S5: Scalable Semi-Supervised Semantic Segmentation in Remote Sensing

Liang Lv[1], **Di Wang**[1,2], **Jing Zhang**[1*], **Lefei Zhang**[1*]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2]Zhongguancun Academy
{lianglyu, d_wang, jingzhang.cv, zhanglefei}@whu.edu.cn

Abstract

Semi-supervised semantic segmentation (S4) has advanced remote sensing (RS) analysis by leveraging unlabeled data through pseudo-labeling and consistency learning. However, existing S4 studies often rely on small-scale datasets and models, limiting their practical applicability. To address this, we propose S5, the first scalable framework for semi-supervised semantic segmentation in RS, which unlocks the potential of vast unlabeled Earth observation data typically underutilized due to costly pixel-level annotations. Built upon existing large-scale RS datasets, S5 introduces a data selection strategy that integrates entropy-based filtering and diversity expansion, resulting in the RS4P-1M dataset. Using this dataset, we systematically scale up S4 into a new pretraining paradigm, S4 pre-training (S4P), to pretrain RS foundation models (RSFMs) of varying sizes on this extensive corpus, significantly boosting their performance on land cover segmentation and object detection tasks. Furthermore, during fine-tuning, we incorporate a Mixture-of-Experts (MoE)-based multi-dataset fine-tuning approach, which enables efficient adaptation to multiple RS benchmarks with fewer parameters. This approach improves the generalization and versatility of RSFMs across diverse RS benchmarks. The resulting RSFMs achieve state-of-the-art performance across all benchmarks, underscoring the viability of scaling semi-supervised learning for RS applications.

Code — https://github.com/MiliLab/S5

Introduction

Remote sensing (RS) semantic segmentation is a key task in RS image understanding, aiming to accurately classify each pixel to enable automatic recognition and analysis of land cover information (Zhang and Zhang 2022). However, traditional fully supervised segmentation methods heavily rely on pixel-level annotations, making the acquisition of high-quality training samples extremely costly. This dependency also limits model generalization in diverse scenarios. To reduce the burden of manual labeling and lower costs, semi-supervised semantic segmentation (S4) (Ouali, Hudelot, and Tami 2020) has gained increasing attention. S4 enhances RS image segmentation performance by combining a small amount of labeled images with a large number of unlabeled images during the training stage.

Early S4 research explored GAN-based methods (Souly, Spampinato, and Shah 2017). Subsequently, data augmentation was recognized as a crucial factor (French et al. 2019). Recent approaches mainly rely on pseudo-labeling and consistency regularization. ST++ (Yang et al. 2022) demonstrates that strong data augmentation significantly boosts self-training performance, though multi-stage pipelines often reduce efficiency. UniMatch (Fan et al. 2023) revisits consistency regularization using weak-to-strong augmentation, a strategy originally generalized by FixMatch (Sohn et al. 2020) for semi-supervised classification. FixMatch generates pseudo-labels on weakly augmented images and enforces consistency on their strongly augmented counterparts, using a confidence threshold to ensure pseudo-label reliability. Thanks to its simplicity and efficiency, FixMatch has become a key baseline in S4, inspiring many follow-up methods, such as UniMatch (Fan et al. 2023), RankMatch (Mai et al. 2024), and CorrMatch (Sun et al. 2024). In RS, methods like RanPaste (Wang et al. 2021), WSCL (Lu et al. 2023), and SegMind (Li et al. 2023) also adopt the FixMatch framework, enhancing it with domain-specific augmentations such as random copy-paste, dual-view augmentation, and random masking for complex RS scenes.

However, current S4 research still relies on small-scale models and datasets. As illustrated in Figure 1 (a), a common strategy involves splitting the training set of standard segmentation datasets (e.g., iSAID (Waqas Zamir et al. 2019)) into labeled and unlabeled subsets. By leveraging S4 methods in combination with unlabeled images, model performance under limited supervision can be significantly improved compared to purely supervised training (SupTrain). However, such settings are typically constrained to a single dataset, which limits the exploration of S4's potential in harnessing large-scale Earth observation data.

Meanwhile, RS foundational models (RSFMs) have made significant progress, benefiting from large datasets like MillionAID (Long et al. 2021) and SAMRS (Wang et al. 2023), alongside extensive exploration of different pre-training strategies. Self-supervised learning methods, such as contrastive learning (Tian et al. 2020) and masked image modeling (MIM) (He et al. 2022), extract generalizable fea-

*Corresponding author.

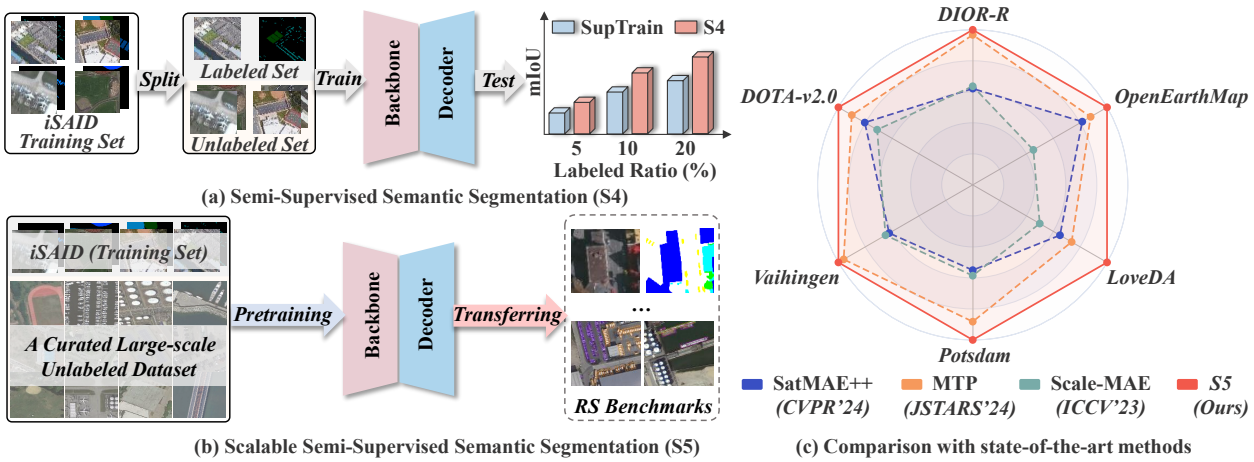

Figure 1: (a) Traditional S4 workflow: splitting the dataset into labeled and unlabeled subsets to improve model performance with few labeled samples. (b) The proposed S5 workflow: perform semi-supervised segmentation pretraining on both labeled and large-scale unlabeled datasets, followed by fine-tuning on RS benchmarks. (c) Comparison of performance across four RS segmentation and two object detection benchmarks.

tures without relying on labeled data. In contrast, supervised pre-training better aligns upstream and downstream tasks and domains, enhancing transferability. For instance, RSP (Wang et al. 2022a) applies supervised pre-training on the MillionAID dataset, producing RSFMs that perform well across diverse downstream tasks. The SAMRS (Wang et al. 2023) dataset is used to explore segmentation pre-training (SEP) and multitask pre-training (MTP) (Wang et al. 2024a), which further enhance generalization by narrowing the gap between pre-training and target tasks. However, SAMRS relies on the Segment Anything Model (SAM) (Kirillov et al. 2023) to generate segmentation masks from bounding-box annotations. This dependence, along with its limited annotation scale constrains scalability. In particular, this mask generation process resembles the self-training paradigm of S4. Furthermore, given that pre-training is most effective when tasks and domains are well aligned, a key question arises: *Can we scale up S4 to pre-train RSFMs on massive RS imagery, and thereby enhance their performance across diverse RS applications?*

To answer this question, we introduce Scalable Semi-supervised Semantic Segmentation (S5), the first framework to pre-train RSFMs using large-scale unlabeled RS images via semi-supervised learning.

As illustrated in Figure 1 (b), S5 curates a large-scale unlabeled RS dataset using a sample selection strategy that combines entropy-based filtering with diversity expansion. Building on this dataset, we systematically perform S4 pre-training (S4P) on RSFMs of varying capacities, all initialized with MAE (He et al. 2022) pretrained weights. Experiments show that S4P not only enhances the general representation ability learned from MAE pre-training, but also significantly improves downstream performance on land cover segmentation and object detection tasks. Furthermore, during fine-tuning, S5 introduces a multi-dataset fine-tuning strategy based on Mixture-of-Experts (MoE), which incorporates both task-shared and task-specific Feedforward Networks (FFNs). This design allows the RSFM to adapt efficiently to multiple RS benchmarks with minimal additional parameters. Extensive experiments demonstrate that RSFMs pre-trained under the S5 framework achieve state-of-the-art (SOTA) performance across multiple segmentation and detection benchmarks (see Figure 1 (c)), validating the potential and promise of large-scale semi-supervised pre-training for RSFMs development.

Our main contributions are summarized as follows:

- We introduce the S5 framework for RS, addressing the limitations of traditional S4 methods that rely on small-scale datasets and models. S5 establishes a new paradigm for leveraging vast amounts of unlabeled RS images to develop RSFMs.
- We propose a low-entropy filtering and diversity expansion strategy to curate RS4P-1M, a dataset of one million reliable RS images covering diverse geospatial scenes. This dataset enables effective S4P of RSFMs and lays a solid foundation for future research in scalable RSFMs pre-training.
- We design a MoE-based multiple dataset fine-tuning (MoE-MDF) approach, which combines task-shared and task-specific FFNs to enable efficient joint adaptation across multiple RS benchmarks with minimal parameter overhead, significantly enhancing RSFMs' generalization and transferability.

Related Work

Semi-supervised Semantic Segmentation

S4 aims to train semantic segmentation models using a small amount of labeled data and a large pool of unlabeled data. Early S4 approaches relied on consistency regularization and pseudo-labeling, ensuring stable predictions under perturbations or leveraging the model's own outputs as labels.

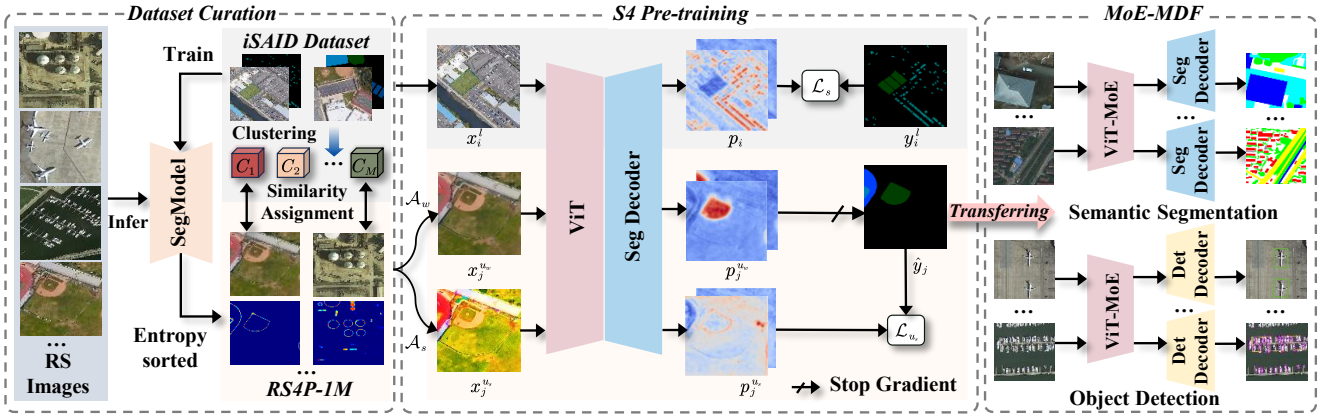

Figure 2: The overall pipeline of the proposed S5 framework. It starts with the construction of the RS4P-1M dataset, followed by training RSFMs based on the S4P. The pre-trained weights are then fine-tuned on semantic segmentation and object detection benchmarks through the MoE-based multiple dataset fine-tuning (MoE-MDF) scheme. ViT-MoE indicates the integration of the FFN-MoE modules into the ViT backbones.

Recent deep learning-based S4 methods have significantly improved performance. UniMatch (Fan et al. 2023) enforces weak-to-strong consistency at both image and feature levels, AugSeg (Zhao et al. 2023b) enhances robustness through data augmentation, and iMAS (Zhao et al. 2023a) introduces instance-specific, model-adaptive supervision. CorrMatch (Sun et al. 2024) propagates labels via correlation matching, AllSpark (Wang et al. 2024b) employs a Transformer-based approach leveraging labeled features, SemiVL (Hoyer et al. 2024) integrates vision-language guidance for better pseudo-labeling, and UniMatchV2 (Yang, Zhao, and Zhao 2025), built on DINOV2 (Oquab et al. 2024), further improves S4 performance.

S4 methods in RS address domain-specific challenges. For instance, RanPaste (Wang et al. 2021) exploits unlabeled data through consistency and pseudo-labeling, WSCL (Lu et al. 2023) transitions from weak to strong labels, SegMind (Li et al. 2023) integrates mask image modeling with contrastive learning, and DWL (Huang et al. 2024) enhances segmentation via decoupling and weighting of different components. Unlike existing S4 methods, which are often constrained by small datasets, we introduce S5, a scalable semi-supervised framework designed to leverage large-scale unlabeled RS data for pre-training RSFMs.

Remote Sensing Foundation Models

RSFMs have gained attention for their ability to learn generalizable and transferable features. Pre-training approaches are typically supervised or self-supervised. RSP (Wang et al. 2022a) pioneered this approach by pre-training CNNs and vision transformers on Million-AID (Long et al. 2021), while MTP (Wang et al. 2024a) aligns multiple downstream tasks for enhanced representation learning. SAMRS (Wang et al. 2023) improves segmentation-focused FMs using a 100,000-sample dataset built with SAM (Kirillov et al. 2023). However, these methods rely on labeled data, which are scarce and hard to scale. Recent work favors self-supervised pre-training, either contrastive—forming positive-negative pairs from image augmentations (Li et al. 2022), multimodal images (Stojnic and Risojevic 2021; Jain, Schoen-Phelan, and Ross 2022), geographic priors (Li et al. 2021), or temporal imagery (Mall, Hariharan, and Bala 2023)—or generative methods like MIM (He et al. 2022) that reconstruct masked regions to capture structural features. RingMo (Sun et al. 2022) employs incomplete masking for dense small objects in RS scenes, while RVSA (Wang et al. 2022b), initialized with MIM weights, introduces rotational window attention to enhance target representation with lower computational cost. GFM (Mendieta et al. 2023) refines MIM pre-training by leveraging ImageNet-pre-trained FMs, while SatMAE (Cong et al. 2022) and Scale-MAE (Reed et al. 2023) incorporate multi-temporal and multi-scale features, respectively. SkySense (Guo et al. 2024) proposes a billion-scale multimodal RSFM that leverages multi-level contrastive learning and geo-context learning for large-scale pre-training on RS images. In contrast, our work explores S4P for RSFMs and introduces S5: the first scalable framework for semi-supervised semantic segmentation in RS. Leveraging the newly established RS4P-1M dataset in this work, we successfully pre-train RSFMs with up to 600M parameters, achieving SOTA performance across multiple benchmarks.

Methodology

As illustrated in Figure 2, the proposed S5 framework consists of three key components: dataset curation, S4 pre-training (S4P), and MoE-based multi-dataset fine-tuning (MoE-MDF). These components will be introduced in detail in the following text.

Pre-training Dataset Curation

To explore a suitable S4P paradigm for large-scale RS images, we begin by analyzing existing public datasets. In RS, the MillionAID dataset (Long et al. 2021) is one of the most widely used resources for pre-training RSFMs, due to its

large scale and diverse scene types. As such, it serves as our primary source of unlabeled imagery. To enhance dataset diversity, we further incorporate SAMRS (Wang et al. 2023) and STAR (Li et al. 2024a) datasets. As these datasets are primarily collected from Google Earth, we adopt iSAID (Waqas Zamir et al. 2019), an object-level segmentation dataset with similar resolution and imaging style, as our labeled data source.

Based on the above datasets, we combine a certain amount of labeled data with large-scale unlabeled images, and leverage S4P to transfer annotation knowledge to broader RS scenarios. This enhances the model's generalization ability and bridges the gap between pre-training and downstream tasks. However, using all unlabeled data may degrade pseudo-label quality due to distribution mismatch.

To mitigate this, we propose an entropy-based filtering and semantic diversity expansion strategy. Treating the labeled set as a curated subset, we aim to extend it with unlabeled samples that are both reliable and diverse. We first train an initial segmentation model (ViT-H + UperNet (Xiao et al. 2018)) on the labeled data and apply it to cropped unlabeled patches. Considering that the model may exhibit high uncertainty on low-quality or out-of-distribution samples, we adopt the pixel-level average entropy as a confidence measure for the pseudo-labels. For each unlabeled image $x \in \mathbb{R}^{H\times W\times 3}$, the average entropy is defined as:

$$E(x) = -\frac{1}{H\times W}\sum_{i=1}^{H\times W}\sum_{k=1}^{K} P^k(x^i)\log P^k(x^i), \quad (1)$$

where $P^k(x^i)$ is the probability of pixel x^i belonging to class k (excluding background). Patches are then ranked by entropy:

$$\begin{gathered}\mathcal{D} = \{(x_j, E_j)\,,\ j = 1,\ldots,N\}\,,\\ \text{with } E_1 \le E_2 \le \cdots \le E_N.\end{gathered} \quad (2)$$

where N is the total number of cropped unlabeled image patches. High-entropy samples are likely noisy or out-of-distribution, so we prioritize low-entropy patches. Nevertheless, low entropy selection may yield semantic redundancy, with the chosen images concentrated in a few common scenes. To promote diversity, as illustrated in the similarity assignment step of Figure 2, we extract features from labeled images using the trained segmentation model's backbone and apply K-Means to cluster them into M clusters, denoted as $C_1, C_2, \ldots, C_M$. For each unlabeled image, we compute its cosine similarity to these cluster prototypes and assign it to the nearest cluster. Let B^u denote the target number of selected unlabeled samples, where $N > B^u$. Each cluster is allocated a quota:

$$B^u_m = B^u \cdot \frac{N^l_m}{B^l}, \quad (3)$$

N^l_m denotes the number of labeled samples in the m-th cluster, and B^l is the total number of labeled samples. Once a cluster reaches its quota, we exclude further similar images to avoid semantic redundancy.

We set B^u to a million to construct the RS4P-1M dataset, which balances pseudo-label quality and semantic diversity, greatly improving model generalization and transferability.

S4 Pre-training

Through the above process, we construct the RS4P-1M dataset for S4P. We then introduce a general RSFMs pre-training framework that is compatible with existing S4 methods. Considering that more complex algorithms may suffer from low training efficiency on large-scale images, we adopt FixMatch (Sohn et al. 2020), an efficient and widely used consistency regularization method that applies weak-to-strong data augmentations for pre-training.

Consider a dataset consisting of labeled image pairs from the iSAID dataset $\{(x^l_i, y^l_i)\}_{i=1}^{B_l}$ and unlabeled images from our curated RS4P-1M collection $\{x^u_j\}_{j=1}^{B_u}$. Each labeled image $x^l_i \in \mathbb{R}^{H\times W\times 3}$ has corresponding pixel-level annotations $y^l_i \in \mathbb{R}^{H\times W\times K}$, where K is the number of classes in the pre-training phase. Each unlabeled image $x^u_j \in \mathbb{R}^{H\times W\times 3}$ has no annotations. The numbers of labeled and unlabeled images are B_l and B_u, respectively.

FixMatch employs distinct transformation strategies for data augmentation: *Weak augmentation* $\mathcal{A}_w$ involves random scaling, cropping, rotation, and flipping, while *Strong augmentation* $\mathcal{A}_s$ applies more aggressive transformations, such as CutMix (Yun et al. 2019), color jitter, grayscale conversion, and Gaussian blur.

For each unlabeled image x^u_j, we generate two augmented views using sequential transformations:

$$x^{u_w}_j = \mathcal{A}_w(x^u_j), \quad x^{u_s}_j = \mathcal{A}_s(\mathcal{A}_w(x^u_j)). \quad (4)$$

The overall training objective function for both labeled and unlabeled images is given by:

$$\mathcal{L} = \mathcal{L}_s + \lambda\,\mathcal{L}_{u_s}. \quad (5)$$

Here, the supervised and unsupervised loss terms are denoted as $\mathcal{L}_s$ and $\mathcal{L}_{u_s}$, respectively, with λ as a hyperparameter that controls the weight of the unsupervised loss. These losses are defined as:

$$\mathcal{L}_s = \frac{1}{B_l}\sum_{i=1}^{B_l}\mathcal{L}_{ce}(y^l_i, p_i), \quad (6)$$

$$\mathcal{L}_{u_s} = \frac{1}{B_u}\sum_{j=1}^{B_u}\mathbb{1}(\max(p^{u_w}_j) \ge \tau)\mathcal{L}_{ce}(\hat{y}_j, p^{u_s}_j), \quad (7)$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss, $p_i = f(x^l_i)$, $p^{u_w}_j = f(x^{u_w}_j)$, and $p^{u_s}_j = f(x^{u_s}_j)$ are the predicted probability maps for labeled and unlabeled images. The pseudo-label $\hat{y}_j$ is determined by $\hat{y}_j = \arg\max(p^{u_w}_j)$, with τ being the confidence threshold for the pseudo-labels. The indicator function $\mathbb{1}$ ensures that only high-confidence predictions contribute to the unsupervised loss. During semi-supervised learning, the segmentation network $f(\cdot)$ is optimized across both supervised and unsupervised branches, allowing it to learn more representative and discriminative features.

MoE-based Multiple Dataset Fine-tuning

During fine-tuning, existing RSFMs often follow a "one dataset, one model" paradigm, requiring a separate model to be trained for each downstream dataset. This approach leads

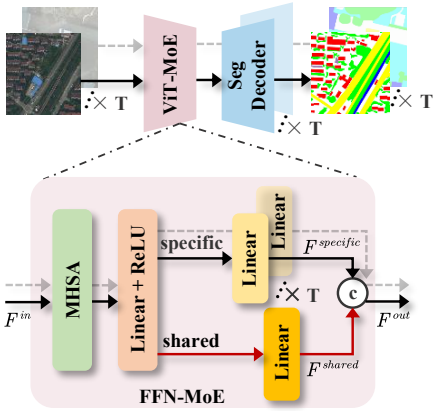

Figure 3: The workflow of MoE-MDF. ViT-MoE refers to the incorporation of the FFN-MoE module into the standard ViT. The black solid arrows and dashed arrows represent the forward propagation paths for different datasets, while the red arrows indicate the shared forward propagation.

to significant parameter redundancy and lacks generalization across datasets, making it inefficient for unified deployment and scalability.

To address these issues, our goal is to develop general-purpose RSFMs that support multiple datasets for the same task (e.g., land cover segmentation), enabling unified fine-tuning and deployment. As shown in Figure 3, we first adopt a shared backbone combined with dataset-specific decoders, and train the model via multiple-dataset joint fine-tuning (MDF) to handle T datasets simultaneously. However, due to the considerable differences in data distributions, label spaces, and annotation styles across datasets, naively sharing the backbone may lead to feature interference, thereby hindering performance and convergence. To mitigate this, we further propose the MoE-MDF approach, which decouples shared and dataset-specific knowledge within the model. Specifically, we integrate the MoE modules into the ViT by splitting the FFN into multiple branches: a shared expert for learning general representations, and T dataset-specific experts for modeling dataset-specific features. This design results in the FFN-MoE, as illustrated in Figure 3.

Taking a standard Transformer block as an example, the computation of FFN-MoE begins with the input feature F^{in}, it first passes through multi-head self-attention (MHSA), followed by a shared linear layer and a ReLU activation to obtain the intermediate feature:

$$F_{\text{FFN}} = \text{ReLU}(\text{Linear}(\text{MHSA}(F^{\text{in}})), \tag{8}$$

$F_{\text{FFN}} \in \mathbb{R}^{N \times D}$, where N is the number of tokens and D is the intermediate dimensionality. Next, this intermediate representation is passed through two parallel branches: one shared expert independent of the task, and one task-specific expert. The computation is defined as:

$$\begin{aligned} F^{\text{shared}} &= \text{Linear}_{\text{shared}}^{D \to (1-\alpha)C}(F^{\text{FFN}}), \\ F^{\text{specific}} &= \text{Linear}_{\text{specific}}^{D \to \alpha C}(F^{\text{FFN}}), \end{aligned} \tag{9}$$

where C is the output channel dimension, and $\alpha \in [0, 1]$ is a hyperparameter controlling the partition ratio between shared and task-specific capacities. The shared linear layer is updated using data from all tasks, while $\text{Linear}_{\text{specific}}^{(t)}$ is exclusively trained with samples from task t. Finally, the two outputs are concatenated along the channel dimension to form the final output of the FFN:

$$F^{\text{out}} = \text{Concat}\left(F^{\text{shared}},\ F^{\text{specific}}\right). \tag{10}$$

Overall, the proposed MoE-MDF approach enables efficient and unified processing of multiple datasets, greatly reducing memory use and computational cost compared to maintaining separate models. It adds no extra parameters or inference delay, making it an efficient and practical universal backbone for RS segmentation. This method can also be extended to object detection, allowing one model to generalize across multiple datasets using the same fine-tuning and deployment pipeline. Overall, this unified framework provides a solid foundation for building general-purpose RSFMs that handle diverse downstream tasks effectively.

Experiments

In this section, we comprehensively evaluate the proposed method on RS semantic segmentation and object detection tasks. S5 successfully pre-trains ViT RSFMs ranging from ViT-B to ViT-H (up to 600 million parameters) and achieves state-of-the-art performance on multiple RS benchmarks with fewer parameters after fine-tuning. We also validate the effectiveness of S5's core components through ablation studies. More details of dataset, pre-training, and fine-tuning implementations can be found in the appendix.

Comparison with SOTA Methods

To thoroughly assess our method, we fine-tune it on diverse RS benchmarks covering semantic segmentation and object detection. We compare S5 with state-of-the-art RSFMs including RVSA (Wang et al. 2022b), GFM (Mendieta et al. 2023), Scale-MAE (Reed et al. 2023), SAMRS (Wang et al. 2023), SatMAE++(Noman et al. 2024), MTP(Wang et al. 2024a), MA3E(Li et al. 2024b), BillionFM (Cha, Seo, and Lee 2023), OREOLE (Dias et al. 2024), and SelectiveMAE (Wang et al. 2025). As summarized in Table 1, the results clearly demonstrate the superiority of S5 across a wide range of datasets.

Object Detection. We evaluate the detection performance of the proposed method on two representative RS object detection datasets: DIOR-R and DOTA-v2. As shown in Table 1, under the Oriented R-CNN (Xie et al. 2021) detection framework, S5 consistently delivers superior performance across various backbone scales. From ViT-B, ViT-L to the larger ViT-H, S5 outperforms SOTA methods with comparable parameter sizes, and even matches or surpasses larger models while using fewer parameters. For example, with the ViT-L backbone, S5 achieves better performance than methods like MTP and SelectiveMAE, while requiring only about half the number of parameters when handling multiple datasets. These results demonstrate the strong scalability and generalization ability of S5 in RS object detection tasks.

Method	Backbone	Params Det (M)		**Object Detection**		Params Seg (M)		**Semantic Segmentation**			
		Single	Multiple	DIOR-R	DOTA-v2	Single	Multiple	Vaihingen	Potsdam	LoveDA	OpenEarthMap
RVSA (Wang et al. 2022b)	ViT-B + RVSA	111.2	222.4	68.06	55.22	103.2	412.8	78.49	91.58	52.44	66.63
GFM (Mendieta et al. 2023)	Swin-B	104.1	208.2	67.67	59.15	96.9	387.6	79.61	91.85	54.98	67.78
Scale-MAE (Reed et al. 2023)	ViT-L	334.6	669.2	66.47	56.97	327.4	1309.6	78.64	91.54	53.67	68.54
SAMRS (Wang et al. 2023)	ViT-B + RVSA	-	-	-	-	103.2	412.8	78.73	91.69	53.04	67.37
SatMAE++ (Noman et al. 2024)	ViT-L	334.6	669.2	66.82	55.60	327.4	1309.6	78.80	91.64	52.82	65.62
BillionFM (Cha, Seo, and Lee 2023)	ViT-G	996.9	1993.9	73.62	58.69	990.9	-	-	92.58	54.40	-
OREOLE (Dias et al. 2024)	ViT-G	996.9	-	71.31	-	990.9	-	-	92.20	54.00	-
MTP (Wang et al. 2024a)	ViT-L + RVSA	334.6	669.2	74.54	58.41	327.4	1309.6	80.62	92.47	54.16	69.04
MA3E (Li et al. 2024b)	ViT-B	111.2	-	71.82	-	103.2	-	-	91.50	-	-
SelectiveMAE (Wang et al. 2025)	ViT-L	334.6	669.2	71.75	57.84	327.4	1309.6	80.45	92.78	54.31	69.30
S5	ViT-B	111.2	138.3	72.95	57.20	103.2	160.4	79.85	92.40	54.02	68.65
S5	ViT-L	334.6	377.8	75.21	59.71	327.4	435.0	80.72	92.78	**55.67**	69.66
S5	ViT-H	671.7	730.0	**75.30**	**59.89**	663.4	824.5	**80.85**	**92.97**	55.65	**70.02**

Table 1: Comparison with existing RSFMs across multiple RS benchmarks. The evaluation metric for object detection is mAP, while for semantic segmentation it is mIoU, except for Potsdam, which uses mF1. Params Det (M) and Params Seg (M) indicate the number of parameters used for detection and segmentation models, respectively. Single and Multiple refer to the parameters required for handling a single dataset and multiple datasets.

Semantic Segmentation. We further validate the generalization capability of S5 on four challenging RS semantic segmentation datasets: Vaihingen, Potsdam, LoveDA, and OpenEarthMap. Built upon the UperNet (Xiao et al. 2018) segmentation model, S5 brings consistent and significant performance improvements across all benchmarks. With the ViT-B backbone, S5 already outperforms methods such as RVSA and SAMRS with similar parameter sizes. As the backbone scales up to ViT-L and ViT-H, S5 continues to set new SOTA results across multiple datasets. Moreover, S5 demonstrates outstanding parameter efficiency in multi-dataset settings.

For example, with the ViT-L backbone, it uses less than one-third of the segmentation parameters compared to Scale-MAE, SatMAE++, and SelectiveMAE, while delivering superior performance. This further highlights the scalability and generalization strength of S5 in RS semantic segmentation tasks.

Ablation Studies

In this section, we perform ablation studies to assess the effectiveness of each component in the S5 framework. Experiments are conducted on two semantic segmentation benchmarks (Vaihingen and LoveDA) and one object detection benchmark (DIOR-R). As LoveDA's test set requires online evaluation, all results on this dataset are reported on the official validation set.

Effectiveness of pre-training Dataset Curation As shown in Table 2, we use the iSAID training set as the labeled data and take the MAE pre-trained ViT-B from the MillionAID dataset as our baseline. We then conduct S4P experiments with three different unlabeled datasets. We evaluate the generalization ability of S4P on three representative downstream benchmarks: semantic segmentation on Vaihingen and LoveDA, and object detection on DIOR-R. In addition, we assess performance on the iSAID validation set under the conventional S4 setting to evaluate the quality of pseudo-labels indirectly.

We first use the curated SAMRS dataset as the unlabeled data. The results show that introducing this large-scale RS images into S4P significantly improves performance on downstream segmentation and detection tasks. For example, we observe mIoU of 79.61 on Vaihingen, 53.66 on LoveDA and mAP of 69.13 on DIOR-R, both clearly surpassing the results obtained from MAE pre-training alone. This confirms that S4P serves as an effective complementary pre-training strategy for enhancing the performance of RSFMs on downstream tasks. Next, we randomly sample a subset of 100k images from MillionAID, denoted as MillionAID-random, to match the scale of SAMRS. Under this setting, the model shows a slight performance drop across all downstream tasks, with a notable decrease on LoveDA from 53.66 to 53.20. This suggests that a randomly sampled unlabeled set may suffer from data distribution mismatch and insufficient diversity, leading to lower pseudo-label quality and thus limiting the effectiveness of training. Meanwhile, the performance gain on the iSAID validation set is also less pronounced compared to SAMRS, rising only from 65.93 to 66.32, which further supports this observation.

To improve the reliability of pseudo-labels and the diversity of the unlabeled data, we propose a data selection strategy based on entropy filtering and diversity expansion, resulting in a curated subset named MillionAID*. Experimental results show that MillionAID* achieves the best performance across all evaluation tasks: 67.66 on the iSAID validation set, outperforming both SAMRS (67.59) and MillionAID-random (66.32); and 79.77, 53.81, and 69.65 on Vaihingen, LoveDA, and DIOR-R, respectively. These results clearly demonstrate that the proposed method effectively selects high-quality and diverse unlabeled images. Compared with the existing SAMRS dataset, our method offers better scalability, thereby further enhancing the generalization and transferability of S4P models.

Scalability of S4P We investigate the scalability of S4P with respect to both the model size and the scale of unlabeled data, as illustrated in Figure 4. Across all three downstream benchmarks, S4P consistently outperforms the MAE

Labeled Data	Unlabeled Data	Images	Pre-training	Backbone	S4	S4P (Pre-train → Finetune)		
					iSAID (Val)	Vaihingen	LoveDA	DIOR-R
-	-	-	MAE	ViT-B	65.93	78.27	52.47	68.02
iSAID (Train)	SAMRS	100k	MAE + S4P	ViT-B	67.59	79.61	53.66	69.13
	MillionAID-random	100k	MAE + S4P	ViT-B	66.32	79.49	53.20	69.02
	MillionAID*	100k	MAE + S4P	ViT-B	**67.66**	**79.77**	**53.81**	**69.65**

Table 2: Comparison of S4P using different pre-training datasets.

Pre-training	Backbone	Fine-tuning	Params (M)	GFLOPs	Resolution	Vaihingen	Potsdam	OpenEarthMap	LoveDA	Average
MAE	ViT-B	SDF	412.8	178.3	512 × 512	78.27	91.58	66.23	52.47	72.14
MAE + S4P	ViT-B	SDF	412.8	178.3	512 × 512	**79.93**	92.24	67.35	54.51	73.51
	ViT-B	MDF	132.1			79.82	92.25	68.41	54.53	73.75
	ViT-B-MoE ($\alpha = 1/8$)	MoE-MDF	146.2			79.76	92.30	68.52	54.62	73.80
	ViT-B-MoE ($\alpha = 1/4$)	MoE-MDF	160.4			79.85	**92.40**	**68.80**	54.57	**74.15**
	ViT-B-MoE ($\alpha = 1/2$)	MoE-MDF	188.7			79.84	92.39	68.66	**54.64**	73.88

Table 3: Comparison of different fine-tuning strategies. SDF (single-dataset fine-tuning): an independent segmentation model per dataset. MDF (multi-dataset fine-tuning): a shared backbone with dataset-specific decoders.

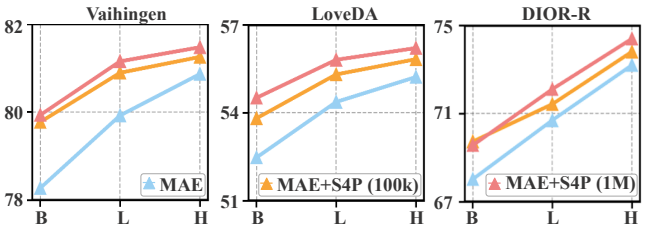

Figure 4: Fine-tuning results on three RS benchmarks with varying pre-training dataset sizes and backbones. "100K" and "1M" indicate the number of images used for S4P.

baseline, with performance further improving as model capacity or dataset scale increases. For instance, on the Vaihingen dataset, upgrading the backbone from ViT-B to ViT-H results in a significant performance boost for all pre-training setups. More importantly, expanding the unlabeled set from 100k to 1M images using our selection strategy leads to further improvements under each backbone. Similar trends are observed on the LoveDA and DIOR-R benchmarks. These results highlight two key scalability strengths of S4P: (1) effective utilization of large-scale unlabeled data when carefully selected for quality and diversity; and (2) consistently achieving higher performance with increasing model size. This underscores S4P's potential as a general and scalable semi-supervised pre-training framework for RSFMs.

The Ratio of Shared and Specific Experts Based on the results in Table 3, we evaluate various pre-training and fine-tuning strategies across four RS segmentation benchmarks, highlighting the advantages of MDF and the effect of the shared-to-specific expert ratio (α) in the MoE-FFN design. The analysis is conducted using the average accuracy (Average) across the four datasets.

First, under the SDF paradigm, S4P effectively enhances the MAE backbone, improving performance across all four downstream tasks with the average accuracy rising from 72.14 to 73.51. Compared with SDF, MDF achieves consistently better results on all tasks, increasing the average accuracy to 73.75 while reducing parameters by nearly four times. This demonstrates that MDF enables the model to learn more general and robust feature representations under diverse data distributions. Finally, we introduce MoE-FFN to expand model capacity and analyze the impact of the specific expert ratio (α). As α increases from 1/8 to 1/4, the average accuracy steadily improves from 73.80 to 74.15 across multiple datasets.

However, when α further increases to 1/2, performance gains plateau or slightly decline, likely because a high proportion of specific experts weakens the model's generalization across datasets and reduces its adaptability to diverse data characteristics.

Balancing performance and parameter efficiency, we choose $\alpha = 1/4$ as the optimal specific expert ratio in the MoE architecture. This setting delivers the best overall results across four benchmarks with only a slight increase in model size, demonstrating the value of a balanced design between shared and task-specific experts.

Conclusion

In this paper, we propose S5, a scalable semi-supervised semantic segmentation framework designed to pre-train RSFMs by leveraging vast unlabeled RS image data. S5 overcomes the limitations of prior S4 approaches constrained by small datasets and models, enabling large-scale semi-supervised pre-training. We introduce RS4P-1M, a million-scale curated dataset created via low-entropy filtering and diversity expansion, which serves as a solid basis for effective S4 pre-training of RSFMs with varying sizes. Additionally, S5 incorporates a MoE-MDF strategy to efficiently adapt RSFMs across multiple RS benchmarks with minimal parameter overhead. Extensive experiments demonstrate that RSFMs pre-trained under S5 achieve SOTA performance on diverse segmentation and detection benchmarks, highlighting the promise of scalable semi-supervised learning in advancing RS applications.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62431020 and 624B2109, the National Key Research and Development Program of China under Grant 2024YFE0111800. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Cha, K.; Seo, J.; and Lee, T. 2023. A billion-scale foundation model for remote sensing images. *arXiv preprint arXiv:2304.05215*.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D.; and Ermon, S. 2022. Satmae: Pretraining transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35: 197–211.

Dias, P.; Tsaris, A.; Bowman, J.; Potnis, A.; Arndt, J.; Yang, H. L.; and Lunga, D. 2024. OReole-FM: successes and challenges toward billion-parameter foundation models for high-resolution satellite imagery. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, 597–600.

Fan, Y.; Kukleva, A.; Dai, D.; and Schiele, B. 2023. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, 131(3): 626–643.

French, G.; Laine, S.; Aila, T.; Mackiewicz, M.; and Finlayson, G. 2019. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*.

Guo, X.; Lao, J.; Dang, B.; Zhang, Y.; Yu, L.; Ru, L.; Zhong, L.; Huang, Z.; Wu, K.; Hu, D.; et al. 2024. Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27672–27683.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009.

Hoyer, L.; Tan, D. J.; Naeem, M. F.; Van Gool, L.; and Tombari, F. 2024. SemiVL: semi-supervised semantic segmentation with vision-language guidance. In *In Proceedings of the European Conference on Computer Vision*, 257–275. Springer.

Huang, W.; Shi, Y.; Xiong, Z.; and Zhu, X. X. 2024. Decouple and weight semi-supervised semantic segmentation of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212: 13–26.

Jain, P.; Schoen-Phelan, B.; and Ross, R. 2022. Self-supervised learning for invariant representations from multi-spectral and SAR images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15: 7797–7808.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, H.; Li, Y.; Zhang, G.; Liu, R.; Huang, H.; Zhu, Q.; and Tao, C. 2022. Global and local contrastive self-supervised learning for semantic segmentation of HR remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Li, W.; Chen, K.; Chen, H.; and Shi, Z. 2021. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Li, Y.; Wang, L.; Wang, T.; Yang, X.; Luo, J.; Wang, Q.; Deng, Y.; Wang, W.; Sun, X.; Li, H.; et al. 2024a. Star: A first-ever dataset and a large-scale benchmark for scene graph generation in large-size satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Z.; Chen, H.; Wu, J.; Li, J.; and Jing, N. 2023. SegMind: Semisupervised remote sensing image semantic segmentation with masked image modeling and contrastive learning method. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–17.

Li, Z.; Hou, B.; Ma, S.; Wu, Z.; Guo, X.; Ren, B.; and Jiao, L. 2024b. Masked angle-aware autoencoder for remote sensing images. In *In Proceedings of the European Conference on Computer Vision*, 260–278. Springer.

Long, Y.; Xia, G.-S.; Li, S.; Yang, W.; Yang, M. Y.; Zhu, X. X.; Zhang, L.; and Li, D. 2021. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 4205–4230.

Lu, X.; Jiao, L.; Li, L.; Liu, F.; Liu, X.; Yang, S.; Feng, Z.; and Chen, P. 2023. Weak-to-strong consistency learning for semisupervised image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Mai, H.; Sun, R.; Zhang, T.; and Wu, F. 2024. RankMatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3391–3401.

Mall, U.; Hariharan, B.; and Bala, K. 2023. Change-aware sampling and contrastive learning for satellite images. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5261–5270.

Mendieta, M.; Han, B.; Shi, X.; Zhu, Y.; and Chen, C. 2023. Towards geospatial foundation models via continual pretraining. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16806–16816.

Noman, M.; Naseer, M.; Cholakkal, H.; Anwer, R. M.; Khan, S.; and Khan, F. S. 2024. Rethinking transformers pre-training for multi-spectral satellite imagery. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27811–27819.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research Journal*, 1–31.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-supervised semantic segmentation with cross-consistency training. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12674–12684.

Reed, C. J.; Gupta, R.; Li, S.; Brockman, S.; Funk, C.; Clipp, B.; Keutzer, K.; Candido, S.; Uyttendaele, M.; and Darrell, T. 2023. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4088–4099.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C. A.; Cubuk, E. D.; Kurakin, A.; and Li, C.-L. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33: 596–608.

Souly, N.; Spampinato, C.; and Shah, M. 2017. Semi supervised semantic segmentation using generative adversarial network. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5688–5696.

Stojnic, V.; and Risojevic, V. 2021. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1182–1191.

Sun, B.; Yang, Y.; Zhang, L.; Cheng, M.-M.; and Hou, Q. 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3097–3107.

Sun, X.; Wang, P.; Lu, W.; Zhu, Z.; Lu, X.; He, Q.; Li, J.; Rong, X.; Yang, Z.; Chang, H.; et al. 2022. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–22.

Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; and Isola, P. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33: 6827–6839.

Wang, D.; Zhang, J.; Du, B.; Xia, G.-S.; and Tao, D. 2022a. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–20.

Wang, D.; Zhang, J.; Du, B.; Xu, M.; Liu, L.; Tao, D.; and Zhang, L. 2023. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36: 8815–8827.

Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. 2024a. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; and Zhang, L. 2022b. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–15.

Wang, F.; Wang, H.; Wang, D.; Guo, Z.; Zhong, Z.; Lan, L.; Yang, W.; and Zhang, J. 2025. Harnessing Massive Satellite Imagery with Efficient Masked Image Modeling. *arXiv preprint arXiv:2406.11933*.

Wang, H.; Zhang, Q.; Li, Y.; and Li, X. 2024b. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3627–3636.

Wang, J.-X.; Chen, S.-B.; Ding, C. H.; Tang, J.; and Luo, B. 2021. RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

Waqas Zamir, S.; Arora, A.; Gupta, A.; Khan, S.; Sun, G.; Shahbaz Khan, F.; Zhu, F.; Shao, L.; Xia, G.-S.; and Bai, X. 2019. isaid: A large-scale dataset for instance segmentation in aerial images. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 28–37.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified perceptual parsing for scene understanding. In *In Proceedings of the European Conference on Computer Vision*, 418–434.

Xie, X.; Cheng, G.; Wang, J.; Yao, X.; and Han, J. 2021. Oriented R-CNN for object detection. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3520–3529.

Yang, L.; Zhao, Z.; and Zhao, H. 2025. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yang, L.; Zhuo, W.; Qi, L.; Shi, Y.; and Gao, Y. 2022. St++: Make self-training work better for semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4268–4277.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *In Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032.

Zhang, L.; and Zhang, L. 2022. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geoscience and Remote Sensing Magazine*, 10(2): 270–294.

Zhao, Z.; Long, S.; Pi, J.; Wang, J.; and Zhou, L. 2023a. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23705–23714.

Zhao, Z.; Yang, L.; Long, S.; Pi, J.; Zhou, L.; and Wang, J. 2023b. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11350–11359.