FreeMix: Open Vocabulary Domain Generalization of Remote Sensing Images for Semantic Segmentation

Jingi Wu , Jingye Shi , Zeyong Zhao , Ziyang Liu , Ruicong Zhi *

Posted Date: 8 January 2025

doi: 10.20944/preprints202501.0448.v1

Keywords: Open vocabulary; Semantic segmentation; Domain generalization; Self-Supervised Learning

Article

FreeMix: Open Vocabulary Domain Generalization of Remote Sensing Images for Semantic Segmentation

Jinyi Wu [1,2], Jingye Shi [3], Zeyong Zhao[1,2], Ziyang Liu [1,2] and Ruicong Zhi [1,2,*]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] Beijing Key Laboratory of Knowledge Engineering for Material Science, Beijing 100083, China
[3] Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China
* Correspondence: zhirc@ustb.edu.cn

Abstract: In this study, we present a novel concept termed open vocabulary domain generalization (OVDG), which we investigate within the context of semantic segmentation. OVDG presents greater difficulty compared to conventional domain generalization, yet it offers greater practicality. It jointly considers 1) recognizing both base and novel classes and 2) generalizing to unseen domains. In OVDG, only the labels of base classes and the images from source domains are available to learn a robust model. Then the model could be generalized to images from novel classes and target domains directly. In this paper, we propose a dual-branch FreeMix module to implement the OVDG task effectively in a universal framework: the Base Segmentation Branch (BSB) and the Entity Segmentation Branch (ESB). First, the entity mask is proposed for the first time for segmentation generalization, and the semantic logits are learned for both base mask and entity mask, so that to enhance the diversity and completeness of masks for both base classes and novel classes. Second, the FreeMix utilizes pre-trained self-supervised learning on large-scale remote sensing data (RS_SSL) to extract domain-agnostic visual features for decoding masks and semantic logits. Third, a training tactic called dataset-aware sampling (DAS) is introduced for multi-source domain learning, aimed at improving the overall performance. In summary, RS_SSL, ESB and DAS can significantly improve the generalization ability of model on both class-level and domain-level. Experiments demonstrate that our method produces state-of-the-art results on several remote sensing semantic segmentation datasets, including Potsdam, GID5, DeepGlobe, and URUR, for OVDG.

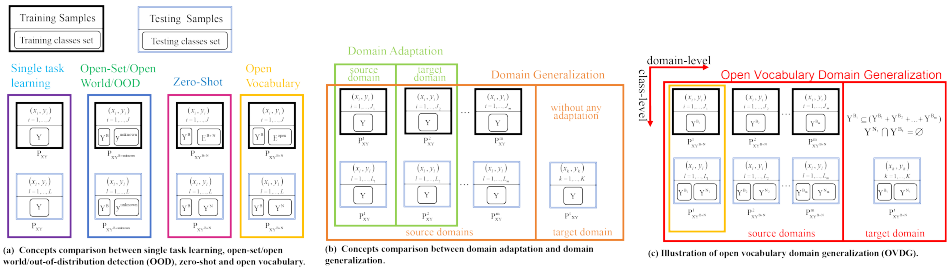
Keywords: open vocabulary; semantic segmentation; domain generalization; self-supervised learning

1. Introduction

Remote sensing images (RSI) are typically obtained from satellites, aerial platforms, or drones and provide valuable information about the Earth's surface and its features. Semantic segmentation of these images, where the goal is to classify each pixel in an image into predefined categories, is crucial for various applications including urban planning, environmental monitoring, agriculture, disaster management, and military intelligence. However, classes that are not labeled and not visible during the training process cannot be recognized during the inference stage, which greatly limits the scope of application. For instance, the Potsdam dataset [1], which is widely used for benchmarking semantic segmentation algorithms, only includes 5 classes. We call them base classes. But in reality, RSI often have more than 5 different types of objects. The classes that fall outside the 5 classes are called novel classes, which are presented as clutter or background in Potsdam dataset. To identify novel classes, researchers have introduced different concepts like open-set learning [2–4], open world learning [5,6], out-of-distribution detection (OOD) [7,8], zero-shot learning (ZSL) [9–16],and open vocabulary learning (OVL) [17–27]. The difference of these concepts are shown in Figure 1(a). The open-set, open world, and OOD tasks only need to identify novel classes and set them as one label, named 'unknown'. They do not need to identify the specific class categories in such settings. But

 doi:10.20944/preprints202501.0448.v1

zero-shot must classify novel classes into specific categories by using predefined word embeddings [28–30]. During training, zero-shot learning model is strictly trained on base classes. In the open vocabulary setting, the model can classify novel classes with the help of pretrained vision language models, which has large language vocabulary knowledge but is not strictly required to contain base classes and novel classes. Compared with zero-shot learning, open vocabulary learning can further extend models' generalizability on class-level [31]. Despite the resounding success of open vocabulary learning in the computer vision field, its potential application in the context of RS imagery remains relatively unexplored. In addition, open vocabulary learning assumes that samples come from a single known domain, resulting in limited applications in the real world.



(a) Concepts comparison between single task learning, open-set/open world/out-of-distribution detection (OOD), zero-shot and open vocabulary. (b) Concepts comparison between domain adaptation and domain generalization. (c) Illustration of open vocabulary domain generalization (OVDG).

Figure 1. Concepts comparison between different setting. In single-task learning, both training samples and test samples come from the same distribution P_{XY}. In the open-set/Open World/OOD settings, the model only needs to identify novel classes and mark them as 'unknown'. In the zero-shot setting, a model must classify novel classes into specific classes by using predefined word embedding. In the open vocabulary settings, the model only trains on base classes B and can classify novel classes N with the help of large language vocabulary knowledge instead of word embedding. In the domain adaption setting, the model is trained on a data distribution P^1_{XY} and is adapted to a unseen distributions P^2_{XY} by accessing unseen data. In domain generalization, the model is evaluated directly on data from unseen distribution P^t_{XY} without adaptation, while in open vocabulary domain generalization, the model needs to be performant on an any unseen distribution $P^1_{XY^{B+N}}$, where Y contains both base classes B_t and novel classes N_t.

Due to the influence of spatial resolution, shooting time, shooting equipment, and geographical location, RSI have different styles. Therefore, there is a significant domain shift between different remote sensing datasets. The model trained on a specific domain exhibits poor performance on the target domains, which have a different but related distribution. To this end, it is necessary to develop a model that can generalize to unseen distributions. As illustrated in Figure 1(b), domain adaptation (DA) focuses on adapting a model from one domain to another, while domain generalization aims to develop models that can generalize well across multiple domains without explicit domain-specific data during training. Besides, domain generalization (DG) involves addressing a difficult scenario where one or multiple distinct yet interconnected domains are provided, with the objective being to train a model capable of generalizing effectively to an unseen test domain. However, existing techniques for domain generalization presuppose the presence of identical classes across all domains, which limits the practical applicability of these methods. Our endeavor seeks to alleviate this constraint by enabling unseen test domains to contain novel classes absent in any training domain. As indicated in Figure 1(c), we introduce this more challenging setting as Open Vocabulary Domain Generalization (OVDG) for semantic segmentation, which, to the best of our understanding, represents the inaugural attempt at tackling this issue. Compared with DG, our OVDG demands not only considering the generalization performance on unseen domains but also identifying novel classes. The closest match to our setting is universal domain adaptation [32–34], and the primary distinction of our OVDG lies in involving source domains and identifying novel classes rather than simply labeling them as 'unknown'.

Existing semantic segmentation models [35–38] lack the capacity to handle OVDG task. For example, conventional semantic segmentation models are trained on closed set, which limits their ability to generalize at the class and domain levels. While there are existing methods [39–42] that focus on the model's capacity to generalize to new visual concepts, such as zero-shot learning (ZSL)

 doi:10.20944/preprints202501.0448.v1

or generalized zero-shot learning (GZSL), they are dependent on semantic information from novel classes, such as visual attributes [43,44] or Word2Vec vector representations [41,44].

In OVDG, the source data and target data remain isolated from each other. In this setting, aligning the domain distributions becomes unfeasible, unlike traditional methods [9,45,46] that rely on the simultaneous presence of source and target data. This paper introduces FreeMix, an efficient framework for Open Vocabulary Domain Generalization (OVDG). FreeMix includes a dual-branch segmentation module with a Base Segmentation Branch (BSB) and an Entity Segmentation Branch (ESB), along with a CLIP-based recognition module. The dual-branch design generates diverse masks and visual-semantic features for class- and domain-level generalization. The BSB integrates the CMID model [47], adapted for mask and semantic logit decoding as RS_SSL, while the ESB incorporates entity masks from a pre-trained extractor and a custom feature extractor. Outputs from both branches are fused for open-category recognition using CLIP. By freezing key components, we reduce trainable parameters and computation. To enhance performance across domains, we propose Dataset-Aware Sampling (DAS), ensuring uniform domain sampling in each mini-batch. Our contributions are as follows:

1. We introduce a new setting for semantic segmentation, i.e., open vocabulary domain generalization (OVDG), which is an important yet unstudied problem. In addition, we propose an effective framework FreeMix for solving OVDG, which focuses on learning a generalized model by integrating entity mask to enhance the diversity and completeness of masks for both base classes and novel classes.
2. We propose a dual-branch universal segmentation module by unifying the base segmentation branch (BSB) and the entity segmentation branch (ESB) in an end-to-end trainable framework, where the BSB leverages a self-supervised pre-trained model, CMID, to extract domain-agnostic visual features for decoding masks and semantic logits.
3. To integrate and leverage information from various source domains, we propose a simple yet effective training strategy, called dataset-aware sampling (DAS). Extensive experiments on four benchmark datasets reveal that our proposed method outperforms the state-of-the-art methods on the OVL and the OVDG benchmark.

In the following sections of this paper, we will first review the relevant literature in Section 2, which discusses related work in the field. Section 3 introduces our proposed FreeMix method, detailing its innovative approach and underlying principles. In section 4, we outline the experimental setup, datasets used, and present the results of our experiments, providing a comprehensive analysis of the performance of the FreeMix method. Finally, section 5 concludes the paper, summarizing our findings and suggesting directions for future research.

2. Related works

2.1. Open Vocabulary Semantic Segmentation

With the development of Visual Language Pre-training Models (VLPMs), such as CLIP [48] and ALIGN [49], models can now localize and recognize classes beyond the annotated label space, no longer confined to identifying predefined classes present in the training set. Many studies have successfully transferred their robust class generalization capabilities to pixel-level classification tasks, such as semantic segmentation. Based on a pretrained CLIP model, Chen *et al.* [50] use a conditional Unet model to predict segmentation masks and use text descriptions and annotations from OpenStreetMap as auxiliary supervision. Exploration of open vocabulary semantic segmentation in the remote sensing field is relatively limited, while it is more prevalent in natural images. ZegFormer [16], as a simple yet effective zero-shot semantic segmentation model, decomposes the problem into a class-agnostic segmentation task and mask classification tasks. Furthermore, it transfers semantic knowledge from seen classes to unseen classes solely with the assistance of VLPMs. Similarly, ZSSeg [51] proposes a two-stage semantic segmentation framework where the first stage extracts generic candidate masks, and the second stage utilizes CLIP model for open-vocabulary classification of the mask images

 doi:10.20944/preprints202501.0448.v1

generated in the first stage. To circumvent the time-consuming process of clipping image patches and computing features from an external pre-trained CLIP model, MaskCLIP [19] designs a Relative Mask Attention (RMA) module, treating segmentation as additional tokens for ViT CLIP models.

Unlike the two-stage segmentation models discussed earlier, SAN [52] attaches a lightweight side network to a pretrained VLM to predict candidate masks and classification outputs. To enhance the versatility of the framework, FreeSeg [24] jointly learns multiple related segmentation tasks, including Open Vocabulary Semantic Segmentation (OVSS), Open Vocabulary Instance Segmentation (OVIS), and Open Vocabulary Panoramic Segmentation (OVPS). Open vocabulary methods achieve impressive results on natural images, but remain relatively unexplored in the field of remote sensing. Existing visual language pre-training models lack participation of remote sensing images in training, resulting in limited generalization capabilities in the field of remote sensing. To address this limitation, some recent studies have specifically constructed large-scale image-text pairs datasets in the remote sensing domain for training visual language models. For example, RS5M [53], SkyScript [54], MMRS [55].

RemoteCLIP transforms heterogeneous annotations of detection boxes and segmentation masks into a unified image-caption data format through Box-to-Caption (B2C) and Mask-to-Box (M2B) strategies for training purposes. In addition, GeoRSCLIP [53] proposes an image-text paired dataset RS5M in the remote sensing domain, consisting of 5 million RS images with English descriptions. Furthermore, building upon the alignment of ground images with text, GRAFT [56] constructs pairs of ground images and remote sensing images to train VLP models, thereby enabling the training of remote sensing image vision-language models without using any text annotations.

2.2. Domain Generalization

Domain generalization (DG) task is one of the key challenges for deep learning models, facing domain shift between training and testing distributions. When the source domain consists of only one dataset, it simplifies to single-source domain generalization [57,58]. Existing domain generalization methods can be categorized into three main types. The first type [57,58] involves utilizing data augmentation to assist the model in learning universal representations, such as randomizing, transforming, or generating diverse data inputs to enhance the dataset. The second type [59–61] focuses on representation learning, aiming to learn domain-invariant representations or decompose representations into domain-shared and domain-specific components for better generalization, often incorporating feature alignment techniques to minimize discrepancies between different domains. The third type of methods [62–65] leverage general learning strategies to improve generalization capabilities, including ensemble learning, meta-learning, gradient manipulation, distributed robust optimization, and self-supervised learning. These three types of methods can complement each other and may be combined to achieve higher performance. Specifically, in the first type of methods, CCDR [57] performs texture and style randomization for simple yet effective auxiliary domain generation to improve the reliability of classification in arbitrary unseen target domains. Recently, several methods [66,67] have utilized Mixup [68] for domain generalization, generating new samples by applying Mixup directly in the original space.

The second category of methods has received considerable attention. A Maximum Mean Discrepancy Depth Reconstruction Classification Network (MMD-DRCN) [60] is proposed for detecting oil palm trees from multi-source high-resolution satellite images in a new environment. The core idea is to utilize the Maximum Mean Discrepancy (MMD) module to learn invariant features across different source domains. Moreover, to acquire domain-invariant features, the frequency-based optimal style mix (FOSMix) [59] model randomizes the styles of images in the source domain. Additionally, language-aware domain generalization network (LDGnet) [61] is proposed to learn cross-domain-invariant representation from cross-domain shared prior knowledge.

The third type of methods leverage general learning strategies to improve generalization capabilities. The approach of leveraging general learning strategies to improve generalization capabilities is simple yet effective, hence it has gained popularity. Segu *et al.*[69] maintain domain-specific batch normalization (BN) parameters for different source domains while sharing other parameters. Li *et*

 doi:10.20944/preprints202501.0448.v1

al.[62] proposes Meta-Learning for Domain Generalization (MLDG), applying meta-learning strategies to domain generalization. MLDG divides the data in the source domain into meta-training and meta-testing sets to simulate domain transfer scenarios for learning universal representations. Recently, Self-supervised Learning (SSL) has emerged as a popular learning paradigm, constructing self-supervised tasks from large-scale unlabeled data. SSL, as a universal paradigm, can be applied to any existing DG method, particularly for unsupervised domain generalization where labeled data is unavailable in the training domain. Bhattacharya *et al*.[70] proposes a self-supervised prompt learning approach for remote sensing images, which preserves domain-invariant feature learning while enhancing the representation of visual features. Our approach falls into this category by introducing a self-supervised backbone network to achieve domain generalization.

2.3. Self-Supervised Learning in Remote Sensing

Self-supervised learning (SSL) has garnered significant attention in the remote sensing community and has undergone initial exploration in this field. SSL methods leverage large amounts of unlabeled data to learn generic representations, which can enhance the performance of downstream tasks. Contrastive self-supervised learning involves constructing contrastive learning tasks to train models by leveraging the inherent positive and negative samples within remote sensing data. For instance, SauMoCo [71] utilizes the semantic similarity between nearby geographical locations and the inherent diversity within land cover concepts to train the model. Moreover, Kumar *et al*.[72] utilizes temporally aligned images as positive sample data and introduce an auxiliary task of predicting the source of the images to enhance pretraining effectiveness. Similarly, Oscar *et al*.[73] exploits seasonal information inherent in the data to construct tasks for SSL. In addition to temporal information, Dilxat *et al*.[74] proposes IndexNet, which learns spatiotemporal invariant features by combining image-level contrast and pixel-level contrast.

Compared to contrastive self-supervised learning, Masked Image Modeling (MIM) self-supervised learning has become more popular. Studies have focused on collecting data from various sources such as satellites or aerial platforms to build large-scale datasets covering multiple scenes worldwide for MIM self-supervised pretraining. RingMo [75] optimizes mask strategies for small objects in remote sensing images and employs the MAE model for self-supervised representation learning on a dataset comprising 3 million unlabeled remote sensing images. To better represent robust remote sensing data across various spatial scales, Scale-MAE [76] explicitly learns the relationships between data at different known scales throughout the entire pretraining process. In addition, to handle large-sized images and objects of different orientations in RS images, RVSA [77] introduces a new Rotated Varied-Size Window Attention (RVSA) mechanism, which significantly reduces computational costs and memory usage. Recently, a Transformer-based geographic spatial foundation model named Prithvi [78] has been proposed, pretraining it on over 1TB of multispectral satellite images from the Harmonized Landsat-Sentinel 2 (HLS) dataset.

Despite achieving some success, these methods have prerequisites in terms of underlying architectures; namely, contrastive self-supervised learning relies on CNNs, while most MIM methods are restricted to ViTs. In contrast, the CMID [47] method is not only agnostic to architectures but also combines contrastive and masked generation approaches to learn representations with global semantic separability and local spatial awareness. This ensures that the learned representations have sufficient generalization to meet the requirements of various downstream tasks in remote sensing. What's more, SMLFR [79] has constructed a large dataset named GeoSense, comprising approximately 9 million diverse remote sensing images, to enhance the robustness and generalization capabilities of foundation models during the pretraining phase. Additionally, they implemented masked image modeling (MIM) based on CNNs architectures.

 doi:10.20944/preprints202501.0448.v1

3. Proposed Method

3.1. Problem Definition

Let X represent visual space, and Y represent label space. We define a domain as the joint distribution of data space, it can be represented as: $D = \{(x_i, y_i) \in X_J \times Y_J\}_{i=1}^{J} \sim P_{XY}$, where x_i represents the i^{th} sample, y_i is the corresponding label and J is the size of the domain. We define Y^B as label space with base classes, Y^N as label space with novel classes, and Y^{B+N} as label space with both base and novel classes. As illustrated in Figure 1(c), we have m source domains $\mathbf{D_{src}} = \left\{ \left(D_i^{train}, D_i^{test}\right) \sim P^i_{XY^{B+N}} \right\}_{i=1}^{m}$ and a target domain $\mathbf{D_{tar}} = \{(x_k, y_k) \in X_K \times Y_K^{B_t+N_t}\}_{k=1}^{K} \sim P^t_{XY^{B_t+N_t}}$ with K samples. The target domain can be more than one, but for the sake of simplicity in explanation, only one is presented here. Here $D_i^{train} = \left\{ (x_j, y_j) \in X_J \times Y_J^{B_i} \right\}_{j=1}^{J} \sim P^i_{XY^{B+N}}$ is the training set of the i^{th} source domain, $D_i^{test} = \{(x_l, y_l) \in X_L \times Y_L^{B_i+N_i}\}_{l=1}^{L} \sim P^i_{XY^{B+N}}$ is the testing set of the i^{th} source domain, J is the size of the training set of the source domain, and L is the size of the testing set of the source domain. In the open vocabulary domain generalization (OVDG) setting, the base class labels of the target domain are a subset of the aggregated base class labels from the m source domains: $Y^{B_t} \subseteq \left(Y^{B_1} + Y^{B_2} + ... + Y^{B_m}\right)$. Additionally, novel class labels in the target domain do not intersect with base class labels, formally expressed as $Y^{N_t} \bigcap Y^{B_t} = \emptyset$. The main objective of OVDG setting is to train a model on the m source domain, and perform well on both target domain and source domains.

3.2. Overview

In this section, we describe our proposed framework for OVDG, named FreeMix. Figure 2 shows the overall architecture of our proposed FreeMix, which comprises two main module: a universal segmentation module and an open vocabulary recognition module. Furthermore, the universal segmentation module consists of a base segmentation branch (BSB) and an entity segmentation branch (ESB). In the OVDG task, our attention is required to encompass not just the OV aspect, but also the DG aspect. When tackling domain generalization challenges, we incorporate a self-supervised base model, CMID, within the BSB branch and efficiently adapt it using mask2former, thereby achieving seamless preservation of pre-trained knowledge. Additionally, to tackle the challenge of open vocabulary, in the ESB, we integrate a diverse range of entity masks and design an extractor specifically for extracting the visual features of these masks. This endeavor seeks to enhances the model's ability to segment and recognize both base and novel classes while maintaining generalizability. Finally, we utilize an open vocabulary recognition module to classify these masks based on CLIP [48] model. The classification method entails comparing the semantic logits of the masks with the cosine similarity of text embeddings derived from a contrastive language-image pre-trained model.

Next, we first introduce our universal segmentation module (Sec. 3.3), and than describe the proposed the dataset-aware sampling training tactic on multi-source domain (Sec. 3.4) thoroughly.

 doi:10.20944/preprints202501.0448.v1

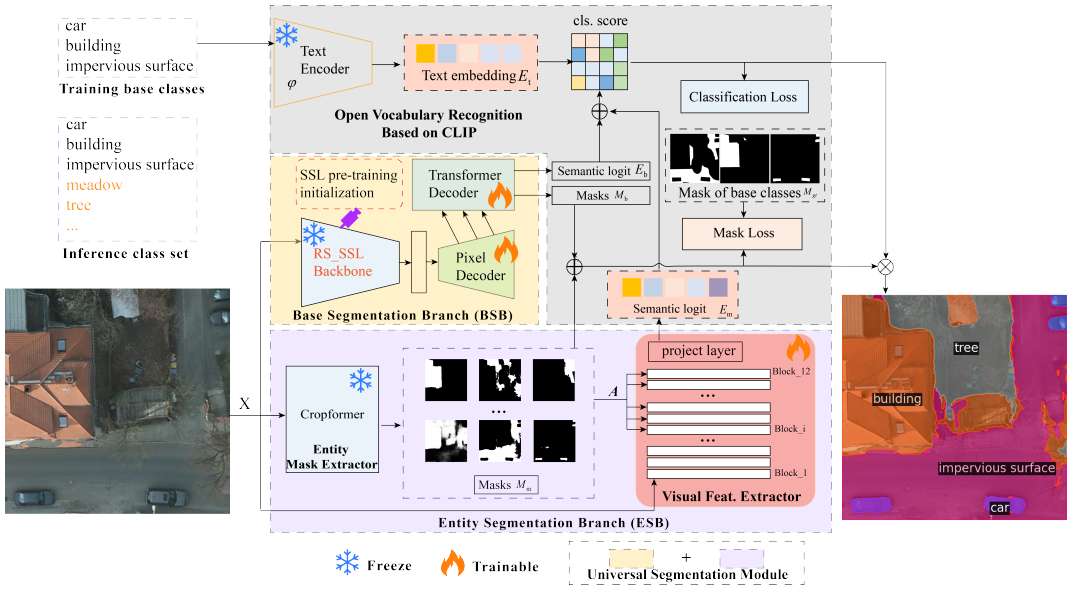


Figure 2. Overview of the proposed FreeMix. It consists of two branches: a base segmentation branch (BSB) and an entity segmentation branch (ESB). In order to enhance the extraction and recognition effectiveness of masks for both base and novel classes, we introduce the ESB to extract universal masks and their corresponding visual features. To maintain the model's generalization performance, we employ self-supervised learning initialization (RS_SSL) to BSB and freeze the image encoder of the BSB and the mask extractor of the ESB.

3.3. Universal Segmentation Module

The proposed universal segmentation module comprises two branches: a base segmentation branch (**BSB**) and an entity segmentation branch (**ESB**). Firstly, the diversity and completeness of extracted masks are pivotal factors influencing model performance in semantic segmentation. However, segmentation models trained on base classes often exhibit category bias, as novel classes are typically perceived as background and thus suppressed. Consequently, the mask proposal generator fails to truly achieve class-agnostic mask extraction, tending instead to extract masks corresponding to base classes. Even when unmatched candidate masks do not contribute to loss computation during training, this tendency persists. Moreover, due to disparities in data style between the target and source domains used for training, segmentation models trained on a specific data domain may exhibit sensitivity to stylistic differences in the target domain, leading to suboptimal outcomes. Hence, we endeavor to enhance model robustness and reduce sensitivity to data styles by incorporating dual branches (BSB and ESB). The BSB excels at segmenting the base classes in supervised learning. However, the single-branch BSB extracts very few masks for novel classes, resulting to low recall rates. Therefore, integrating an ESB is necessary to bolster mask extraction capabilities, especially for novel classes, by generating more universal masks.

1) Base Segmentation Branch (BSB) To make the model applicable across various data domains, we propose adapting self-supervised pre-trained backbone for remote sensing semantic segmentation, named RS_SSL. This backbone has been trained on large-scale remote sensing data. Consequently, the architecture of the BSB comprises a self-supervised pre-trained backbone network, a pixel decoder, and a Transformer decoder as illustrated in Figure 2. Given an image $X \in \mathbb{R}^{H\times W\times C}$, the BSB will output mask proposals $M_\text{b} \in \mathbb{R}^{N\times H\times W}$ and theirs semantic logits $E_\text{b} \in \mathbb{R}^{N\times d}$, where N is the number of mask proposals, d represents the dimensionality of semantic logits, and H, W and C correspond to image height, image width and number of channels. The backbone network of popular open-vocabulary segmentation models [16,23,24,51] is typically trained in a supervised manner on the ImageNet dataset [80] or trained from scratch. However, the major distinction lies in the fact that our backbone network is initialized using a self-supervised pre-trained model CMID [47] and remains frozen during training, with only the decoders being trainable. CMID combines contrastive learning and masked image modeling to learn robust representations in a self-distillation way. Moreover, it is architecture-agnostic,

 doi:10.20944/preprints202501.0448.v1

and is compatible with both convolutional neural networks (CNN) and vision transformers (ViT), allowing it to be easily adapted to our BSB. Compared to mask proposal generator of other models [52], we do not modify the foundational model, allowing us to leverage the powerful generalization capabilities of self-supervised learning in remote sensing images to extract domain-agnostic universal features.

2) Entity Segmentation Branch (ESB) To achieve the elimination of sensitivity across disparate data domains, the key is extracting class-agnostic masks and their corresponding visual features from images. In order to achieve this goal, we first employ CropFormer [81], a tiny entity segmentation model, as an entity mask extractor to generate high-quality entity masks. We then establish a visual feature extractor for these masks, leveraging the Transformer architecture.

Entity masks extractor. An entity refers to each semantically coherent region within an image. Entity segmentation is an emerging task that focuses on open-world, class-agnostic dense image segmentation. It is designed to have superior generalization capabilities for segmenting novel classes [82]. CropFormer [81], trained on a large-scale, high-quality entity segmentation dataset that includes images from various domains, including remote sensing images, is highly suitable for extracting class-agnostic masks. In entity masks extractor, we use N K-dimensional queries $Q \in \mathbb{R}^{N\times K}$ to generate entity masks $M_{\mathrm{m}} \in \mathbb{R}^{N\times H\times W}$.

Visual feature extractor. The visual feature extractor is based on vision Transformer (ViT) architecture which consists of 12 Transformer block. We denote these blocks as $B = \{b_1, b_2, \ldots, b_i, \ldots, b_{12}\}$. Each block comprises a multi-head attention layer followed by two MLP layers with GELU [83] non-linearity. Layer normalization is applied before each layer, and residual connections are added after each layer.

In the first k Transformer blocks, the visual feature extractor initially encodes the entire image $X \in \mathbb{R}^{H\times W\times C}$ to obtain a representation $E \in \mathbb{R}^{(hw+1)\times d}$. Here h and w represent the height and width of the attention map in the ViT, the additional 1 corresponds to the semantic logits for the entire image, and d denotes the dimensionality of the features. In the remaining $12-k$ Transformer blocks, to extract the semantic logits $E_{\mathrm{m}} \in \mathbb{R}^{N\times d}$ for entity masks $M_{\mathrm{m}} \in \mathbb{R}^{N\times H\times W}$, we assign independent classification queries to each entity mask by by repeating the semantic logits of the entire image N times. We then utilize the entity masks as the attention bias (A) in the Multihead Attention mechanism:

$$E_m^{b+1} = \mathrm{softmax}\left(\frac{\mathrm{Q}\left(E_m^b\right)\mathrm{K}\left(E^b\right)^{\mathrm{T}}}{\sqrt{d}} + A\right)\mathrm{V}\left(E^b\right) \tag{1}$$

where b indicates the block number, $K\left(E^b\right) = W_k E^b$ and $V\left(E^b\right) = W_v E^b$ are the key and value embeddings of the representation, and $Q\left(E_m^b\right) = W_q E_m^b$ is the query embedding of entity masks. Here, W_q , W_k , and W_v are the weights of the query, key and value embedding layer, respectively. To ensure the updating of semantic logits E_m^b for entity masks, we exclusively consider the representation corresponding to entity mask and its own semantic logits, without referencing the semantic logits of other entity masks. We construct a self-attention bias matrix $A \in \mathbb{R}^{N\times(N+hw)}$ as follows:

$$A(x,y) = \begin{cases} 0 & ,\text{if } \overline{M}(x,y) = 1 \\ -\infty & ,\text{if } \overline{M}(x,y) = 0 \end{cases} \tag{2}$$

$$\overline{M} = \mathrm{concat}[\mathrm{I}(N,N), f(M_m)] \tag{3}$$

where (x,y) is the feature location, $\mathrm{I}(N,N)$ denotes the (N,N) identity matrix, $f(M_m)$ represents resizing M_m to (h,w) and then flattening it, and $\overline{M} \in \{0,1\}^{N\times(N+hw)}$ is the binarized output (thresholded at 0.5) of the flattened mask.

In open vocabulary recognition module, given the base classes Y^B during training, classes prompts V_y are generated using the template: {"semantic segmentation" + learnable vectors + c }, where c

 doi:10.20944/preprints202501.0448.v1

represents the filled-in class names. Then the text prompts are then embedded using the pre-trained CLIP text encoder φ:

$$E_t = \varphi(V_y), y \in Y^B \tag{4}$$

To predict the class of masks $P \in \mathbb{R}^{C\times N}$, we compare the similarity between the semantic logits $E = \textbf{concat}[\psi(E_{\mathrm{b}}), \psi(E_{\mathrm{m}})]$ of mask groups $M = \mathrm{concat}[M_{\mathrm{b}}, M_{\mathrm{m}}]$ and the text embedding $E_t \in \mathbb{R}^{C\times r}$. Here $\psi()$ represents the normalization process, C is the number of classes and r is the dimension of text embeddings. Finally, we compute the semantic segmentation map $S = M \times P^T$.

3.4. Train Tactics: Dataset-Aware Sampling

In order to effectively improve the generalization ability of the model, it is necessary to train on multi-source datasets $D = \{D_1, D_2, ..., D_i, ...\}$ which exhibit different styles [84,85]. Each dataset has its own distinct label space $Y = \{Y_1, Y_2, ...Y_i, ...\}$. A straightforward approach to train on multiple datasets is to combine all annotations from these datatsets into a larger dataset $D = D_1 \cup D_2 \cup ...$, and then relabel and merge their label spaces into $Y = Y_1 \cup Y_2 \cup ....$ The model is optimized with the same loss on a larger dataset. Fortunately, the number of classes for labeling remote sensing semantic segmentation data is relatively small, making manual label mapping feasible. However, significant variations exist in dataset sizes: GID5 [86] contains 3× more images than Potsdam [1], DeepGlobe [87] is 8× larger than Potsdam [1], and URUR [88] is 79× larger. This imbalance in class distributions and dataset sizes virtually ensures that a mere concatenation of datasets will not work.

To effectively utilize multi-source datasets, we propose a simple yet efficient training strategy: Dataset-aware Sampling (**DAS**). Re-sampling is a widely used strategy in addressing class-imbalanced or long-tail learning scenarios [89]. Inspired by this, we extend the concept of sampling to the data domain level, introducing dataset-aware sampling. Specifically, we uniformly sample instances from each dataset within every mini-batch. For each dataset, we then compute the loss individually for its samples. Finally, we aggregate these per-dataset losses by averaging them before performing backpropagation:

$$L_{\mathrm{overall}} = \frac{1}{T}\sum_i L_i \tag{5}$$

where T is the number of datasets, and L_i represents the loss from the i^{th} dataset within a mini-batch. During training, L_i comprises both mask losses and classification loss. Specifically, the mask losses include Dice loss L_{dic} and binary cross-entropy loss L_{bce}, while the classification loss L_{cls} uses cross-entropy loss.

$$L_i = \lambda_1 L_{\mathrm{dic}} + \lambda_2 L_{\mathrm{bce}} + \lambda_3 L_{\mathrm{cls}} \tag{6}$$

where λ_1, λ_2, λ_3 are set to 20, 5 and 2, respectively, in our setting. In the proposed DAS, we do not need to design a separate head for each dataset, because we are addressing a single task across multiple datasets [90]. Consequently, there is no need to design distinct loss functions for each dataset or employ complex algorithms to search for the optimal weights of different losses. The experimental results in Section 4 demonstrate that our method, which employs a single model and a unified type of loss function, is both simple and effective for the OVDG task.

4. Experiments

4.1. Experimental Datasets and Processing

We evaluate the proposed methods on four well-known remote sensing datasets: (1) **Potsdam** [1] dataset is a widely used benchmark in remote sensing, comprising 38 images of size 6,000×6,000 pixels. It includes 6 classes, with 'background' as one of them. (2) **GID5** [86] contains 150 pixel-level annotated GF-2 images of 6,800×7,200, sampled from various cities in china. It has a total of 5 classes. (3) **DeepGlobe** [87] is a large-scale land-cover dataset which contains 803 images (2,448×2,448 pixels). It includes 7 classes of landscape regions, including "unknown" region. (4) **URUR** [88] features a substantial number of high-resolution images (3,008 images of size 5,120×5,120 pixels) covering a

 doi:10.20944/preprints202501.0448.v1

wide range of complex scenes from 63 cities. In open vocabulary setting, we first remove the "clutter" class from Potsdam, the "background" class from DeepGlobe, and the "other" category from URUR by setting the corresponding labels to 255, marking them as invalid classes. Next, we merge similar classes across datasets. For instance, categories such as "built up", "building", and "urban land" are unified and renamed as "building". Finally, we randomly partition the classes into base and novel classes across all datasets. We combine the four datasets into a single dataset, referred as GPDU, comprising a total of 12 classes. These include 7 base classes (building, farmland, forest land, impervious surface, car, range land, greenhouse) and 5 novel classes (meadow, water, tree, bare land, road). The details of class partitioning and the class mapping relationships can be found in Table 1. The class frequency distributions on the training and testing sets of the four datasets are illustrated in Figure 3.

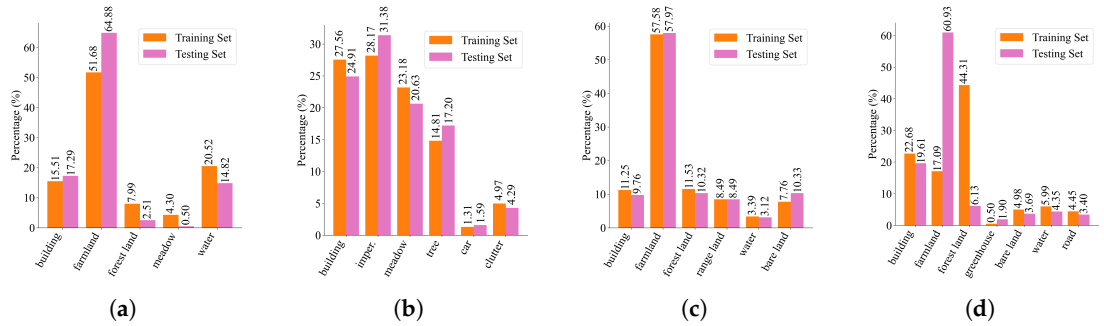


Figure 3. The class frequency distributions on the training and testing sets of the four datasets: (**a**) GID5, (**b**) Potsdam, (**c**) DeepGlobe, and (**d**) URUR.

Table 1. The mapping of classes for each dataset, as well as the partitioning into basic and novel classes. 'Original' indicates the annotated labels from the original dataset. 'Mapped' represents the labels after class mapping.

Dataset	Type	Base classes				Novel classes		
Potsdam	Original	impervious surface	building	car		low vegetation	tree	
	Mapped	impervious surface	building	car		meadow	tree	
GID5	Original	built up	farmland	forest		meadow	water	
	Mapped	building	farmland	forest land		meadow	water	
DeepGlobe	Original	urban land	agriculture land	range land	forest land	water	barren land	
	Mapped	building	farmland	range land	forest land	water	bare land	
URUR	Original	building	farmland	greenhouse	wood land	bare land	water	road
	Mapped	building	farmland	greenhouse	forest land	bare land	water	road

4.2. Implementation Details

In open vocabulary recognition module, we adopt the pre-trained vision-language model CLIP, employing the ViT-B backbone as the text encoder. For the base segmentation branch, we use ResNet50 [91] as the backbone. In the Entity Segmentation Branch (ESB), we employ CropFormer [81] with a Swin-Tiny backbone for entity mask extraction. To stabilize training and leverage pre-trained knowledge, we freeze several components: 1) The text and image encoders of CLIP to preserve the learned multimodal representations. 2) The entity mask extractor in ESB to maintain the robustness of the extracted masks. 3) The backbone of the base segmentation branch after initializing it with self-supervised pre-training weights. During training, we focus on fine-tuning the decoders of BSB and the visual feature extractor of ESB. We conduct training on an NVIDIA GTX 3090 GPU using a mini-batch size of 2 images. Note that, we use $N = 100$ queries in both branches and set $k = 8$ in visual feature extractor. For a fair comparison, we adopt the same training settings as FreeSeg [24]: The optimizer AdamW is adopted with an initial learning rate of 10^{-4} and weight decay of 10^{-4}. To avoid over-fitting on training set, the learning rate of image encoder is multiplied by a factor of $\lambda = 0.01$. There are totally 40k training iterations. Each dataset is separately divided into training and testing sets. During training, input images are cropped to 512×512 pixels.

 doi:10.20944/preprints202501.0448.v1

4.3. Evaluation Metrics

We evaluate our model and the baselines with Mean Intersection over Union (mIoU), Frequency Weighted IoU (fwIoU), Pixel Accuracy (pACC) and Mean Pixel Accuracy (mAcc). First, mIoU is a widely-used metric for evaluating semantic segmentation models, providing an average measure of segmentation accuracy across all classes. fwIoU adjusts for class imbalance by assigning more weight to frequently occurring classes during evaluation. pACC measures the percentage of correctly classified pixels in the entire image, while mACC calculates the average proportion of correctly classified pixels across all classes.

$$mIoU = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{7}$$

$$fwIoU = \frac{1}{\sum_{c=1}^{C} TP_c + FP_c + FN_c}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{8}$$

$$pACC = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} TP_c + FP_c + FN_c + TN_c} \tag{9}$$

$$mACC = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c} \tag{10}$$

where C is the number of classes, and for each class c: TP_c is the number of true positive pixels. FP_c is the number of false positive pixels. FN_c is the number of false negative pixels. TN_c is the number of true negative pixels. Further, we compute mIoU and pACC separately on the base and novel classes, denoted as follows: $mIoU_s$: Mean Intersection over Union for base classes, $mIoU_{us}$: Mean Intersection over Union for novel classes, $pACC_s$: Pixel Accuracy for base classes, $pACC_{us}$: Pixel Accuracy for novel classes.

4.4. Comparison with SOTA Methods

1) Results of open vocabulary semantic segmentation: Our proposed FreeMix is compared against several existing open vocabulary methods, including ZSSeg [51], ZegFormer [16], MaskCLIP [19], SAN [52], OVSeg [23] , FC-CLIP [92] and FreeSeg [24]. In the open vocabulary setting, the model is trained using only base classes and their corresponding masks. During testing, both base classes and novel classes are evaluated to assess the model's generalization capability. Table 2 shows the results on Potsdam dataset. For a fair comparison, all models use ResNet50 as the image encoder, with the exception of OVSeg, which uses ResNet101, and FC-CLIP, which employs a large version of ConvNeXt. The proposed FreeMix significantly outperforms other methods across multiple evaluation metrics. Specifically, it achieves the highest mean Intersection over Union (mIoU) of 63.44% and the highest mean Pixel Accuracy (mACC) of 73.87%, both of which are the best among all compared methods. Notably, FreeMix excels in segmenting base classes, achieving an mIoU of 86.46%. This high score underscores its proficiency in accurately delineating familiar object categories. Additionally, for novel classes, FreeMix attains an mIoU of 28.92%. Although this is lower than the performance on base classes, it still represents a commendable achievement considering the inherent challenges of recognizing previously unseen categories. These results collectively demonstrate that FreeMix not only maintains high accuracy for known classes but also generalizes well to novel classes, highlighting its robustness and adaptability in open vocabulary semantic segmentation tasks.

Compared to the second-best method, ZSSeg, our FreeMix achieves significant improvements across all metrics. This is primarily because ZSSeg relies on existing mask proposal networks without any optimization, leading to lower-quality mask generation. In contrast, our approach enhances mask quality through tailored optimizations, resulting in superior segmentation and recognition performance. Notably, as shown in Table 2, our FreeMix outperforms FC-CLIP, which uses a larger image encoder. While FC-CLIP achieves the highest Pixel Accuracy (pACC) on base classes, this is due to its frozen CNN-based CLIP backbone and supervised fine-tuning specifically for base classes,

 doi:10.20944/preprints202501.0448.v1

causing it to be more biased toward these categories. Our FreeMix incorporates a frozen self-supervised backbone in the base segmentation branch, contributing to strong performance on both base and novel classes. This design, facilitated by the two-branch architecture of FreeMix, helps maintain robust segmentation capabilities across all categories. In terms of novel classes, FreeMix demonstrates substantial improvements. Specifically, It gains an 11.11% IoU on the tree class and a 46.73% IoU on the meadow class. These improvements are particularly noteworthy because certain models, such as ZSSeg and ZegFormer, struggle to distinguish between similar classes like trees and meadows in remote sensing images. The Entity Segmentation Branch (ESB) of FreeMix plays a crucial role by generating universal masks and extracting corresponding visual features, thereby enhancing the model's ability to generalize to novel classes. Overall, the proposed FreeMix can generalize well to novel classes while maintaining strong performance on base classes. This balanced approach ensures that FreeMix has the best overall performance, as evidenced by its superior results across multiple evaluation metrics.

In addition, The qualitative results on the Potsdam testing set are depicted in Figure 4, our FreeMix obtains accurate semantic segmentation for both base and novel classes. Notably, when segmenting novel classes, models often struggle due to confusion between similar categories. However, as shown in the top row and third row of Figure 4, our segmentation module excels at distinguishing between challenging classes such as "meadow" and "tree," outperforming any off-the-shelf model. This demonstrates the strong open-vocabulary segmentation ability of our model.

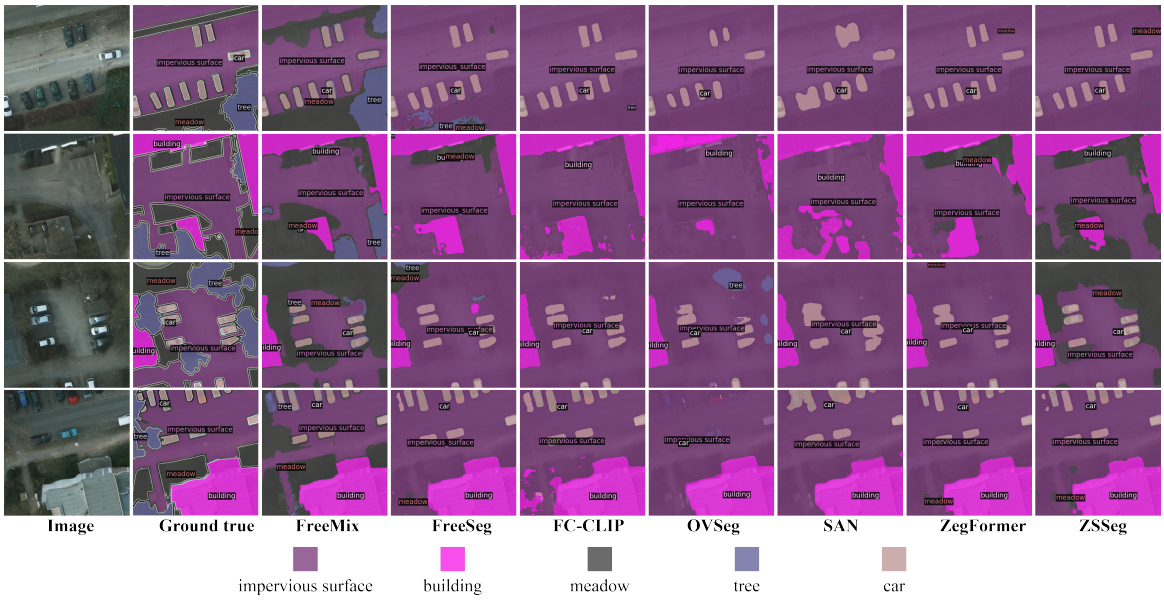


Figure 4. Qualitative results of semantic segmentation Potsdam dataset. The images are taken from the Potsdam testing set.

2) Results of OVDG: We also verify the generalization ability of our FreeMix across different datasets. Specifically, the model is trained on one dataset and directly evaluated on another dataset without fine-tuning. This setup presents a challenging task where the model must handle both novel classes and address the domain gap between different datasets. We report the mean Intersection over Union (mIoU) for both base and novel classes. As shown in Table 3, FreeMix consistently achieves the highest average mIoU and mACC across both the source domain and the three target domains, regardless of the training dataset. Specifically, when trained on the GID5 dataset, FreeMix achieves the highest average mIoU of 25.46%, and when trained on the URUR dataset, it attained the highest average mAcc of 39.98%. FreeMix demonstrates strong single-source domain generalization capabilities, with robust performance across different source domains. While the overall performance remains consistently high, we observe minor variations depending on the specific training dataset used. These variations likely stem from the differing characteristics of each domain, which can impact

 doi:10.20944/preprints202501.0448.v1

the ease of generalization. Overall, the GID5 dataset proves to be a particularly strong choice for achieving better generalization performance. The consistent high performance across multiple domains underscores the robustness and adaptability of FreeMix in open vocabulary domain generalization tasks.

Notably, when trained on the Potsdam dataset, FreeMix generalizes well to GID5 but underperforms when tested on DeepGlobe and URUR. A similar trend can be observed in other models. This phenomenon may be attributed to the relatively smaller size of the Potsdam dataset compared to others, which renders it more susceptible to overfitting given an equivalent number of training iterations. In contrast, the GID5 dataset—closer in size to Potsdam—demonstrates superior performance relative to both DeepGlobe and URUR. However, this also presents a significant limitation. Training on datasets with a narrow range of classes, such as Potsdam—which predominantly contains urban categories—can lead to suboptimal performance when FreeMix is tested on datasets like DeepGlobe. DeepGlobe includes natural and semi-natural classes that are not present in Potsdam. This mismatch in class distributions underscores the critical importance of dataset composition in domain generalization tasks. Additionally, our FreeMix not only enhances generalization performance but also excels in open-vocabulary segmentation capabilities. When trained on GID5 and tested on other dataset, it attains $mIoU_{us}$ of 15.83% on Potsdam, 12.38% on DeepGlobe and 9.79% on URUR. Besides, $mIoU_{us}$ reaches 20.26% on GID5 and 6.49% on URUR when FreeMix only trained on DeepGlobe dataset, which outperforms FreeSeg with +4.30% and +3.78%, respectively. These results demonstrate that our proposed FreeMix not only performs excellently within the domain but also exhibits strong generalization capabilities across other data domains. When trained and tested on the single source domain URUR, FreeMix surpasses FreeSeg, achieving mIoU and mACC that are 8.73% and 7.74% higher, respectively. Moreover FreeMix generalizes well to other target domain datasets, with the exception of GID5. Through analysis, we found that the IoU for the easily confused classes, **farmland** from the base classes and **meadow** from the novel classes, is lower in FreeMix compared to FreeSeg. Therefore, enhancing the model's ability to accurately recognize similar classes in cross-domain scenarios presents a more challenging and promising direction for future work.

MaskCLIP shares some similarities with our ESB, as both methods utilize a class-agnostic mask proposal network and extract features from a pretrained CLIP ViT model. However, FreeMix's design incorporates several enhancements that lead to higher-quality masks and richer semantic features: the Entity Mask Extractor and BSB generate more accurate masks compared to MaskCLIP. And the ESB in FreeMix includes an independent Visual Feature Extractor, which contributes to the richness of the extracted features. This analysis reveals that both mask quality and the semantic richness of the features play a critical role in determining the final segmentation performance. The approach most akin to ours BSB is FC-CLIP [24], which involves directly substituting the backbone network of the mask generator with CLIP's image encoder and subsequently freezing it during training. This strategy enables maximal retention of the model's original generalization capabilities while concurrently reducing the memory requirements during training. Despite using a larger image encoder, FC-CLIP is outperformed by FreeMix across all datasets. Notably, FreeMix achieves these results without employing multi-scale test-time augmentation, which other models use. Intuitively, the improved modules in FreeMix can more effectively generate universal masks for both base and novel classes, thereby enhancing its segmentation generalization ability on both class and domain levels. These findings underscore the effectiveness of the proposed method in open domains.

 doi:10.20944/preprints202501.0448.v1

Table 2. Open-vocabulary semantic segmentation on Potsdam dataset. The proposed FreeMix demonstrates better performances than prior arts. "imper." indicates the class of impervious surface, and "VLM" stands for Vision Language Model. The best results are marked in bold, while the second best results are underlined. (Values in %).

Method	Year	Image encoder	VLM	Potsdam								IoU of base classes			IoU of novel classes	
				mIoU	$mIoU_s$	$mIoU_{us}$	*fwIoU*	*mACC*	*pACC*	$pACC_s$	$pACC_{us}$	imper.	building	car	tree	meadow
ZSSeg	2021	ResNet50	CLIP-B/16	54.27	78.49	17.94	51.02	66.71	66.98	88.05	34.75	59.74	85.31	90.42	0.00	35.88
ZegFormer	2022	ResNet50	CLIP-B/16	49.20	71.99	15.01	45.24	61.73	62.27	84.67	27.99	53.93	75.52	86.51	0.00	30.03
MaskCLIP	2023	ResNet50	CLIP-L/16	15.58	21.84	6.19	21.50	28.54	39.46	60.16	7.78	32.24	33.29	0.00	**11.23**	1.16
SAN	2023	ResNet50	CLIP-B/16	38.56	60.25	6.02	38.82	59.80	60.71	96.01	6.70	52.94	69.04	58.77	2.12	9.92
OVSeg	2023	ResNet101	CLIP-B/16	31.56	50.43	3.25	35.07	43.49	54.28	87.44	3.54	41.72	74.62	34.96	0.21	6.28
FC-CLIP	2023	ConvNeXt_L	CLIP-RN50	44.78	73.74	1.32	39.03	59.12	59.76	**97.85**	1.48	48.12	81.32	91.79	0.05	2.60
FreeSeg	2023	ResNet50	CLIP-B/16	51.25	75.89	14.29	46.57	64.12	65.10	95.89	18.00	53.99	81.86	**91.82**	3.54	25.05
FreeMix(ours)	2024	ResNet50	CLIP-B/16	**63.44**	**86.46**	**28.92**	**64.45**	**73.87**	**75.87**	89.92	**54.37**	**83.89**	**90.16**	85.32	11.11	**46.73**

Table 3. Generalization performance of the open-vocabulary domain generalization for semantic segmentation. All models are trained on Single dataset. "MS" indicates multi-scale testing, "SS" indicates single-scale testing. The best results are marked in bold, while the second best results are underlined. (Values in %)

Training dataset	Model	Testing type	Testing dataset: Potsdam				Testing dataset: GID5				Testing dataset: DeepGlobe				Testing dataset: URUR				avg. mIoU	avg. mAcc
			$mIoU_s$	$mIoU_{us}$	*mIoU*	*mAcc*	$mIoU_s$	$mIoU_{us}$	*mIoU*	*mAcc*	$mIoU_s$	$mIoU_{us}$	*mIoU*	*mAcc*	$mIoU_s$	$mIoU_{us}$	*mIoU*	*mAcc*		
Potsdam	ZSSeg	MS	78.49	17.94	54.27	66.71	1.73	13.66	6.50	34.53	0.15	10.83	3.71	18.85	0.15	3.13	1.43	14.99	16.47	33.77
	ZegFormer	MS	71.99	15.01	49.20	61.73	0.58	4.87	2.30	20.33	**12.56**	2.13	**9.08**	15.95	**14.17**	1.61	**8.79**	13.85	17.34	27.96
	MaskCLIP	MS	21.84	6.19	15.58	28.54	13.85	0.35	8.45	23.68	7.07	0.00	4.71	14.68	7.08	0.00	4.04	12.12	8.19	19.75
	FC-CLIP	MS	73.74	1.32	44.78	59.12	12.51	0.00	7.51	16.64	5.55	0.00	3.70	13.44	7.87	0.00	4.50	8.39	15.12	24.39
	FreeSeg	MS	75.89	14.29	51.25	64.12	2.99	**18.80**	9.31	33.35	2.96	**11.87**	5.93	**22.65**	1.56	**10.34**	5.32	**21.59**	17.95	35.42
	FreeMix(ours)	SS	**86.46**	**28.92**	**63.44**	**73.87**	**15.73**	16.31	**15.96**	**43.90**	3.41	8.93	5.25	19.95	3.53	3.30	3.43	15.76	**22.02**	**38.37**
GID5	ZSSeg	MS	0.00	10.77	4.31	20.00	33.15	0.63	20.14	37.60	3.03	5.85	3.97	18.61	9.45	3.13	6.74	19.43	8.79	23.91
	ZegFormer	MS	6.35	12.30	8.73	23.93	28.16	4.18	18.57	38.21	**28.46**	0.27	19.06	29.75	6.35	**12.30**	8.73	23.93	13.77	28.95
	MaskCLIP	MS	21.06	8.60	16.08	29.19	16.45	0.66	10.13	20.78	10.71	0.00	7.14	16.49	9.89	0.00	5.65	9.98	9.75	19.11
	FC-CLIP	MS	**22.78**	10.00	**17.67**	**36.48**	6.12	0.13	3.72	19.66	3.59	0.00	2.40	16.54	3.87	0.01	2.21	10.55	6.50	20.80
	FreeSeg	MS	3.32	**17.59**	9.02	23.46	73.36	22.22	52.91	61.88	19.05	8.81	15.64	26.30	15.86	1.72	9.80	15.51	21.84	31.78
	FreeMix(ours)	SS	8.33	15.83	11.33	26.36	**76.47**	**22.55**	**54.90**	**65.44**	23.01	**12.38**	**19.47**	**35.81**	**20.95**	9.79	**16.17**	**26.48**	**25.46**	**38.52**
DeepGlobe	ZSSeg	MS	5.43	11.51	7.86	23.29	14.59	9.85	12.69	32.53	0.85	5.60	2.44	17.18	0.93	5.32	2.81	20.31	6.45	23.32
	ZegFormer	MS	0.00	12.27	4.91	20.95	0.17	0.25	0.20	20.10	7.20	5.70	6.70	19.71	0.01	1.13	0.49	14.28	3.07	18.76
	MaskCLIP	MS	16.43	7.51	12.86	23.85	6.51	0.00	3.90	20.36	9.95	0.00	6.63	26.26	5.59	0.00	3.19	14.53	6.64	21.25
	FC-CLIP	MS	**24.89**	5.92	17.30	37.53	5.77	0.00	3.46	19.72	2.71	0.00	1.80	14.25	3.74	0.00	2.14	8.73	6.17	20.05
	FreeSeg	MS	17.62	**22.61**	**19.61**	37.10	**41.40**	15.96	**31.22**	44.04	9.44	7.03	8.63	23.16	8.44	2.71	5.99	17.81	16.36	30.52
	FreeMix(ours)	SS	17.89	17.14	17.59	**39.37**	32.89	**20.26**	27.84	**49.37**	**24.97**	**9.35**	**19.76**	**33.88**	**19.12**	**6.49**	**13.71**	**24.97**	**19.72**	**36.89**
URUR	ZSSeg	MS	2.92	11.12	6.20	21.64	7.53	8.36	7.86	33.10	5.52	6.39	5.81	20.30	5.18	1.87	3.76	16.37	5.90	22.85
	ZegFormer	MS	0.59	5.34	2.49	22.72	0.76	0.25	0.56	20.47	10.56	0.00	7.04	22.31	0.02	1.13	0.50	14.30	2.64	19.95
	MaskCLIP	MS	12.94	12.63	12.82	28.19	15.39	0.44	9.41	21.48	10.39	0.00	6.93	17.01	12.24	0.00	6.99	13.17	9.03	19.96
	FC-CLIP	MS	**26.37**	8.89	**19.38**	5.17	5.74	0.00	3.44	19.89	2.97	0.77	2.24	16.36	3.78	0.00	2.16	10.19	6.80	20.40
	FreeSeg	MS	12.95	**22.56**	16.79	32.22	**43.92**	**21.93**	**35.12**	**57.84**	21.25	8.05	16.85	32.12	21.00	5.81	14.49	24.71	20.81	36.72
	FreeMix(ours)	SS	15.70	21.73	18.11	**36.12**	33.06	16.97	26.62	54.02	**28.39**	**13.08**	**23.28**	**37.33**	**33.29**	**9.80**	**23.22**	**32.45**	**22.80**	**39.98**

 doi:10.20944/preprints202501.0448.v1

4.5. Experiments on Multi-Source Domain

To evaluate the performance of FreeMix across multiple source domains, we train it on a combined dataset named GPDU. This setup allows us to compare FreeMix with other state-of-the-art models under fair conditions, where all models are trained for 40K iterations. It is important to note that existing models are typically limited to training on a single dataset at a time. To enable a comprehensive comparison, we created the GPDU dataset by performing label mapping for the training of these models. In contrast, FreeMix can be trained on multiple datasets simultaneously, demonstrating its scalability in handling multi-dataset training.

As depicted in Table 4, FreeMix consistently outperforms other models across all datasets. The implementation of the universal segmentation module with a dual-branch architecture and dataset-aware sampling (DAS) has helped FreeMix set a new benchmark, achieving the highest average mIoU. In comparison, FreeSeg and SAN exhibit significantly lower performance, underscoring the superiority of FreeMix in semantic segmentation tasks. Specifically, the use of DAS contributes to an overall improvement in FreeMix's average mIoU. However, this enhancement reveals a trade-off between base and novel class performance. Significant improvements are observed on datasets like Potsdam and URUR. There is a noticeable decrease, particularly on GID5 and DeepGlobe. This trade-off highlights the need for further optimization to balance the performance between base and novel classes when applying DAS. Addressing this issue could lead to more consistent and robust performance across all classes and datasets.

Table 4. Semantic segmentation result on testing set of joint dataset GPDU. "MS" indicates multi-scale testing, "SS" indicates single-scale testing. FreeMix† indicates that DAS is not used during the training phase. The best results are marked in bold, while the second best results are underlined. (Values in %).

Model	Training dataset	Testing type	Potsdam			GID5			DeepGlobe			URUR			*avg.mIoU*
			mIoU	$mIoU_s$	$mIoU_{us}$	*mIoU*	$mIoU_s$	$mIoU_{us}$	*mIoU*	$mIoU_s$	$mIoU_{us}$	*mIoU*	$mIoU_s$	$mIoU_{us}$	
ZSSeg	GPDU	MS	9.84	10.85	8.32	3.66	0.02	9.13	2.38	0.00	7.16	1.94	1.65	2.32	4.46
ZegFormer			4.31	0.00	10.77	0.50	0.83	0.00	8.11	11.59	1.15	0.48	0.00	1.13	3.35
MaskCLIP			11.72	18.78	1.13	13.62	22.70	0.00	9.20	13.80	0.00	6.87	12.02	0.00	10.35
SAN			23.84	23.99	23.61	35.10	57.67	1.25	29.38	42.86	2.41	30.44	48.39	6.49	29.69
OVSeg			9.32	14.96	0.86	15.58	22.15	5.73	25.87	37.57	2.47	19.56	31.86	3.16	17.58
FC-CLIP			21.02	28.76	9.40	2.87	4.79	0.00	1.81	2.72	0.00	1.48	2.60	0.00	6.80
FreeSeg			17.58	14.75	21.83	25.26	33.94	12.24	24.55	31.58	10.49	23.71	38.99	3.34	22.78
FreeMix†(ours)	GPDU	SS	19.98	17.25	24.06	**57.26**	**75.91**	**29.27**	32.03	41.17	**13.75**	29.3	42.41	**11.83**	34.64
FreeMix(ours)	GPDU	SS	**47.03**	**69.54**	13.26	43.13	67.91	5.97	**35.14**	**52.69**	0.04	**35.72**	**60.85**	2.22	**40.26**

4.6. Ablation Experiments

In this section, we conduct ablation studies on our FreeMix using the joint GPDU dataset. All experiments utilize a ResNet50 backbone and the CLIP-B/16 VLM, with FreeSeg serving as our baseline for comparison. Table 5 summarizes the effectiveness of three key components: the proposed initialization method with self-supervised learning in remote sensing (RS_SSL), the entity segmentation branch (ESB) and the dataset-aware sampling (DAS) training tactic. These results clearly indicate that each proposed component consistently enhances the overall performance of FreeMix, highlighting their individual and combined contributions to improved semantic segmentation outcomes. Using RS_SSL initialization provides a substantial improvement over the baseline, indicating the value of leveraging self-supervised learning specifically tailored for remote sensing tasks. The introduction of ESB yields consistent performance boosts across all datasets, suggesting its critical role in capturing more accurate and detailed entity representations. DAS not only elevates mIoU and mACC but also demonstrates its ability to balance performance across diverse datasets, thereby ensuring robust generalization.

When using supervised pre-training on ImageNet1K to initialize the image encoder and fine-tuning it on the GPDU dataset, we observe a degradation in performance. However, with our RS_SSL initialization method and freezing the image encoder during training, the model can extract domain-agnostic universal features, thereby accelerating decoder optimization and improving learning efficiency. Furthermore, incorporating the ESB branch to construct a dual-branch network further enhances performance. The ESB effectively extracts entity masks and captures accurate semantic features for classification, contributing significantly to better segmentation results. Additionally, employing DAS improves overall performance, especially on datasets with relatively small amounts

 doi:10.20944/preprints202501.0448.v1

of data. For instance, on the Potsdam dataset, mIoU increased from 19.98% to 47.03%, demonstrating substantial improvement. Even on larger datasets like URUR, there was an increase in mIoU from 29.30% to 35.72%. The above findings indicate that our proposed methods—RS_SSL initialization, ESB, and DAS—are highly effective in enhancing the performance of FreeMix. The qualitative results of FreeMix are visualized in Figure 5, providing a clear demonstration of its superior performance in semantic segmentation tasks.

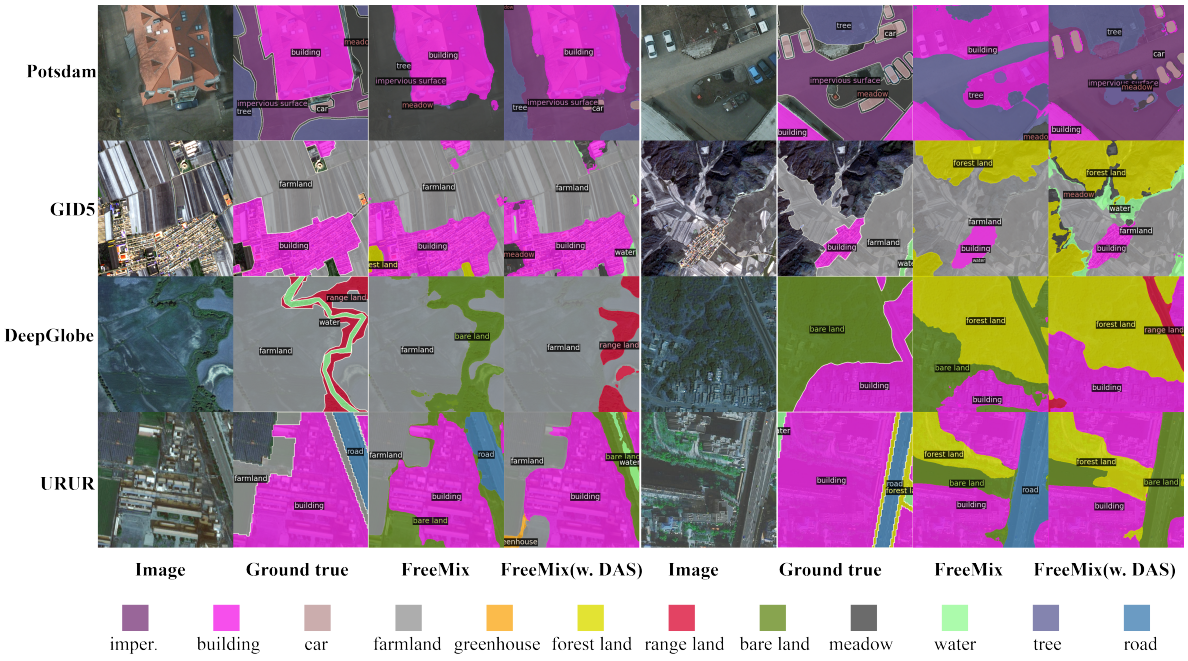


Figure 5. Qualitative results on Potsdam, GID5, DeepGlobe and URUR dataset.

Table 5. Ablations on RS_SSL, Entity Segmentation Branch and Dataset-aware Sampling training tactic. The best results are marked in bold, while the second best results are underlined. (Values in %).

Training Dataset	RS_SSL	ESB	DAS	Potsdam		GID5		DeepGlobe		URUR		*avg.mIoU*	*avg.mACC*	*ΔmIoU*	*ΔmACC*
				mIoU	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*				
GPDU				17.58	32.58	25.26	32.92	24.55	39.75	23.71	30.98	22.78	34.05		
	√			19.73	<u>37.32</u>	38.53	57.22	31.97	<u>47.06</u>	26.46	34.49	29.17	44.02	+6.39	+9.96
	√	√		<u>19.98</u>	35.92	**57.26**	**71.10**	<u>32.03</u>	46.49	<u>29.30</u>	<u>37.85</u>	<u>34.64</u>	<u>47.84</u>	+5.47	+3.81
	√	√	√	**47.03**	**62.56**	<u>43.13</u>	<u>58.44</u>	**35.14**	**47.37**	**35.72**	**45.75**	**40.26**	**53.53**	+5.62	+5.69

4.7. Additional Experimental Results

1) Performance on different image encoder of BSB: To further demonstrate the adaptability of the proposed RS_SSL adaptation strategy, we conduct experiments on both CNN and ViT architectures. In FreeMix, the backbone of BSB is from CMID [47], which is pre-trained in a self-supervised manner on the remote sensing dataset MillionAID. In contrast, other methods are either trained in a supervised manner on the ImageNet1K dataset or train the backbone network from scratch. Furthermore, we report comparative results under different training tactic: random sampling and the proposed DAS. According to the Table 6 , using ResNet50 as backbone and employing RS_SSL initialization to BSB, our FreeMix achieves average mIoU scores of 34.64% and 40.25% when using random sampling and DAS training tactic, respectively. It should be noticed that, without introducing RS_SSL initialization to BSB, the performance will drop significantly to average mIoU scores of 29.17% (random sampling) and 23.71% (DAS). Besides, applying RS_SSL initialization to Transformer-based backbone network Swin-B, the model gains average mACC of 61.00% and 64.29% when using random sampling and DAS, respectively. Compared to training BSB from scratch, RS_SSL initialization improved average mACC across the four datasets by 1.27% (random sampling) and 7.08% (DAS). These results indicate that RS_SSL initialization is effective for both CNN and ViT architectures. Meanwhile, it demonstrates that

 doi:10.20944/preprints202501.0448.v1

FreeMix is architecture-agnostic, allowing further performance improvement by leveraging superior pre-trained models of both convolutional neural networks and vision transformers.

Table 6. Ablation study on different image encoders and pre-training methods. All methods utilize CLIP with a ViT-B backbone. The training dataset used is GPDU, which integrates four remote sensing datasets. "In1K" denotes the ImageNet-1k dataset. The best results within each group are highlighted in bold. (Values in %).

Backbone	Pre-train Type	Pre-train Dataset	Training tactic	Potsdam		GID5		DeepGlobe		URUR		*avg.mIoU*	*avg.mACC*
				mIoU	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*		
ResNet50	Supervised	In1K	random	19.73	**43.66**	38.53	76.23	31.97	**63.66**	26.46	70.45	29.17	63.50
ResNet50	Self-Supervised	MillionAID	random	**19.98**	40.43	**57.26**	**87.65**	**32.03**	63.24	**29.30**	**71.70**	**34.64**	**65.75**
ResNet50	Supervised	In1K	DAS	39.49	61.35	41.18	78.04	7.81	22.71	6.39	19.02	23.71	45.28
ResNet50	Self-Supervised	MillionAID	DAS	**47.03**	**62.47**	**43.13**	**81.90**	**35.14**	**69.38**	**35.72**	**78.45**	**40.25**	**73.05**
Swin-B	–	–	random	**11.79**	30.92	33.76	73.99	22.20	**62.06**	26.05	71.96	23.45	59.73
Swin-B	Self-Supervised	MillionAID	random	11.54	**31.09**	**39.31**	**78.63**	**23.22**	61.91	**27.71**	**72.38**	**25.44**	**61.00**
Swin-B	–	–	DAS	36.26	60.70	34.29	72.58	19.78	48.39	15.05	47.18	26.34	57.21
Swin-B	Self-Supervised	MillionAID	DAS	**43.85**	**66.57**	**40.11**	**79.49**	**23.81**	**56.59**	**18.44**	**54.53**	**31.55**	**64.29**

2) Performance on scaling model size of ESB: To evaluate the impact of scaling the model size of the ESB, we train 6 FreeMix models, sweeping over backbone of ESB (Swin-Tiny, Swin-Large, Hornet-Large) and training tactics (random sampling, DAS). The main results are illustrated in Table 7. Swin-Tiny (Swin-T) consistently shows strong performance across the datasets, particularly when paired with the DAS training tactic. It achieves the highest average mIoU (40.25%) and mACC (73.05%). This confirms that smaller backbones like Swin-T perform more efficiently in terms of both segmentation accuracy and generalization across domains. Swin-Large (Swin-L), while larger in size, shows a drop in where its average mIoU falls to 32.51% and mACC to 49.67%. This performance drop may be attributed to Swin-L requiring more data to train effectively, as larger models often need more data to fully capture patterns and avoid overfitting. Hornet-Large (Hornet-L) performs well in certain metrics, such as achieving the highest mACC of 67.61% with random sampling and 66.17% on GID5 with DAS, but it underperforms in terms of average mIoU, especially on DeepGlobe and URUR. The mixed performance indicates that while Hornet-L excels in certain conditions, it may not generalize as effectively across all domains. In conclusion, these results underscore that smaller backbones like Swin-T not only maintain better overall performance but also generalize more effectively across datasets, particularly when using the DAS training tactic. Conversely, larger models like Swin-L and Hornet-L show diminishing returns, especially in terms of mIoU, suggesting that adding a lightweight ESB branch can lead to greater efficiency without sacrificing accuracy.

Table 7. Scalability of Entity Segmentation Branch. All methods utilize CLIP with a ViT-B backbone. The training dataset is GPDU, which combines four remote sensing datasets. The best results within each group are highlighted in bold. (Values in %).

Backbone of ESB	Training tactic	Potsdam		GID5		DeepGlobe		URUR		*avg.mIoU*	*avg.mACC*
		mIoU	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*	*mIoU*	*mACC*		
Swin-T	random	19.98	40.43	**57.26**	**87.65**	**32.03**	63.24	**29.30**	71.70	**34.64**	65.75
Swin-L		**23.31**	**45.90**	55.07	88.71	24.85	50.92	23.40	59.67	31.65	61.30
Hornet-L		21.86	45.65	53.14	88.63	30.40	**64.15**	27.75	**72.03**	33.28	**67.61**
Swin-T	DAS	47.03	62.47	43.13	81.90	**35.14**	**69.38**	**35.72**	**78.45**	**40.25**	**73.05**
Swin-L		47.39	57.09	57.11	**85.90**	10.89	31.00	14.66	24.70	32.51	49.67
Hornet-L		**53.68**	**66.17**	**53.40**	82.55	12.91	36.81	15.67	30.10	33.91	53.90

3) Comparison of the extracted proposal masks: We visualize and compare the proposal masks extracted by FreeSeg and our FreeMix, as shown in Figure 6. FreeSeg processes 100 queries to generate 100 proposal masks. For visualization purposes, we randomly select 10 of these masks. Besides, FreeMix generates 100 proposal masks for each branch (BSB and ESB). And we randomly select 10 masks from each branch for visualization and comparison, indicated in yellow (BSB) and purple (ESB) in Figure 6. Based on the results, it is evident that the proposal masks generated by FreeSeg exhibit low distinguishability and low predicted confidence, as indicated by their gray appearance. In contrast, the masks decoded by FreeMix's BSB, with the assistance of RS_SSL, exhibit higher confidence and greater distinguishability. Additionally, the ESB of FreeMix offers a richer set of entity masks, capturing more detailed and accurate representations of entities within the images. By combining these high-quality

 doi:10.20944/preprints202501.0448.v1

universal masks from both branches, FreeMix demonstrates superior performance on OVDG task compared to other models.

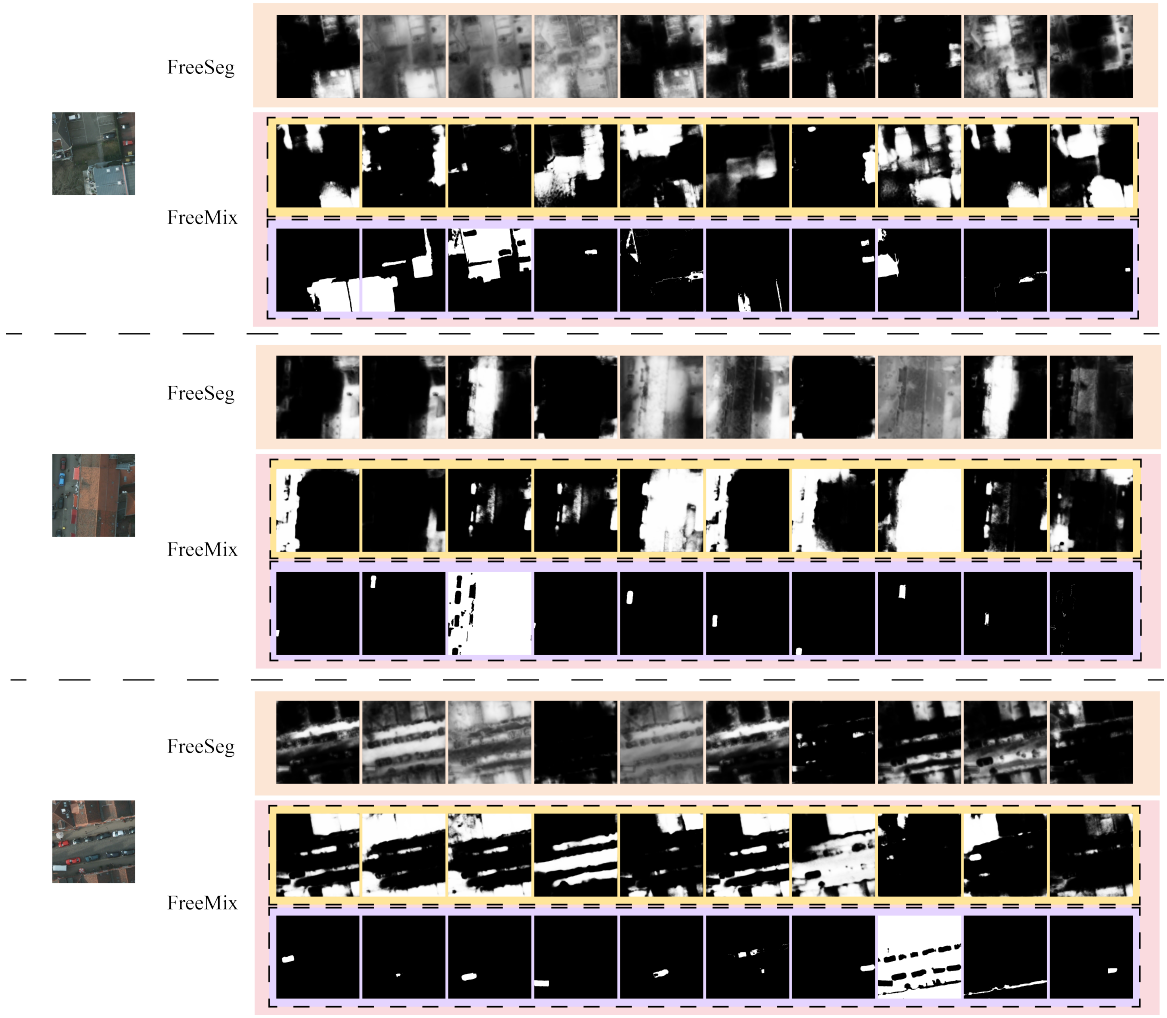


Figure 6. Comparison of proposal masks between FreeSeg and our FreeMix. In FreeMix, the base segmentation branch is highlighted in yellow, while the entity segmentation branch is depicted in purple.

5. Conclusion

In this work, we introduce a new setting for remote sensing image semantic segmentation called open vocabulary domain generalization (OVDG). This novel setting holds great potential in the remote sensing community by addressing the challenge of generalizing models to unseen domains and classes without retraining. To tackle this challenging problem, we propose an effective framework to train a robust model. Firstly, the proposed universal segmentation module has dual-branch: base segmentation branch (BSE) and entity segmentation branch (ESB). Moreover, the remote sensing self-supervised learning (RS_SSL) initialization and adaptation method is introduced to extract domain-agnostic visual feature for decoding masks and semantic logits. Additionally, ESB is proposed to generate entity masks for enhancing the segmentation and recognition of bath base and novel classes. Furthermore, a dataset-aware sampling (DAS) training tactic is designed for multi-source domain learning, aiming to enhance the overall performance of the model. Extensive experiments demonstrate the effectiveness of our proposed universal segmentation module, incorporating RS_SSL initialization, BSB and ESB as well as the DAS training tactic. Our FreeMix achieves state-of-the-art results on open vocabulary benchmarks and OVDG task. Although good semantic segmentation results have been achieved on OVDG, FreeMix relies on class names provided during testing. Future work will focus on designing a more effective training strategy for multi-data set learning and extending the FreeMix

 doi:10.20944/preprints202501.0448.v1

framework by incorporating multimodal large language models to generate class predictions and localization.

Acknowledgments: This work is supported in part by the Key Laboratory Funding of China, grant number 2022-JCJQ-LA-001-080, and in part by the National Key Research and Development Program of China, grant number 2018YFC0823002.

References

1. Rottensteiner, F.; Sohn, G.; Gerke, M.; Wegner, J.D. ISPRS semantic labeling contest. *ISPRS: Leopoldshöhe, Germany* **2014**, *1*, 4.
2. Hong, J.; Li, W.; Han, J.; Zheng, J.; Fang, P.; Harandi, M.; Petersson, L. Goss: Towards generalized open-set semantic segmentation. *The Visual Computer* **2024**, *40*, 2391–2404.
3. Nunes, I.; Laranjeira, C.; Oliveira, H.; dos Santos, J.A. A systematic review on open-set segmentation. *Computers & Graphics* **2023**.
4. Nunes, I.M.; Poggi, M.; Oliveira, H.; Pereira, M.B.; Dos Santos, J.A. Deep open-set segmentation in visual learning. In Proceedings of the 2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). IEEE, 2022, Vol. 1, pp. 314–319.
5. Joseph, K.; Khan, S.; Khan, F.S.; Balasubramanian, V.N. Towards open world object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 5830–5840.
6. Bendale, A.; Boult, T. Towards open world recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1893–1902.
7. Yang, J.; Zhou, K.; Li, Y.; Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334* **2021**.
8. Liu, J.; Shen, Z.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624* **2021**.
9. Zhang, H.; Ding, H. Prototypical matching and open set rejection for zero-shot semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6974–6983.
10. He, S.; Ding, H.; Jiang, W. Semantic-promoted debiasing and background disambiguation for zero-shot instance segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19498–19507.
11. Baek, D.; Oh, Y.; Ham, B. Exploiting a joint embedding space for generalized zero-shot semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 9536–9545.
12. Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; Zhang, L. Context-aware feature generation for zero-shot semantic segmentation. In Proceedings of the Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 1921–1929.
13. Zheng, Y.; Wu, J.; Qin, Y.; Zhang, F.; Cui, L. Zero-shot instance segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2593–2602.
14. He, S.; Ding, H.; Jiang, W. Primitive generation and semantic-related alignment for universal zero-shot segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11238–11247.
15. Bucher, M.; Vu, T.H.; Cord, M.; Pérez, P. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems* **2019**, *32*.
16. Ding, J.; Xue, N.; Xia, G.S.; Dai, D. Decoupling zero-shot semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11583–11592.
17. Ma, C.; Yang, Y.; Wang, Y.; Zhang, Y.; Xie, W. Open-vocabulary semantic segmentation with frozen vision-language models. *arXiv preprint arXiv:2210.15138* **2022**.
18. Chen, X.; Li, S.; Lim, S.N.; Torralba, A.; Zhao, H. Open-vocabulary panoptic segmentation with embedding modulation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1141–1150.
19. Ding, Z.; Wang, J.; Tu, Z. Open-Vocabulary Panoptic Segmentation MaskCLIP. *arXiv preprint arXiv:2208.08984* **2022**.

20. Ghiasi, G.; Gu, X.; Cui, Y.; Lin, T.Y. Scaling open-vocabulary image segmentation with image-level labels. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 540–557.
21. Zhou, C.; Loy, C.C.; Dai, B. Extract free dense labels from clip. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 696–712.
22. Huynh, D.; Kuen, J.; Lin, Z.; Gu, J.; Elhamifar, E. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7020–7031.
23. Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; Marculescu, D. Open-vocabulary semantic segmentation with mask-adapted clip. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7061–7070.
24. Qin, J.; Wu, J.; Yan, P.; Li, M.; Yuxi, R.; Xiao, X.; Wang, Y.; Wang, R.; Wen, S.; Pan, X.; et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19446–19455.
25. Ren, S.; Zhang, A.; Zhu, Y.; Zhang, S.; Zheng, S.; Li, M.; Smola, A.J.; Sun, X. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. *Advances in Neural Information Processing Systems* **2024**, *36*.
26. Zhang, H.; Li, F.; Zou, X.; Liu, S.; Li, C.; Yang, J.; Zhang, L. A simple framework for open-vocabulary segmentation and detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 1020–1031.
27. Wu, J.; Li, X.; Xu, S.; Yuan, H.; Ding, H.; Yang, Y.; Li, X.; Zhang, J.; Tong, Y.; Jiang, X.; et al. Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* **2018**.
29. Ding, H.; Cohen, S.; Price, B.; Jiang, X. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, 2020, pp. 417–435.
30. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* **2013**, *26*.
31. Zhu, C.; Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**.
32. You, K.; Long, M.; Cao, Z.; Wang, J.; Jordan, M.I. Universal domain adaptation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 2720–2729.
33. Saito, K.; Kim, D.; Sclaroff, S.; Saenko, K. Universal domain adaptation through self supervision. *Advances in neural information processing systems* **2020**, *33*, 16282–16292.
34. Kundu, J.N.; Venkat, N.; Babu, R.V.; et al. Universal source-free domain adaptation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4544–4553.
35. Li, R.; Zheng, S.; Duan, C.; Su, J.; Zhang, C. Multistage attention ResU-Net for semantic segmentation of fine-resolution remote sensing images. *IEEE Geoscience and Remote Sensing Letters* **2021**, *19*, 1–5.
36. Niu, X.; Zeng, Q.; Luo, X.; Chen, L. FCAU-Net for the Semantic Segmentation of Fine-Resolution Remotely Sensed Images. *Remote Sensing* **2022**, *14*, 215.
37. Wang, L.; Li, R.; Wang, D.; Duan, C.; Wang, T.; Meng, X. Transformer meets convolution: A bilateral awareness network for semantic segmentation of very fine resolution urban scene images. *Remote Sensing* **2021**, *13*, 3065.
38. He, X.; Zhou, Y.; Zhao, J.; Zhang, D.; Yao, R.; Xue, Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–15.
39. Gui, R.; Xu, X.; Wang, L.; Yang, R.; Pu, F. A generalized zero-shot learning framework for PolSAR land cover classification. *Remote Sensing* **2018**, *10*, 1307.
40. Jia, X.; Khandelwal, A.; Nayak, G.; Gerber, J.; Carlson, K.; West, P.; Kumar, V. Incremental dual-memory lstm in land cover prediction. In Proceedings of the Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, 2017, pp. 867–876.
41. Li, A.; Lu, Z.; Wang, L.; Xiang, T.; Wen, J.R. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 4157–4167.
42. Sumbul, G.; Cinbis, R.G.; Aksoy, S. Fine-grained object recognition and zero-shot learning in remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *56*, 770–779.

 doi:10.20944/preprints202501.0448.v1

43. Luo, C.; Li, Z.; Huang, K.; Feng, J.; Wang, M. Zero-shot learning via attribute regression and class prototype rectification. *IEEE Transactions on Image Processing* **2017**, *27*, 637–648.
44. Long, Y.; Shao, L. Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In Proceedings of the 2017 IEEE winter conference on applications of computer vision (WACV). IEEE, 2017, pp. 907–915.
45. Tsai, Y.H.; Hung, W.C.; Schulter, S.; Sohn, K.; Yang, M.H.; Chandraker, M. Learning to adapt structured output space for semantic segmentation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7472–7481.
46. Zheng, Z.; Yang, Y. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision* **2021**, *129*, 1106–1120.
47. Muhtar, D.; Zhang, X.; Xiao, P.; Li, Z.; Gu, F. Cmid: A unified self-supervised learning framework for remote sensing image understanding. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
48. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
49. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 4904–4916.
50. Chen, Y.; Bruzzone, L. Toward Open-World Semantic Segmentation of Remote Sensing Images. In Proceedings of the IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2023, pp. 5045–5048.
51. Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 736–753.
52. Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; Bai, X. Side adapter network for open-vocabulary semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2945–2954.
53. Zhang, Z.; Zhao, T.; Guo, Y.; Yin, J. Rs5m: A large scale vision-language dataset for remote sensing vision-language foundation model. *arXiv preprint arXiv:2306.11300* **2023**.
54. Wang, Z.; Prabha, R.; Huang, T.; Wu, J.; Rajagopal, R. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 5805–5813.
55. Zhang, W.; Cai, M.; Zhang, T.; Zhuang, Y.; Mao, X. Earthgpt: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *arXiv preprint arXiv:2401.16822* **2024**.
56. Mall, U.; Phoo, C.P.; Liu, M.K.; Vondrick, C.; Hariharan, B.; Bala, K. Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv preprint arXiv:2312.06960* **2023**.
57. Liang, C.; Li, W.; Dong, Y.; Fu, W. Single Domain Generalization Method for Remote Sensing Image Segmentation via Category Consistency on Domain Randomization. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
58. Wang, M.; Liu, J.; Luo, G.; Wang, S.; Wang, W.; Lan, L.; Wang, Y.; Nie, F. Smooth-Guided Implicit Data Augmentation for Domain Generalization. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
59. Iizuka, R.; Xia, J.; Yokoya, N. Frequency-based Optimal Style Mix for Domain Generalization in Semantic Segmentation of Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing* **2023**.
60. Zheng, J.; Wu, W.; Yuan, S.; Fu, H.; Li, W.; Yu, L. Multisource-domain generalization-based oil palm tree detection using very-high-resolution (vhr) satellite images. *IEEE Geoscience and Remote Sensing Letters* **2021**, *19*, 1–5.
61. Zhang, Y.; Zhang, M.; Li, W.; Wang, S.; Tao, R. Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–12.
62. Li, D.; Yang, Y.; Song, Y.Z.; Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2018, Vol. 32.
63. Balaji, Y.; Sankaranarayanan, S.; Chellappa, R. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems* **2018**, *31*.
64. Li, Y.; Yang, Y.; Zhou, W.; Hospedales, T. Feature-critic networks for heterogeneous domain generalization. In Proceedings of the International Conference on Machine Learning. PMLR, 2019, pp. 3915–3924.

 doi:10.20944/preprints202501.0448.v1

65. Shankar, S.; Piratla, V.; Chakrabarti, S.; Chaudhuri, S.; Jyothi, P.; Sarawagi, S. Generalizing across domains via cross-gradient training. *arXiv preprint arXiv:1804.10745* **2018**.
66. Wang, Y.; Li, H.; Kot, A.C. Heterogeneous domain generalization via domain mixup. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 3622–3626.
67. Shu, Y.; Cao, Z.; Wang, C.; Wang, J.; Long, M. Open domain generalization with domain-augmented meta-learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 9624–9633.
68. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* **2017**.
69. Segu, M.; Tonioni, A.; Tombari, F. Batch normalization embeddings for deep domain generalization. *Pattern Recognition* **2023**, *135*, 109115.
70. Bhattacharya, A.; Singha, M.; Jha, A.; Banerjee, B. C-SAW: Self-Supervised Prompt Learning for Image Generalization in Remote Sensing. In Proceedings of the Proceedings of the Fourteenth Indian Conference on Computer Vision, Graphics and Image Processing, 2023, pp. 1–10.
71. Kang, J.; Fernandez-Beltran, R.; Duan, P.; Liu, S.; Plaza, A.J. Deep unsupervised embedding for remotely sensed images based on spatially augmented momentum contrast. *IEEE Transactions on Geoscience and Remote Sensing* **2020**, *59*, 2598–2610.
72. Ayush, K.; Uzkent, B.; Meng, C.; Tanmay, K.; Burke, M.; Lobell, D.; Ermon, S. Geography-aware self-supervised learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10181–10190.
73. Manas, O.; Lacoste, A.; Giró-i Nieto, X.; Vazquez, D.; Rodriguez, P. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9414–9423.
74. Muhtar, D.; Zhang, X.; Xiao, P. Index your position: A novel self-supervised learning method for remote sensing images semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–11.
75. Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing* **2022**.
76. Reed, C.J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; Darrell, T. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4088–4099.
77. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *61*, 1–15.
78. Jakubik, J.; Roy, S.; Phillips, C.; Fraccaro, P.; Godwin, D.; Zadrozny, B.; Szwarcman, D.; Gomes, C.; Nyirjesy, G.; Edwards, B.; et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660* **2023**.
79. Dong, Z.; Gu, Y.; Liu, T. Generative ConvNet Foundation Model with Sparse Modeling and Low-Frequency Reconstruction for Remote Sensing Image Interpretation. *IEEE Transactions on Geoscience and Remote Sensing* **2024**.
80. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
81. Qi, L.; Kuen, J.; Shen, T.; Gu, J.; Guo, W.; Jia, J.; Lin, Z.; Yang, M.H. High Quality Entity Segmentation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4024–4033.
82. Qi, L.; Kuen, J.; Wang, Y.; Gu, J.; Zhao, H.; Torr, P.; Lin, Z.; Jia, J. Open world entity segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
83. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* **2016**.
84. Shi, B.; Zhang, X.; Xu, H.; Dai, W.; Zou, J.; Xiong, H.; Tian, Q. Multi-dataset pretraining: A unified model for semantic segmentation. *arXiv preprint arXiv:2106.04121* **2021**.
85. Chen, Y.; Wang, M.; Mittal, A.; Xu, Z.; Favaro, P.; Tighe, J.; Modolo, D. ScaleDet: A scalable multi-dataset object detector. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7288–7297.

 doi:10.20944/preprints202501.0448.v1

86. Tong, X.Y.; Xia, G.S.; Lu, Q.; Shen, H.; Li, S.; You, S.; Zhang, L. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment* **2020**, *237*, 111322.
87. Demir, I.; Koperski, K.; Lindenbaum, D.; Pang, G.; Huang, J.; Basu, S.; Hughes, F.; Tuia, D.; Raskar, R. Deepglobe 2018: A challenge to parse the earth through satellite images. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 172–181.
88. Ji, D.; Zhao, F.; Lu, H.; Tao, M.; Ye, J. Ultra-high resolution segmentation with ultra-rich context: A novel benchmark. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23621–23630.
89. Shi, J.X.; Wei, T.; Xiang, Y.; Li, Y.F. How Re-sampling Helps for Long-Tail Learning? *Advances in Neural Information Processing Systems* **2023**, *36*.
90. Zhou, X.; Koltun, V.; Krähenbühl, P. Simple multi-dataset detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 7571–7580.
91. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
92. Yu, Q.; He, J.; Deng, X.; Shen, X.; Chen, L.C. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems* **2024**, *36*.