S5: Scalable Semi-Supervised Semantic Segmentation in Remote Sensing

Liang Lv
Wuhan University
Wuhan, China
lianglyu@whu.edu.cn

Di Wang
Wuhan University
Wuhan, China
d_wang@whu.edu.cn

Jing Zhang*
Wuhan University
Wuhan, China
jingzhang.cv@gmail.com

Lefei Zhang*
Wuhan University
Wuhan, China
zhanglefei@whu.edu.cn

Bo Du
Wuhan University
Wuhan, China
dubo@whu.edu.cn

Liangpei Zhang
Wuhan University
Wuhan, China
zlp62@whu.edu.cn

Abstract

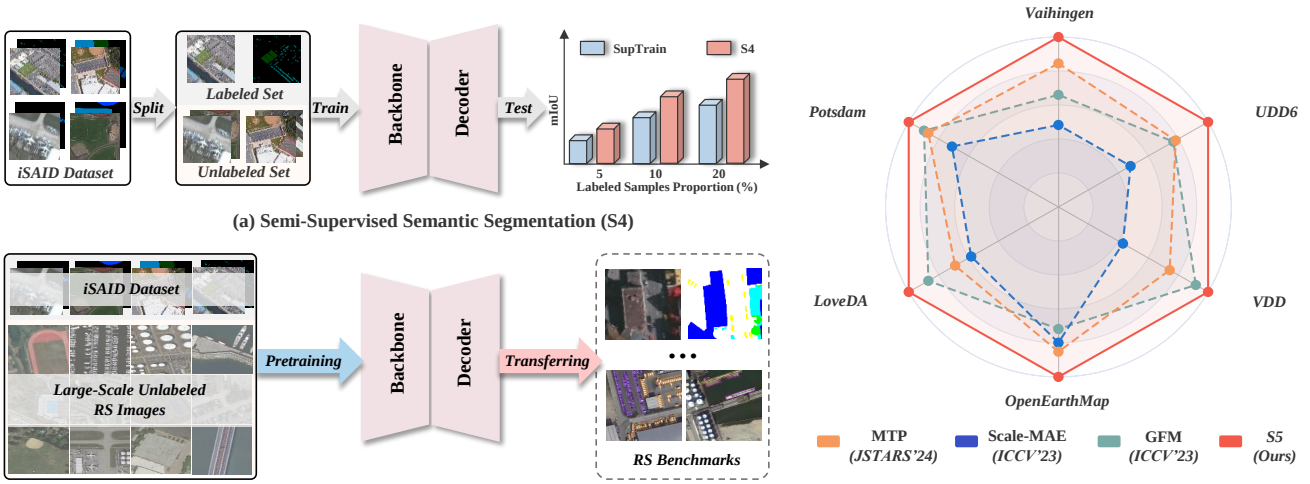
Semi-supervised semantic segmentation (S4) has advanced remote sensing (RS) analysis by leveraging unlabeled data through pseudo-labeling and consistency learning. However, existing RS studies often rely on small-scale datasets and models, limiting their practical applicability. To address this, we propose 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. Our approach introduces MillionSeg, a novel dataset comprising over one million unlabeled RS images spanning diverse geospatial scenes, and systematically scales S4 methods by pre-training RS foundation models of varying sizes on this extensive corpus, generating high-quality pseudo-labels as a byproduct. Through experiments on extensive benchmarks of diverse RS tasks involving semantic segmentation, object detection, and change detection, we show that S5 effectively learns generalizable representations, and scales model capacity to enhance the performance of various downstream RS tasks. The resulting foundation models achieve state-of-the-art performance across all benchmarks, underscoring the viability of scaling semi-supervised learning for RS applications. All datasets, code, and models will be released at S5.

1. Introduction

Remote sensing (RS) semantic segmentation is a crucial field in RS image understanding, focusing on the precise classification of each pixel to achieve automatic recognition and analysis of land cover information [60]. In recent years, significant progress has been made in this field, largely due to the powerful feature extraction capabilities of deep learning models. However, training an effective land cover segmentation model requires a large amount of densely annotated data. Acquiring accurate pixel-level annotations is both time-consuming and costly, which limits the development of RS image segmentation. To reduce the burden of manual labeling and lower costs, semi-supervised semantic segmentation (S4) [34] 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 [38] explores generative adversarial networks (GAN) based methods, while data augmentation has been identified as a key factor [12]. Recent S4 methods rely on pseudo-labeling and consistency regularization techniques. ST++ [56] shows that applying strong data augmentation techniques in self-training significantly improves results, though self-training typically involves multiple stages, making it less efficient. UniMatch [11] revisits consistency regularization from weak to strong augmentation, a method originally simplified and generalized by FixMatch [37] in semi-supervised classification tasks. FixMatch generates pseudo-labels for weakly augmented images and supervises the predictions of strongly augmented images, ensuring the reliability of the pseudo-labels through a confidence threshold. Due to its simplicity and efficiency, FixMatch becomes a key baseline in the S4 field, with many subsequent works (*e.g.*, the classic Match series such as UniMatch [11], RankMatch [29], and CorrMatch [40]) building upon its framework. In RS, works like RanPaste [49], WSCL [28], and SegMind [24] also adopt the FixMatch framework, designing more effective data augmentation strategies, such as random copy-paste, dual-view augmentation, and random masking for complex RS scenes.

*Corresponding author



(a) Semi-Supervised Semantic Segmentation (S4)

(b) Scalable Semi-Supervised Semantic Segmentation (S5)

(c) Comparison with state-of-the-art methods

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 segmentation performance across six RS segmentation benchmarks.

Recent S4 research in RS still relies on small-scale models and datasets. As illustrated in Fig. 1 (a), a common strategy is to split segmentation benchmark datasets 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 [27] and SAMRS [45], alongside extensive exploration of different pre-training strategies. Self-supervised learning methods, such as contrastive learning [42] and masked image modeling (MIM) [14], extract generalizable features without relying on labeled data. In contrast, supervised pre-training better aligns upstream and downstream tasks and domains, enhancing transferability. For instance, RSP [43] applies supervised pre-training on the MillionAID dataset, producing RSFMs that perform well across diverse downstream tasks. Multi-task pre-training (MTP) [46] further strengthens generalization by bridging the gap between pre-training and target tasks. Notably, SAMRS [45] utilizes the Segment Anything Model (SAM) [19] to generate 100,000 segmentation masks from box annotations, improving land cover segmentation through segmentation pre-training (SEP). However, its reliance on box annotations and limited labeling scale constrains scalability. Given that pre-training is most effective when tasks and domains are well-aligned—as is the case for S4—a key question arises: *Can S4 be scaled to pre-train foundational models on large RS imagery to advance land cover segmentation?*

To answer this question, we introduce Scalable Semi-supervised Semantic Segmentation (S5), the first framework designed to leverage vast amounts of unlabeled RS data for pre-training RSFMs through semi-supervised learning. As shown in Fig. 1 (b), S5 begins by introducing MillionSeg, a novel dataset comprising over one million unlabeled RS images covering a wide range of geospatial scenarios. By integrating MillionSeg with FixMatch [37], a leading semi-supervised learning framework, we systematically explore S5's potential for pre-training RSFMs at various scales. It enhances the transferability of feature representations learned from MAE or ImageNet pre-trained weights. Extensive experiments on multiple RS benchmarks including ISPRS Potsdam, ISPRS Vaihingen, and LoveDA demonstrate that RSFMs trained with S5 achieve state-of-the-art (SOTA) performance. Further evaluations on object detection and change detection tasks confirm that S5 pre-trained weights also lead to substantial performance gains in these tasks, establishing semi-supervised learning as a promising pathway for RSFM development. Additionally, the trained models generate high-quality pixel-level pseudo-labels for MillionSeg, which further improve fully supervised base model training. This self-reinforcing cycle opens new avenues for exploring more advanced RSFM pre-training approaches.

Our 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 imagery to develop RSFMs.

- We curate MillionSeg, a large-scale dataset comprising over one million RS images spanning diverse geospatial scenes. This dataset enables effective pre-training of RSFMs, which in turn produce high-quality segmentation pseudo-labels, paving the way for future research in scalable RSFM pre-training.
- Extensive experiments on extensive datasets demonstrate that RSFMs developed using S5 performance set new SOTA performance across all benchmark evaluations.

2. Related Work

2.1. 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. Recent deep learning-based S4 methods have significantly improved performance. UniMatch [11] enforces weak-to-strong consistency at both image and feature levels, AugSeg [62] enhances robustness through data augmentation, and iMAS [61] introduces instance-specific, model-adaptive supervision. CorrMatch [40] propagates labels via correlation matching, AllSpark [47] employs a Transformer-based approach leveraging labeled features, SemiVL [15] integrates vision-language guidance for better pseudo-labeling, and UniMatchV2 [57], built on DINOV2 [33], further improves S4 performance.

S4 methods in RS address domain-specific challenges. For instance, RanPaste [49] exploits unlabeled data through consistency and pseudo-labeling, WSCL [28] transitions from weak to strong labels, SegMind [24] integrates mask image modeling with contrastive learning, and DWL [16] 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.

2.2. 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 [43] pioneered this approach by pre-training CNNs and vision transformers on Million-AID [27], while MTP [46] aligns multiple downstream tasks for enhanced representation learning. SAMRS [45] improves segmentation-focused FMs using a 100,000-sample dataset built with SAM [19]. 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 [20], multimodal images [17, 39], geographic priors [22], or temporal imagery [30]—or generative methods like MIM [14] that reconstruct masked regions to capture structural features. RingMo [41] employs incomplete masking for dense small objects in RS scenes, while RVSA [44], initialized with MIM weights, introduces rotational window attention to enhance target representation with lower computational cost. GFM [31] refines MIM pre-training by leveraging ImageNet-pretrained FMs, while SatMAE [5] and Scale-MAE [35] incorporate multi-temporal and multi-scale features, respectively. In contrast, our work explores S4 pre-training for RSFMs and introduces S5—the first scalable framework for semi-supervised semantic segmentation in RS. Leveraging the newly established MillionSeg dataset in this work, we successfully pretrain RSFMs with up to 600M parameters, achieving SOTA performance across multiple benchmarks.

3. Method

3.1. Pre-training Dataset

Earth observation satellites generate massive volumes of remote sensing imagery daily. However, the high cost of annotation limits their effective use, particularly for land cover segmentation requiring pixel-level labels. To explore S4 in the context of large-scale remote sensing images, we first analyze existing publicly available datasets.

In RS, the MillionAID dataset [27], with 1 million scene-level labeled samples across 62 categories, is widely used for pre-training RSFMs. Its large-scale and diverse scenes make it our primary source of unlabeled images. Since MillionAID primarily comprises satellite images from Google Earth, we also include iSAID [51], a similar instance-level RS segmentation dataset, as the accompanying labeled set.

To facilitate the study of using MillionAID for scaling S4, we first train a fully supervised segmentation model (ViT-L [9] + UperNet [54]) on iSAID and use it to generate pseudo-labels for MillionAID. To evaluate their quality, we sample a subset of MillionAID matching the SAMRS dataset [45] in scale and conduct pre-training with UperNet using a Swin-T [26] backbone. The model is then fine-tuned on ISPRS Vaihingen[1]. We compare three pre-training strategies: ImageNet Pre-training (IMP), IMP + Segmentation Pre-training (SEP) from SAMRS [45], and IMP + SEP trained with our pseudo-labeled MillionAID subset.

The results listed in Table 1 suggest that the pseudo-labels obtained from MillionAID do not significantly improve performance compared to SAMRS labels. To explore this further, we examine MillionAID samples and find that many scenes, such as beaches and bare land (Fig. 3), lack

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

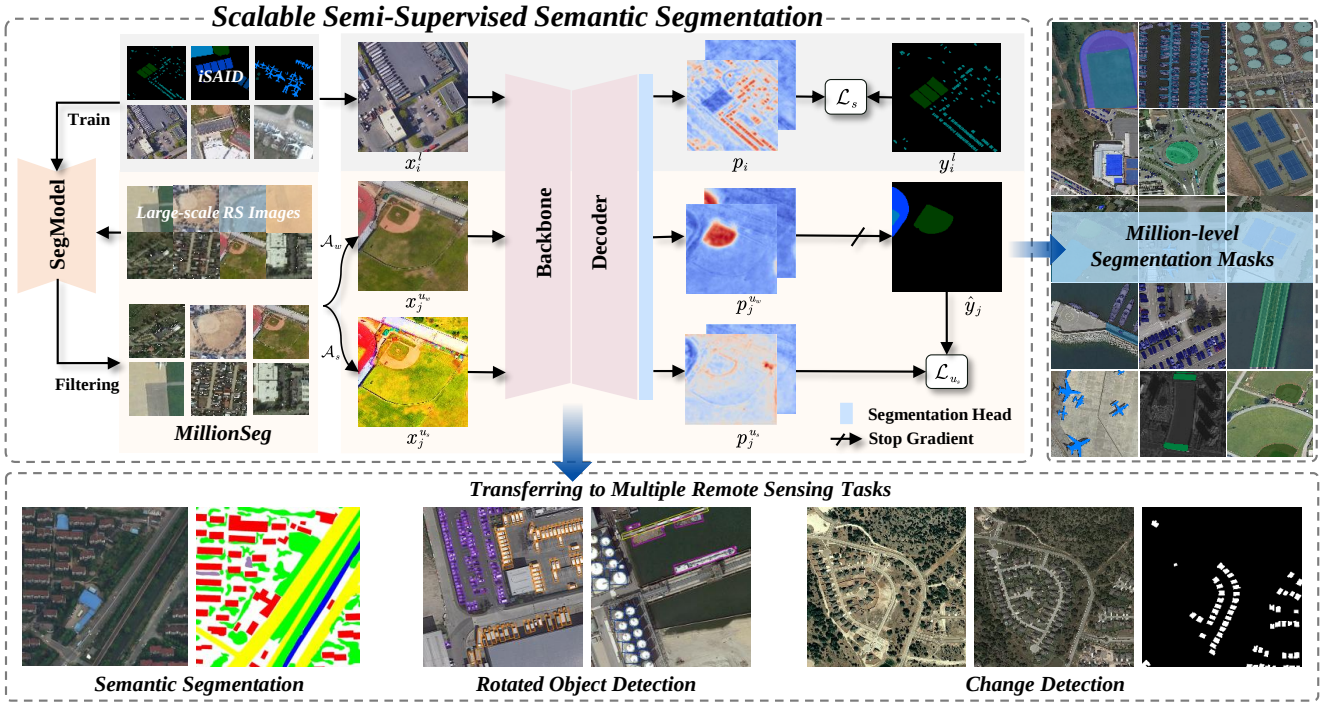


Figure 2. The overall framework of the proposed S5. First, the MillionSeg dataset is constructed for pre-training, where segmentation models with various backbones are trained. Then, the pretrained weights are transferred to different RS benchmarks for fine-tuning. The SegModel refers to a segmentation model trained on the iSAID dataset, designed to filter and retain only the images that contain foreground objects.

(a) (b)

Figure 3. **Visualization of candidate samples.** (a) Examples of discarded images. (b) Retained images with corresponding foreground object masks.

distinct foreground objects. This misalignment with the iSAID dataset, which labels only instance-level foreground objects, may hinder the segmentation model's ability to generate discriminative labels.

To validate this hypothesis, we remove samples without foreground objects in their pseudo-labels, resulting in a new MillionAID dataset, denoted as "MillionAID*". We then repeat the SEP and fine-tuning processes, ensuring that the image scale used for SEP remains consistent with SAMRS for a fair comparison. The results in Table 1 show an accuracy improvement over SAMRS, confirming the effectiveness of MillionAID. This suggests that pre-training with segmentation pseudo-labels generated by on-site models is effective, *even without the use of expensive large-scale visual foundation models like SAM [19] for labeling, as done in SAMRS*. Moreover, using the full MillionAID dataset for pre-training, instead of the 100K images we currently use, is expected to yield a more powerful segmentation model.

Table 1. **Dataset Analysis**. We assess the effectiveness of the MillionAID dataset for SEP and compare it with SAMRS. * denotes the filtered MillionAID subset.

Method	SEP Dataset	# Images	mIoU (%)
IMP	-	-	77.65
IMP + SEP	SAMRS	100K	78.05
IMP + SEP	MillionAID	100K	77.72
IMP + SEP	MillionAID*	100K	78.13
S4 Pre-training	MillionAID*	100K	78.75

3.2. S4 Pre-training

The above process typically involves three stages: (1) training a seed segmentation model with a limited labeled dataset, (2) generating pseudo-labels for unlabeled images

through inference, and (3) retraining the model using these pseudo-labeled data. This workflow is a standard self-training approach [56], commonly used in S4. However, self-training often requires multiple iterations of steps (2) and (3) to refine pseudo-labels and improve the model. As the dataset size grows, this iterative process becomes less efficient. To improve the scalability of S4 for large datasets, we adopt the weak-to-strong consistency regularization for pre-training, such as FixMatch [37]. This method establishes a dual-view learning mechanism, where pseudo-labels from weakly augmented images guide strongly augmented counterparts. A confidence-based thresholding strategy is employed to retain high-confidence predictions while discarding uncertain ones. Specifically, we introduce a universal pre-training framework for RSFM that is compatible with existing S4 methods, exemplified by FixMatch [37]. *Note that this paper does not propose a new semi-supervised learning algorithm for pre-training RSFMs. Instead, it uses FixMatch as a representative approach for implementing S4. Other approaches have also been shown to be compatible with our framework and effective (Table 5).*

Consider a dataset comprising labeled image pairs $\{(x_i^l, y_i^l)\}_{i=1}^{B_l}$ and unlabeled images $\{x_j^u\}_{j=1}^{B_u}$. Each labeled image $x_i^l \in \mathbb{R}^{H\times W\times 3}$ has corresponding pixel-level annotations $y_i^l \in \mathbb{R}^{H\times W\times K_p}$, where K_p is the number of classes in the pre-training phase. Each unlabeled image $x_j^u \in \mathbb{R}^{H\times W\times 3}$ 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 [58], color jitter, grayscale conversion, and Gaussian blur.

For each unlabeled image x_j^u, we generate two augmented views using sequential transformations:

$$x_j^{u_w} = \mathcal{A}_w(x_j^u), \quad x_j^{u_s} = \mathcal{A}_s(\mathcal{A}_w(x_j^u)). \tag{1}$$

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

$$\mathcal{L} = \mathcal{L}_s + \lambda\, \mathcal{L}_{u_s}. \tag{2}$$

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:

$$\mathcal{L}_s = \frac{1}{B_l}\sum_{i=1}^{B_l} \mathcal{L}_{ce}(y_i^l, p_i), \tag{3}$$

$$\mathcal{L}_{u_s} = \frac{1}{B_u}\sum_{j=1}^{B_u} \mathbb{1}(\max(p_j^{u_w}) \geq \tau)\mathcal{L}_{ce}(\hat{y}_j, p_j^{u_s}), \tag{4}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss, $p_i = f(x_i^l)$, $p_j^{u_w} = f(x_j^{u_w})$, and $p_j^{u_s} = f(x_j^{u_s})$ are the predicted probability maps for labeled and unlabeled images. The pseudo-label $\hat{y}_j$ is determined by $\hat{y}_j = \arg\max(p_j^{u_w})$, 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. As a result, the S4 pre-trained model is expected to outperform the SEP pre-trained model, as demonstrated in the last row of Table 1.

3.3. S5: A Scalable Semi-Supervised Semantic Segmentation Framework

The findings above demonstrate the effectiveness of S4 pre-training on large datasets. Building on this, we propose the Scalable Semi-Supervised Semantic Segmentation (S5) framework, as illustrated in Fig. 2, which scales efficiently to millions of samples using the MillionSeg dataset. By incorporating S4 methods, S5 successfully pre-trains RSFMs with up to 600M parameters, achieving SOTA performance in RS segmentation.

To create MillionSeg, we enlarge the dataset by including images from several large-scale RS datasets, such as DOTA-v2.0 [8], SIOR [45], FAST [45], and STAR [23], alongside MillionAID and iSAID. Since some datasets (*e.g.*, DOTA-v2.0, STAR, and MillionAID) contain large images (over 1024 $\times$ 1024), we crop them to a uniform 512 $\times$ 512, following standard practices in the RS segmentation community [43]. Meanwhile, iSAID images are cropped to 1024 $\times$ 1024 following [50]. By combining these datasets, MillionSeg is a large-scale, segmentation-focused pre-training resource that covers a wide range of geospatial scenes and object types. Note that the current version of MillionAID includes slightly fewer than one million samples due to filtering. A detailed dataset composition is provided in Table 2.

After constructing the large-scale MillionSeg dataset, we apply the proposed S5 framework to develop RSFMs for segmentation tasks. To assess the impact of scaling, we experiment with various segmentation models that incorporate different backbone architectures, from convolutional networks to vision transformers, with parameter counts ranging from 20M to 600M. During pre-training, the entire encoder-decoder-head structure is optimized to learn rich features from MillionSeg. For downstream tasks, we transfer the pre-trained encoder-decoder and adapt the segmentation head to suit the task-specific categories, ensuring seamless application across various scenarios.

Implementation Details: The experiments are conducted using PyTorch on 8 NVIDIA RTX 3090 GPUs.

Table 2. Dataset breakdown in MillionSeg. We collect and process images from various open-source RS datasets to compile one million samples.

	Dataset	# Images	Image Size
Labeled Images	iSAID [51]	15,031	1024 × 1024
Unlabeled Images	DOTA-v2.0 [8]	35,151	512 × 512
	SIOR [45]	23,463	800 × 800
	FAST [45]	64,147	600 × 600
	STAR [23]	66,234	512 × 512
	MillionAID [27]	812,200	512 × 512

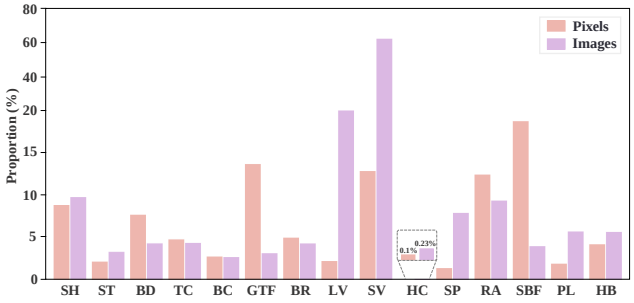


Figure 4. Pixel and image statistics for each category in MillionSeg: SH (ship), ST (storage tank), BD (baseball diamond), TC (tennis court), BC (basketball court), GTF (ground track field), BR (bridge), LV (large vehicle), SV (small vehicle), HC (helicopter), SP (swimming pool), RA (roundabout), SBF (soccer ball field), PL (plane), and HB (harbor).

We use two widely adopted segmentation baselines: (1) DeepLabV3+ [3] with ResNet-101 [13] backbone, and (2) UperNet [54] with Swin-T [26], ViT-B [9], ViT-L [9], ViT-B [9] + RVSA [44], ViT-L [9] + RVSA [44] and ViT-H [9] backbones. The ViT-based backbones are initialized with weights pre-trained on MillionAID, while ResNet-101 and Swin-T are initialized with weights pre-trained on ImageNet. Models are trained for 120,000 iterations with the AdamW optimizer, a weight decay of 0.01, and a cosine learning rate schedule. The base learning rate is set to 2e-5 for ViT-B, ViT-L, ViT-B + RVSA, ViT-L + RVSA and ViT-H, while other models use 6e-5. The batch size is 48 for ViT-H and 96 for the rest. Input images are cropped to 512 × 512 during pre-training. To optimize memory usage and efficiency, we employ mixed precision training and a checkpointing technique for ViTs.

Million-level Labeled Segmentation Dataset: After pre-training RSFMs with S5, we obtain more powerful segmentation models and, as a byproduct, high-quality pseudo labels for MillionSeg. To better understand MillionSeg, we analyze its category distribution at both the pixel and image levels. As shown in Fig. 4, small vehicles (SV) and large vehicles (LV) are the most prevalent categories, while helicopters (HC) have the lowest representation. The remaining categories are more evenly distributed, reflecting real-world patterns. Additionally, Fig. 5 illustrates the pseudo-labels generated by the S5-trained ViT-H, which accurately capture object shapes. Overall, we believe MillionSeg serves as a valuable large-scale dataset for RSFM pre-training.

Table 3. Fine-tuning results of various methods and backbones on the ISPRS Vaihingen dataset. IMP denotes pre-training on ImageNet [7], while MAE refers to Masked Autoencoder pre-training [14] on MillionAID [27].

Method	Pretrain	Backbone	Vaihingen	LoveDA	UDD6
Convolutional Networks					
DeepLabV3+ [3]	IMP	ResNet-101 [13]	77.95	52.67	71.81
DeepLabV3+ [3]	S5	ResNet-101 [13]	**78.98**	**53.74**	**72.14**
Vision Transformers					
UperNet [54]	IMP	Swin-T [26]	77.65	51.94	71.65
UperNet [54]	S5	Swin-T [26]	**79.21**	**53.95**	**72.41**
UperNet [54]	MAE	ViT-B [9]	78.27	52.55	72.52
UperNet [54]	S5	ViT-B [9]	**79.96**	**54.28**	**74.83**

4. Experiments

4.1. Fine-tuning Datasets and Implementations

For the fine-tuning experiments, we adopt six RS semantic segmentation datasets: ISPRS Vaihingen, ISPRS Potsdam, LoveDA [48], OpenEarthMap [53], UDD6 [4], and VDD [1]. To ensure a fair comparison, we use the mean F1 score (mF1) and mean Intersection over Union (mIoU) as evaluation metrics. Additional details about these datasets and the implementations are provided in the appendix.

4.2. Fine-tuning Results and Analyses

4.2.1. Effectiveness of S5

We begin by validating the effectiveness of S5 by fine-tuning various segmentation models on the ISPRS Vaihingen dataset, as presented in Table 3. The results demonstrate that S5 consistently enhances performance across both CNN and vision transformer networks, regardless of model size. This improvement holds true whether the backbones are pre-trained on remote sensing images (MillionAID) or natural scenes (ImageNet). These findings highlight S5 as a highly effective pre-training approach for land cover segmentation tasks.

4.2.2. Comparison with SOTA Methods

To comprehensively compare our method with existing SOTA methods—including RVSA [44], GFM [31], Scale-MAE [35], SAMRS [45], SatMAE++ [32], and MTP [46]—we further fine-tune various pre-trained models on additional public RS segmentation datasets. Specifically, we evaluate these pre-trained models using larger vision transformer networks, ViT-L and ViT-H. The experimental results, presented in Table 4, demonstrate that our method consistently outperforms the previous approaches across all

Figure 5. Visualization of generated pseudo-labels in MillionSeg.

Table 4. Fine-tuning results of various methods and backbones across multiple RS benchmarks. The best and second-best scores are highlighted in **bold** and blue, respectively.

Method	Pretrain	Backbone	#Parameter (M)	Vaihingen (mIoU)	Potsdam (mF1)	LoveDA (mIoU)	OpenEarthMap (mIoU)	VDD (mIoU)	UDD6 (mIoU)
Comparison Methods									
RVSA [44]	MAE [14]	ViT-B + RVSA [44]	86	78.49	91.58	52.44	66.63	73.22	71.13
GFM [31]	GFM [31]	Swin-B [26]	88	79.61	92.55	54.98	67.78	79.16	75.38
Scale-MAE [35]	Scale-MAE [35]	ViT-L [9]	307	78.64	92.02	53.67	68.54	74.30	72.63
SAMRS [45]	SEP [45]	ViT-B + RVSA [44]	86	78.73	91.69	53.04	67.37	74.79	71.39
SatMAE++ [32]	SatMAE++ [32]	ViT-L [9]	307	78.80	91.64	52.82	65.62	73.31	72.38
MTP [46]	MTP [46]	ViT-L + RVSA [46]	307	80.62	92.47	54.16	69.04	77.42	75.53
Our Methods									
DeepLabV3+ [3]	S5	ResNet-101 [13]	45	78.98	91.70	53.74	66.14	76.50	73.57
UperNet [54]	S5	Swin-T [26]	28	79.21	92.38	53.95	67.21	76.64	72.41
UperNet [54]	S5	ViT-B [9]	86	79.96	92.23	54.12	68.46	77.07	74.81
UperNet [54]	S5	ViT-B + RVSA [44]	86	79.94	92.14	54.14	67.60	78.16	75.15
UperNet [54]	S5	ViT-L [9]	307	81.15	92.73	55.15	70.08	79.76	76.54
UperNet [54]	S5	ViT-L + RVSA [44]	307	80.90	92.70	54.99	69.95	79.85	76.84
UperNet [54]	S5	ViT-H [9]	632	**81.47**	**92.84**	**55.58**	**70.44**	**79.98**	**77.61**

datasets and model sizes. Notably, the S5 pre-trained ViT-L enables UperNet to surpass all existing advanced methods. Moreover, when utilizing the larger ViT-H model, performance improves even further, achieving the best results across all datasets. This highlights the scalability of our approach. Additionally, we provide a qualitative comparison of segmentation results on the ISPRS Vaihingen and OpenEarthMap datasets, as shown in Fig. 7. The visualizations, generated using UperNet with a ViT-H backbone, illustrate the superior segmentation capability of our S5-pretrained model, particularly for small objects (see red boxes in the first row). Furthermore, our model exhibits fewer misclassifications (see red boxes in the second row).

4.2.3. Scalability of S5

Fig. 6 presents the fine-tuning results of UperNet on the ISPRS Vaihingen dataset, using different scales of the MillionSeg dataset and various backbones. As shown in Fig. 6 (a), increasing the amount of unlabeled data from MillionSeg leads to significant performance improvements. For example, when using ViT-B, pre-trained with MAE on MillionAID, as the backbone, the mIoU rises from 78.27% to 79.35% after S5^1 (100,000 images). Further scaling up the dataset to 500,000 images in S5^2 and the full dataset in S5^3 results in additional gains, with mIoU reaching 79.76% and 79.96%, respectively. These findings highlight the effectiveness of S5 in enabling the model to leverage larger amounts of unlabeled data for enhanced performance.

In Fig. 6 (b), a similar trend is observed when using ViT-L and ViT-H as backbones. The mIoU for ViT-L improves from 79.92% to 81.15%, while ViT-H sees an increase from 80.86% to 81.47%. These results demonstrate that S5 consistently enhances performance across backbone networks of different sizes (from ViT-B to ViT-H), with its benefits becoming more pronounced as the dataset scale increases. Overall, this highlights the effectiveness of S5 in improving the segmentation performance of RSFMs across various backbones and underscores its potential in building RSFMs.

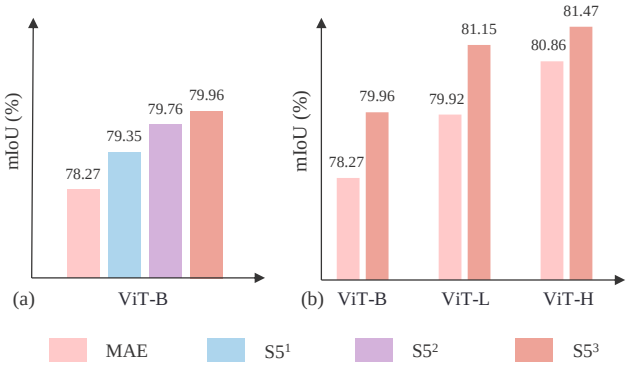


Figure 6. Fine-tuning results of UperNet on the ISPRS Vaihingen dataset with different MillionSeg dataset sizes and backbone configurations. (a) Bar charts in various colors show results using different amounts of unlabeled MillionSeg images: S5^1 (100K), S5^2 (500K), and S5^3 (full dataset). (b) Fine-tuning performance with ViT-B, ViT-L, and ViT-H backbones.

Table 5. Compatibility of S5 with representative S4 methods. We compare pre-training costs across different S4 methods and evaluate fine-tuning results on the ISPRS Vaihingen dataset. GPU memory and training time are measured on NVIDIA RTX 3090 GPUs with a batch size of 6.

Method	Memory (G) / GPU	Time (h) / Epoch	mIoU
FixMatch [37]	8.45	3.52	78.98
UniMatch [11]	15.88	4.71	79.21
CorrMatch [40]	22.33	6.21	79.40
WSCL [28]	8.90	3.63	78.86

4.2.4. Comparison of Various S4 Methods

We investigate the impact of using various S4 pre-training methods in our S5 framework on fine-tuning performance. For comparison, we follow the official implementations of UniMatch [11] and CorrMatch [40], using DeepLabV3+ with ResNet-101 as the backbone network. The experimental results are presented in Table 5. Among the methods evaluated, FixMatch stands out as the most efficient, requiring only 8.45 GB of GPU memory and 3.52 hours per epoch while achieving a competitive fine-tuned mIoU of 78.98%. In contrast, although UniMatch and CorrMatch yield slightly higher accuracy, they come at the cost of significantly increased GPU memory footprint and training time. Compared to FixMatch, the method proposed in [28] incurs slightly higher computational costs while delivering marginally lower performance. Taking both fine-tuning performance and pre-training costs into account, we select FixMatch as default S4 pre-training method in our S5 framework, as it offers the best trade-off.

4.2.5. Pretraining with Other Labeled Datasets

Based on the experimental results presented in Table 6, pre-training with various labeled datasets followed by fine-tuning generally yields effective performance improvements. Nonetheless, the results are influenced by factors such as dataset size and annotation quality. Among the evaluated datasets, LoveDA shows the lowest performance, likely due to its smaller scale or inferior annotation quality; OpenEarthMap yields moderate results, while iSAID consistently outperforms the others across most benchmarks, demonstrating its superiority as a labeled pre-training source in our experimental setting.

4.2.6. Fine-Tuning Evaluation on Other Tasks

We evaluate the transferability of the S5 pre-trained backbones on object detection and change detection tasks. Surprisingly, S5 demonstrates strong generalization ability across both tasks. We validate object detection performance on the DIOR-R [21] and DOTA-v2.0 [8] datasets, and assess change detection performance on the WHU [18] and LEVIR-CD [2] datasets, the fine-tuning experimental settings all follow MTP [46] and CDMamba [59].

Table 7 presents the fine-tuning performance of different methods on object detection datasets. Across all backbone variants from ViT-B to ViT-H, models initialized with S5 pre-trained weights consistently outperform those initialized with MAE. Table 8 shows the fine-tuning results on the change detection benchmarks WHU [18] and LEVIR-CD [2]. At all model scales, S5 yields performance gains over MAE. These results demonstrate that the feature representations learned by S5 not only excel in semantic segmentation but also transfer effectively to downstream tasks such as object detection and change detection, highlighting its strong generalization capability.

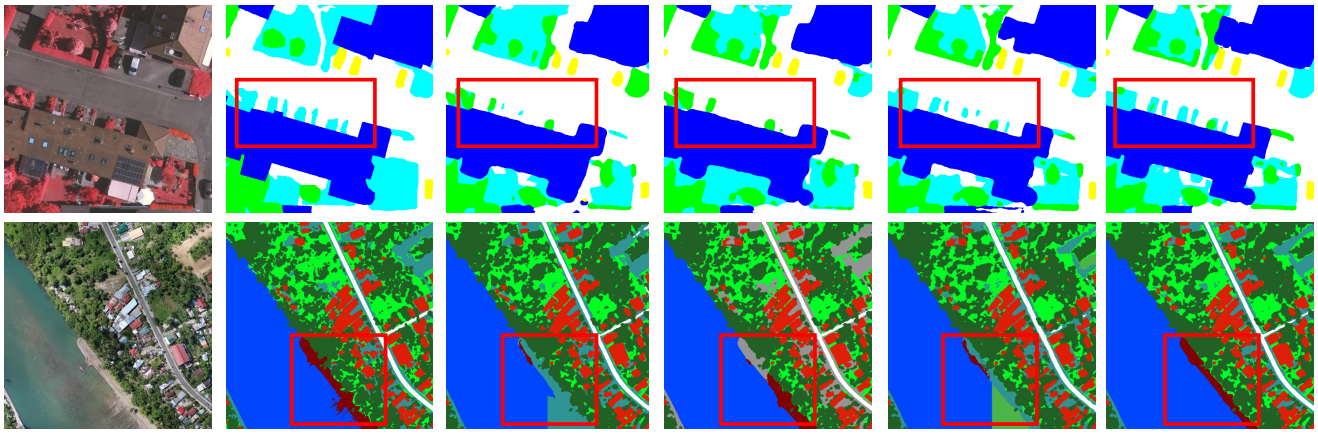
4.2.7. Evaluation on Natural Image Segmentation Benchmarks

To further validate the effectiveness of the proposed S5 framework on natural images, we conduct experiments following the UniMatch V2 [57] setting. We use ADE20K [63] as the labeled dataset, which contains 150 semantic categories and covers diverse indoor and outdoor scenes. In addition, we utilize unlabeled data from COCO [25], Cityscapes [6], and Pascal VOC [10] (excluding their validation and test splits) for pre-training.

Table 9 summarizes the segmentation results on Cityscapes [6], COCO [25], and Pascal VOC [10] using different pre-training strategies. Compared to MAE*, S5 yields consistent improvements across all datasets, achieving gains of +1.17%, +2.08%, and +2.86% on Cityscapes, COCO, and Pascal VOC, respectively, when using the ViT-B backbone with UperNet. These results indicate that S5 not only enhances the representation capability of models in

Table 6. Comparison of fine-tuning results after pre-training on different labeled datasets using the UperNet framework with ViT-B backbone.

Labeled Dataset	Method	Pretrain	Backbone	Vaihingen	Potsdam	LoveDA	OpenEarthMap	VDD	UDD6
-	UperNet	MAE	ViT-B	78.27	91.85	52.55	66.32	73.56	70.86
iSAID	UperNet	S5	ViT-B	**79.96**	92.23	**54.12**	**68.46**	77.07	**74.81**
LoveDA	UperNet	S5	ViT-B	79.18	92.01	-	66.13	76.61	74.27
OpenEarthMap	UperNet	S5	ViT-B	79.49	**92.37**	54.01	-	**77.25**	74.43



(a) (b) (c) (d) (e) (f)

Figure 7. Comparison of segmentation results on the ISPRS Vaihingen (top row) and OpenEarthMap (bottom row) datasets. From left to right: (a) Image, (b) Ground Truth, (c) MTP [46], (d) Scale-MAE [35], (e) SatMAE++ [32], and (f) our S5.

Table 7. Comparison of fine-tuning results (mAP) using different pre-training strategies (MAE vs. S5) on object detection benchmarks (DIOR-R [21] and DOTA-v2.0 [8]).

Method	Pretrain	Backbone	DIOR-R	DOTA-V2.0
ORCN [55]	MAE [14]	ViT-B [9]	68.02	55.51
ORCN [55]	S5	ViT-B [9]	**69.02**	**56.75**
ORCN [55]	MAE [14]	ViT-B + RVSA [44]	68.06	55.22
ORCN [55]	S5	ViT-B + RVSA [44]	**71.22**	**57.18**
ORCN [55]	MAE [14]	ViT-L [9]	70.70	58.91
ORCN [55]	S5	ViT-L [9]	**72.11**	**58.93**
ORCN [55]	MAE [14]	ViT-L + RVSA [44]	70.54	58.96
ORCN [55]	S5	ViT-L + RVSA [44]	**73.30**	**60.33**
ORCN [55]	MAE [14]	ViT-H [9]	73.20	58.97
ORCN [55]	S5	ViT-H [9]	**74.40**	**59.63**

RSFMs, but also improves generalization in the natural image domain, further demonstrating its strong transferability across diverse vision tasks.

4.2.8. High-quality Pseudo-labels of S5

As a byproduct of S5, the pre-trained FMs generate a vast number of pseudo-labels (see Fig. 5), which can be directly leveraged for segmentation pre-training (SEP). To evaluate their quality, we conduct a series of experiments. First, we pre-train a large segmentation model using UperNet with a ViT-H backbone through S5. We then use this model to annotate MillionSeg with pseudo-labels and subsequently perform SEP on a new UperNet model with a ViT-B backbone. The fine-tuning results on the ISPRS Vaihingen dataset, presented in Table 10, show that this approach outperforms ViT-B pre-trained with S5 alone. These findings demon-

Table 8. Comparison of fine-tuning results (F1 score) using different pre-training strategies (MAE vs. S5) on change detection benchmarks (WHU [18] and LEVIR-CD [2]).

Method	Pretrain	Backbone	WHU	LEVIR-CD
UNet [36]	MAE [14]	ViT-B [9]	94.39	91.92
UNet [36]	S5	ViT-B [9]	**94.97**	**92.14**
UNet [36]	MAE [14]	ViT-B + RVSA [44]	94.49	92.21
UNet [36]	S5	ViT-B + RVSA [44]	**94.87**	**92.21**
UNet [36]	MAE [14]	ViT-L [9]	94.92	92.26
UNet [36]	S5	ViT-L [9]	**95.26**	**92.37**
UNet [36]	MAE [14]	ViT-L + RVSA [44]	94.91	92.52
UNet [36]	S5	ViT-L + RVSA [44]	**95.29**	**92.68**
UNet [36]	MAE [14]	ViT-H [9]	95.36	92.70
UNet [36]	S5	ViT-H [9]	**95.66**	**92.75**

Table 9. Segmentation performance (mIoU) on natural image benchmarks using different pre-training methods. Results are reported on Cityscapes [6], COCO [25], and Pascal VOC [10] using ViT-B [9] as the backbone with UperNet [54]. Where MAE* refers to the weights pre-trained on ImageNet [7].

Method	Pretrain	Backbone	Cityscapes	COCO	Pascal VOC
UperNet [54]	MAE* [14]	ViT-B [9]	77.81	54.13	77.97
UperNet [54]	S5	ViT-B [9]	**79.01**	**56.21**	**80.83**

Table 10. mIoU (%) on the Vaihingen dataset using pseudo-labels generated by S5 pre-trained ViT-H models.

Model	Pre-training	Pseudo-labels	Vaihingen
ViT-B + UperNet	MAE	-	78.27
	S5	-	79.96
	SEP	S5-trained ViT-H	**80.45**

strate that the pseudo-labels generated by ViT-H after S5 pre-training are of high quality and can be effectively utilized for SEP, further enhancing supervised pre-training for RSFMs.

Furthermore, it is worth noting that S5 pre-training involves three input branches—supervised, weakly augmented, and strongly augmented—leading to high computational costs. In contrast, directly utilizing high-quality pseudo-labels for SEP significantly reduces computational demands—a byproduct gift from our S5.

4.2.9. Discussions

Our study does not aim to develop a new S4 algorithm. Instead, we introduce S5, a novel and scalable framework for investigating the impact of S4 pre-training on RSFM construction. To support this, we present MillionSeg, a large-scale dataset used to systematically explore pre-training strategies within S5. While FixMatch serves as a representative S4 method, we demonstrate S5's compatibility with other approaches. Using MillionSeg and S5, we scale the model from 20M to 600M parameters, validating the effectiveness of generated pseudo-labels and achieving SOTA performance across multiple RS segmentation benchmarks. In the future, we will enhance S5 by exploring advanced S4 methods that leverage the multi-temporal and multimodal characteristics of RS data.

5. Conclusion

In this paper, we introduce S5, a scalable semi-supervised semantic segmentation framework for building RSFMs. S5 effectively leverages the vast amount of unlabeled RS images, overcoming the limitations of existing S4 methods that rely on small datasets, and unlocking their potential for training RSFMs. We also contribute MillionSeg, a large-scale pre-training dataset for RS segmentation, enabling S5 to pretrain RSFMs with 20M to 600M parameters, setting new state-of-the-art performance across representative benchmarks. In addition, S5 offers high-quality pseudo-labels for MillionSeg as a byproduct, paving the way for future research in RSFM pre-training. We believe our work will inspire continued progress and the development of even more powerful RSFMs.

References

[1] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 6, 1

[2] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote sensing*, 12(10):1662, 2020. 8, 9

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 6, 7

[4] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Pattern Recognition and Computer Vision: First Chinese Conference, PRCV 2018, Guangzhou, China, November 23-26, 2018, Proceedings, Part I 1*, pages 347–359. Springer, 2018. 6, 1

[5] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022. 3

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 8, 9

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6, 9

[8] Jian Ding, Nan Xue, Gui-Song Xia, Xiang Bai, Wen Yang, Michael Ying Yang, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE transactions on pattern analysis and machine intelligence*, 44(11): 7778–7796, 2021. 5, 6, 8, 9

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 6, 7, 9

[10] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 8, 9

[11] Yue Fan, Anna Kukleva, Dengxin Dai, and Bernt Schiele. Revisiting consistency regularization for semi-supervised learning. *International Journal of Computer Vision*, 131(3): 626–643, 2023. 1, 3, 8

[12] Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 3, 6, 7, 9, 1

[15] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2024. 3

[16] Wei Huang, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Decouple and weight semi-supervised semantic segmentation of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212:13–26, 2024. 3

[17] Pallavi Jain, Bianca Schoen-Phelan, and Robert Ross. 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, 2022. 3

[18] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 8, 9

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3, 4

[20] Haifeng Li, Yi Li, Guo Zhang, Ruoyun Liu, Haozhe Huang, Qing Zhu, and Chao Tao. 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, 2022. 3

[21] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS journal of photogrammetry and remote sensing*, 159:296–307, 2020. 8, 9

[22] Wenyuan Li, Keyan Chen, Hao Chen, and Zhenwei Shi. Geographical knowledge-driven representation learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 3

[23] Yansheng Li, Linlin Wang, Tingzhu Wang, Xue Yang, Junwei Luo, Qi Wang, Youming Deng, Wenbin Wang, Xian Sun, Haifeng Li, et al. 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*, 2024. 5, 6, 3

[24] Zhenghong Li, Hao Chen, Jiangjiang Wu, Jun Li, and Ning Jing. 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, 2023. 1, 3

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 8, 9

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3, 6, 7

[27] Yang Long, Gui-Song Xia, Shengyang Li, Wen Yang, Michael Ying Yang, Xiao Xiang Zhu, Liangpei Zhang, and Deren Li. On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid. *IEEE Journal of selected topics in applied earth observations and remote sensing*, 14:4205–4230, 2021. 2, 3, 6

[28] Xiaoqiang Lu, Licheng Jiao, Lingling Li, Fang Liu, Xu Liu, Shuyuan Yang, Zhixi Feng, and Puhua Chen. Weak-to-strong consistency learning for semisupervised image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. 1, 3, 8

[29] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3391–3401, 2024. 1

[30] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5261–5270, 2023. 3

[31] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16806–16816, 2023. 3, 6, 7

[32] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27811–27819, 2024. 6, 7, 9, 5

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 1

[35] Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Brian Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, and Trevor Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, pages 4088–4099, 2023. 3, 6, 7, 9, 5

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 9

[37] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 1, 2, 5, 8

[38] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696, 2017. 1

[39] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1182–1191, 2021. 3

[40] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3097–3107, 2024. 1, 3, 8

[41] Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, et al. Ringmo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2022. 3

[42] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839, 2020. 2

[43] Di Wang, Jing Zhang, Bo Du, Gui-Song Xia, and Dacheng Tao. An empirical study of remote sensing pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–20, 2022. 2, 3, 5

[44] Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2022. 3, 6, 7, 9

[45] Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023. 2, 3, 5, 6, 7, 1

[46] Di Wang, Jing Zhang, Minqiang Xu, Lin Liu, Dongsheng Wang, Erzhong Gao, Chengxi Han, Haonan Guo, Bo Du, Dacheng Tao, et al. Mtp: Advancing remote sensing foundation model via multi-task pretraining. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024. 2, 3, 6, 7, 8, 9, 5

[47] Haonan Wang, Qixiang Zhang, Yi Li, and Xiaomeng Li. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2024. 3

[48] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 6, 1

[49] Jia-Xin Wang, Si-Bao Chen, Chris HQ Ding, Jin Tang, and Bin Luo. Ranpaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2021. 1, 3

[50] Libo Wang, Rui Li, Ce Zhang, Shenghui Fang, Chenxi Duan, Xiaoliang Meng, and Peter M Atkinson. Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 190:196–214, 2022. 5

[51] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 28–37, 2019. 3, 6

[52] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 3

[53] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 6, 1

[54] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3, 6, 7, 9

[55] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3520–3529, 2021. 9

[56] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022. 1, 5

[57] Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 3, 8

[58] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[59] Haotian Zhang, Keyan Chen, Chenyang Liu, Hao Chen, Zhengxia Zou, and Zhenwei Shi. Cdmamba: Incorporating local clues into mamba for remote sensing image binary change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 8

[60] Lefei Zhang and Liangpei Zhang. Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.*, 10(2):270–294, 2022. 1

[61] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23705–23714, 2023. 3

[62] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. Augmentation matters: A simple-yet-effective approach to semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11350–11359, 2023. 3

[63] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 8

S5: Scalable Semi-Supervised Semantic Segmentation in Remote Sensing

Supplementary Material

6. Additional Experimental Studies

Stage-wise Pre-training. S5 is based on existing pre-trained backbones and aims to improve the pixel-level segmentation capability of RSFMs. The experimental results shown in Table 11 indicate that applying pre-trained weights significantly enhances the model's performance. Specifically, without pre-trained initialization weights, the segmentation model only reaches a mIoU of 65.72% on the Vaihingen dataset. When using MAE [14] pre-training weights, the mIoU increases to 78.27%, demonstrating a clear performance gain. It is worth noting that when S5 is used for pre-training from scratch, the model's mIoU reaches 76.28%, which is lower than when using MAE alone. In our consideration, initializing with pre-trained weights enables the model to have an image representation ability, while directly performing pixel-level S4 pre-training cannot provide a holistic prior of image understanding, which may not be favorable for model convergence. Finally, when both MAE and S5 are used together, the model achieves the highest mIoU of 79.96%. This result proves that S5, as a complementary pre-training strategy, can effectively enhance the land-cover segmentation capability of RS foundation models.

Table 11. The mIoU (%) on the Vaihingen dataset using pre-trained weights from different stages of the ViT-B backbone.

Model	MAE	S5	mIoU
ViT-B + UperNet	-	-	65.72
	✓	-	78.27
	-	✓	76.28
	✓	✓	**79.96**

7. Fine-tuning Datasets and Implementation

7.1. Datasets

OpenEarthMap [53] is a benchmark dataset for global high-resolution land cover mapping. It features satellite and aerial images with a ground sampling distance between 0.25 and 0.5 meters. The dataset is manually annotated with nine semantic classes—background, bareland, rangeland, developed space, road, tree, water, agricultural land, and building—plus a background category. Its wide geographic coverage spans 97 regions across 44 countries on six continents, and for evaluation, only the validation subset (excluding xBD data) is used.

LoveDA [48] is designed for domain-adaptive semantic segmentation in RS. It consists of 5,987 high-resolution images (0.3 m) collected from both urban and rural areas in cities such as Nanjing, Changzhou, and Wuhan. The dataset labels seven categories: building, road, water, barren, forest, agriculture, and background.

Potsdam[2] and **Vaihingen** [3] datasets are established benchmarks for urban semantic segmentation. Potsdam offers images at an ultra-high resolution of 5 cm, while Vaihingen provides 9 cm resolution imagery. Both datasets are annotated with six classes: typically, impervious surfaces (e.g., roads and parking lots), buildings, low vegetation, trees, cars, and clutter. In the experiments, we follow [45] and ignore the clutter class.

UDD [4] is a UAV-based dataset captured by a DJI Phantom 4 at altitudes ranging from 60 to 100 meters. It focuses on urban environments and is annotated with six primary classes: other, facade, road, vegetation, vehicle, and roof.

VDD [1] consists of 400 high-resolution RGB images (4000×3000 pixels) captured by the DJI MAVIC AIR II at altitudes ranging from 50 to 120 meters. The dataset labels 7 categories: other, whall, road, vegetation, vehicle, roof, and water.

7.2. Implementations

In the fine-tuning experiments, we evaluate seven different backbone networks, ResNet-101, Swin-T, ViT-B, ViT-B + RVSA, ViT-L, ViT-L + RVSA, and ViT-H, on six RS segmentation benchmark datasets. All experiments employ the AdamW optimizer, with an initial learning rate of 5e-5 for the ViT series and 1e-4 for the other networks. The weight decay is set to 0.01, and a cosine schedule is applied for dynamic learning rate adjustment. To reduce memory consumption, mixed-precision training is utilized. Additionally, data augmentation strategies include random rotation, flipping, resizing, cropping, color jittering, random grayscale transformation, and image blurring. More detailed hyperparameter settings are provided in Table 12.

8. Datasheet

8.1. Motivation

1. For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

A1: MillionSeg is designed to advance research in the field of remote sensing (RS) segmentation. Due to the complexity of pixel annotation in remote sensing images

[2] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-potsdam.html

[3] http://www2.isprs.org/commissions/comm3/wg4/2d-sem-label-vaihingen.html

Table 12. Detailed hyperparameter settings for fine-tuning pre-trained models on different datasets.

Dataset	Vaihingen	Potsdam	LoveDA	OpenEarthMap	VDD	UDD6
Training Image Number	1324	7776	4191	2303	360	106
Training Epoch Number	75	75	120	100	300	600
Batch Size	24	24	24	24	24	24
Training Image Size	512	512	512	512	512	512
Class Number	5	5	7	9	7	6

(RSIs), the field still lacks large-scale RS segmentation datasets, hindering the implementation of RS segmentation pre-training. As a large-scale RS segmentation pre-training dataset with a capacity reaching the million level, MillionSeg, when combined with semi-supervised semantic segmentation (S4), can effectively bridge this gap.

2. Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

A2: MillionSeg is created by the authors.

3. Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

A3: N/A.

8.2. Composition

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances(e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

A1: MillionSeg consists of 16 categories, with each sample comprising a remote sensing image and its corresponding pixel-level semantic labels.

2. How many instances are there in total (of each type, if appropriate)?

A2: MillionSeg has 1,016,226 images.

3. Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

A3: MillionSeg is a real-world sample dataset of global ground objects, containing pixel-level classification annotations. It is the largest dataset in the field of high-resolution remote sensing segmentation, with a scale reaching the million level compared to other high-resolution remote sensing segmentation datasets.

4. What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

A4: Each instance consists of one land object with its pixel-level semantic annotations and the unprocessed image data.

5. Is there a label or target associated with each instance? If so, please provide a description.

A5: Yes. Each target is associated with pixel-level semantic labels lying in the corresponding *.png image.

6. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

A6: Yes. A limited number of instances may exhibit incomplete masks, as the labels are obtained by S5 pre-trained UperNet with ViT-H backbone.

7. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

A7: Yes. The instances' information is stored in the *.png images, and different instances can be clearly by pixel positions and filenames.

8. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

A8: Yes. We use 5,297 annotated images as the validation set to monitor the training process, while the remaining data is used as the training set.

9. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A9: Since most of the images in MillionSeg are raw and unlabeled, we used the S5-pre-trained UperNet to generate pixel-level pseudo-labels for them, which may contain some noisy annotations.

10. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses,

fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

A10: The MillionSeg dataset is composed of several publicly available datasets, including DOTA-V2.0 [52], MillionAID [27], STAR [23], SIOR [45], and FAST [45]. These datasets are publicly accessible and can be downloaded from their respective websites. We sincerely appreciate the significant contributions of their authors to the research community.

11. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

A11: No.

12. Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

A12: No.

8.3. Collection Process

1. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A1: The data associated with each instance are directly observable, as they are stored in the common png format and can be easily viewed via Python Imaging Library or Open Source Computer Vision Library.

2. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

A2: The images in MillionSeg come from dataset publicly available datasets described above, which can be directly downloaded from their websites.

3. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

A3: No.

4. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

A4: The authors.

5. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

A5: Data collection took approximately 3 days, and annotation took around 5 days. Independent processing via programming was required, including clipping, standardizing filenames and formats, and converting label formats. Finally, we used the S5-trained UperNet [54] with the ViT-H [9] backbone to generate labels.

8.4. Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

A1: For the iSAID dataset, we crop the images to a size of 1024 × 1024. For the STAR, MillionAID, and DOTA-V2.0 datasets, we crop the images to 512 × 512. Additionally, we use the UperNet model to filter the images in the STAR, MillionAID, and DOTA-V2.0 datasets, retaining only those that contain foreground objects.

2. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

A2: No.

3. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

A3: We use the Python Imaging Library for cropping.

8.5. Uses

1. Has the dataset been used for any tasks already? If so, please provide a description.

A1: No.

2. Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

A2: A link will be provided upon acceptance of this paper.

3. What (other) tasks could the dataset be used for?

A3: MillionSeg can be used for the research of supervised or self-supervised pre-training of RS semantic segmentation models.

4. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might

need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

A4: No.

5. Are there tasks for which the dataset should not be used? If so, please provide a description.

A5: No.

8.6. Distribution

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

A1: Yes. The dataset will be publicly available.

2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

A2: It will be publicly available on the project website.

3. When will the dataset be distributed?

A3: The dataset will be distributed once the paper is accepted after peer review.

4. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

A4: It will be distributed under the MIT license.

5. Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

A5: No.

6. Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

A6: No.

8.7. Maintenance

1. Who will be supporting/hosting/maintaining the dataset?

A1: The authors.

2. How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

A2: They can be contacted via email available on the project website.

3. Is there an erratum? If so, please provide a link or other access point.

A3: No.

4. Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

A4: No.

5. Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

A5: N/A.

6. If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

A6: N/A.

9. Additional Visualization Results

We present more visual comparison results in Fig. 8.

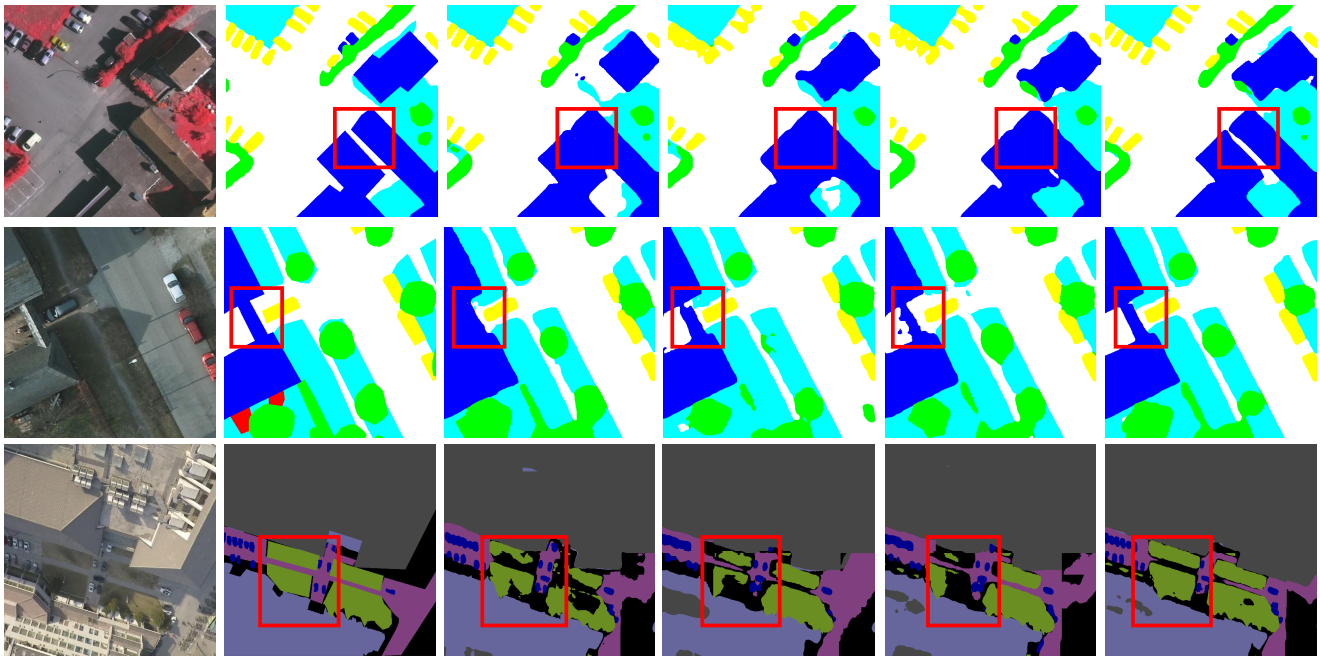


Figure 8. Comparison of segmentation results on the ISPRS Vaihingen (top row), ISPRS Potsdam (middle row) and UDD6 (bottom row) datasets. From left to right: (a) Image, (b) Ground Truth, (c) MTP [46], (d) Scale-MAE [35], (e) SatMAE++ [32], and (f) our S5.