Parameter Efficient Self-Supervised Geospatial Domain Adaptation

Linus Scheibenreif[1] Michael Mommert[2,1] Damian Borth[1]
[1]University of St. Gallen, Switzerland [2]Stuttgart University of Applied Sciences, Germany
{linus.scheibenreif, damian.borth}@unisg.ch michael.mommert@hft-stuttgart.de

Abstract

As large-scale foundation models become publicly available for different domains, efficiently adapting them to individual downstream applications and additional data modalities has turned into a central challenge. For example, foundation models for geospatial and satellite remote sensing applications are commonly trained on large optical RGB or multi-spectral datasets, although data from a wide variety of heterogeneous sensors are available in the remote sensing domain. This leads to significant discrepancies between pre-training and downstream target data distributions for many important applications. Fine-tuning large foundation models to bridge that gap incurs high computational cost and can be infeasible when target datasets are small. In this paper, we address the question of how large, pre-trained foundational transformer models can be efficiently adapted to downstream remote sensing tasks involving different data modalities or limited dataset size. We present a self-supervised adaptation method that boosts downstream linear evaluation accuracy of different foundation models by 4-6% *(absolute) across 8 remote sensing datasets while outperforming full fine-tuning when training only* 1-2% *of the model parameters. Our method significantly improves label efficiency and increases few-shot accuracy by* 6-10% *on different datasets*[1].

1. Introduction

Remote sensing data, such as satellite imagery and aerial photographs, have become ubiquitously available in recent years. Governmental programs such as Landsat [44] and Copernicus [11] produce vast amounts of high quality data and make them publicly available. Important environmental and societal problems can now be addressed by applying methods from computer vision to remote sensing data [41]. These include the monitoring of biodiversity [27], extraction of socioeconomic indicators [45] or the estimation of greenhouse gas emissions [31]. Public remote sensing archives provide high quality data with global coverage, while private fleets of smaller satellites already provide data at high spatial and temporal resolutions, which enables applications that depend on timely observations or high resolution imagery, such as mapping of natural disasters [18], monitoring of marine traffic [12], or precision agriculture [34].

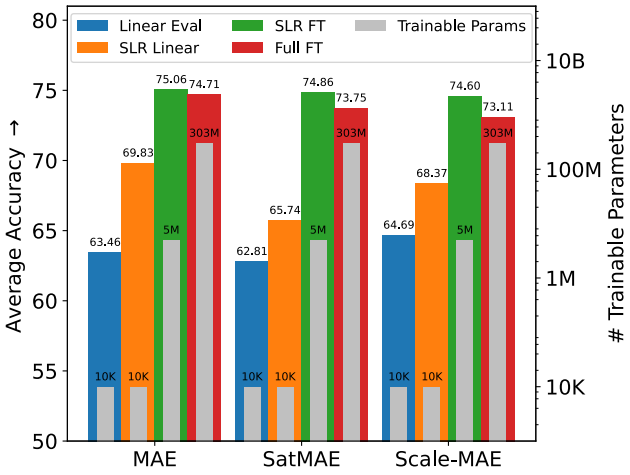


Figure 1. Average performance (colored bars) and number of trainable parameters (gray bars) for different visual foundation models across 8 remote sensing datasets. Our Scaled Low-Rank (SLR) adapter method achieves significant performance improvements across datasets and models with no (SLR Linear) and as little as 1-2% (SLR fine-tuned) additional trainable parameters.

The remote sensing field is characterized by a high heterogeneity of available data sources. Most satellite instruments are custom-designed to monitor specific phenomena enabling the satellite's scientific or commercial mission. Commonly collected data modalities include optical data such as RGB images, multi-spectral data (e.g., near-infrared or short-wave infrared), hyperspectral data, or synthetic aperture radar (SAR). Accordingly, computer vision approaches for remote sensing data are highly fragmented into specialized sub-fields defined by the different modalities or the application of interest (see [49] for a review).

[1]Code available at github.com/HSG-AIML/GDA

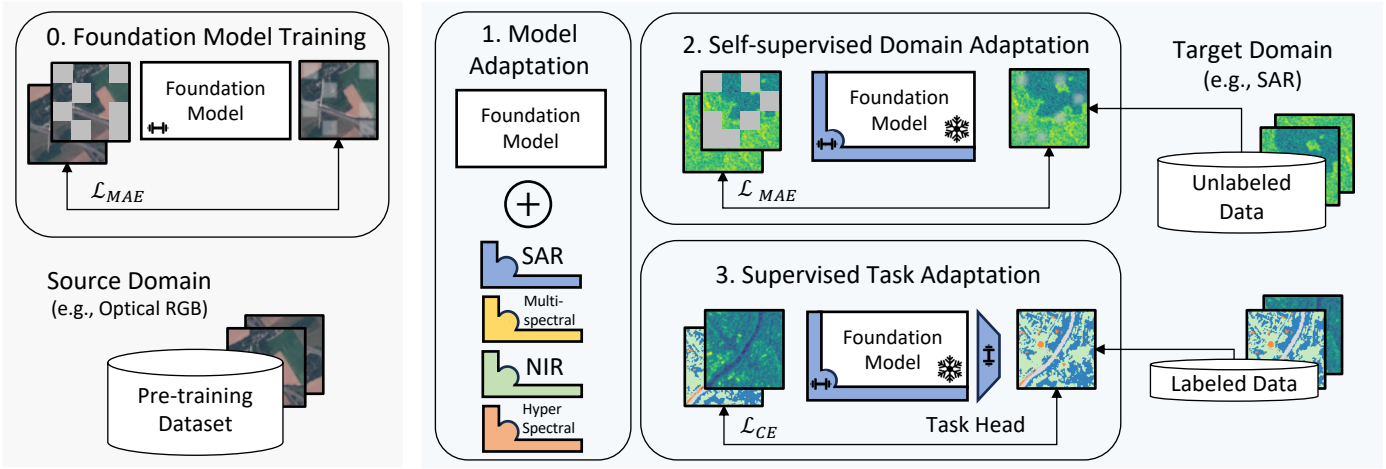


Figure 2. Overview of our parameter-efficient continual pre-training framework. An existing pre-trained visual foundation model (0) is modified with SLR adapters (1). Then, the adapters are trained in a self-supervised way on the target *domain* (2) before the model is fine-tuned for the target *task* in a supervised manner (3). The contributions of this work lie in steps 1–3.

With the success of self-supervised learning, the idea of general foundation models that can be leveraged for many different tasks has gained traction in the computer vision field [4]. Large pre-trained models have become a standard component in computer vision pipelines for classification [13], segmentation [19] or object-detection [47]. In large part, the success of these approaches is due to the ubiquitous nature of optical RGB data in computer vision. Deep neural networks trained on large optical RGB datasets such as ImageNet [29] have shown to be robust to lower-level differences between individual optical camera sensors. This robustness enables the transfer of the pre-trained models to unseen imagery with similar characteristics. Similar results are possible in natural language processing, where unsupervised pre-training on large amounts of textual data yields strong performance on diverse tasks [5].

A number of foundation models for remote sensing data haven recently been proposed [8, 23, 28]. Most of these models follow the computer vision approach and are pre-trained on optical RGB imagery, albeit collected from satellites or airborne observing platforms. However, optical RGB data corresponds only to a small fraction of the commonly used data modalities in the remote sensing domain [38]. Foundation models promise large benefits in remote sensing, where labeled datasets are small and the acquisition of labels can be very expensive. To date, significant potential for remote sensing foundation models remains for data modalities beyond optical RGB data. First steps haven been made to include other data modalities such as multi-spectral [8] or SAR [30] data in the pre-training progress, but remote sensing foundation models remain limited in their ability to adapt to downstream tasks utilizing unseen modalities. Without zero- or few-shot capabilities on modalities other than optical data, expensive fine-tuning protocols have to be employed, resulting in the re-training of large foundation models for datasets involving new modalities. This requires large amounts of labeled samples to adapt the model and comes with high computational cost.

In this paper, we present a new approach for adapting large remote sensing foundation models to novel tasks and modalities in a computationally efficient way. Our method introduces Scaled Low-Rank (SLR) adapters with a small number of parameters to add new data modalities to a pre-trained foundation model. Through self-supervised learning on unlabeled data of the target domain, these additional parameters allow the model to adapt to the characteristics of the new data modality, while the pre-trained parameters are kept fixed. This approach helps to generalize remote sensing foundation models beyond their pre-training data modalities while fully leveraging their existing capabilities. The SLR adapters enable parameters-efficient and label-efficient supervised training for new downstream tasks.
The contributions of our work are as follows:

- We present SLR adapters, a parameter efficient domain adaptation method to utilize visual foundation models on new data modalities.
- We introduce a self-supervised continual pre-training framework to optimize SLR adapters on unlabeled data from new domains.
- Our empirical results demonstrate strong performance of the proposed method. SLR adapters drastically reduce the memory footprint, outperform fine-tuning of all model parameters, and significantly improve performance in few-shot scenarios.

2. Related Work

Computer Vision for Remote Sensing Large amounts of remote sensing data are routinely collected by a wide variety of heterogeneous sensors aboard satellites and airborne vehicles [38]. Much of this data, e.g., optical or radar observations, can be represented as imagery and be processed with computer vision methods [41]. This enables applications ranging from monitoring of biodiversity [27] or wildlife conservation [40] to the estimation of demographic parameters [45] or industrial air pollution [31]. While unlabeled remote sensing data is automatically acquired everyday, producing high quality labeled datasets is difficult for many important applications [3].

Geospatial Foundation Models Foundational models are large general purpose models that are typically trained with self-supervised learning [4]. To solve specific problems, the pre-trained models are then adapted to the target task with supervised fine-tuning. A common pre-training technique for foundation models is masked-autoencoding (MAE) [13], which is used to train transformer models by reconstructing masked portions of the input data. After extensive training on large datasets, MAE produces strong general-purpose visual features [13]. A number of recent works adapt the MAE framework to the characteristics of remote sensing data: SatMAE [8] introduces specialized encodings for the temporal or multi-spectral dimension of satellite imagery and explores different masking schemes for self-supervised training. Scale-MAE [28] explicitly addresses the problem of varying ground sampling distance (GSD) of different remote sensing sensors, and proposes a GSD-aware positional encoding scheme. Masked-autoencoding has also been used with other remote sensing modalities such as SAR [43] or hyperspectral data [32]. Other approaches for geospatial foundation models have used contrastive learning between satellite observations at different points in time [22, 23] or data modalities [30]. Another line of work emphasizes the advantages of hierarchical pre-training approaches [28] for remote sensing foundation models [24].

Parameter Efficient Finetuning General purpose foundation models can be adapted for specific tasks by supervised fine-tuning on labeled datasets [4]. This approach leverages the model's pre-trained representations and adapts them to the characteristics of the target data and task. However, the standard fine-tuning process is very costly, as every parameter of the large foundation model has to be re-trained, which is computationally expensive, has a large memory footprint, and requires a sizeable labeled target dataset. Recently, this process has been simplified by limiting the number of parameters that are trained during the fine-tuning stage [16]. Parameter efficient fine-tuning approaches train only the bias parameters [48], add additional trainable adapter modules [14, 16] such as low-rank matrices [17] between the pre-trained layers, or rescale activations with learnable vectors [20]. Originally proposed for supervised fine-tuning of large-language models, these techniques can reach performance on-par with fine-tuning of the entire model while drastically reducing the memory requirements for individual downstream tasks.

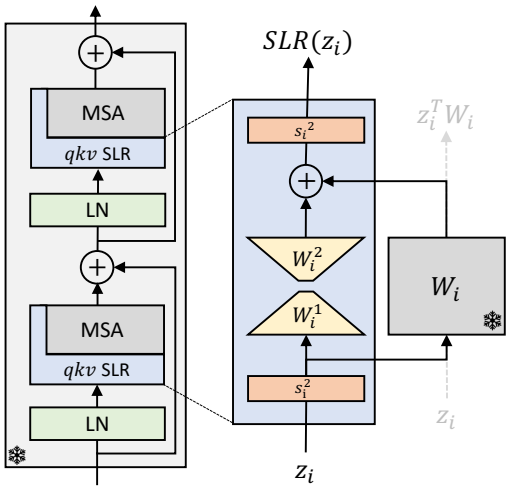


Figure 3. Transformer block (left) with SLR adapters (right). We add individual SLR adapters to linear transformations in the qkv projection and mlp layers of the transformer block. SLR freezes the original transform $\boldsymbol{W}_i$ and introduces trainable scaling parameters $\boldsymbol{s}_i^{1,2}$ and low-rank matrices $\boldsymbol{W}_i^{1,2}$.

Geospatial Domain Adaptation Domain shifts due to changes in acquisition region, time, sensor or environmental conditions are a common problem in geospatial machine learning (see [39] for a review). Unlike most of the general domain adaptation literature, we focus on adapting unsupervised foundation models to new modalities and targets.

3. Method

This section describes our proposed method for efficiently adapting visual foundation models for new remote sensing tasks. We first introduce the proposed continual pre-training framework and then the scaled low-rank adapters. Together, these components facilitate a data-efficient and compute-efficient adaptation of the foundation model to the target domain and task.

Preliminaries We investigate the scenario when a reconstruction-based visual foundation model $f_\theta \colon \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{C\times H\times W}$, trained on a dataset from the source domain $\mathcal{D}_s$, should be transferred to a dataset from the target domain $\mathcal{D}_t$. We focus on vision transformer (ViT) [10] based foundation models trained with masked autoencoding (MAE) [13]. For self-supervised training, the unlabeled input imagery $\boldsymbol{x}_s \sim \mathcal{D}_s$ with $\boldsymbol{x}_s \in \mathbb{R}^{C\times H\times W}$

is split into n non-overlapping patches $\boldsymbol{p} \in \mathbb{R}^{n \times P^2 C}$ of patch size P. The patches are then embedded with a linear transform $f_t \colon \mathbb{R}^{n \times P^2 C} \rightarrow \mathbb{R}^{n \times D}$ to create tokens of dimension D. For each image $\boldsymbol{x}_s$, a binary mask $\boldsymbol{m}$ randomly drops a fraction d of the tokens. The remainder of the tokens is then processed by a ViT encoder $f_{\theta_e} \colon \mathbb{R}^{(1-d) \cdot n \times D} \rightarrow \mathbb{R}^{(1-d) \cdot n \times D}$. After the encoder, MAE models introduce learnable `[MASK]` tokens at the positions of dropped tokens to recover the original sequence length n. A transformer [42] decoder maps the tokens back into pixel space $f_{\theta_d} \colon \mathbb{R}^{n \times D} \rightarrow \mathbb{R}^{C \times H \times W}$ to compute the reconstruction loss $\mathcal{L}_{MAE}$:

$$\mathcal{L}_{MAE} = \frac{1}{d \cdot n} \sum_i \boldsymbol{m}_i \cdot (\boldsymbol{x}_i - f_\theta(\boldsymbol{x}_i))^2 \tag{1}$$

3.1. Parameter Efficient Continual Pre-training

To adapt the model to data $\boldsymbol{x}_t \sim \mathcal{D}_t$ from the unseen target domain, we introduce scaled low-rank (SLR) adapters $f_{\theta_{\text{ada}}}$ to the foundation model f_θ. The resulting model $f_\theta \circ f_{\theta_{\text{ada}}}$ is then trained with unlabeled samples $\boldsymbol{x}_i \in \mathcal{D}_t$ from the target dataset. During that process, the pre-trained parameters θ are kept fixed and only the parameters of the adapters θ_{ada} are optimized with stochastic gradient descent. In practice, we use a masked autoencoding objective $\mathcal{L}_{MAE}$ to train the adapter parameters in a self-supervised fashion.

This makes it possible to leverage all available unlabeled data samples $\boldsymbol{x}_i$ of the target domain. Masked autoencoding reduces the computational costs for continual pre-training as the majority of the patches from each image is dropped in the forward pass, reducing the input sequence length by the factor d. This has significant advantages for transformer models, as their computational complexity is quadratic in the number of input tokens. During the backward pass, we only perform gradient updates for the adapters, as all other parameters are fixed. Training of the adapters primes the foundation model for new types of remote sensing data, even new data modalities, and facilitates successive training for a target task with limited labeled samples. After adaptation to the target *domain*, the foundation model is then adapted to the target *task* (e.g., classification) by supervised fine-tuning of the pre-trained adapter parameters. Based on our setup and experience, the proposed method reduces the number of parameters that are trained in the continual pre-training an fine-tuning stages by about two orders of magnitude for commonly used MAE visual foundation models when compared to the standard fine-tuning approach.

3.2. Scaled Low Rank Adapters

The scaled low-rank (SLR) adapters are designed to augment the linear transformations $\boldsymbol{W}_i$ in a pre-trained transformer foundation model in a parameter efficient way, while maintaining as much capacity as possible. ViTs stack multiple blocks b of multi-head self-attention (MSA) and multi-layer perceptrons (MLP) with layer normalization (LN) [2]:

$$\begin{aligned} \boldsymbol{z}'_b &= \text{MSA}_b(\text{LN}^1_b(\boldsymbol{z}_b)) + \boldsymbol{z}_b \\ \boldsymbol{z}_{b+1} &= \text{MLP}_b(\text{LN}^2_b(\boldsymbol{z}'_b)) + \boldsymbol{z}'_b \end{aligned} \tag{2}$$

Notably, the MSA computes query, key and value representations of the inputs through a linear projection $f_{\text{qkv}} \colon \mathbb{R}^D \rightarrow \mathbb{R}^{3D}$. Similarly, the feed-forward layers of the MLP consist of linear projections $f_{\text{mlp}} \colon \mathbb{R}^D \rightarrow \mathbb{R}^D$. These operations contain the majority of trainable parameters in ViT models.

We propose SLR adapters that scale activations $\boldsymbol{z}_{i-1}$ from the preceding layer element-wise with a learnable vector $\boldsymbol{s}^1_i \in \mathbb{R}^D$. The resulting rescaled feature vector is then passed through the original linear transform $(\boldsymbol{s}^1_i \odot \boldsymbol{z}_i)\boldsymbol{W}_i$. Inspired by low-rank adaptation methods [17], SLR uses symmetric low-rank matrices $\boldsymbol{W}^1_i \in \mathbb{R}^{D \times r}$ and $\boldsymbol{W}^2_i \in \mathbb{R}^{r \times D}$ to process the scaled input features $((\boldsymbol{s}^1_i \odot \boldsymbol{z})\boldsymbol{W}^1_i)\boldsymbol{W}^2_i$. Finally, the adapter adds the features from the original transform and those from the low-rank matrices and multiplies them with a second scaling vector $\boldsymbol{s}^2_i$. Using an SLR adapter (see Fig. 3), the linear transform $f(\boldsymbol{z}_i) = \boldsymbol{z}_i \boldsymbol{W}_i$ turns into:

$$f_{\text{ada}}(\boldsymbol{z}_i) = \boldsymbol{s}^2_i \left[(\boldsymbol{s}^1_i \odot \boldsymbol{z}_i)\boldsymbol{W}_i + ((\boldsymbol{s}^1_i \odot \boldsymbol{z}_i)\boldsymbol{W}^1_i)\boldsymbol{W}^2_i \right] \tag{3}$$

We add instances of this adapter to the linear transforms of the MLP and MSA layers throughout the MAE encoder and decoder. The scaling parameters $\boldsymbol{s}^{1,2}_i$ are initialized with vectors of ones, and the low-rank matrices $\boldsymbol{W}^1_i \sim \mathcal{N}$ and $\boldsymbol{W}^2_i$ with zeros [17]. The SLR adapter consist of $2 \cdot D \cdot r + 2 \cdot D$ parameters. We choose $r \ll D$, making the adapter significantly smaller than the original linear transform $\boldsymbol{W}_i$ with D^2 parameters. Depending on model architecture and bottleneck size r, this introduces $\approx 1\%$ of the original model size as additional parameters (see Sec. 4).

3.3. Supervised Task Adaptation

SLR adapters are added throughout the pre-trained foundation model and trained with masked autoencoding on the target *domain* $\mathcal{D}_t$. We then transfer the model to the target *task* of interest $\mathcal{T}$. The model encoder f_{θ_e} and its adapter parameters are combined with a task-specific head $f_{\theta_{task}}$. The resulting model can then be trained in a supervised way with labeled data $(\boldsymbol{x}_t, \boldsymbol{y}_t)$ from the target domain (e.g., for a k-way classification task with a linear head $\boldsymbol{W}^h \in \mathbb{R}^{D \times k}$ and cross-entropy objective). We investigate supervised task adaptation settings with different efficiency and performance trade-offs (see Sec. 4).

4. Experiments & Results

Datasets We evaluate our experiments on 8 different remote sensing datasets (see Tab. 2): EuroSAT [15], RESISC45 [6], UCMerced[46], FireRisk [33], TreeSatAI [1],

Dataset	MAE				SatMAE				ScaleMAE			
	Linear	SLR Lin.	SLR FT	FT	Linear	SLR Lin.	SLR FT	FT	Linear	SLR Lin.	SLR FT	FT
EuroSAT	93.27	96.61	98.66	98.82	92.00	94.53	98.21	97.79	94.52	95.69	98.65	98.73
RESISC45	77.90	87.08	93.84	95.16	81.80	84.02	92.57	93.39	86.28	87.40	92.87	95.12
FireRisk	37.89	41.78	52.23	49.17	38.27	38.14	50.80	51.21	41.05	41.86	52.63	51.45
TreeSatAI	23.05	38.69	57.66	53.78	21.33	29.55	55.56	50.99	23.48	37.15	53.97	52.58
EuroSAT-SAR	77.95	84.22	87.00	86.46	71.83	79.66	87.17	86.86	73.53	82.73	86.44	78.49
BENGE-S1-C	34.81	42.14	42.80	46.82	35.23	35.60	45.14	44.77	35.35	37.02	45.59	45.07
BENGE-S1-S	68.06	69.74	69.84	69.00	66.57	68.45	70.63	68.27	68.10	68.44	69.05	67.23
UCMerced	94.74	98.35	98.43	98.50	95.44	95.95	98.81	96.65	95.18	96.67	97.60	96.20
Average	63.46	69.83	75.06	74.71	62.81	65.74	74.86	73.74	64.69	68.37	74.60	73.11

Table 1. Classification and segmentation accuracy of different visual foundation models across 8 remote sensing datasets. **Linear**: Linear evaluation of the pre-trained model. **SLR Linear**: Linear evaluation after self-supervised training of SLR adapters. **SLR FT**: Supervised training of the SLR adapters after self-supervised pre-training of SLR adapters. **FT**: Fine-tuning of the full foundation model.

Dataset	# Samples	Modality	GSD
EuroSAT [15]	27k	Multispec.	10m
RESISC45 [6]	31k	RGB	0.2-30m
FireRisk [33]	91k	RGB	1m
TreeSatAI [1]	50k	multiple	0.2-10m
EuroSAT-SAR [43]	27k	SAR	10m
BENGE-8k [25]	8k	multiple	10m
UCMerced [46]	2.1k	RGB	0.3m

Table 2. Overview of the remote sensing datasets used in this work. Each dataset is used to learn a supervised classification and/or segmentation task. The datasets have between 2-90k samples and combine observations from different modalities that vary in their ground-sampling distance (GSD).

BENGE-8k [25] classification, BENGE-8k segmentation and EuroSAT-SAR [43]. These datasets contain different data modalities such as RGB and multi-spectral imaging, SAR polarimetry and others at ground-sampling distances between 30 cm and 30 m per pixel. Detailed information on the specific downstream task trained on each dataset (classification or segmentation) and the corresponding target are presented in the supplemental material. When available, we use the dataset splits defined by the `torchgeo` library [35].

Implementation Details To use foundation models for the downstream classification and segmentation tasks, we append a linear layer or a convolutional layer, respectively, to the model encoder. We use the AdamW [21] optimizer for training and reduce the learning rate by a factor of 10 when the validation loss plateaus. The number of training steps is fixed for each dataset, and we report the test performance of the checkpoint with the lowest validation loss. Images are re-sized to 224^2 pixels and 75% of tokens are masked during the self-supervised SLR adapter pre-training. For few-shot experiments we report the average performance and standard-deviation across three training runs on few-shot samples chosen with different random seeds. Samples are chosen with replacement if k is greater than the total number of samples for a class in the dataset. Further details on datasets and training procedures can be found in the supplemental material.

Foundation Models The SLR adapter method can be applied to any neural network model with dense layers for continual pre-training and efficient fine-tuning. In our experiments, we use ViT-L models with pre-trained weights from three visual foundation models: **MAE**, the vanilla masked autoencoder [13], pre-trained on ImageNet [29]. **SatMAE**, a geospatial foundation model [8] with specialized positional embeddings for remote sensing data modalities. And **Scale-MAE**, a geospatial foundation model [28] with positional embeddings that are invariant to changes in the ground-sampling distance of remote sensing data. Both geospatial foundational models were pre-trained on optical remote sensing data [7].

Evaluation Settings We evaluate the quality of self-supervised domain adaptation with our method with different supervised downstream training settings on the target domain. These approaches offer different trade-offs between label-efficiency and computational cost ranging from standard **linear evaluation** to **fine-tuning**, which serve as our baselines.

SLR Linear: To evaluate the quality of representations from a foundation model after adaptation to the target domain with self-supervised SLR adapter training, we fix all model parameters and train only the (linear) task head. This incurs the same training cost as conventional linear probing.

SLR Scaling: The design of the SLR adapters facilitates a parameter-efficient fine-tuning method. Instead of re-training all parameters of the model, we fix the original foundation model, as well as the parameters of the low-rank matrices, after the self-supervised domain adaptation stage.

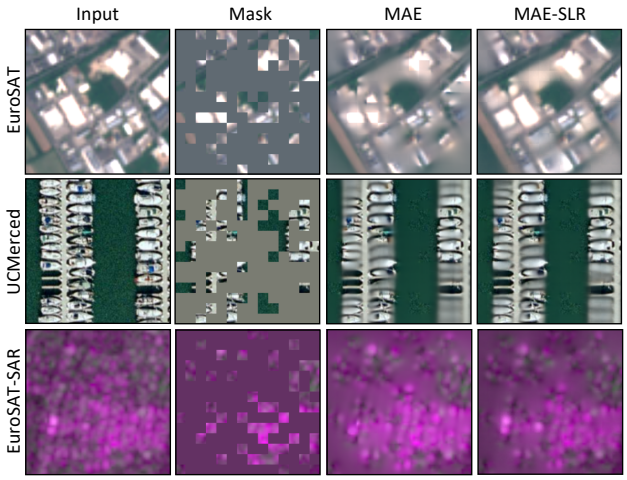


Figure 4. Masked autoencoder reconstruction examples. The first two columns show the original image and the masked version, respectively. The last two columns show the reconstructions from a masked autoencoder trained on ImageNet and after self-supervised domain adaptation with SLR adapters. Training with adapters reduces the reconstruction loss from 1.35 to 0.87 in the top row, 0.62 to 0.59 in the middle row, and 0.17 to 0.14 in the bottom row.

We then train all scaling vectors $\boldsymbol{s}_{\{i\}}^{1,2}$ and normalization parameters together with the task head.

SLR Fine-tuning: In this setting, all parameters of the SLR adapter and normalization parameters are trained along with the task head, while the parameters of the original foundation model remain fixed.

Self-supervised Domain Adaptation We evaluate the performance of our method when transferring visual foundation models to different new modalities (see Tab. 1). To that end, a set of SLR adapters for each dataset is added to the foundation models, as detailed in Section 3. Self-supervised training on the target dataset improves the models' reconstruction capacities and recovers data modality-specific details (see Fig. 4). In supervised downstream experiments, we find that SLR adaptation improves the resulting data representations across datasets and modalities (see Tab. 1). For the MAE foundation model, SLR adapters improve the average linear evaluation accuracy from 63.46% to 69.83% (+6.37% absolute improvement) across all datasets (see Fig. 1). Similarly, for SatMAE and Scale-MAE linear evaluation accuracy improves $+2.93\%$ and $+3.68\%$, respectively. In the fine-tuning setup, our method outperforms fine-tuning of the entire model on most dataset and foundation model combinations. On average, SLR adapter fine-tuning improves model accuracy by $+0.35\%$ for MAE, $+1.12\%$ for SatMAE and $+1.49\%$ for Scale-MAE over fine-tuning of the full model.

To evaluate the degree to which self-supervised training of SLR adapters captures the benefits that continual pre-training can provide [24], we perform continual pre-training of all model parameters for the RESISC45 and EuroSAT-SAR datasets. On both datasets, SLR adapters reach $>98\%$ of the accuracy achieved by full continual pre-training (see Sec. 9 in the Supplementary Material).

Method	Params	$k = 10$	$k = 100$
Linear Eval.	10k	75 ± 0.5	89 ± 0.5
SLR Linear	10k	74 ± 0.2	92 ± 0.5
SLR Scale	0.5M	87 ± 0.6	$\mathbf{96 \pm 0.1}$
SLR FT	7.3M	$\mathbf{88 \pm 2.0}$	$\mathbf{96 \pm 0.1}$
Fine-tune	304M	82 ± 2.0	95 ± 0.4

Table 3. Few-shot results with SatMAE on EuroSAT.

Method	Params	$k = 10$	$k = 100$
Linear Eval.	10k	63 ± 0.8	63 ± 0.2
SLR Linear	10k	71 ± 2.9	75 ± 0.1
SLR Scale	0.5M	$\mathbf{74 \pm 3.0}$	77 ± 0.3
SLR FT	7.3M	72 ± 3.0	$\mathbf{82 \pm 1.0}$
Fine-tune	303M	64 ± 1.6	77 ± 3.0

Table 4. Few-shot results with MAE on EuroSAT-SAR.

Few-shot Learning We investigate the label efficiency of our method with few-shot experiments on the EuroSAT (see Tab. 3) and EuroSAT-SAR (see Tab. 4) datasets. Fine-tuning of the SLR adapters outperforms full-model fine-tuning on both modalities, as well as for different labeled dataset sizes. In a k-shot experiment, we randomly select k samples for every class from the training set. With k=10, SLR adapter fine-tuning improves land-cover classification performance by $+6\%$ and $+8\%$ on EuroSAT and EuroSAT-SAR, respectively, compared with fine-tuning of the full model. To further reduce the number of trainable parameters, we also fix the low-rank matrices $\boldsymbol{W}_i^{1,2}$ after self-supervised adapter training (**SLR Scale**). In this setting, only the task head and the scaling parameters are trained with labeled data. This approach results in the best k=10 performance on EuroSAT-SAR over all tested approaches. It outperforms full fine-tuning on EuroSAT by $+5\%$ and by $+10\%$ on EuroSAT-SAR. Only 0.5M of the 303M parameters of the model ($\approx 0.2\%$) are optimized in this setting, which limits the risk of overfitting and improves label efficiency. Even at k=100 labeled samples per class, SLR scale outperforms full fine-tuning of the model on EuroSAT ($+1\%$) and achieves the same accuracy at lower variance across random seeds on EuroSAT-SAR ($77 \pm 0.3\%$).

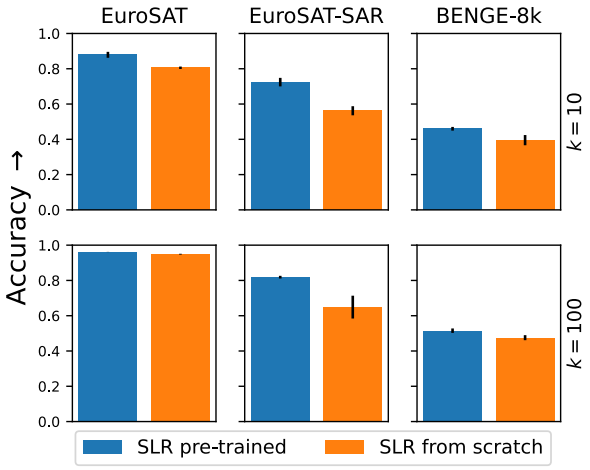


Figure 5. Classification accuracy of the MAE foundation model with SLR adapters. **SLR pre-trained**: Self-supervised pre-training of adapters on the target dataset. **SLR from scratch**: Random initialization of adapter parameters and supervised training on the target data.

Ablation on Self-supervised Adapter Training In this experiment, we investigate the value of self-supervised adapter training on the target domain (see Fig. 5). To that end, we compare the SLR adapters initialized by self-supervised pre-training with adapters trained from random initialization on the supervised target task. We find that self-supervised pre-training helps to stabilize the supervised training phase, facilitates faster convergence, and ultimately leads to higher performance on the target task. This advantage is more pronounced when the dataset for pre-training and the shift from source to target domain is larger. With k=10 labeled samples per class, self-supervised SLR training improves classification performance by $+6\%$ on BENGE-8k and by $+7\%$ on EuroSAT over randomly initialized adapters. On the EuroSAT-SAR dataset, we find an improvement of $+16\%$. For a fair evaluation, the same number of samples for self-supervised and supervised training of the SLR adapters is used in this experiment. However, our self-supervised adapter training method is not limited by the number of available labeled samples. It also has lower computational costs than supervised fine-tuning, as 75% of each input image is dropped for reconstruction and not processed by the model encoder.

Segmentation We fine-tune SLR adapters on the BENGE-8k Sentinel-1 SAR dataset with land-cover masks (see Fig. 6). For SatMAE, using SLR adapters improves performance from 66.57% to 68.45% with a frozen encoder ($+1.88\%$) and up to 70.63% when training the SLR adapters along with the segmentation head. This corresponds to a $+2.36\%$ improvement over fine-tuning of the full model.

Adapter Design We compare our SLR adapters with other parameter-efficient training methods: LoRA [17] adds low-rank matrices to the model, $(\text{IA})^3$ [20] re-scales intermediate activations, BitFit [48] only trains bias parameters, and NormTuning [9] modulates the normalization layers. When necessary, we slightly adapt the original methods to ensure a fair evaluation. LoRA matrices are added to the same linear transforms where we place our SLR adapters (i.e., to both, the MSA and MLP blocks of the transformer). $(\text{IA})^3$ activation scaling is applied to the query, key and value projections of the foundation model. The results indicate that SLR adapters achieve the best performance when adapting a visual foundation model from RGB to the SAR data modality (see Tab. 5). Additionally, SLR is designed to combine domain adaptation and few-shot capabilities through its scaling parameters.

Method	Accuracy
BitFit [48]	80.34 ± 2.4
$(\text{IA})^3$ [20]	76.72 ± 3.5
Norm tuning [9]	79.00 ± 3.1
LoRA [17]	85.86 ± 0.3
SLR (ours)	$\mathbf{87.14 \pm 0.1}$

Table 5. Performance of different parameter efficient fine-tuning methods when adapting an ImageNet MAE to SAR data (EuroSAT-SAR).

Parameter Efficiency In our experiments, we use ViT-L encoder models, which consist of 303M parameters in the standard setting. For linear evaluation, we add a linear classifier, introducing $D \cdot \mathrm{n}_c + n_c$ additional parameters, where D is the model's embedding dimension and n_c is the number of classes for the task at hand. Our SLR adapters introduce $2 \cdot D \cdot r + 2 \cdot D$ parameters for each linear projection in the model. For ViT-L, we add 194 adapters throughout the model. Using different bottleneck values $r \in \{8, 16\}$ based on the target data modality, gives an additional 3.9M or 7.1M parameters in total. This corresponds to $\approx$1% and $\approx$2% of the model's total parameters, respectively. Using SLR adapters thus drastically reduces the storage requirements when training models for multiple data modalities compared to individual fine-tuning of the full model for each data modality. In the linear evaluation setting, the SLR adapters significantly improve performance across data modalities, while introducing negligible computational overhead for training and inference. In our experiments, we find no significant difference in training time between standard linear evaluation and SLR linear evaluation (using a single NVIDIA Tesla V100 GPU).

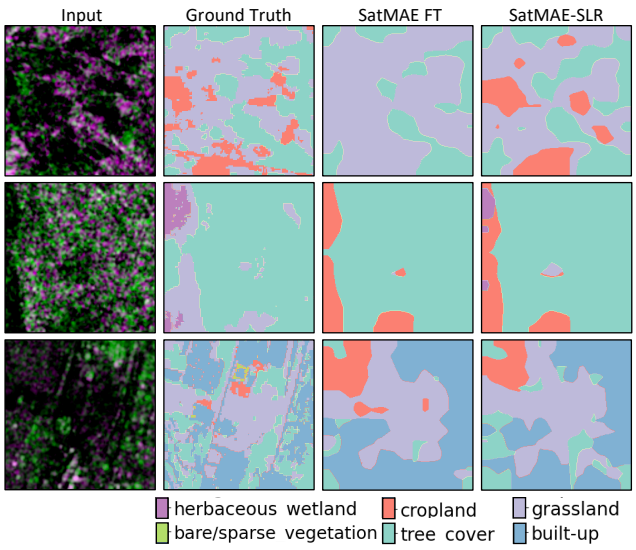


Figure 6. Segmentation examples on Sentinel-1 SAR imagery from the BENGE-8k dataset. Predictions for 8 different land-cover classes obtained with SatMAE combined with a fully convolutional segmentation head. Comparison between full SatMAE fine-tuning and fine-tuning of SLR adapters.

5. Discussion

We investigate the problem of adapting visual foundation models to remote sensing downstream tasks involving different data modalities. The large number of different heterogeneous data modalities in the remote sensing domain makes it difficult to design foundation models that can readily adapt to any of them. Furthermore, domain adaptation might also be necessary within a single data modality to accommodate different sensor types, due different spatial or spectral resolutions and other effects. Recently, a number of data modality-specific foundation models have been presented. In this work, we take an orthogonal approach and propose to explicitly adapt an existing foundation model for the data modality of interest with a self-supervised reconstruction objective. To alleviate the computational cost and to enable supervised training on small target datasets, we reduce the trainable parts of the model to a fraction of its total parameter count for the adaptation process. We focus on ViT foundation models trained with masked-autoencoding, but the proposed approach is generally applicable to neural networks with dense layers and not specific to geospatial or computer vision data modalities. For example, our method could easily be applied in other domains with multi-modal data, as, for instance, encountered in autonomous driving or robotics scenarios.

Parameter efficient training methods make it possible to train and store large numbers of modality-specific or task-specific models that are derived from the same foundation model with small memory footprint. More generally, utilizing exchangable adapters to introduce specialized knowledge into a general foundation is a first step towards a more modular deep learning framework where models for specific problems are created by combining individual pre-trained building blocks. Such a framework would be well suited for the geospatial computer vision domain, where training independent foundation models for each data modality of interest would incur huge computational cost.

Broader Impact and Limitations Advances in geospatial foundation models can improve our understanding of geophysical variables and improve estimates of socioeconomic indicators. This contributes to fields such as environmental sciences or public policy. In particular, applications where little labeled data is available stand to benefit from these developments. As our abilities to collect and analyse remote sensing data improve, we need to be mindful of implications on surveillance technology and individual privacy rights. This work focuses on static remote sensing data without direct observations of people or their individual activities. The proposed method is limited by public access to a pre-trained foundation model and performance might degrade when the difference between source and target data distribution gets very large (e.g., adaptation of a vision foundation model to audio data).

6. Conclusion

Geospatial foundation models promise to simplify the analysis of remote sensing imagery by providing a strong, task agnostic starting point for building specialized deep learning models. We still face significant challenges when applying foundation models on data modalities that were not seen during the pre-training stage. The standard solution to this problem, i.e., supervised fine-tuning on the target task, incurs high computational cost and fails altogether on small datasets. In this work, we show that self-supervised training of a small number of additional adapter parameters suffices to adapt foundation models to new remote sensing data modalities. This provides a resource-efficient way to apply existing large visual models on new remote sensing tasks with small labeled datasets, or in settings where computational constraints prevent fine-tuning of the full model. The presented method represents a memory-efficient improvement over fine-tuning of the full model. We demonstrate improved performance across different data modalities and target tasks and strongly outperform existing approaches in few-shot learning scenarios. We believe that these results will be also valuable beyond the analysis of remote sensing data in any setting where visual foundation models are applied across different data modalities.

References

[1] Steve Ahlswede, Christian Schulz, Christiano Gava, Patrick Helber, Benjamin Bischke, Michael Förster, Florencia Arias, Jörn Hees, Begüm Demir, and Birgit Kleinschmit. TreeSatAI Benchmark Archive: A Multi-sensor, Multi-label Dataset for Tree Species Classification in Remote Sensing. *Earth System Science Data Discussions*, 2022:1–22, 2022. 4, 5, 1

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[3] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 3

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the Opportunities and Risks of Foundation Models. *arXiv preprint arXiv:2108.07258*, 2021. 2, 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2

[6] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 4, 5, 1

[7] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional Map of the World. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018. 5

[8] 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. 2, 3, 5

[9] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. Modulating Early Visual Processing by Language. *Advances in Neural Information Processing Systems*, 30, 2017. 7

[10] 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. In *International Conference on Learning Representations*, 2020. 3

[11] Ferran Gascon, Catherine Bouzinac, Olivier Thépaut, Mathieu Jung, Benjamin Francesconi, Jérôme Louis, Vincent Lonjou, Bruno Lafrance, Stéphane Massera, Angélique Gaudel-Vacaresse, et al. Copernicus Sentinel-2A calibration and products validation status. *Remote Sensing*, 9(6):584, 2017. 1

[12] Harm Greidanus and Naouma Kourti. Findings of the DECLIMS Project—Detection and Classification of Marine Traffic from Space. *Proceedings of SEASAR 2006*, pages 23–26, 2006. 1

[13] 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, 5

[14] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient Model Adaptation for Vision Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 817–825, 2023. 3

[15] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 4, 5, 1

[16] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient Transfer Learning for NLP. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 3

[17] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. 3, 4, 7

[18] Karen E Joyce, Stella E Belliss, Sergey V Samsonov, Stephen J McNeill, and Phil J Glassey. A Review of the Status of Satellite Remote Sensing and Image Processing Techniques for Mapping Natural Hazards and Disasters. *Progress in Physical Geography*, 33(2):183–207, 2009. 1

[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. *arXiv preprint arXiv:2304.02643*, 2023. 2

[20] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot Parameter-efficient Fine-tuning is Better and Cheaper than In-context Learning. *Advances in Neural Information Processing Systems*, 35:1950–1965, 2022. 3, 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2018. 5

[22] 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

[23] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal Contrast: Unsupervised Pre-training from Uncurated Remote Sensing Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9414–9423, 2021. 2, 3

[24] 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

[25] Michael Mommert, Nicolas Kesseli, Joëlle Hanna, Linus Scheibenreif, Damian Borth, and Begüm Demir. Ben-Ge: Extending BigEarthNet with Geographical and Environmental Data. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 1016–1019. IEEE, 2023. 5, 1

[26] Maxim Neumann, Andre Susano Pinto, Xiaohua Zhai, and Neil Houlsby. In-domain Representation Learning for Remote Sensing. *arXiv preprint arXiv:1911.06721*, 2019. 1

[27] Omiros Pantazis, Gabriel J Brostow, Kate E Jones, and Oisin Mac Aodha. Focus on the Positives: Self-supervised Learning for Biodiversity Monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10583–10592, 2021. 1, 3

[28] Colorado J Reed, Xiangyu Yue, Ani Nrusimha, Sayna Ebrahimi, Vivek Vijaykumar, Richard Mao, Bo Li, Shanghang Zhang, Devin Guillory, Sean Metzger, et al. Self-supervised Pretraining Improves Self-supervised Pretraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2584–2594, 2022. 2, 3, 5

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 2, 5

[30] Linus Scheibenreif, Joëlle Hanna, Michael Mommert, and Damian Borth. Self-supervised Vision Transformers for Land-cover Segmentation and Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1422–1431, 2022. 2, 3

[31] Linus Scheibenreif, Michael Mommert, and Damian Borth. Toward Global Estimation of Ground-Level NO2 Pollution with Deep Learning and Remote Sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022. 1, 3

[32] Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked Vision Transformers for Hyperspectral Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2165–2175, 2023. 3

[33] Shuchang Shen, Sachith Seneviratne, Xinye Wanyan, and Michael Kirley. FireRisk: A Remote Sensing Dataset for Fire Risk Assessment with Benchmarks Using Supervised and Self-supervised Learning. *arXiv preprint arXiv:2303.07035*, 2023. 4, 5, 1

[34] Rajendra P Sishodia, Ram L Ray, and Sudhir K Singh. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing*, 12(19):3136, 2020. 1

[35] Adam J Stewart, Caleb Robinson, Isaac A Corley, Anthony Ortiz, Juan M Lavista Ferres, and Arindam Banerjee. Torchgeo: Deep Learning with Geospatial Data. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–12, 2022. 5

[36] Gencer Sumbul, Marcela Charfuelan, Begüm Demir, and Volker Markl. BigEarthNet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904. IEEE, 2019. 1

[37] Gencer Sumbul, Arne De Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mario Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large-scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):174–180, 2021. 1

[38] Charles Toth and Grzegorz Jóźków. Remote Sensing Platforms and Sensors: A Survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:22–36, 2016. 2, 3

[39] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, 4(2):41–57, 2016. 3

[40] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank van Langevelde, Tilo Burghardt, et al. Perspectives in Machine Learning for Wildlife Conservation. *Nature Communications*, 13(1):792, 2022. 3

[41] Devis Tuia, Konrad Schindler, Begüm Demir, Gustau Camps-Valls, Xiao Xiang Zhu, Mrinalini Kochupillai, Sašo Džeroski, Jan N van Rijn, Holger H Hoos, Fabio Del Frate, et al. Artificial Intelligence to Advance Earth Observation: A Perspective. *arXiv preprint arXiv:2305.08413*, 2023. 1, 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[43] Yi Wang, Hugo Hernández Hernández, Conrad M Albrecht, and Xiao Xiang Zhu. Feature Guided Masked Autoencoder for Self-supervised Learning in Remote Sensing. *arXiv preprint arXiv:2310.18653*, 2023. 3, 5, 1

[44] Michael A Wulder, David P Roy, Volker C Radeloff, Thomas R Loveland, Martha C Anderson, David M Johnson, Sean Healey, Zhe Zhu, Theodore A Scambos, Nima Pahlevan, et al. Fifty Years of Landsat Science and Impacts. *Remote Sensing of Environment*, 280:113195, 2022. 1

[45] Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. 1, 3

[46] Yi Yang and Shawn Newsam. Bag-of-visual-words and Spatial Extensions for Land-use Classification. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 270–279, 2010. 4, 5, 1

[47] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new Foundation Model for Computer Vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[48] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1–9, 2022. 3, 7

[49] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 1