively ignoring other regions. This design enables the use of partially labelled images without confusing the network, a feature particularly advantageous in the context of RS where the availability of comprehensive datasets is limited. Indeed, we can aggregate multiple datasets for training a single model. This two-stage approach combines SAM's generalized mask generation capabilities with the classification capabilities of the adapted Remote FC-CLIP model. By avoiding training the mask extractor, we aim to overcome the biases observed in our previous solution while maintaining a good segmentation quality in RS images.

3.5 Custom Ensemble Datasets

As explained in Section 3.4, our proposed model, SAM-FC-CLIP, can be trained on a mixture of various instance segmentation datasets with partial annotations. For example, we could fuse a dataset in which only roads are annotated with a dataset containing just building annotations. During training, the network will align the textual class embeddings with the corresponding object mask class embeddings, effectively ignoring unlabeled regions and not being misled by missing annotations. This characteristic is particularly advantageous in the domain of high-resolution RS, in which there are various datasets with partial annotations (e.g. [56, 55, 66, 67]). For our experiments, we constructed two mixture datasets, both derived from iSAID and OEM. The first, denoted as $iSAID{+}OEM_{building}$, encompasses all classes present in iSAID along with the building class from OEM. The second, referred to as *iSAID+OEM*, incorporates the complete class sets from both datasets. The first one is an instance-only dataset, meanwhile the second encompasses also *stuff* classes. Table 3.1 displays the complete set of classes contained in each dataset. Both mixture datasets contain 14506 images, 11506 derived from the patchified iSAID and 3000 from the OEM train partition. The iSAID dataset, being inherently designed for instance segmentation, is directly employable for our training procedure since it already provides individual masks for each instance object. In contrast, the OEM dataset is designed for semantic segmentation. While it would be feasible to extract and utilize entire class masks for training, this approach could potentially lead to suboptimal inference performance, as the mask classifier would be conditioned to expect semantic segmentation masks. This becomes problematic especially when employing SAM as mask generator, given its tendency to produce compact, globular masks rather than comprehensive class-wide masks. To adapt OEM to this purpose, given a segmentation mask for a specific class, we decompose it into independent instances by separating connected components. This process could lead to generating a substantial number of masks, some of which may be excessively small, lacking sufficient semantic information and potentially introducing noise into the model. To mitigate this issue, we filter and discard masks below a certain size threshold. We adopt two distinct strategies, resulting in two versions of the *iSAID+OEM* dataset: *iSAID+OEM*$_{1k}$ and *iSAID+OEM*$_{fine}$. For the first version, we eliminate all masks smaller than 1000 pixels. Meanwhile, for the second one, we apply class-specific thresholds based on the desired granularity level for each class. The chosen pixel area thresholds are the following: (i) 100

Dataset	**Source**	**Classes**
iSAID+OEM$_{building}$	iSAID	plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, swimming pool
	OEM	Buildings
iSAID+OEM	iSAID	plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, swimming pool
	OEM	Bareland, Grass, Pavement, Road, Tree, Water, Cropland, Buildings

Table 3.1: Composition of each mixture dataset, specifying the source of each class.

pixels for Bareland, Road, Water and Cropland classes, (ii) 200 pixels for the Building class, (iii) 500 pixels for the Grass, Pavement and Tree classes. These threshold values are empirically derived based on a comprehensive analysis of the data distribution, aiming to mitigate the class imbalance. Analyzing both Table 3.2 and Figure 3.6 it is possible to have insights about instance frequency and size across different classes. Specifically, the Grass, Pavement and Tree classes exhibit a high frequency of small, fragmented instances, while Bareland, Road, Water, Cropland, and Building classes tend to form fewer instances of cohesive and semantically meaningful structures. The impact of these thresholding strategies on instance counts is quantified in Table 3.2, while Figure 3.6 illustrates the resulting changes in average instance area across classes. This methodology for transforming a semantic segmentation dataset into a pseudo-instance dataset is generalizable. It can be applied to other segmentation datasets that one might wish to incorporate into the training of SAM-FC-CLIP. Figure 3.7 shows some samples of the resulting *iSAID+OEM$_{1k}$* and *iSAID+OEM$_{fine}$* datasets. They can be compared with the original segmentation masks in Section 2.3.2.

Class	$OEM_{\mathbf{w/o\ th}}$	$OEM_{\mathbf{fine}}$	$OEM_{\mathbf{1k}}$
Bareland	3,872	2,998	1,431
Buildings	242,557	193,933	89,119
Cropland	11,230	8,664	7,065
Grass	303,105	85,892	55,160
Pavement	242,943	75,960	52,086
Road	17,848	15,298	9,713
Tree	550,214	109,666	60,191
Water	11,548	7,964	2,529

Table 3.2: Number of instances for each class across OEM dataset variants. $OEM_{\text{w/o th}}$ indicates that no filtering has been applied in the instance conversion.

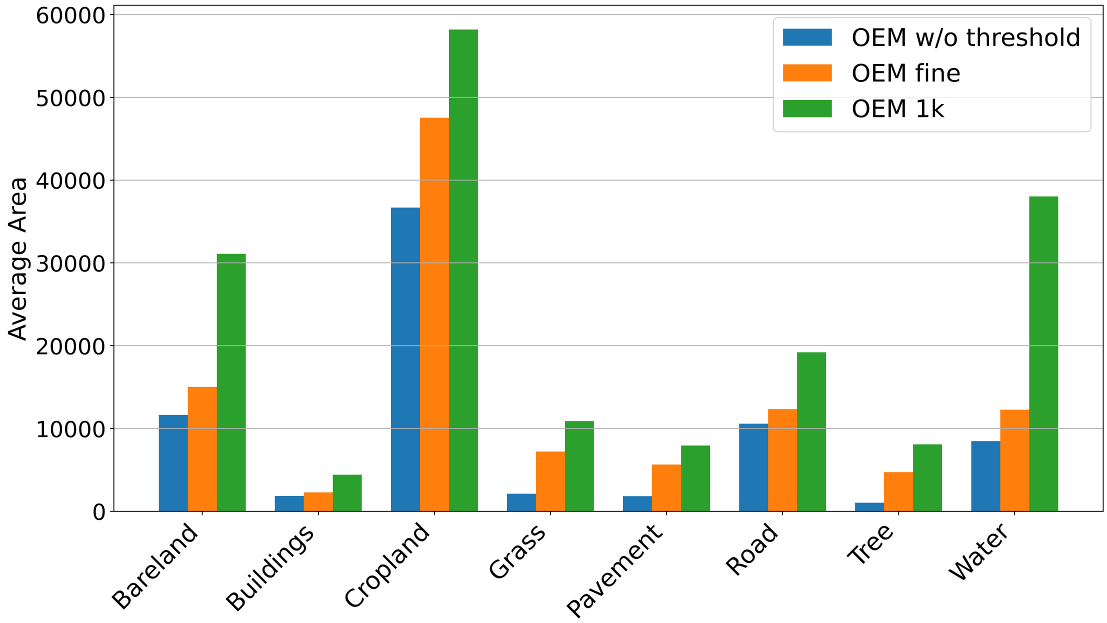


Figure 3.6: Average area for each class across OEM dataset variants. Areas are measured in number of pixels. The OEM w/o threshold indicates that no filtering has been applied in the instance conversion.

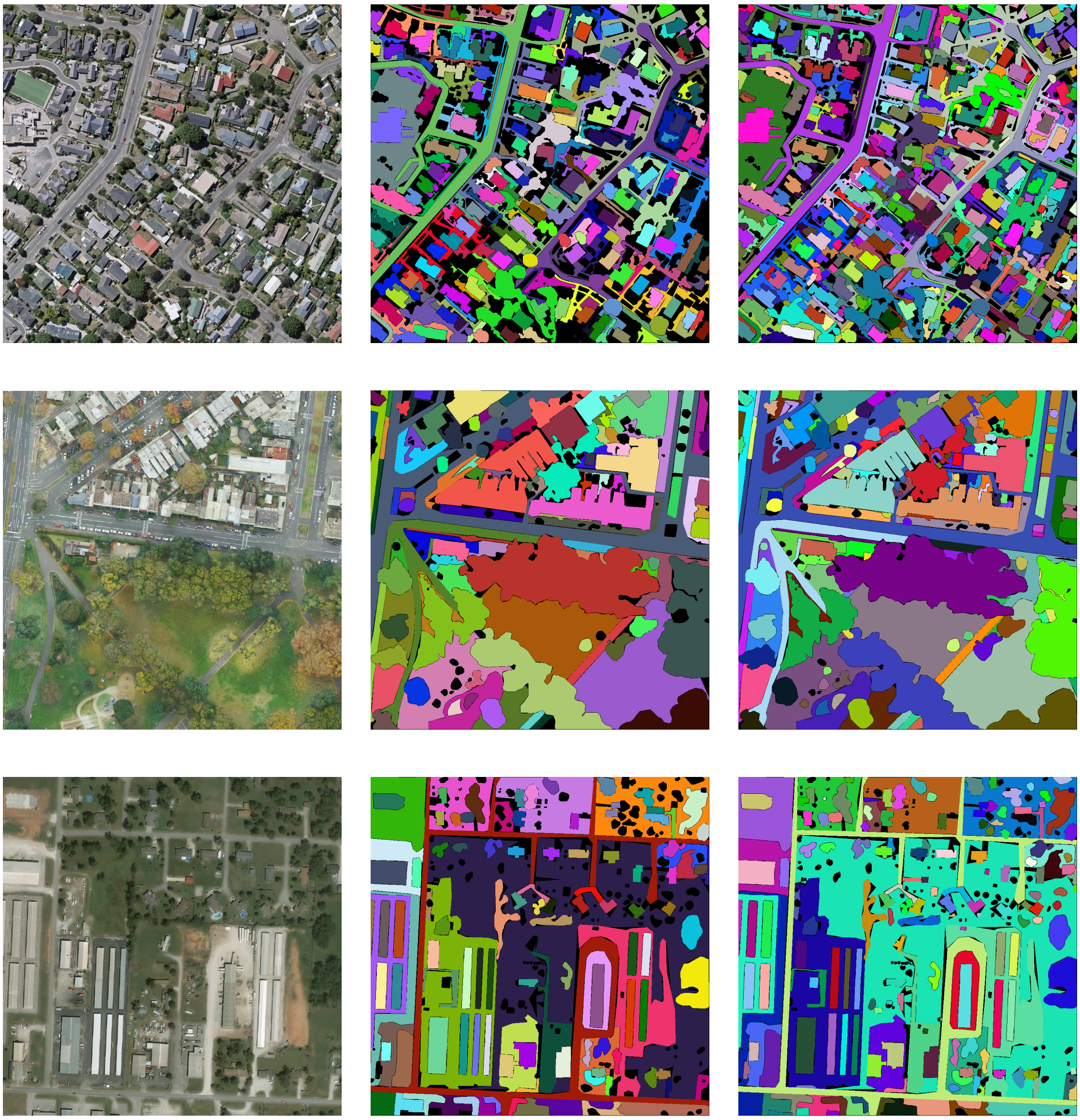


Figure 3.7: Three images from the OEM dataset variants. For each row, from left to right: natural image, OEM_{1k}, and OEM_{fine}. Black regions denote areas where no object instances are extracted. Notice how the *fine* partition presents more masks and fewer black areas.

Chapter 4

Experiments

4.1 Implementation Details

This section provides a comprehensive overview of the technical aspects involved in computing the baselines and implementing our two proposed approaches: Remote FC-CLIP and SAM-FC-CLIP. We detail training and inference choices along with hardware and software configurations.

4.1.1 Baselines

We establish baselines for OVS in RS field by evaluating state-of-the-art models designed for natural images. Specifically, we evaluate the four promising models already described in Section 2.2 and 3.2.1: OVSeg [39], MaskCLIP [41], MasQCLIP [53] and FC-CLIP [42]. For OVSeg, we employed the largest available model, which combines a Swin-Base backbone with a CLIP-ViT-L/14 vision transformer. In our evaluation of MaskCLIP, we utilized the only publicly available weights. For MasQCLIP, we examined two distinct checkpoints: the base-novel setting and the cross-dataset setting. The former was trained using instance segmentation annotations from COCO, while the latter leveraged panoptic segmentation labels. FC-CLIP is available in multiple checkpoints that primarily differ in the size of their CLIP backbone architecture. We tested both the smallest ResNet50-based and the largest, ConvNeXt-L-based, checkpoints. For evaluation, we selected two diverse and challenging RS datasets: OEM [55] and LoveDA [57]. To ensure consistency and to adhere to the standard practices in OVS, we employed class names as text queries for all models across both datasets. We follow this methodology even for the LoveDA's *background* class, whose name does not convey specific semantic information. However, in Section 4.1.2 we demonstrate an alternative way to account for it.

4.1.2 Remote FC-CLIP

As described in Section 3.3, our first solution consisted in fine-tuning an FC-CLIP model in which RemoteCLIP has substituted the CLIP backbone. Both training and inference have been conducted using Detectron2 [68], an open-source object detection and segmentation framework developed by Meta AI. Built on PyTorch, Detectron2 provides a robust environment for implementing and reproducing experiments with state-of-the-art computer vision models, making it suitable for our research objectives.

Architecture - The fine-tuning process was initialized using the pre-trained FC-CLIP RN50 weights, which were trained on the COCO dataset. The backbone is replaced with the RemoteCLIP RN50 version, ensuring seamless integration with the rest of the architecture. For this reason, the mask generator part remains consistent with the base model, comprising nine mask decoders and 250 object queries. Both mask and text embeddings maintain a dimensionality of 1024.

Training strategy - The training configuration is equivalent to that used in FC-CLIP. We employed the AdamW optimizer [69] with a constant learning rate of 1e-4 and a weight decay of 0.05. The model was trained with a batch size of 8 for a maximum of 10,000 iterations. To monitor the impact of fine-tuning duration, we evaluated and saved checkpoints at iterations 100, 500 and then every 1000. Given the limited number of iterations, we implemented a streamlined augmentation strategy. This consisted of random horizontal and vertical flips, each with a 50% probability. To address memory constraints, input images were resized to maintain the shortest side at 800 pixels.

Inference strategy - The α and β parameters, utilized for merging predictions from both branches, remain consistent with the baseline model. As for training, images are resized to have the shortest dimension to 800 pixels. Some datasets in our study include a *background* class, which lacks explicit semantic meaning. We employ FC-CLIP's panoptic post-processing to handle background classification. This method distinguishes between stuff and thing classes, inherently treating unclassified regions as background without requiring explicit text queries. We set it to process all semantically meaningful classes as stuff. The result is a series of unified segmentation masks where objects are not divided by instances and any non-extracted area is automatically classified as background. All training and inference procedures were conducted on NVIDIA A100 GPUs using MIG devices with 20GB of memory.

4.1.3 SAM-FC-CLIP

Architecture - As denoted in Section 3.4, this two-stage model uses SAM as mask extractor and the pixel decoder along with the in and out-vocabulary branches of FC-CLIP for classification. In particular, for the first stage, we employ a SAM variant called EfficientViT-SAM [15]. For the second one, we do not modify any component from the architecture of the base RN50 version, except by using the RemoteCLIP backbone instead of the CLIP one.

Training Strategy - For training, we utilized both custom datasets (*iSAID+ OEM*$_{1k}$ and *iSAID+OEM*$_{\mathit{fine}}$). Since the images in OEM are almost a quarter of those in iSAID, we implemented a balanced sampling strategy to ensure an equal probability of selecting from either dataset during training. For data augmentation, we adopted techniques similar to those employed in [22]. These include random horizontal and vertical flips, as well as a multi-scale approach where a random resize scale is uniformly sampled from the range [0.1, 2]. Subsequently, we applied a random crop of fixed dimensions (1024x1024) to the resized images. For the classification task, we employed the Cross Entropy (CE) loss function. Because of the class imbalance present in our dataset (as illustrated in Table 3.2), we conducted experiments using both uniform and weighted CE loss. The class-specific weight (w_c) for the weighted CE loss was computed as follows:

$$w_c = \frac{\sum_{i=1}^{n} |C_i|}{n \cdot |C_c|} \tag{4.1}$$

where n denotes the total number of classes and C_i represents the set of annotations for class i. This weighting scheme effectively increases the loss contribution of classes with fewer instances, thereby mitigating the impact of class imbalance. We used a batch size of 6 images. We employed the WarmupCosineLR learning rate scheduler, provided by Detectron2, which combines a warm-up phase with a cosine annealing phase. The learning rate starts at a low value ($\text{base}_{\text{lr}} \cdot \text{warmup}_{\text{factor}}$) to increase reaching base_{lr}, subsequentially it smoothly decreases to zero following a cosine curve. The base_{lr} was set to 2.5e-4, the warmup factor to 0.001 and the warm-up iterations to 200. This schedule allows for a stable training start and helps in finding a better optimum by gradually decreasing the learning rate. For optimization, we chose AdamW with weight decay set to 5e-5. The model was trained for 10,000 iterations, with evaluating and saving checkpoints every 1000.

Inference Strategy - For inference on OEM, LoveDA and FMARS we used two distinct SAM configurations found via random search. We conducted separate searches due to the significant disparity in image size between datasets, which we hypothesized could lead to divergent optimal values for certain parameters, such as points per side. Table 4.1 shows the hyperparameter search spaces. To

Hyperparameter	**Space 1**	**Space 2**
`points_per_side`	[32, 128]	[100, 128]
`pred_iou_thresh`	[0.0, 1.0]	[0.0, 1.0]
`stability_score_thresh`	[0.0, 1.0]	[0.0, 1.0]
`box_nms_thresh`	[0.0, 1.0]	[0.5, 1.0]
`crop_n_layer`	0 (*fixed*)	0 (*fixed*)
`min_mask_region_area`	0 (*fixed*)	0 (*fixed*)

Table 4.1: Ranges of the two random search hyperparameter spaces used respectively in stage 1 and 2.

mitigate their breadth, we constrained certain parameters. Specifically, we fixed `crop_n_layer` to zero, as our empirical observations indicated that applying the process to sub-crops of the image requires a significant amount of time without yielding notable improvements. Similar considerations led us to maintain the default value for `min_mask_region_area`. Given the dimensionality of the search space, for the LoveDA dataset we implemented a two-stage random search strategy. The initial stage, exploring space 1, involved 10 images and aimed to rapidly eliminate suboptimal parameter configurations. The subsequent stage, utilizing 40 images to explore space 2, allowed for a more granular analysis. For the FMARS dataset, given its relatively modest size, we employed a single-stage search in space 1 with 10 images. To justify our decision to use such small image sets for hyperparameter optimization, we note that certain configurations resulted in processing times up to 2 minutes per photo. Clearly, the objective function for the search was the maximization of mIoU obtained from classifying the extracted masks. Overall, we performed around 250 runs for LoveDA and 400 for FMARS. The generated instance masks, after classification, undergo a hierarchical merging process, where they are composited in descending order of area to preserve the visibility of smaller instances. This composition strategy ensures that fine-grained semantic details are retained in the final segmentation output. The semantic segmentation mask is constructed through a dual-thresholding mechanism for confidence scores. Specifically, we employ distinct confidence thresholds for in-vocabulary and out-of-vocabulary classes, set at 0.8 and 0.4 respectively. Regions that either lack instance masks or fall below these confidence thresholds are assigned to the background class. Following the baseline implementation, we utilize weighting coefficients $\alpha = 0.4$ and $\beta = 0.8$ for balancing the contributions of in and out-vocabulary classification branches. During inference, images are processed at their original resolution without any preprocessing transformations.

4.2 Results

This section presents the experimental results of our novel solutions for OVS in the remote sensing domain. We begin by evaluating several state-of-the-art models on the OEM and LoveDA datasets, establishing baseline performance metrics that reveal how models originally designed for natural images translate to satellite imagery. We then examine the results of our two proposed approaches: Remote FC-CLIP and SAM-FC-CLIP. For each model, we provide comprehensive quantitative metrics supported by qualitative examples.

4.2.1 Baseline Results

To establish a foundation for our research, we first evaluated the performance of existing OVS models on remote sensing data without any domain-specific adaptations. These baseline results serve as benchmarks for assessing our proposed improvements. We tested the following architectures: OVSeg [39], MaskCLIP [41], MasQCLIP [53] and FC-CLIP [42]. Their performance on the OEM and LoveDA datasets is summarized in Tables 4.2 and 4.3, respectively, with percentage per-class IoU and mIoU metrics. Analysis of the OEM dataset results (Table 4.2) reveals significant performance variations across models and land cover classes. OVSeg leads the performance metrics (mIoU: 26.40%), demonstrating particular strength in segmenting common land cover types: water bodies (58.27%), grass areas (33.22%), buildings (28.47%) and trees (27.73%). Both FC-CLIP variants achieve comparable overall performance (ConvNeXt-L: 25.55%, ResNet50: 25.70%), with the ResNet50 variant particularly excelling in tree segmentation (42.32%). MaskQCLIP's variant

Model	Var.	Bareland	Grass	Pavement	Road	Tree	Water	Cropland	Building	mIoU
OvSeg	-	0.00	**33.22**	**24.51**	15.35	27.73	58.27	23.69	28.47	**26.40**
MaskCLIP	-	0.77	10.65	3.98	11.24	29.22	39.03	27.59	28.91	18.92
MaskQCLIP	A	**1.95**	4.51	15.44	4.78	4.77	16.42	31.92	22.03	12.73
	B	0.27	31.85	8.25	7.06	6.90	59.71	**47.48**	33.37	24.36
FC-CLIP	C	0.00	17.39	3.74	19.91	**42.32**	50.43	35.01	36.80	25.70
	D	0.25	10.60	8.89	**31.47**	15.27	**64.55**	31.31	**42.06**	25.55

Table 4.2: Performance comparison on OEM dataset for various baseline models. MaskQCLIP variants A and B refer to the base-novel and cross-dataset settings, respectively. FC-CLIP variant C employs a ResNet50 backbone, while variant D a ConvNeXt-L one. Results are reported in percentage IoU.

B significantly outperforms version A (mIoU: 24.36% vs 12.73%), excelling in water and cropland segmentation (59.71% and 47.48%, respectively). This improvement can be attributed to variant B's broader training set (80 *thing* and 53 *stuff* classes) compared to variant A's limited scope (48 classes), demonstrating the importance of diverse training data for model generalization. MaskCLIP achieves an overall low mIoU of 18.92% on OEM. Waterbody segmentation consistently achieved high metrics. Two factors can explain this robust performance. First, water bodies exhibit minimal domain shift between natural and satellite imagery, as their appearance remains relatively consistent across these modalities. Second, water is included as a *stuff* class in the COCO dataset, which is commonly used for pretraining. Having established baseline performance on the OEM dataset, we next examined in Table 4.3 models behaviour on the LoveDA dataset, which presents different challenges in urban and rural scene understanding. FC-CLIP variant D emerges as the top performer (mIoU: 33.83%), surpassing MaskQCLIP version B by 5 percentage points. Unlike its balanced performance on OEM, the ConvNeXt-L backbone shows clear superiority over ResNet50 (33.83% vs. 22.69%) on LoveDA, particularly excelling in building and road segmentation (49.88% and 40.36%). MaskQCLIP's B variant metrics remain high (mIoU 28.48%), particularly excelling in the agricultural class (IoU 44.36%). As before, the A variant of MaskQCLIP shows significantly lower performance (mIoU 12.02%). OVSeg and MaskCLIP show similar overall results (23.13% and 24.01%) but with distinct strengths. OVSeg dominates water body segmentation (42.69%), while MaskCLIP maintains balanced metrics across urban categories. While agricultural lands show consistent quality across models (IoU: 37.82-48.47%), barren lands and background areas remain

Model	Var.	Background	Building	Road	Water	Barren	Forest	Agricultural	mIoU
OVSeg	-	0.72	29.94	12.82	42.69	7.63	26.00	42.14	23.13
MaskCLIP	-	7.68	30.72	24.90	32.35	8.27	24.15	40.01	24.01
MaskQCLIP	A	**35.09**	3.07	7.84	0.20	0.01	0.09	37.82	12.02
	B	19.46	29.92	27.30	42.81	10.70	24.78	44.36	28.48
FC-CLIP	C	5.77	26.19	12.42	39.54	2.90	29.71	42.29	22.69
	D	5.29	**49.88**	**40.36**	**47.50**	**11.10**	**34.24**	**48.47**	**33.83**

Table 4.3: Performance comparison on LoveDA dataset for various baseline models. MaskQCLIP variants A and B refer to the base-novel and cross-dataset settings, respectively. FC-CLIP variant C employs a ResNet50 backbone, while variant D a ConvNeXt-L one. Results are reported in percentage IoU.

particularly challenging for segmentation. These baseline results highlight both the potential and limitations of applying general-purpose OVS models to remote sensing tasks. While some classes transfer well from natural to satellite imagery, others show significant degradation, motivating our proposed domain-specific adaptations.

4.2.2 Remote FC-CLIP Results

We evaluated the Remote FC-CLIP's performance on the validation partitions of three datasets: OEM, LoveDA, and FMARS. To assess the impact of the fine-tuning duration we save a checkpoint and validate at iterations 100, 500 and then every 1000. We report quantitative and qualitative results for the checkpoints at iteration 100 and 10k, respectively variants C and D. Presenting the two extremes we aim to identify potential overfitting on training classes caused by extended fine-tuning. Also, at the end of the section, we plot the performance trend across training iterations for all three datasets to demonstrate how fine-tuning affects each of them differently.

OEM - Table 4.4 presents a performance comparison on the OEM dataset. Our fine-tuned models demonstrate substantial gains over the baselines (FC-CLIP ConvNeXt-L and RN50) across the majority of classes. Version D achieves the highest overall performance with a mIoU of 65.50%, allowing for an increase of almost 40 percentage points over FC-CLIP (mIoU of 25.55% for ConvNeXt-L and 25.70% for RN50). The C variant also shows considerable enhancement, reaching a mIoU of 47.54%. Analysis of category-wise results reveals that Remote FC-CLIP obtains notable improvements in challenging classes such as *bareland* and *pavement*. For instance, the IoU for *bareland* increases from 0.25% (ConvNeXt-L) and 0.00% (RN50) to 45.54% using our D version. Similarly, *pavement* recognition progresses from 8.89% (ConvNeXt-L) and 3.74% (RN50) to 56.09% through variant D. Despite these advancements, these categories remain the most difficult, even after fine-tuning. On the other hand, labels like *tree*, *water*, *cropland* and *building* exhibit the highest scores. The results indicate that increasing the fine-tuning iterations generally leads to better performance on this dataset. However, it is important to contextualize these improvements within the experimental setup. These enhanced performances were expected given that Remote FC-CLIP has been fine-tuned on images from the same distribution as the validation partition. In contrast, the baseline models were primarily trained on natural images, potentially limiting their generalization to this specific domain. Figure 4.1 provides qualitative results across industrial, rural, and urban areas, offering insights into segmentation capabilities in diverse scenarios. The improvements achieved through fine-tuning are notable. Remote FC-CLIP exhibits greater precision in delineating individual buildings (represented in red), whereas the baseline tends to over-segment. Other notable improvements include better discrimination between roads and pavement (white and grey, respectively) and superior detection of small objects, as evidenced by accurate segmentation of the small blue pond in the second row and consistent identification of individual trees across all three segmentations. These fine-grained details proved challenging for the baseline model, likely due to its training dataset's bias toward larger segmentation masks.

Model	Var.	Bareland	Grass	Pavement	Road	Tree	Water	Cropland	Building	mIoU
FC-CLIP	A	0.00	17.39	3.74	19.91	42.32	50.43	35.01	36.80	25.70
	B	0.25	10.60	8.89	31.47	15.27	64.55	31.31	42.06	25.55
Remote	C	16.74	39.37	34.08	48.78	62.26	51.44	62.96	64.71	47.54
FC-CLIP	D	**45.54**	**57.22**	**56.09**	**63.25**	**71.47**	**76.14**	**75.39**	**78.94**	**65.50**

Table 4.4: Performance comparison on OEM dataset between FC-CLIP baseline and our Remote FC-CLIP. Model variants are: A (ResNet50 backbone), B (ConvNeXt-L backbone), C (ResNet50 with 100 training iterations), and D (ResNet50 with 10k training iterations). Results are reported as IoU% scores.

LoveDA - To better assess our model generalization capabilities, we evaluate the validation partition of the LoveDA dataset. Table 4.5 summarizes the quantitative results, demonstrating that our fine-tuned model outperforms the baseline architecture by approximately 9% in mIoU. This dataset allows us to evaluate three novel (barren, forest, and agricultural) and three seen classes (building, road, and water) coming from a different distribution. Examining the performance across these two sets provides insights into the model's open-vocabulary capabilities. As expected, previously encountered categories exhibited higher IoU. Notably, while the results on unseen classes were suboptimal, they still surpassed the established baseline. We can infer the effect of the extended fine-tuning by comparing the performance of the two Remote FC-CLIP variants. Version D achieves higher metrics on classes present in OEM (49.10% vs. 40.54% for version C of mIoU on *building*, *road*, and *water*), while the variant with fewer training iterations prevails on unseen categories with an average increase of 3.22 percentage points. This pattern suggests that extended fine-tuning may lead to overfitting on seen labels, potentially at the expense of generalization to new ones. However, Remote FC-CLIP still outperforms the baseline on novel classes, indicating retention of some open vocabulary capabilities. Also, Figure 4.4 displays how the improvement in overall performance from extended fine-tuning is less pronounced on the LoveDA compared to OEM. Figure 4.2 shows some inferences in urban and rural areas revealing notable differences between the model predictions and the ground truth annotations. Both the baseline and our Remote FC-CLIP struggle to accurately classify previously unseen classes.

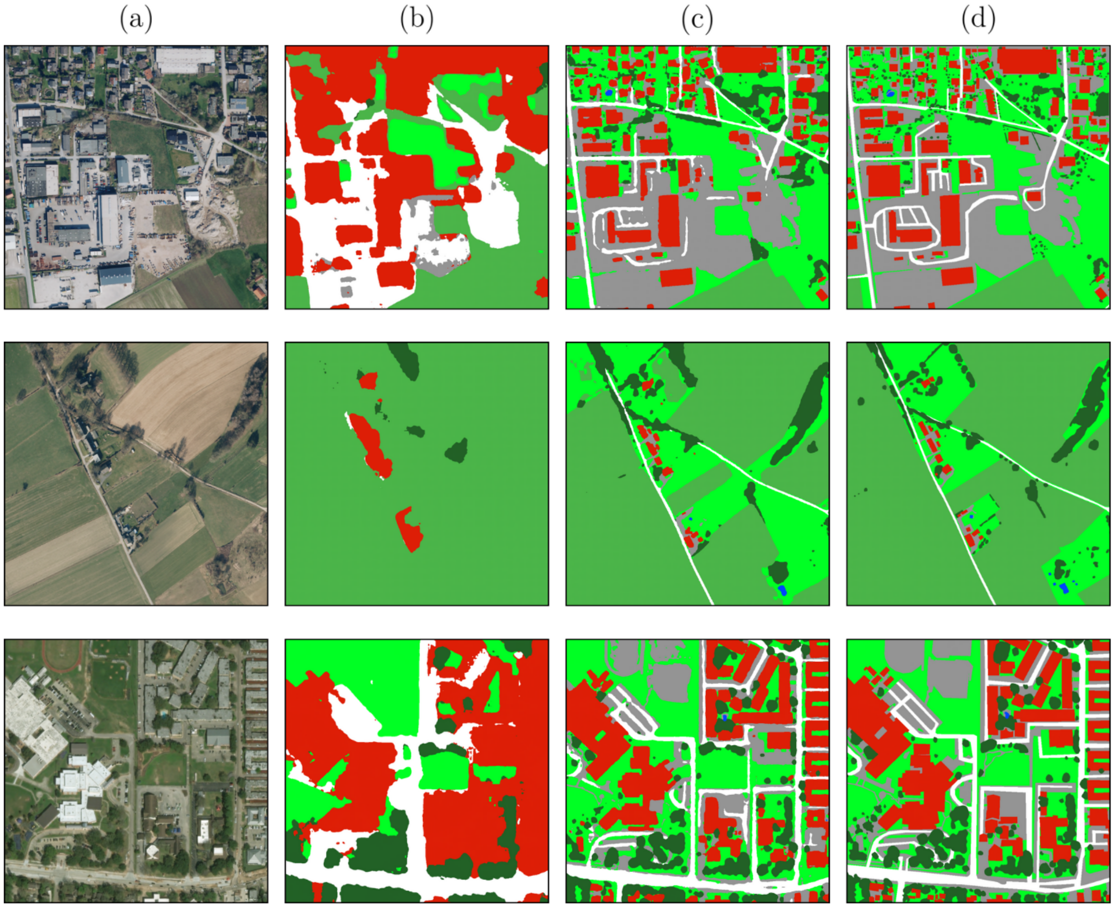


Figure 4.1: Qualitative comparison of semantic segmentation results on three satellite images from the OEM validation partition. From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our Remote FC-CLIP model (10k training iter.), and (d) ground truth segmentation masks. The images showcase industrial, rural and urban scenes, illustrating models performances across diverse areas.

Specifically, in these examples, neither successfully identified the agricultural (in orange) or barren (in purple) land cover types. However, our fine-tuned version demonstrated improved performance in delineating the forest class (in green). For known categories, both models achieved a segmentation precision that appears higher than the provided ground truth annotations, successfully segmenting entire building structures rather than solely their footprints and identifying road networks that were not explicitly labelled in the ground truth data.

FMARS - This dataset contains manually annotated ground truth labels, making

Model	Var.	Background	Building	Road	Water	Barren	Forest	Agricultural	mIoU
As-is	A	35.29	38.79	37.92	28.72	0.00	8.12	0.38	21.32
	B	36.08	25.39	46.48	**51.48**	0.07	11.11	0.16	24.39
Remote FC-CLIP	C	**38.88**	49.86	41.32	30.45	**20.37**	32.04	3.16	30.87
	D	37.83	**59.13**	**47.57**	40.61	5.11	**34.45**	**6.33**	**33.00**

Table 4.5: Performance comparison on LoveDA dataset between FC-CLIP and our Remote FC-CLIP. Model variants are: A (ResNet50 backbone), B (ConvNeXt-L backbone), C (ResNet50 with 100 training iterations), and D (ResNet50 with 10k training iterations). Results are reported as IoU% scores. Note: background class is computed via panoptic inference.

it a particularly reliable benchmark for assessing model performance. These data enable the evaluation of Remote FC-CLIP on known classes from a novel distribution of images. Our model surpasses the baseline, especially on the tree and buildings classes, with an overall mIoU improvement of almost 6%, as seen in Table 4.6. Consistent with previous observations, extended fine-tuning yielded enhanced performance on seen categories, with performance gains plateauing after a few training iterations, as shown in Figure 4.4. Figure 4.3 presents qualitative results across urban and rural scenes. While both architectures demonstrate effective segmentation capabilities, our model exhibits exceptional performance in road delineation, as confirmed by quantitative results (achieving a mIoU of 54.66%). Despite some roads being tiny, our model's predictions closely align with the ground truth labels.

Model	Var.	Road	Tree	Building	Background	mIoU
FC-CLIP	A	27.66	63.09	49.16	27.75	41.91
	B	53.25	61.29	54.88	37.24	51.67
Remote FC-CLIP	C	39.44	**70.84**	58.46	43.80	53.13
	D	**54.66**	62.49	**67.59**	**45.70**	**57.61**

Table 4.6: Performance comparison on FMARS dataset between FC-CLIP and our Remote FC-CLIP. Model variants are: A (ResNet50 backbone), B (ConvNeXt-L backbone), C (ResNet50 with 100 training iterations), and D (ResNet50 with 10k training iterations). Results are reported as IoU% scores.

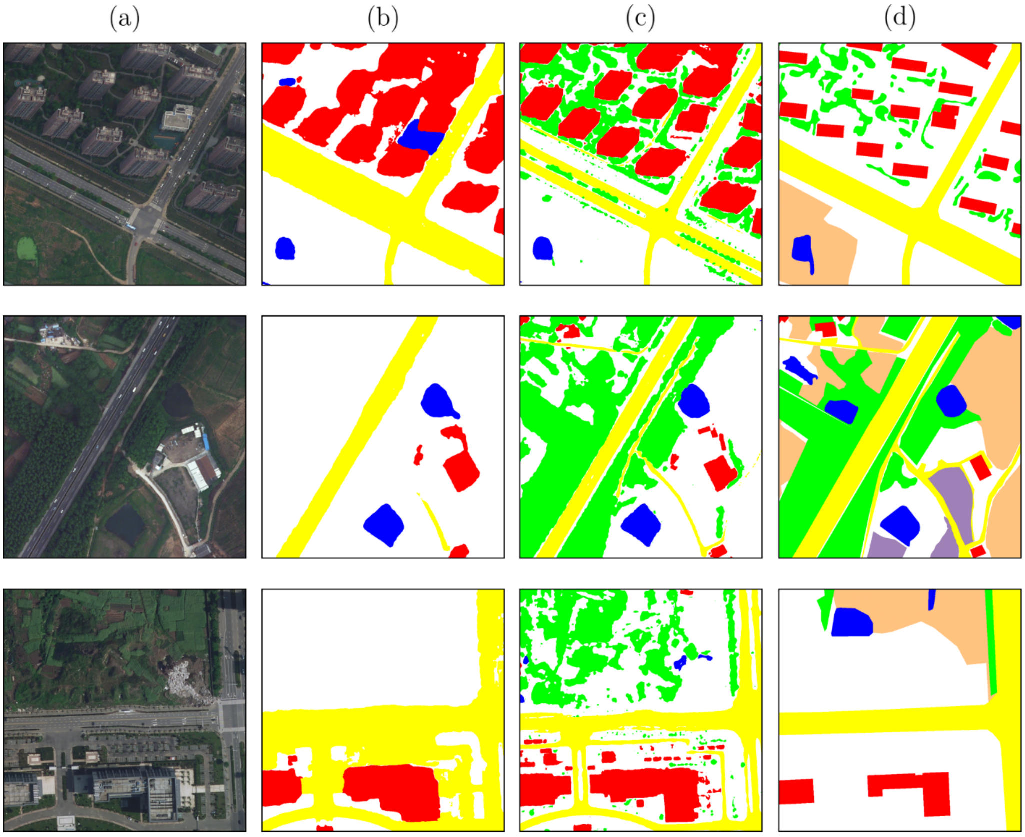


Figure 4.2: From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our Remote FC-CLIP model (10k training iter.), and (d) ground truth segmentation masks.

Despite these promising results, our analysis reveals certain limitations in the current implementation. As described in [39], an important consideration when evaluating systems with separate mask extraction and classification phases is identifying the primary performance bottleneck. While natural image systems typically struggle with mask classification due to the abundance of general segmentation datasets for training mask generators, our findings suggest the opposite in the RS domain. We argue that the main limitation lies in the mask generator component. Fine-tuning on a relatively small semantic segmentation dataset has introduced biases in Remote FC-CLIP's decoder mask queries toward seen shapes, compromising the generalization capabilities of the agnostic mask extractor. This is evident in the model's inability to detect smaller objects absent from the OEM dataset, like vehicles, making it less suitable for true open-vocabulary applications. This limitation partially explains the degradation in open-vocabulary performance, resulting in a model that excels with seen classes but struggles with out-of-domain

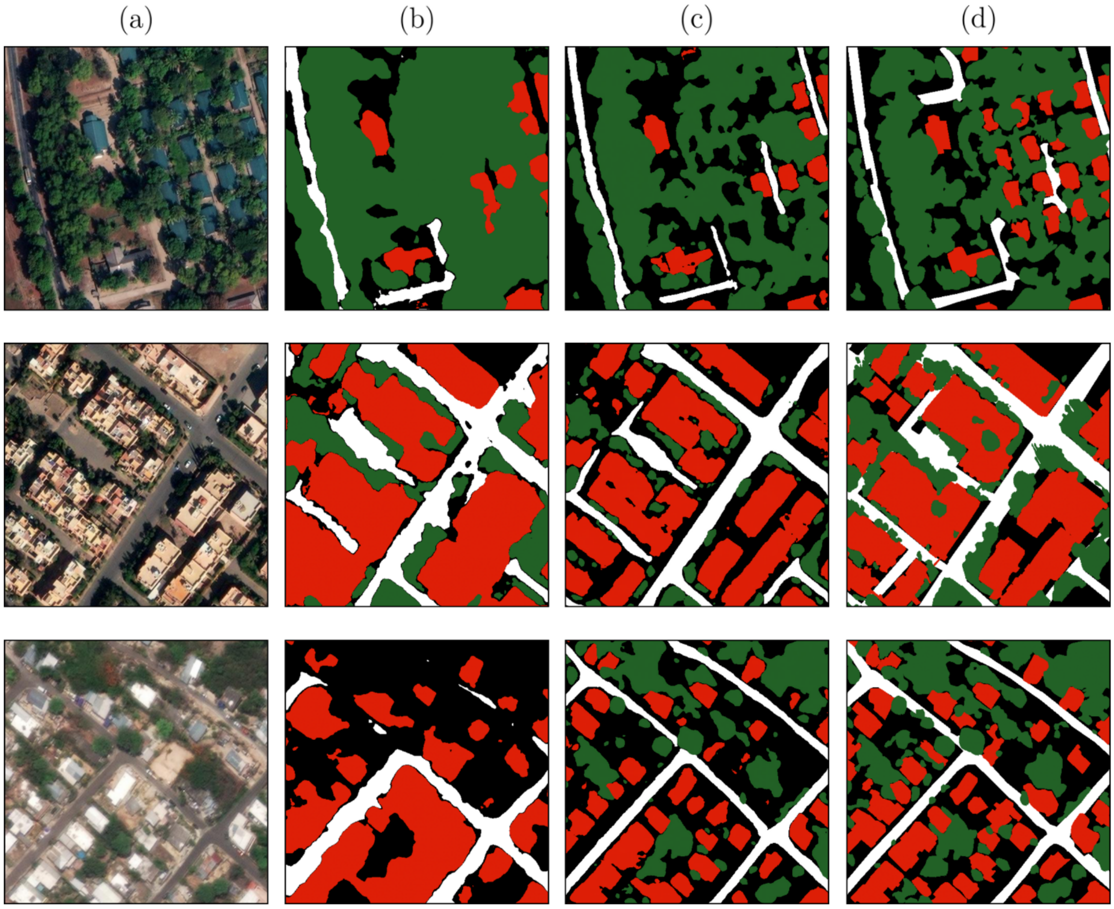


Figure 4.3: From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our Remote FC-CLIP model, and (d) ground truth segmentation masks. Note: background class is computed via panoptic inference.

categories. These findings underscore that the lack of a comprehensive and general dataset is a fundamental challenge in adapting OVS models to remote sensing employing Remote FC-CLIP's strategy. While promising, our proposed adaptation's effectiveness on novel classes is constrained by the current limitations in available training data.

4.2.3 SAM-FC-CLIP Results

This section presents the experimental results obtained with our SAM-FC-CLIP model. We begin by detailing the optimal hyperparameters for SAM, which plays a crucial role in our mask extraction pipeline. We then evaluate our model's

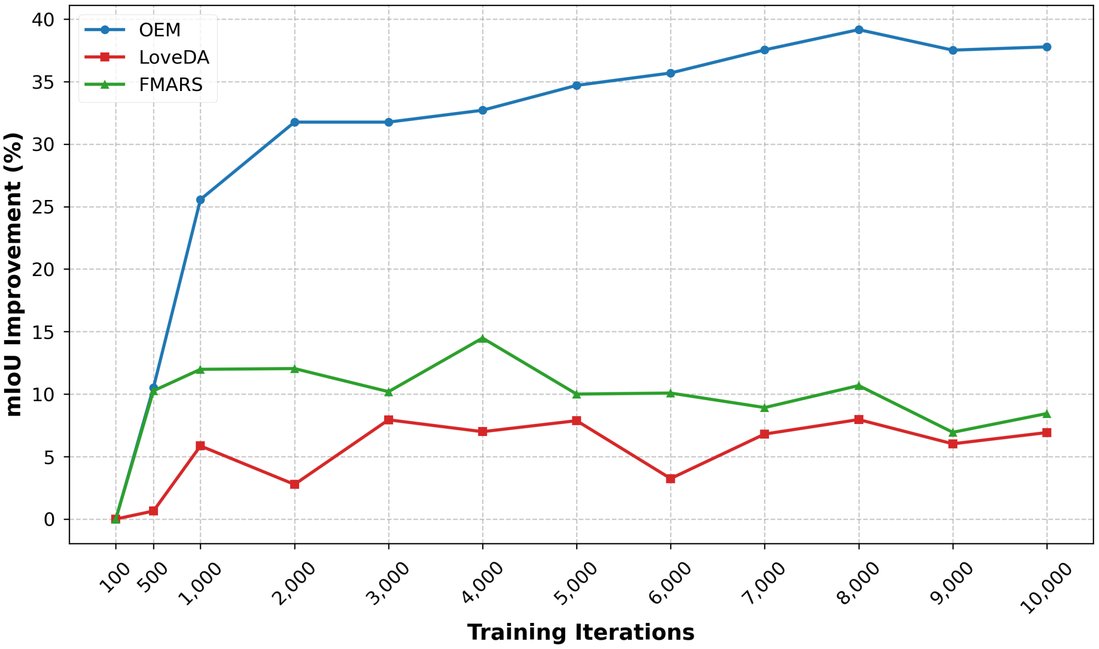


Figure 4.4: Percentage performance improvement of Remote FC-CLIP across training iterations for three different datasets (OEM, LoveDA, and FMARS). The OEM dataset shows the highest gains while LoveDA and FMARS demonstrate moderate progress.

performance across three distinct datasets (i.e. OEM, LoveDA, and FMARS), comparing it against baselines and Remote FC-CLIP. The analysis includes quantitative metrics, focusing on class-wise IoU scores, and qualitative assessments through visual examples. Finally, we explore the model's OVS capabilities by testing its performance on challenging novel classes and out-of-domain images, demonstrating its ability to generalize beyond its training distribution.

SAM's hyperparameters search - The performance of SAM-FC-CLIP is intrinsically tied to the output of the SAM mask extractor. As detailed in Section 4.1.3, we conducted two independent hyperparameter searches for SAM, one on the LoveDA dataset and the other on FMARS. Our empirical analysis revealed that the factors discussed in Section 3.4 affecting segmentation quality operate on opposite ends of the hyperparameter spectrum. On one side, using too many prompt points and applying permissive filtering criteria generates numerous masks but introduces noise. On the other end, sampling fewer points and implementing stricter post-processing filters produces high-quality masks but leaves many areas unsegmented. Our goal was to find an optimal balance that maximizes mask coverage while minimizing both non-segmented areas and segmentation artifacts (such as fragmented objects or overlapping masks of different classes). The random search identified this optimum in using a moderately high number of points per side and non-maximum suppression (NMS) threshold. These parameters achieved comprehensive mask coverage while maintaining quality. These findings align with the characteristics of VHR satellite imagery, where scenes are densely populated with small objects requiring precise segmentation. The optimal SAM configurations identified for both datasets are reported in Table 4.7.

Hyperparameter	**a**	**b**
`points_per_side`	107	100
`pred_iou_thresh`	0.5267	0.1644
`stability_score_thresh`	0.4819	0.6436
`box_nms_thresh`	0.5991	0.3402

Table 4.7: Optimal SAM's hyperparameters for LoveDA (a) and Fmars (b)

We trained four variants of SAM-FC-CLIP: all the combinations of datasets *iSAID+OEM*$_{1k}$ and *iSAID+OEM*$_{fine}$, using instance weighted or uniform CE. We evaluated each of them on various datasets and compared their performances with the baseline model and Remote FC-CLIP.

OEM Results - Quantitative results on the OEM dataset are presented in Table 4.8. Our model demonstrates superior performance compared to both baseline variants. While it does not outperform Remote FC-CLIP on the OEM dataset, this is expected given that the latter was trained end-to-end specifically on this task and dataset. Among the evaluated configurations, variant d (*iSAID+OEM*$_{1k}$ training with uniform CE) exhibits marginally superior performance. This configuration achieves an improvement of approximately 18 percentage points in mIoU relative to the baseline model. The most challenging classes remain *bareland* and *pavement*. Additionally, the model demonstrates suboptimal performance on the *road* class due to SAM's limited efficacy in segmenting elongated structures and objects that may be perceived as background elements. Qualitative results, shown in Figure 4.5, demonstrate our model's ability to generate numerous distinct masks. SAM-FC-CLIP exhibits remarkable sensitivity in segmenting fine-grained objects, successfully identifying even instances not captured in the ground truth annotations. However, this granular segmentation sometimes leads to classification errors, particularly when the model encounters objects whose true semantic classes are not included in the predefined set of textual queries. This limitation is exemplified in the first row (zoom required in the bottom left area of the picture in the first row), where some vehicles are misclassified as *building* due to the absence of an appropriate vehicle class in the query set.

Model	Var.	Bareland	Grass	Pavement	Road	Tree	Water	Cropland	Building	mIoU
FC-CLIP	a	0.00	17.39	3.74	19.91	42.32	50.43	35.01	36.80	25.70
	b	0.25	10.60	8.89	31.47	15.27	64.55	31.31	42.06	25.55
Remote FC-CLIP	c	**45.54**	57.22	**56.09**	**63.25**	**71.47**	**76.14**	**75.39**	**78.94**	**65.50**
SAM-FC-CLIP	d	20.42	59.34	29.88	27.06	41.20	56.73	55.93	61.07	43.95
	e	14.47	58.68	22.53	21.23	40.74	57.04	43.88	62.09	40.08
	f	20.40	30.32	24.14	41.15	58.53	52.26	62.00	59.85	43.58
	g	14.60	**59.89**	20.87	18.57	40.00	57.05	38.54	61.42	38.87

Table 4.8: Performance comparison on the OEM dataset between FC-CLIP, Remote FC-CLIP and SAM-FC-CLIP. Model variants are: a (ResNet50 backbone), b (ConvNeXt-L backbone), c (ResNet50 backbone with 10k training iterations), d and e (trained on *iSAID+OEM*$_{1k}$, e using instance weighted CE), f and g (trained on *iSAID+OEM*$_{fine}$, g using instance weighted CE). Results are reported as IoU% scores.

LoveDA Results - Table 4.9 presents our evaluation on the LoveDA dataset, where SAM-FC-CLIP demonstrates superior performance compared to both the baseline models and Remote FC-CLIP. While *agricultural* and *barren* instances remain challenging to classify, our model achieves more consistent performance across all categories, avoiding the class-specific IoU degradation observed in Table 4.5. This stability suggests enhanced zero-shot generalization capabilities compared to our previous implementation. Among the four variants tested, model *f* achieves the best overall metrics, though the differences between versions are relatively small. Qualitative results, illustrated in Figure 4.6, demonstrate SAM-FC-CLIP's ability to extract numerous detailed masks. This capability is particularly evident in the bottom row, where the model precisely delineates sequences of trees along street boundaries. Our approach also shows superior semantic fidelity to the raw image content compared to the ground truth annotations, especially in areas with fine-grained details. Some examples include the precise segmentation of building structures and minor roadways in the first image, and more accurate delineation of water bodies (shown in blue) in the second and third images. These results are notable given that our model had no prior exposure to the target domain distribution or fine-tuning on the LoveDA dataset. Moreover, several semantic categories present in this evaluation, specifically *forest*, *barren*, and *agricultural*, were absent from the training set, highlighting the model's effective zero-shot generalization capabilities across novel domains and semantic concepts.

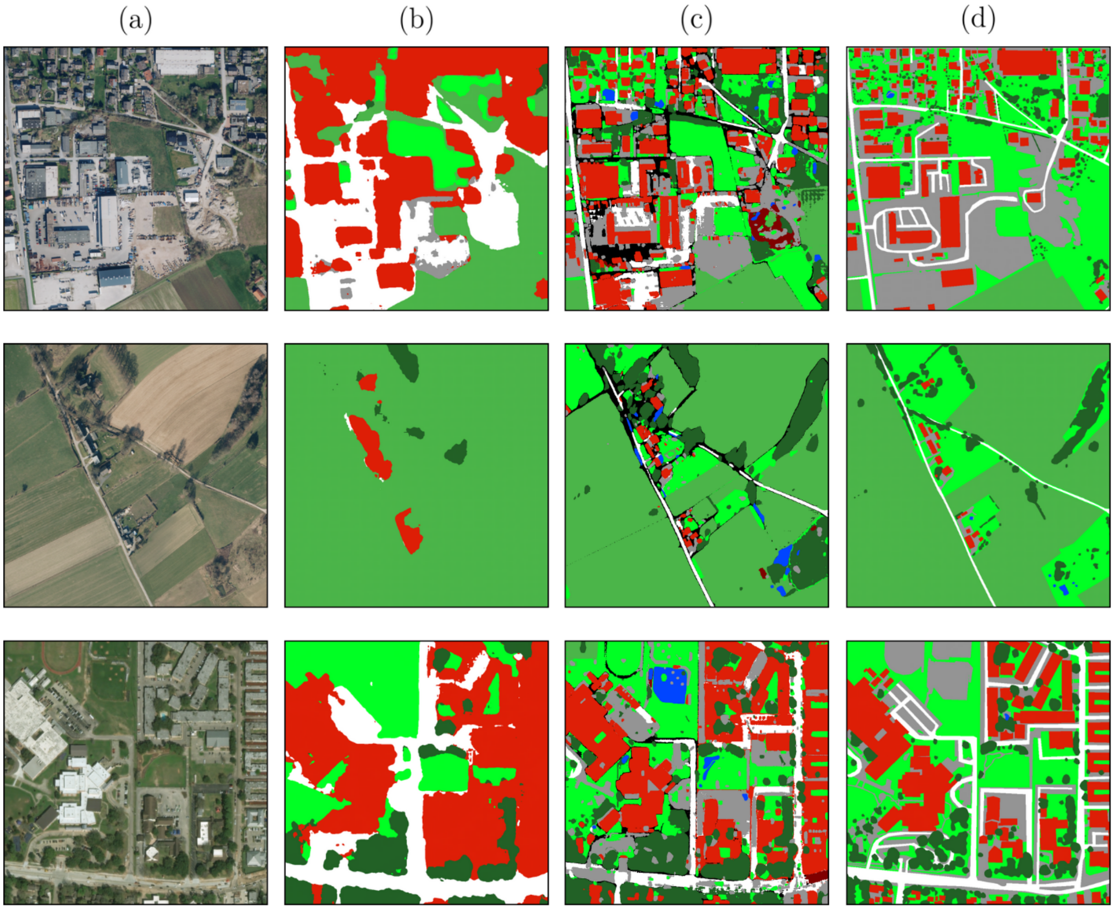


Figure 4.5: Qualitative results on the OEM dataset. From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our SAM-FC-CLIP model, and (d) ground truth segmentation masks. The images showcase industrial, rural and urban scenes, illustrating models performances across diverse areas.

FMARS Results - Table 4.10 presents quantitative evaluation results on the FMARS dataset. Despite being built upon the small baseline architecture, SAM-FC-CLIP achieves metrics comparable to the large baseline variant. It exhibits lower performance compared to Remote FC-CLIP. This gap can be attributed to the dataset's class distribution, which consists entirely of categories present in the Remote FC-CLIP training set. Among the evaluated classes, *road* and *background* proved to be the most challenging, showing consistently lower metrics. Figure 4.7 presents qualitative comparisons across different scenes. While the model struggles to detect all tree areas in the first image, it demonstrates exceptional accuracy in building delineation. The segmentation maps for the second and third images

Model	Var.	Background	Building	Road	Water	Barren	Forest	Agricultural	mIoU
FC-CLIP	a	35.29	38.79	37.92	28.72	0.00	8.12	0.38	21.32
	b	36.08	25.39	46.48	51.48	0.07	11.11	0.16	24.39
Remote FC-CLIP	c	**37.83**	**59.13**	**47.57**	40.61	5.11	**34.45**	6.33	33.00
SAM-FC-CLIP	d	34.18	43.31	42.24	51.93	**21.02**	24.96	15.70	33.34
	e	34.74	48.46	41.08	49.73	20.03	24.90	12.49	33.06
	f	33.91	42.76	43.98	**54.95**	20.73	25.32	**16.12**	**33.97**
	g	35.25	48.54	41.94	49.43	20.56	24.07	12.78	33.22

Table 4.9: Performance comparison on the LoveDA dataset between FC-CLIP, Remote FC-CLIP and SAM-FC-CLIP. Model variants are: a (ResNet50 backbone), b (ConvNeXt-L backbone), c (ResNet50 backbone with 10k training iterations), d and e (trained on *iSAID+OEM*$_{1k}$, e using instance weighted CE), f and g (trained on *iSAID+OEM*$_{fine}$, g using instance weighted CE). Results are reported as IoU% scores.

show particularly strong alignment with the ground truth, highlighting the model's capability to maintain consistency across various urban landscapes.

Model	Var.	Road	Tree	Buildings	Background	mIoU
FC-CLIP	a	27.66	63.09	49.16	27.75	41.91
	b	53.25	61.29	54.88	37.24	51.67
Remote FC-CLIP	c	**54.66**	62.49	**67.59**	**45.70**	**57.61**
SAM-FC-CLIP	d	31.57	66.73	65.76	32.66	49.18
	e	39.38	67.80	63.94	35.12	51.56
	f	32.34	**67.82**	65.14	32.37	49.42
	g	37.14	65.82	61.03	33.63	49.41

Table 4.10: Performance comparison on the FMARS dataset between FC-CLIP, Remote FC-CLIP and SAM-FC-CLIP. Model variants are: a (ResNet50 backbone), b (ConvNeXt-L backbone), c (ResNet50 backbone with 10k training iterations), d and e (trained on *iSAID+OEM*$_{1k}$, e using instance weighted CE), f and g (trained on *iSAID+OEM*$_{fine}$, g using instance weighted CE). Results are reported as IoU% scores.

Exploring OVS Capabilities - We conducted a series of empirical experiments

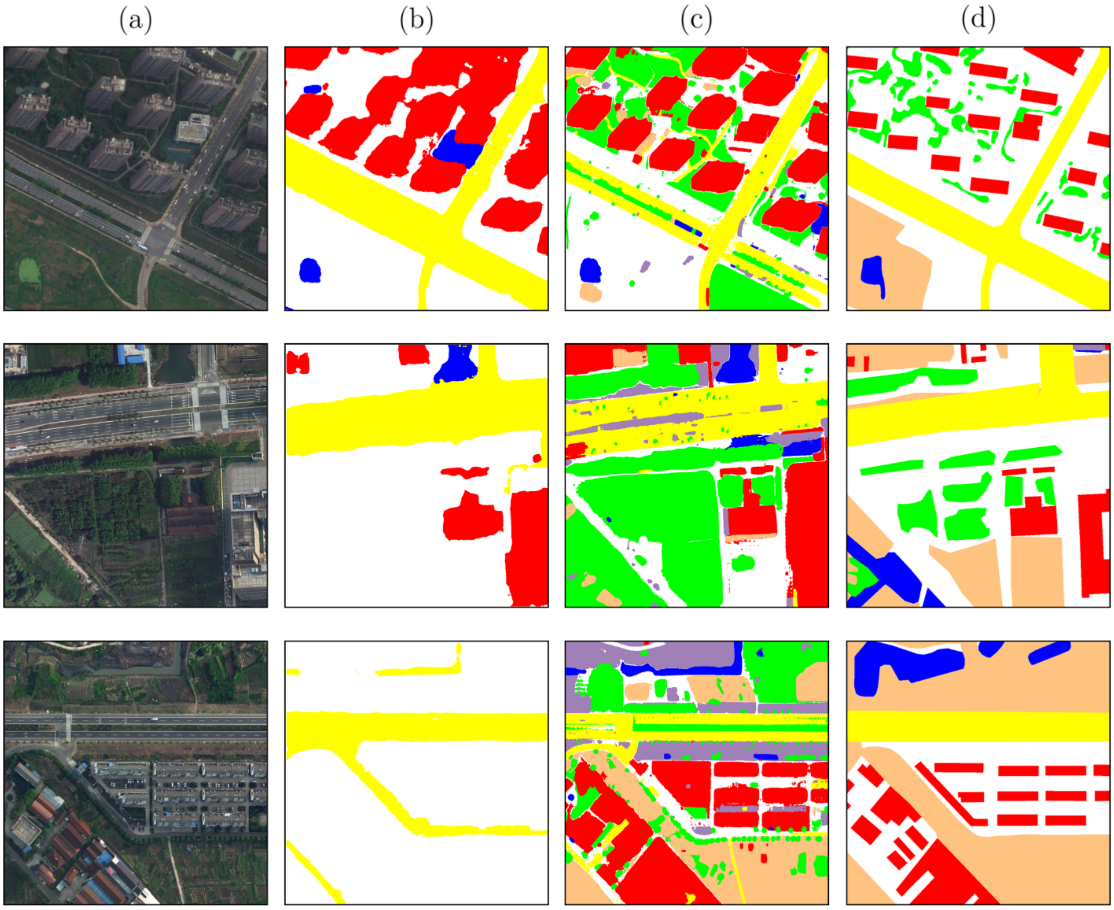


Figure 4.6: Qualitative results on the LoveDA dataset. From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our SAM-FC-CLIP model, and (d) ground truth segmentation masks. The images showcase industrial, rural and urban scenes, illustrating models performances across diverse areas.

to evaluate SAM-FC-CLIP's open vocabulary capabilities. While Section 2.1.5 discussed CDTE as the standard evaluation approach in literature (which we employed in our previous tests) here we focus on qualitative assessment of the model's performance on out-of-domain images with both seen and unseen classes. To deeply test the model's capabilities, we crafted text queries that included both expected classes present in the images and deliberately included absent classes to evaluate the model's discrimination abilities. Figures 4.8, 4.9 and 4.10 showcase these qualitative results. The used textual queries are shown below the images, together with the color used in the visualization. While some instances of mislabeling occurred, the model exhibited satisfactory performance in segmenting

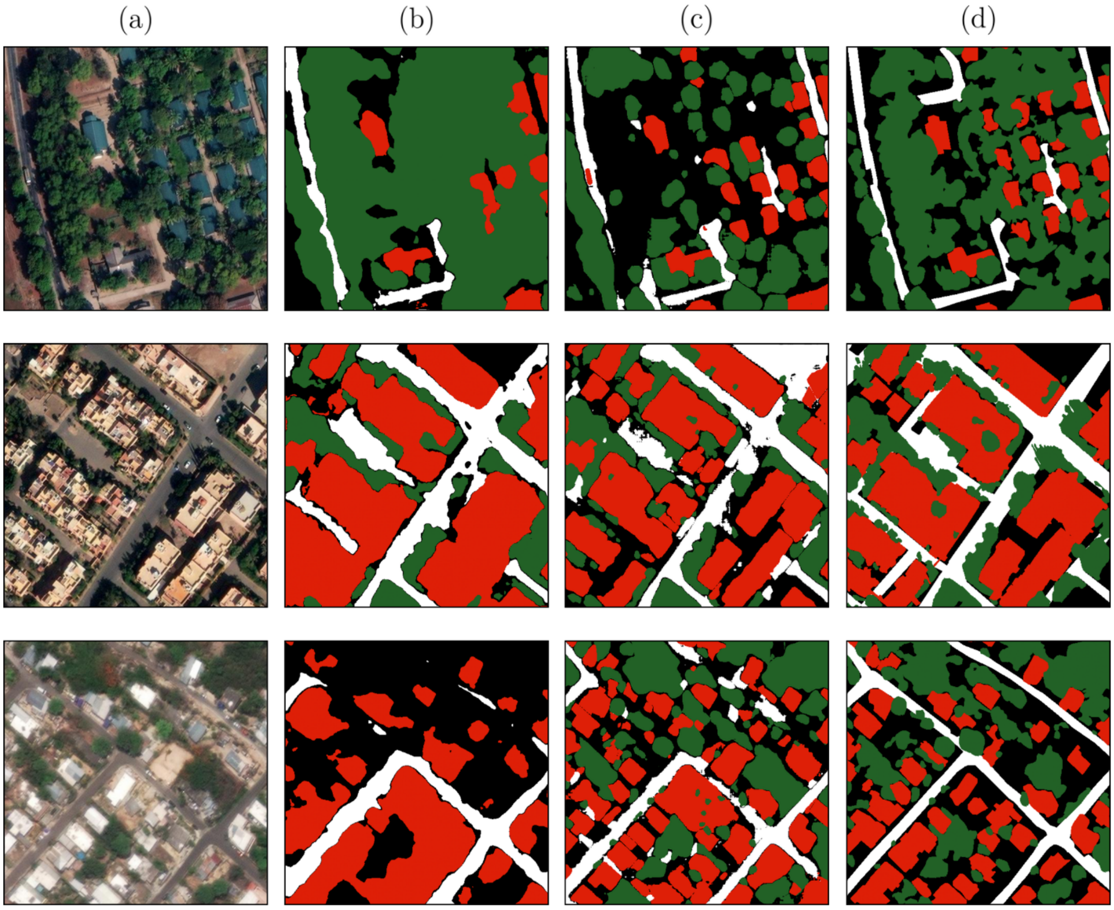


Figure 4.7: Qualitative results on the FMARS dataset. From left to right: (a) original input images, (b) predictions from the baseline FC-CLIP model, (c) predictions from our SAM-FC-CLIP model, and (d) ground truth segmentation masks. The images showcase rural and urban scenes, illustrating models' performances across diverse areas.

small-scale objects, especially vehicles, as evidenced in row 1 of Figure 4.8. Rows 2,3 and 4 highlight the delineation of some iSAID classes like: tennis court, soccer ball field and ground track field. Though we previously demonstrated the model's capability in building and tree segmentation, these images provides additional examples, especially rows 5 and 6. SAM-FC-CLIP demonstrated remarkable versatility in detecting requested classes, successfully identifying complex, out-of-vocabulary objects such as *trains* (visible in the top-left area of image in row 8 of Figure 4.10) and *parking areas* (Figure 4.10, row 7). Notably, neither the baseline model nor Remote FC-CLIP could match these open vocabulary capabilities in our comparative testing. They showed two significant limitations: they struggled to extract fine-grained masks for small objects like vehicles, and they lacked the ability to effectively associate novel semantic concepts with image content. This contrast highlights the enhanced flexibility and broader applicability of SAM-FC-CLIP.

To summarize SAM-FC-CLIP's performance across our experiments, the model demonstrates enhanced open vocabulary capabilities, though at the cost of slightly reduced performance compared to Remote FC-CLIPon previously seen classes. A key strength of our approach lies in the SAM mask extractor which, using optimized hyperparameters, successfully generates both large-scale segmentation masks and numerous fine-grained masks for smaller objects. This multi-scale capability enables the model to identify and classify objects across various sizes, from buildings to individual vehicles. However, the model shows consistent weakness in road segmentation, as SAM struggles to extract coherent masks for extended linear structures. Despite this limitation, the model's overall performance suggests a promising direction for OVS in the RS domain.

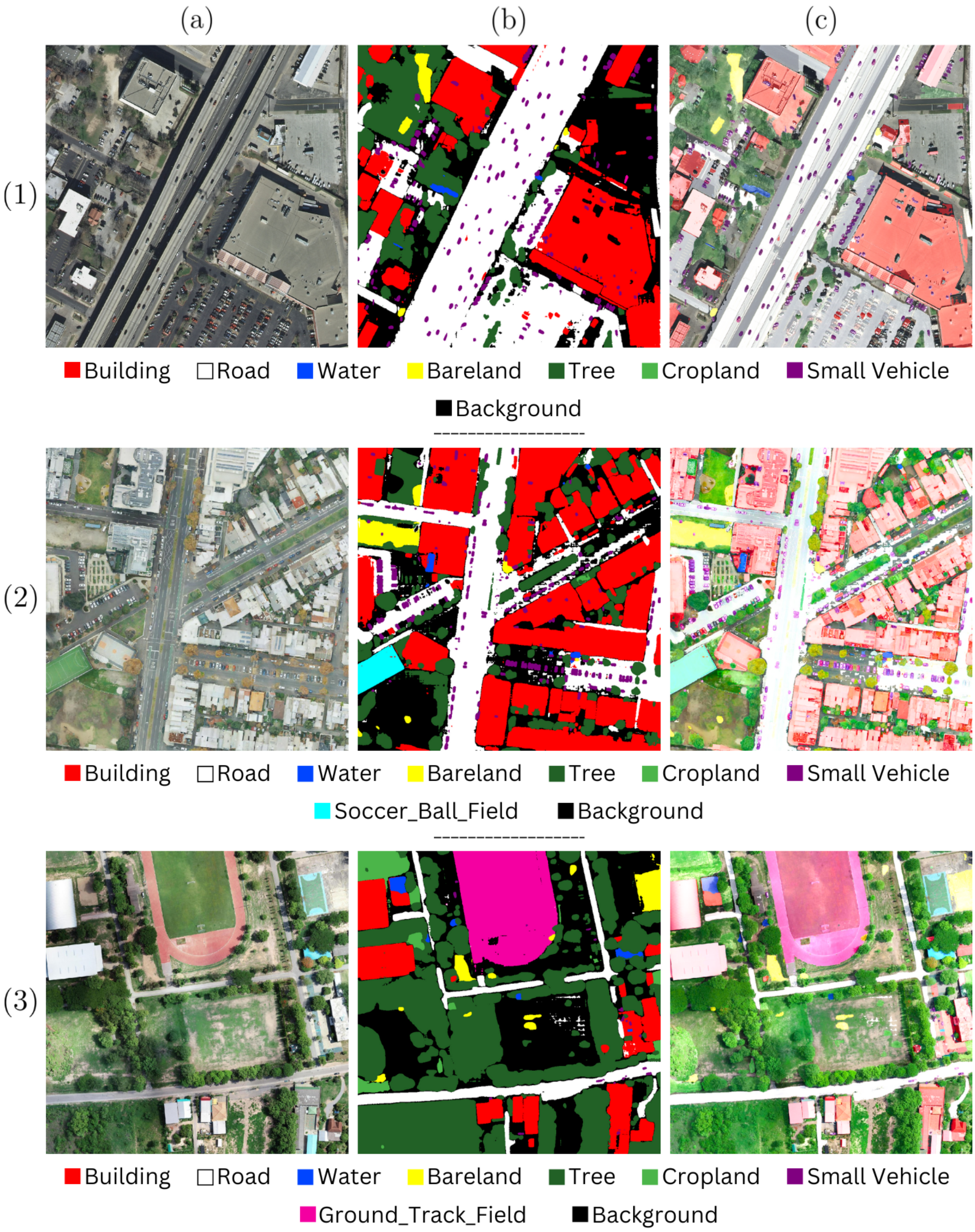

(a)
(b)
(c)
(1)
(2)
(3)
🟥 Building □ Road 🟦 Water 🟨 Bareland 🟩 Tree 🟩 Cropland 🟪 Small Vehicle
■ Background
🟥 Building □ Road 🟦 Water 🟨 Bareland 🟩 Tree 🟩 Cropland 🟪 Small Vehicle
■ Soccer_Ball_Field ■ Background
🟥 Building □ Road 🟦 Water 🟨 Bareland 🟩 Tree 🟩 Cropland 🟪 Small Vehicle
■ Ground_Track_Field ■ Background

Figure 4.8: Qualitative results of SAM-FC-CLIP along with the text queries and the color used in the visualization. Notable delineated classes are the small cars in purple and some rare categories like the soccer ball field and the ground track field.

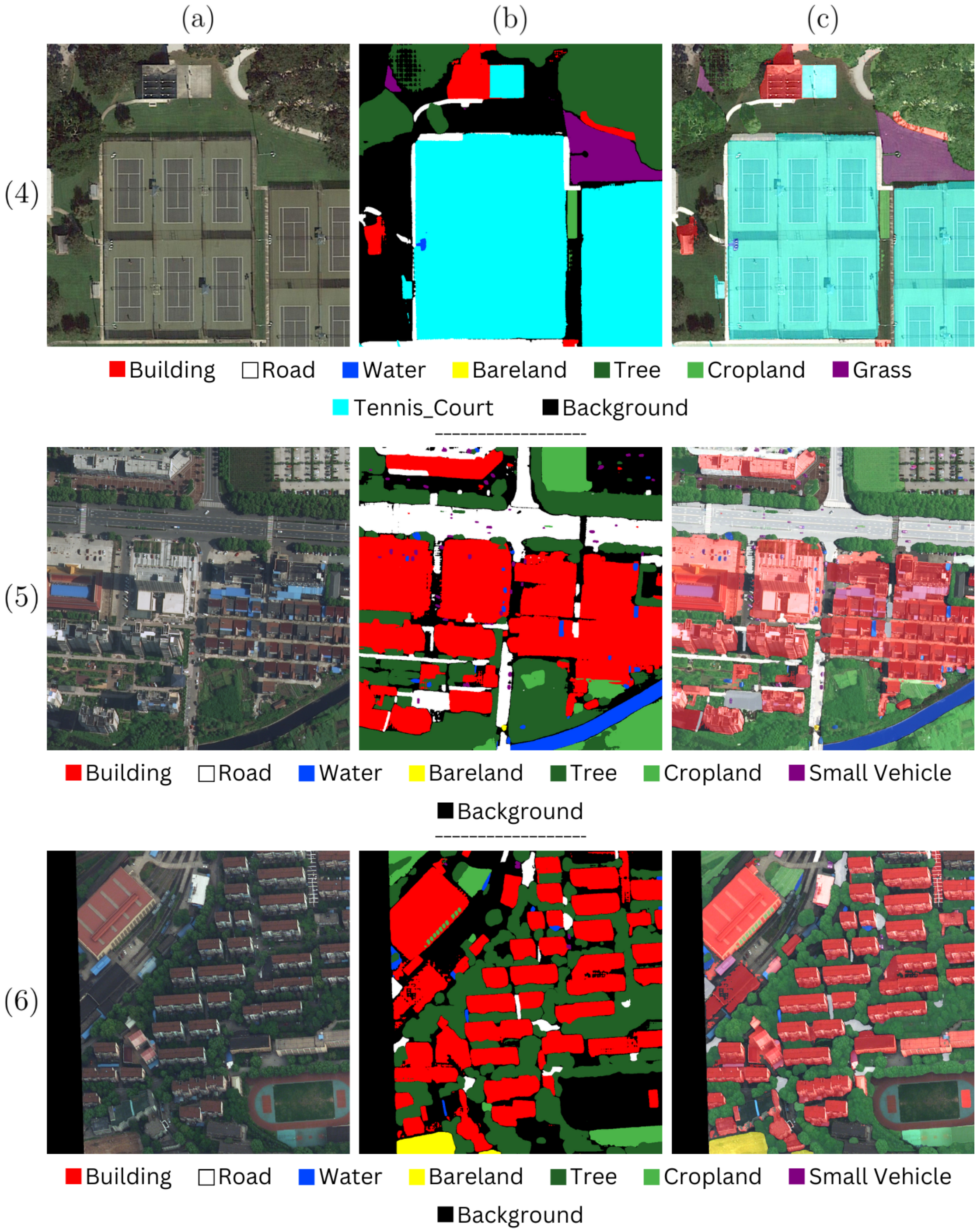


Figure 4.9: Qualitative results of SAM-FC-CLIP along with the text queries and the color used in the visualization.

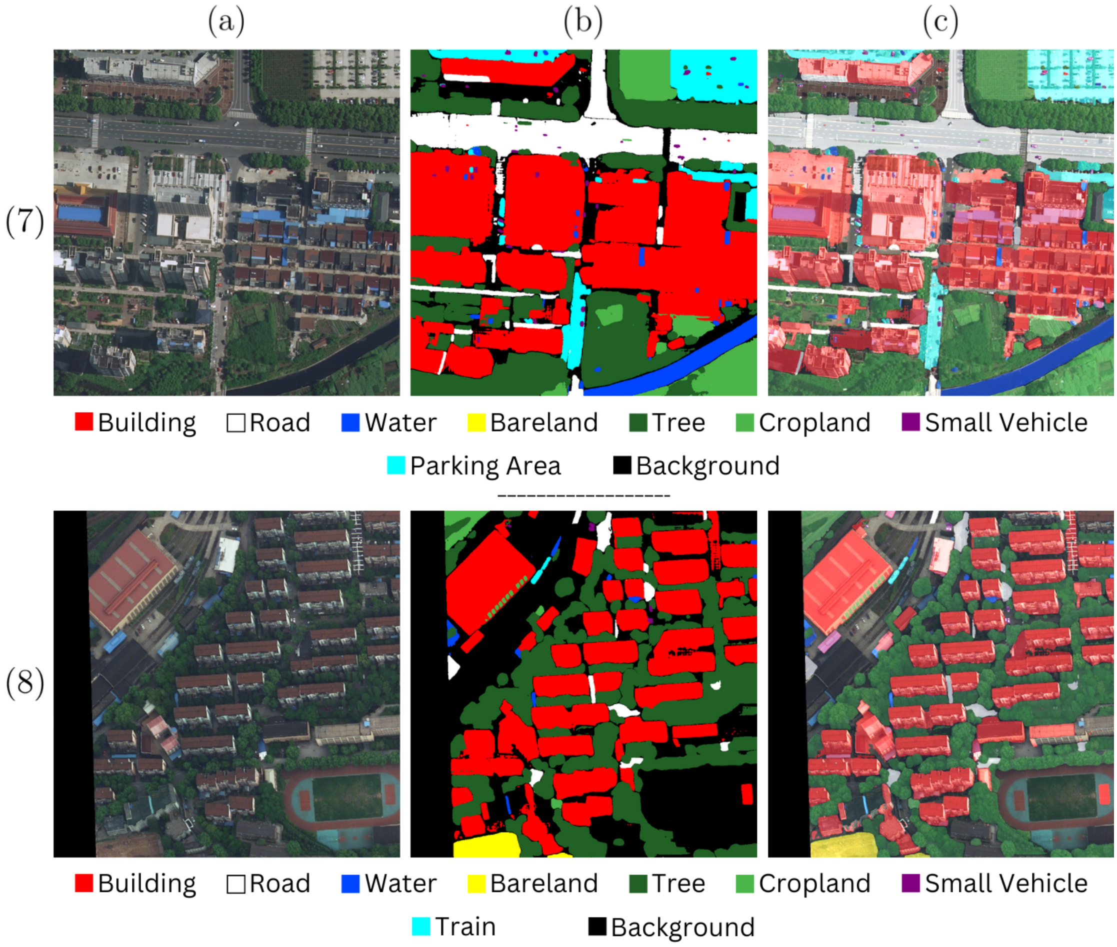


Figure 4.10: Qualitative results of SAM-FC-CLIP along with the text queries and the color used in the visualization. Notable delineated classes are the parking areas and the trains in cyan. Notice how we included some queries about instances not present in the photos to test model's discriminatory ability.

Chapter 5

Conclusions

This thesis has addressed the challenging task of OVS for VHR satellite imagery. We presented two novel approaches to overcome existing methods' limitations, enabling flexible, language-guided segmentation without being constrained to predefined categories.

Our Contributions - We demonstrated the difficulties of directly applying natural image-based open vocabulary models to satellite imagery, highlighting the unique challenges of these aerial photos and the main factors that cause an extreme domain shift. We developed Remote FC-CLIP, which successfully adapted the FC-CLIP architecture by incorporating Remote CLIP, a vision-language foundation model fine-tuned for remote sensing. This approach demonstrated excellent performance on previously seen categories, validating the effectiveness of domain-specific pre-training. Then, we introduced SAM-FC-CLIP, an innovative solution that combines the SAM's powerful mask extraction capabilities with a modified Remote FC-CLIP architecture. To address the critical challenge of limited available data, we trained our model on a unified training dataset we created by merging OEM and iSAID, encompassing 23 distinct classes. This architecture demonstrated superior open vocabulary capabilities, particularly in identifying and segmenting previously unseen objects.

Limitations and Potentials - Remote FC-CLIP exhibits a bias towards seen classes, primarily due to its mask extractor which tends to segment only objects with similar shapes and sizes as those in the limited training set. This constraint particularly affects fine-grained objects (e.g., vehicles), making their delineation and classification infeasible. Although the model demonstrates strong performance on seen categories, its open vocabulary capabilities are reduced when fine-tuned on a small dataset. Turning to SAM-FC-CLIP, its main limitations derive from its two-stage architecture. The model has not been trained end-to-end and its performance is bound by SAM's output, which has not been fine-tuned on satellite imagery beyond hyper-parameter optimization. Furthermore, this design leads to feature duplication, as they are extracted independently by both the SAM model and the classification component. Our empirical evaluation reveals that the model occasionally generates false positives, particularly when processing a limited number of queries or queries that significantly deviate from the image content. Also, we expect that task-specific models trained for particular classes may achieve superior performance, however this trade-off is characteristic of most OVS approaches. Despite these limitations, both proposed approaches demonstrate significant potential. Notably, Remote FC-CLIP and SAM-FC-CLIP surpass the baseline models in quantitative performance metrics. SAM-FC-CLIP, in particular, exhibits remarkable qualitative results, successfully identifying and segmenting object categories that were previously not obtainable with other OVS models. These achievements are especially noteworthy considering the inherent complexity of open vocabulary segmentation in satellite imagery and the constraints imposed by limited training data availability.

Future Works - SAM-FC-CLIP can be further improved by addressing some of its limitations. One advancement would be the development of a unified, single-stage model through knowledge distillation of SAM's capabilities into the FC-CLIP mask extractor. This integration could be enhanced by increasing the number of its learnable queries to better handle the intrinsic complexity of satellite imagery. The current implementation's capacity could be significantly expanded through architectural improvements. The ResNet50 backbone could be replaced with a larger encoder (e.g. ConvNext), while on the decoder side, we could incorporate a transformer-based solution. These architectural enhancements would provide the model with greater capacity. Also, building upon our successful merge of OpenEarthMap and iSAID, future work could focus on incorporating additional satellite imagery datasets. Expanding the base class vocabulary would enhance the model's general understanding of aerial scenes. Furthermore, the model's capabilities could be extended through post-processing techniques to support panoptic segmentation. In particular, we could implement a system where users specify whether prompted classes should be treated as *things* or *stuff*, maintaining instance-level detail for the former while merging segments belonging to the latter category.

In conclusion, this research represents a step forward in making satellite imagery more accessible and interpretable via natural language by bridging the gap between the vast amounts of satellite data being collected and the diverse needs of end-users. As satellite technology advances and the volume of available imagery grows exponentially, these flexible approaches become increasingly relevant. Even if the scarcity of comprehensive annotated domain-specific datasets remains a fundamental challenge, looking ahead, the continuing evolution of this field suggests a promising future where satellite imagery analysis becomes more democratic, efficient, and applicable to a wide range of global challenges, from climate monitoring to disaster response.

Bibliography

[1] European Union. *Copernicus Programme*. 2014. URL: `https://www.copernicus.eu` (visited on 01/05/2024) (cit. on p. 1).

[2] Maxar. *Maxar Open Data Program*. 2017. URL: `https://www.maxar.com/open-data` (visited on 01/05/2024) (cit. on p. 1).

[3] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. «Panoptic Segmentation». In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 4, 11).

[4] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. «Backpropagation Applied to Handwritten Zip Code Recognition». In: *Neural Computation* 1.4 (1989), pp. 541–551. DOI: `10.1162/neco.1989.1.4.541` (cit. on p. 5).

[5] Evan Shelhamer, Jonathan Long, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2016. arXiv: `1605.06211 [cs.CV]`. URL: `https://arxiv.org/abs/1605.06211` (cit. on p. 5).

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: `1505.04597 [cs.CV]`. URL: `https://arxiv.org/abs/1505.04597` (cit. on p. 5).

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*. 2017. arXiv: `1606.00915 [cs.CV]`. URL: `https://arxiv.org/abs/1606.00915` (cit. on p. 5).

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017. arXiv: `1706.05587 [cs.CV]`. URL: `https://arxiv.org/abs/1706.05587` (cit. on p. 5).

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762 (cit. on pp. 5, 24).

[10] Sixiao Zheng et al. *Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers*. 2021. arXiv: 2012.15840 [cs.CV]. URL: https://arxiv.org/abs/2012.15840 (cit. on p. 5).

[11] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. *SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers*. 2021. arXiv: 2105.15203 [cs.CV]. URL: https://arxiv.org/abs/2105.15203 (cit. on p. 5).

[12] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. *Masked-attention Mask Transformer for Universal Image Segmentation*. 2022. arXiv: 2112.01527 [cs.CV]. URL: https://arxiv.org/abs/2112.01527 (cit. on pp. 5–7, 14, 24).

[13] Alexander Kirillov et al. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: https://arxiv.org/abs/2304.02643 (cit. on pp. 5, 12).

[14] Yunyang Xiong et al. *EfficientSAM: Leveraged Masked Image Pretraining for Efficient Segment Anything*. 2023. arXiv: 2312.00863 [cs.CV]. URL: https://arxiv.org/abs/2312.00863 (cit. on p. 5).

[15] Zhuoyang Zhang, Han Cai, and Song Han. *EfficientViT-SAM: Accelerated Segment Anything Model Without Accuracy Loss*. 2024. arXiv: 2402.05008 [cs.CV]. URL: https://arxiv.org/abs/2402.05008 (cit. on pp. 5, 28, 37).

[16] Nikhila Ravi et al. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV]. URL: https://arxiv.org/abs/2408.00714 (cit. on p. 5).

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*. 2014. arXiv: 1311.2524 [cs.CV]. URL: https://arxiv.org/abs/1311.2524 (cit. on pp. 6, 28).

[18] Ross Girshick. *Fast R-CNN*. 2015. arXiv: 1504.08083 [cs.CV]. URL: https://arxiv.org/abs/1504.08083 (cit. on p. 6).

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497 (cit. on p. 6).

[20] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. *YOLACT: Real-time Instance Segmentation*. 2019. arXiv: 1904.02689 [cs.CV]. URL: https://arxiv.org/abs/1904.02689 (cit. on p. 6).

[21] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. *SOLO: Segmenting Objects by Locations*. 2020. arXiv: 1912.04488 [cs.CV]. URL: https://arxiv.org/abs/1912.04488 (cit. on p. 6).

[22] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. *End-to-End Object Detection with Transformers*. 2020. arXiv: 2005.12872 [cs.CV]. URL: https://arxiv.org/abs/2005.12872 (cit. on pp. 6, 37).

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. *Mask R-CNN*. 2018. arXiv: 1703.06870 [cs.CV]. URL: https://arxiv.org/abs/1703.06870 (cit. on p. 6).

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. *Feature Pyramid Networks for Object Detection*. 2017. arXiv: 1612.03144 [cs.CV]. URL: https://arxiv.org/abs/1612.03144 (cit. on p. 6).

[25] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. *Panoptic Feature Pyramid Networks*. 2019. arXiv: 1901.02446 [cs.CV]. URL: https://arxiv.org/abs/1901.02446 (cit. on p. 6).

[26] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. *UPSNet: A Unified Panoptic Segmentation Network*. 2019. arXiv: 1901.03784 [cs.CV]. URL: https://arxiv.org/abs/1901.03784 (cit. on p. 7).

[27] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. *Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation*. 2020. arXiv: 2003.07853 [cs.CV]. URL: https://arxiv.org/abs/2003.07853 (cit. on p. 7).

[28] Rohit Mohan and Abhinav Valada. *EfficientPS: Efficient Panoptic Segmentation*. 2021. arXiv: 2004.02307 [cs.CV]. URL: https://arxiv.org/abs/2004.02307 (cit. on p. 7).

[29] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. *Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation*. 2020. arXiv: 1911.10194 [cs.CV]. URL: https://arxiv.org/abs/1911.10194 (cit. on p. 7).

[30] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. *MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers.* 2021. arXiv: `2012.00759 [cs.CV]`. URL: `https://arxiv.org/abs/2012.00759` (cit. on p. 7).

[31] Alec Radford et al. «Learning Transferable Visual Models From Natural Language Supervision». In: (2021). arXiv: `2103.00020 [cs.CV]`. URL: `https://arxiv.org/abs/2103.00020` (cit. on pp. 7, 8, 14).

[32] Chaoyang Zhu and Long Chen. *A Survey on Open-Vocabulary Detection and Segmentation: Past, Present, and Future.* 2024. arXiv: `2307.09220 [cs.CV]`. URL: `https://arxiv.org/abs/2307.09220` (cit. on p. 7).

[33] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. *Microsoft COCO Captions: Data Collection and Evaluation Server.* 2015. arXiv: `1504.00325 [cs.CV]`. URL: `https://arxiv.org/abs/1504.00325` (cit. on p. 7).

[34] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. *GroupViT: Semantic Segmentation Emerges from Text Supervision.* 2022. arXiv: `2202.11094 [cs.CV]`. URL: `https://arxiv.org/abs/2202.11094` (cit. on p. 7).

[35] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. *Open-world Semantic Segmentation via Contrasting and Clustering Vision-Language Embedding.* 2022. arXiv: `2207.08455 [cs.CV]`. URL: `https://arxiv.org/abs/2207.08455` (cit. on p. 7).

[36] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H. S. Torr, and Ser-Nam Lim. *Open Vocabulary Semantic Segmentation with Patch Aligned Contrastive Learning.* 2022. arXiv: `2212.04994 [cs.CV]`. URL: `https://arxiv.org/abs/2212.04994` (cit. on p. 7).

[37] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. *SAM-CLIP: Merging Vision Foundation Models towards Semantic and Spatial Understanding.* 2024. arXiv: `2310.15308 [cs.CV]`. URL: `https://arxiv.org/abs/2310.15308` (cit. on p. 8).

[38] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. *A Simple Framework for Open-Vocabulary Segmentation and Detection.* 2023. arXiv: `2303.08131 [cs.CV]`. URL: `https://arxiv.org/abs/2303.08131` (cit. on p. 8).

[39] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. *Open-Vocabulary Semantic Segmentation with Mask-adapted CLIP*. 2023. arXiv: 2210.04150 [cs.CV]. URL: https://arxiv.org/abs/2210.04150 (cit. on pp. 8, 15, 35, 39, 45).

[40] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. *Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models*. 2023. arXiv: 2303.04803 [cs.CV]. URL: https://arxiv.org/abs/2303.04803 (cit. on p. 8).

[41] Xiaoyi Dong et al. *MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraining*. 2023. arXiv: 2208.12262 [cs.CV]. URL: https://arxiv.org/abs/2208.12262 (cit. on pp. 8, 15, 35, 39).

[42] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. *Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP*. 2023. arXiv: 2308.02487 [cs.CV]. URL: https://arxiv.org/abs/2308.02487 (cit. on pp. 8, 24, 35, 39).

[43] Vibashan VS, Shubhankar Borse, Hyojin Park, Debasmit Das, Vishal Patel, Munawar Hayat, and Fatih Porikli. *PosSAM: Panoptic Open-vocabulary Segment Anything*. 2024. arXiv: 2403.09620 [cs.CV]. URL: https://arxiv.org/abs/2403.09620 (cit. on p. 8).

[44] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. «The Pascal Visual Object Classes Challenge: A Retrospective». In: *International Journal of Computer Vision* 111.1 (Jan. 2015), pp. 98–136 (cit. on pp. 11, 15).

[45] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. *COCO-Stuff: Thing and Stuff Classes in Context*. 2018. arXiv: 1612.03716 [cs.CV]. URL: https://arxiv.org/abs/1612.03716 (cit. on p. 11).

[46] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. «Semantic understanding of scenes through the ade20k dataset». In: *International Journal of Computer Vision* 127.3 (2019), pp. 302–321 (cit. on pp. 11, 12, 28).

[47] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV]. URL: https://arxiv.org/abs/1405.0312 (cit. on pp. 12, 15, 17, 28).

[48] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. «Objects365: A Large-Scale, High-Quality Dataset for Object Detection». In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 8429–8438. DOI: 10.1109/ICCV.2019.00852 (cit. on p. 12).

[49] 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.* 2024. arXiv: `2306.11029 [cs.CV]`. URL: `https://arxiv.org/abs/2306.11029` (cit. on pp. 12, 26, 27).

[50] Tianhe Ren et al. *Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks.* 2024. arXiv: `2401.14159 [cs.CV]`. URL: `https://arxiv.org/abs/2401.14159` (cit. on p. 13).

[51] Shilong Liu et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection.* 2024. arXiv: `2303.05499 [cs.CV]`. URL: `https://arxiv.org/abs/2303.05499` (cit. on p. 13).

[52] Bowen Cheng, Alexander G. Schwing, and Alexander Kirillov. *Per-Pixel Classification is Not All You Need for Semantic Segmentation.* 2021. arXiv: `2107.06278 [cs.CV]`. URL: `https://arxiv.org/abs/2107.06278` (cit. on pp. 14, 15).

[53] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. «MasQCLIP for Open-Vocabulary Universal Image Segmentation». In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV).* 2023, pp. 887–898. DOI: `10.1109/ICCV51070.2023.00088` (cit. on pp. 15, 35, 39).

[54] Olga Russakovsky et al. *ImageNet Large Scale Visual Recognition Challenge.* 2015. arXiv: `1409.0575 [cs.CV]`. URL: `https://arxiv.org/abs/1409.0575` (cit. on p. 15).

[55] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. *OpenEarthMap: A Benchmark Dataset for Global High-Resolution Land Cover Mapping.* 2022. arXiv: `2210.10732 [cs.CV]`. URL: `https://arxiv.org/abs/2210.10732` (cit. on pp. 17, 30, 35).

[56] Syed Waqas Zamir et al. «iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images». In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.* 2019, pp. 28–37 (cit. on pp. 18, 19, 26, 30).

[57] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. *LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation.* 2022. arXiv: `2110.08733 [cs.CV]`. URL: `https://arxiv.org/abs/2110.08733` (cit. on pp. 19, 35).

[58] Edoardo Arnaudo, Jacopo Lungo Vaschetti, Lorenzo Innocenti, Luca Barco, Davide Lisi, Vanina Fissore, and Claudio Rossi. *FMARS: Annotating Remote Sensing Images for Disaster Management using Foundation Models.* 2024. arXiv: `2405.20109 [cs.CV]`. URL: `https://arxiv.org/abs/2405.20109` (cit. on p. 20).

[59] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. *Deformable DETR: Deformable Transformers for End-to-End Object Detection.* 2021. arXiv: `2010.04159 [cs.CV]`. URL: `https://arxiv.org/abs/2010.04159` (cit. on p. 24).

[60] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. *Masked-attention Mask Transformer for Universal Image Segmentation.* 2022. arXiv: `2112.01527 [cs.CV]`. URL: `https://arxiv.org/abs/2112.01527` (cit. on p. 24).

[61] Harold W. Kuhn. «The Hungarian method for the assignment problem». In: *Naval Research Logistics (NRL)* 52 (1955). URL: `https://api.semanticscholar.org/CorpusID:9426884` (cit. on p. 25).

[62] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. «Exploring Models and Data for Remote Sensing Image Caption Generation». In: *IEEE Transactions on Geoscience and Remote Sensing* 56.4 (Apr. 2018), pp. 2183–2195. ISSN: 1558-0644. DOI: `10.1109/tgrs.2017.2776321`. URL: `http://dx.doi.org/10.1109/TGRS.2017.2776321` (cit. on p. 26).

[63] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. «Exploring a Fine-Grained Multiscale Method for Cross-Modal Remote Sensing Image Retrieval». In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–19. DOI: `10.1109/TGRS.2021.3078451` (cit. on p. 26).

[64] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. «DOTA: A Large-Scale Dataset for Object Detection in Aerial Images». In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* June 2018 (cit. on p. 26).

[65] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. «Object detection in optical remote sensing images: A survey and a new benchmark». In: *ISPRS Journal of Photogrammetry and Remote Sensing* 159 (Jan. 2020), pp. 296–307. ISSN: 0924-2716. DOI: `10.1016/j.isprsjprs.2019.11.023`. URL: `http://dx.doi.org/10.1016/j.isprsjprs.2019.11.023` (cit. on p. 26).

[66] Volodymyr Mnih. «Machine Learning for Aerial Image Labeling». PhD thesis. University of Toronto, 2013 (cit. on p. 30).

[67] 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.* Version 0.1. Zenodo, Oct. 2023. DOI: `10.1109/TPAMI.2023.3296757`. URL: `https://doi.org/10.1109/TPAMI.2023.3296757` (cit. on p. 30).

[68] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. *Detectron2*. `https://github.com/facebookresearch/detectron2`. 2019 (cit. on p. 36).

[69] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: `1711.05101 [cs.LG]`. URL: `https://arxiv.org/abs/1711.05101` (cit. on p. 36).