This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SMARTIES: Spectrum-Aware Multi-Sensor Auto-Encoder for Remote Sensing Images

Gencer Sumbul Chang Xu Emanuele Dalsasso Devis Tuia
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Abstract

From optical sensors to microwave radars, leveraging the complementary strengths of remote sensing (RS) sensors is crucial for achieving dense spatio-temporal monitoring of our planet. In contrast, recent deep learning models, whether task-specific or foundational, are often specific to single sensors or to fixed combinations: adapting such models to different sensory inputs requires both architectural changes and re-training, limiting scalability and generalization across multiple RS sensors. On the contrary, a single model able to modulate its feature representations to accept diverse sensors as input would pave the way to agile and flexible multi-sensor RS data processing. To address this, we introduce SMARTIES, a generic and versatile foundation model lifting sensor-specific/dependent efforts and enabling scalability and generalization to diverse RS sensors: SMARTIES projects data from heterogeneous sensors into a shared spectrum-aware space, enabling the use of arbitrary combinations of bands both for training and inference. To obtain sensor-agnostic representations, we train a single, unified transformer model reconstructing masked multi-sensor data with cross-sensor token mixup. On both single- and multi-modal tasks across diverse sensors, SMARTIES outperforms previous models that rely on sensor-specific pretraining. Our code and pretrained models are available at `https://gsumbul.github.io/SMARTIES`.

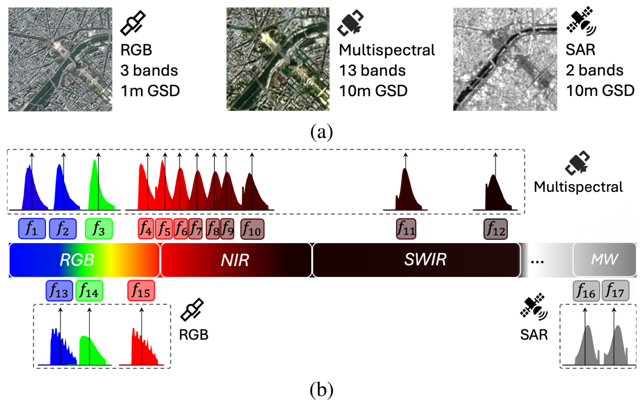

Figure 1. (a) An example of RGB, multispectral, and SAR images representative of the different spectral and spatial properties of RS sensors. (b) The spectral bands' histograms for each sensor are shown as probability density function estimations, aligned with the corresponding wavelength range in the electromagnetic spectrum (shown in log scale). SMARTIES leverages different projection layers $\{f_1, f_2, \ldots, f_{17}\}$ for different spectral ranges that allow a single, unified model independent from sensors specificities.

1. Introduction

Every day, a vast number of airborne and spaceborne sensors generate tens of terabytes of remote sensing (RS) data, empowering Earth observation [7] (Fig. 1a). Unlike RGB natural images, RS data spans a wide spectrum of wavelengths captured by diverse sensors from RGB visible light, through infrared frequencies and all the way to Microwaves (Fig. 1b). RS sensors capture electromagnetic radiation at different frequencies: optical sensors capture (very) high-resolution images mostly in the visible and infrared spectrum, offering rich semantic details, but limited by clouds and daylight; synthetic aperture radars (SAR) are active sensors that can acquire images day and night and independently of weather conditions; thermal sensors measure the energy emitted directly by objects at the surface and can be used to estimate surface temperature. The complementary strengths of these sensors hold the potential for continuous, all-day, and all-weather Earth observation, supporting a wide range of applications in various fields, *e.g.*, agriculture, climate, hydrology, urban planning [14, 42]. Despite the wealth of RS sensors, a significant barrier remains: the lack of unified image representations for multi-modal processing of RS data. A significant obstacle to achieve this goal originates from the highly heterogeneous characteristics of RS sensors, in terms of spectral range, radiometric resolution, and spatial resolution: such diversity forced previous attempts to design restrictive sensor-specific models [9, 12, 19, 23, 32]. To mitigate this issue, several foundation models (FMs) have been proposed by designing sensor-specific backbones, leading to an increase in computational complexity [12, 15, 46]. Adding new sensors at pretraining and finetuning would require modifying the architecture with extra backbones, leading to further computational overhead, and thus limiting the *scalability* of such models. Moreover, models trained on a fixed combination of sensors will develop biases towards them, suffering from limited *generalization* to unseen sensors.

To address these challenges, we propose a sensor-agnostic FM named **S**pectrum-Aware **M**ulti-Sensor **A**uto-Encoder for **R**emo**t**e Sensing **I**mag**es** (**SMARTIES**) that breaks the representation barriers between sensors[1] and enables downstream applications using a single, unified model across diverse sensors, including unseen sensor transfer capabilities. To train a single model on heterogeneous sensors efficiently, we unify sensor representations by projecting data into a shared and divisible space called the spectrum-aware space. This concept is based on the observation that, despite the varying spectral ranges, all the different sensors capture subsets of the full electromagnetic spectrum with well-defined physical properties: an example of three typical RS sensors and of the spectral ranges of their bands is shown in Fig. 1. Given a specific sensor, each one of its bands is projected into the spectrum-aware space by projection layers specific to wavelength ranges (f in Fig. 1). By representing data with this shared space, we eliminate the need to train a separate model for each sensor. Moreover, SMARTIES can generalize to unseen sensors during inference by interpolating the learned projection layers to represent the unseen wavelength ranges when full finetuning is costly to achieve. Leveraging this unified representations, we pretrain a transformer model with a self-supervised objective: we learn with cross-sensor token mixup on paired multi-sensor data and enforce the model to reconstruct randomly masked regions of the representations in the spectrum-aware space. Learning this way, SMARTIES maximizes synergies among diverse sensory inputs, scaling efficiently to large training datasets with multiple sensors and enhancing more generalizable representations.

We perform experiments on 10 datasets composed of various combinations of sensors, all using the same pretrained backbone. Results demonstrate the superior performance of SMARTIES on both single- and multi-modal tasks across diverse sensors, as well as generalization to new unseen sensors during inference. SMARTIES contributes significantly to literature since: (1) SMARTIES is a single, unified FM without sensor-specific pretraining or backbones that can seamlessly tackle diverse sensory inputs (both single- and multi-modal). (2) SMARTIES exhibits *scalability* to diverse sensors with a high pretraining efficiency. Using a ViT encoder, SMARTIES is pretrained with only 500K multispectral, SAR and submeter RGB images and with as little as 300 epochs, showing a smaller computational cost than most RS FMs. (3) SMARTIES transfers not only to different downstream applications, but also to new sensors that were not present during pretraining, demonstrating unprecedented *generalization* capabilities.

[1]We focus only on amplitude-related phenomenology for RS sensors excluding, for instance, phase information recorded by SAR instruments.

2. Related Work

Self-supervised learning (SSL) in RS has been widely used to learn general data representations that can be transferred to various downstream tasks [22, 37, 39, 44]. It allows to reduce the demand of task-specific models and of labeled data, which are often scarce in RS. Among different SSL approaches, masked image modelling through masked autoencoders (MAEs) has recently received increasing attention due to its ability to scale to larger models together with the increasing amount of training data [17, 24].

Single-modal MAEs have demonstrated to learn rich task-transferable representations through a SSL strategy: reconstructing masked parts of images. Several MAE-based FMs have been proposed in RS: SatMAE [9] embeds temporal and spectral information as positional encodings during reconstruction, while SpectralGPT [19] and S2MAE [23] model the spatial-spectral data as 3D cubes and enforce the masked reconstruction in the 3D space. Scale-MAE [34] encodes different spatial scales in positional encodings to learn robust representations across resolutions. Despite the task-agnostic strengths, these models trained on a specific image modality (e.g., multispectral data) struggle to handle data in others (e.g., SAR data), hampering their usability and flexibility for multi-modal processing of RS data.

Multi-modal MAEs extend the MAE framework to develop generalizable representations across different modalities. For example, recent computer vision research focuses on scaling MAE models to accommodate as many image modalities as possible by reconstructing masked multi-modal tokens [3, 30]. In RS, multi-modal FMs also benefit from MAE-based training [12, 15]. We can identify two groups of models: first, models that learn dual (e.g., CROMA [12]) or triple (e.g., SkySense [15]) modal representations with aligned optical and radar sensors; second, models that learn shared features across multiple modalities by blending them into massive training data [31, 46]. The first group of models mostly rely on predefined architectural designs for a set of sensors with sensor-dependent encoders, leading to an increase in computational complexity, and limited generalization towards diverse sensors at pretraining and downstream transfer. Although the second group of models can significantly improve generalization across diverse sensors, they need computationally demanding and complex adjustments to MAEs (e.g., hypernetwork for dynamic weight generation in DOFA [46], sensor encodings and channel-specific tokens in SenPa-MAE [33]). In addition, they require massive pretraining sets for learning rep-

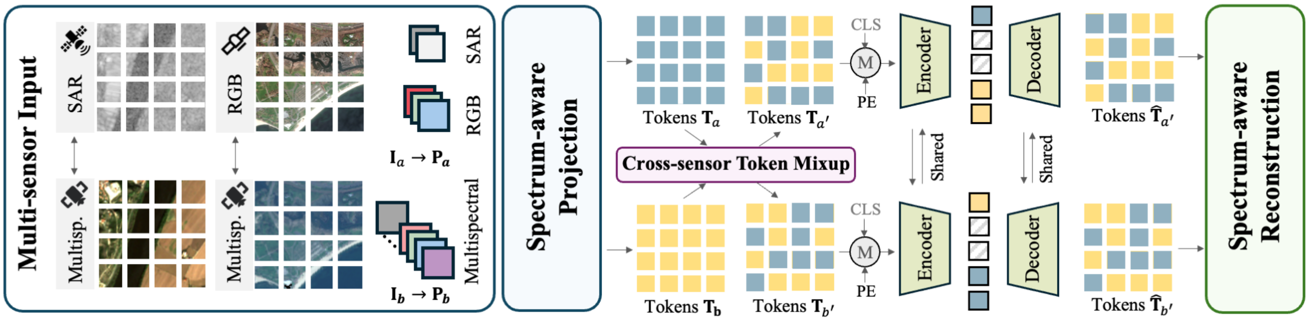

Figure 2. SMARTIES lifts sensor-dependent efforts for multi-sensor RS image representation learning by leveraging: (1) spectrum-aware RS image projection; (2) cross-sensor token mixup; and (3) spectrum-aware RS image reconstruction. PE and Multisp. denote positional encoding and multispectral, respectively.

resentations across diverse sensors (8M images for DOFA), showing lower data efficiency, and thus limited scalability.

To fill these gaps, we propose a unified and versatile model that can seamlessly handle diverse sensors during pretraining. Our model can not only transfer to different downstream tasks on sensors seen during pretraining, but also generalize to new unseen sensors when full finetuning is too expensive.

3. SMARTIES

To leverage the multi-sensor nature of RS, we develop a FM named Spectrum-Aware Multi-Sensor Auto-Encoder for Remote Sensing Images (SMARTIES) that learns image representations transferable to diverse sensors. SMARTIES is designed in a *generic* way (Fig. 2), so that it can accommodate variations in sensor characteristics, *data-efficient* for *scalable* pretraining, and also conceptually *simple* for downstream applications. We realize these properties via the following design decisions:

1. *Spectrum-aware RS Image Projection* (Sec. 3.1): To deal with the variation of spectra in different sensors, we learn spectrum-aware projection layers to tokenize the RS images. These layers depend on the wavelengths of the bands, spanning the continuum of the electromagnetic spectrum covered by RS sensors (Fig. 1 and Fig. 3).
2. *Cross-sensor Token Mixup* (Sec. 3.2): To mitigate the bias specific to sensors or spectrum combinations, we use pairs of aligned images from different sensors as input, and then exchange their tokens. This leads to more scalable and generalizable models, easing pretraining and downstream transfer, respectively.
3. *Spectrum-aware RS Image Reconstruction* (Sec. 3.3): We feed the cross-sensor mixed embeddings into a standard encoder-decoder based transformer, which is easy to deploy for downstream applications. We reproject the decoded images back to the original spectral channels of the RS sensors through spectrum-aware reprojection layers (Fig. 3). Spatial and spectral reasoning is employed through masked image modelling with SSL.
4. *Downstream Transfer to Diverse Sensors* (Sec. 3.4): Thanks to the spectrum-aware image projection and reconstruction, the resulting encoder can generalize to diverse sensors by using either the existing projection layers or adapting them for unseen sensors by interpolation.

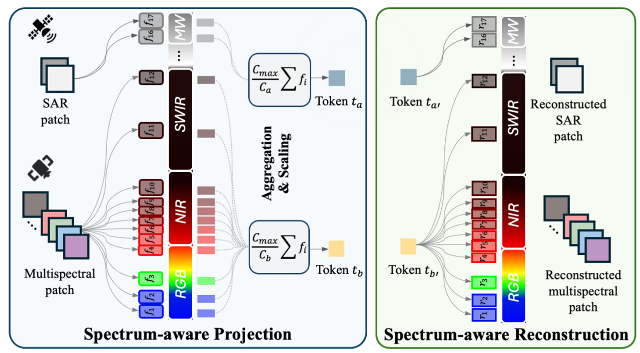

Figure 3. Spectrum-aware RS image projection and reconstruction illustrated on a pair of SAR and multispectral patches.

3.1. Spectrum-aware RS Image Projection

SMARTIES uses different projection layers for different spectral ranges, each one defined by the minimum and maximum wavelengths covered by the corresponding band. These ranges can cover various parts of the electromagnetic spectrum (e.g., visible light, near-infrared, short wave infrared, microwaves etc.). This strategy makes each projection layer learn an embedding specific to a certain spectral range, enforcing physical consistency among embeddings of different sensors covering similar spectral ranges.

By following the official instrument specifications of widely used RS sensors, we define a set of spectrum-aware projection layers $\mathcal{F} = \{f_1, f_2, ..., f_n\}$, where f_i is the mapping of the ith spectral range via a fully-connected layer. Specifically, we associate the spectral range of each band used in pretraining with a projection layer f_i: f_1-f_{12} for

Sentinel-2 (S2), f_{13}-f_{15} for RGB images of Maxar, and f_{16}-f_{17} for Sentinel-1. For instance, f_2 corresponds to the wavelength range 427nm-558nm of Sentinel-2[2]. For bands from different sensors capturing the same band, for instance "red" light, but with different frequency ranges (e.g., S2 vs. Maxar), we keep separate projection layers. For a given RS image $\mathbf{I}_a \in \mathbb{R}^{W_a \times H_a \times C_a}$ (C_a is the number of spectral bands), we first resize it into the input size $W \times H$ of the model, and then divide it into a sequence of N_P non-overlapping patches $\mathbf{P}_a \in \mathbb{R}^{N_W \times N_H \times S^2 C_a}$, where S is the patch size, $N_W = W/S$, $N_H = H/S$ and $N_P = N_W N_H$. Each patch p_a of the image $\mathbf{I}_a$ is then projected to the joint space using the projection layers $\{f_i\}$ in $\mathcal{F}$ corresponding to its bands, where $f_i : \mathbb{R}^{S \times S} \to \mathbb{R}^D$ (D is the size of spectrum-aware embeddings). This leads to C_a embeddings (one per band). For token t_a of this patch, all C_a embeddings are first averaged, and then scaled up by $C_{\text{max}} = 12$, which is the largest number of spectral bands encountered during pretraining. This last operation prevents imbalance between different sensors due to the different number of bands. The whole process is illustrated in the left panel of Fig. 3. We note that thanks to this projection strategy, adding new RS sensors to pretraining simply requires adding new f_is for the additional spectral ranges.

3.2. Cross-sensor Token Mixup

Thanks to the spectrum-aware RS image projection, SMARTIES can operate on RS images acquired by different sensors with a unified approach. However, this would lead to bias towards specific sensors or bands combinations that are more present during pretraining. To alleviate this, we perform cross-sensor token mixup: (1) we first take as input a pair of images acquired by different sensors on the same area, and then (2) we exchange tokens across the images of a pair using mixup, as shown in Fig. 2. This prevents encoding bias towards specific spectral combinations and also enhances the generalization capability of SMARTIES over multiple sensors. For a multi-sensor image pair $(\mathbf{I}_a, \mathbf{I}_b)$, with tokens $\mathbf{T}_a \in \mathbb{R}^{N_W \times N_H \times D}$ and $\mathbf{T}_b \in \mathbb{R}^{N_W \times N_H \times D}$, we define the mixed image tokens $\mathbf{T}_{a'}$ as:

$$\mathbf{T}_{a'} = \mathcal{M} \odot \mathbf{T}_a + (1 - \mathcal{M}) \odot \mathbf{T}_b \tag{1}$$

where $\mathcal{M} \in \mathbb{R}^{N_W \times N_H \times D}$ is a randomly generated binary mask broadcasted along the third dimension. We also perform a mirrored mixup to obtain the mirrored version $\mathbf{T}_{b'}$ of $\mathbf{T}_{a'}$ to avoid losing information during the mixup process:

$$\mathbf{T}_{b'} = (1 - \mathcal{M}) \odot \mathbf{T}_a + \mathcal{M} \odot \mathbf{T}_b. \tag{2}$$

3.3. Spectrum-aware RS Image Reconstruction

For spatial and spectral reasoning, we use a standard encoder-decoder transformer architecture. Given the lack of labels and in order to remain scalable, we employ self-supervised masked image modelling with spectrum-aware reconstruction. As in the vanilla MAE, we follow the **masking**, **encoding**, **decoding**, and **reconstruction** steps. We would like to remind that the encoder and decoder of SMARTIES operate independently from the number of sensors seen during pretraining and also that we do not use sensor-specific encoders/decoders, since data from the different sensors are already projected in the common spectrum-aware space. This allows us to maintain a similar computational complexity to the vanilla MAE.

Masking. Random masking is applied on both $\mathbf{T}_{a'}$ and $\mathbf{T}_{b'}$ with a ratio R. The remaining unmasked tokens are visible tokens: $\mathbf{T}_{a'}^{vis}$ and $\mathbf{T}_{b'}^{vis}$, both of size $(1 - R)N_W N_H \times D$. They are fed into the encoder.

Encoding. We use a ViT architecture (e.g., ViT-B, ViT-L etc.), which processes tokens from $\mathbf{T}_{a'}^{vis}$ and $\mathbf{T}_{b'}^{vis}$ together with the special [CLS] token, and provides latent image representations for all of them. To encode the relative positioning of tokens, we use sinusoidal positional encodings (PE).

Decoding. The decoder receives a hybrid input consisting of the unmasked tokens and special [MASK] tokens leveraged by the positional encodings. The learnable [MASK] token, along with the positional encodings, is exploited to decode the masked patches at specific locations.

Reconstruction. After the decoding phase, we reproject (i.e., reconstruct) the decoded image representations of the pair back to their original spectral channels, leading to reconstructed patches $\hat{\mathbf{P}}_a \in \mathbb{R}^{N_W \times N_H \times S^2 C_a}$ and $\hat{\mathbf{P}}_b \in \mathbb{R}^{N_W \times N_H \times S^2 C_b}$. Similar to the spectrum-aware projection, the spectrum-aware reconstruction uses different reprojection layers for different spectral ranges (i.e., bands). To this end, we first define a set of fully-connected reprojection layers $\mathcal{R} = \{r_1, r_2, ..., r_n\}$, where $r_i : \mathbb{R}^D \to \mathbb{R}^{S \times S}$ is the remapping function for the ith spectral range. There is one reprojection layer r_i corresponding to each projection layer f_i. For the reconstruction of patch p_a from the corresponding decoded token t_a, we use t_a for each band of the patch, and apply the corresponding reprojection layers. The overall process is illustrated in the right panel of Fig. 3. Once this is applied for all the decoded tokens, the reconstructed masked patches $\hat{\mathbf{P}}_{a'}^{mask}$, $\hat{\mathbf{P}}_{b'}^{mask}$ are obtained.

To train our model in a self-supervised way, we use the Mean Squared Error (MSE) loss between the original masked patches $\mathbf{P}_{a'}^{mask}$ and the corresponding reconstructed ones $\hat{\mathbf{P}}_{a'}^{mask}$. For $\mathbf{I}_{a'}$, the reconstruction loss is:

$$\mathcal{L}_{a'} = \frac{\sum(\mathbf{P}_{a'}^{mask} - \hat{\mathbf{P}}_{a'}^{mask})^2}{R N_W N_H} \tag{3}$$

where the denominator denotes the number of masked tokens. The final reconstruction loss is computed over both mixed patches sets: $\mathcal{L} = \mathcal{L}_{a'} + \mathcal{L}_{b'}$.

[2]The detailed ranges of projection layers are provided in the Sec. S1.1 of the supplementary material.

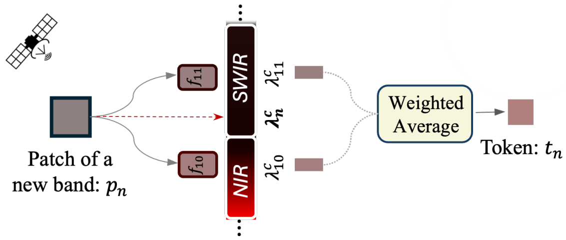

Figure 4. An example of downstream transfer to an unseen spectral band through interpolation. λ^c_{10} and λ^c_{11} denote the centre wavelength of the NIR and SWIR bands seen during pretraining; λ^c_n denotes the centre wavelength of a new, unseen spectral band.

3.4. Downstream Transfer to Diverse Sensors

After SMARTIES has been pretrained, the learned encoder and spectrum-aware projection layers can be used for various downstream tasks (*e.g.*, classification, segmentation) with RS images from diverse sensors. For downstream transfer, there can be three main inference modes: (1) *in-domain, seen sensor inference*, (2) *open-domain, seen sensor inference*, and (3) *open-domain, unseen sensor inference* (cf. Fig. S1 in the supplementary material). For modes (1) and (2), since the sensors for the downstream task have been already seen during pretraining, one needs to select the relevant projection layers among the already learned ones for tokenization. On the contrary, for (3) the downstream transfer uses a sensor unseen during pretraining. This can be achieved in two ways. First, projection layers for the missing ranges can be easily learned during downstream transfer through full finetuning. However, this might not always be feasible, especially when full finetuning is costly to achieve. For such cases, as a second way, we apply *interpolation* to unseen spectral ranges via a weighted average of the result of the closest projection layers. This is illustrated on a synthetic example in Fig. 4, where an unseen sensor has a new band with a different spectral range than all sensors used during pretraining. Since the central wavelength λ^c_n of this new band falls between those of learnt layers λ^c_{10} and λ^c_{11}, its token can be obtained by combining the corresponding projection layers f_{10} and f_{11}, weighted by the normalized distances between the central wavelengths[3]. However, we stress that this can: 1) only operate on the unseen ranges falling inside the minimum and maximum frequencies considered for pretraining; 2) not work for unseen regions out of the pretraining spectra (i.e., *extrapolation*).

4. Experiments

4.1. Pretraining Data

To pretrain SMARTIES, we use paired images from: (1) the Functional Map of the World RGB dataset (fMoW-RGB) [8] and its Sentinel-2 counterpart fMoW-S2 [9]; and (2) the BigEarthNet-MM [38] dataset. To reduce the number of pretraining samples, we randomly selected 60K non-temporal fMoW pairs and 188K non-temporal BigEarthNet pairs, constituting a pretraining set of 496K images in total. This pretraining set is significantly smaller than those used in recent RS models (see Sec. S2 in the supplementary material for a detailed comparison).

fMoW pairs represent various land use classes worldwide. fMoW-RGB [8] includes 470K submeter resolution RGB images of Maxar. fMoW-S2 [9] includes over 882K Sentinel-2 images, containing 13 spectral bands with varying spatial resolutions for different bands (10m, 20m and 60m). The locations and temporal stamps of fMoW-S2 are mostly aligned with those of fMoW-RGB.

BigEarthNet-MM (BEN) [38] includes over 590K pairs of multispectral and SAR images; each pair is acquired by Sentinel-2 and Sentinel-1 satellites on the same geographical area and associated with multi-labels. BEN-S2 images contain 12 spectral bands with 10m, 20m and 60m spatial resolutions, BEN-S1 images include dual-polarized information bands (VV and VH) with 10m spatial resolution.

Preprocessing for Data Harmonization is achieved by first min-max image normalization with 1% and 99% percentile values, and then image standardization with mean and standard deviation values. This allows SMARTIES to be robust towards data distribution differences across multiple sensors (e.g., long-tailed distribution of 12 bit Sentinel-2 images *vs.* short-tailed distribution of 8 bit RGB images).

4.2. Experimental Setup

To pretrain SMARTIES, we follow the same architectural choices and hyperparameters with the vanilla MAE [17], wherever possible. In detail, we pretrain two models with ViT-B and ViT-L [11] backbones for 300 epochs, using the AdamW optimizer [25] with the batch size of 2048 (distributed over 8 A100 GPUs) and the base learning rate of 1.5e-4. The masking ratio R and the input size $W \times H$ are set to 75% and 224×224, respectively, as in the vanilla MAE. The mixup ratio is set to 50%, while the size D of spectrum-aware embeddings is set to 768 and 1024 for ViT-B and ViT-L, respectively (see Sec. S1.2 in the supplementary material for the details of pretraining). To assess the effectiveness of SMARTIES across different sensors, tasks, and scenarios, we consider the following experiments:

- **Multispectral, Radar, and RGB Experiments.** We evaluate the performance of SMARTIES with various single-modal/multi-modal input (Multispectral, Radar, and RGB) to investigate its robustness across different sensors. More specifically, experiments under this group belong to the following two categories: *in-domain, seen sensor* transfer, where performance is assessed on

[3]Projection layer indices refer to Fig. 1b.

Method	Backbone	BEN 10% S1 (LP)	S2 (FT)	MM (LP)
SeCo [26]	RN-50	69.9	**82.6**	76.9
GASSL [2]	RN-50	66.1	80.2	73.2
CACo [27]	RN-50	**70.1**	81.3	**78.5**
SatMAE (S2) [9]	ViT-B	68.4	85.9	77.8
GFM [29]	Swin-B	73.6	86.3	82.0
SatLas (S2) [4]	Swin-B	60.8	82.8	70.1
I-JEPA [1]	ViT-B	N/A	85.9	N/A
SpectralGPT [19]	ViT-B	57.1	85.6	68.5
S2MAE [23]	ViT-B	N/A	85.6	N/A
msGFM [16]	Swin-B	67.5*	86.8	N/A
SMARTIES (Ours)	ViT-B	**78.9**	**86.9**	**85.4**
SatMAE (S2) [9]	ViT-L	67.4	82.1	77.6
CROMA [12]	ViT-B (×2)	79.8	87.6	85.2
SpectralGPT [19]	ViT-L	N/A	86.9	N/A
S2MAE [23]	ViT-L	N/A	86.5	N/A
SatMAE++ (S2) [32]	ViT-L	67.6	85.1	78.1
SMARTIES (Ours)	ViT-L	**80.5**	**87.7**	**86.7**

Table 1. BEN multi-label scene classification results (mAP) when linear probing (LP) or finetuning (FT) is applied with 10% of the training set. N/A indicates *not applicable* due to the lack of publicly available models. The highest results are written in bold, while the second best results are underlined. *We report FT result of msGFM as the LP result is not available in the original paper.

Method	Backbone	LP / FT
SeCo [26]	RN-18	N/A / 93.1
GASSL [2]	RN-18	N/A / 89.5
SeCo [26]	RN-50	95.6 / **97.2**
CACo [27]	RN-50	**95.9** / N/A
SatMAE (S2) [9]	ViT-B	96.6 / 99.2
I-JEPA [1]	ViT-B	95.6 / 99.2
SpectralGPT [19]	ViT-B	N/A / 99.2
S2MAE [23]	ViT-B	N/A / 99.2
SMARTIES (Ours)	ViT-B	**98.4 / 99.4**
SatMAE (S2) [9]	ViT-L	97.7 / 99.0
SatMAE (RGB) [9]	ViT-L	93.0 / 95.7
CROMA [12]	ViT-B (×2)	97.6 / 99.2
SatMAE++ (S2) [32]	ViT-L	N/A / 99.0
SMARTIES (Ours)	ViT-L	**98.9 / 99.6**

Table 2. Top-1 accuracy (%) on EuroSAT. N/A indicates *not available* results in the original papers.

datasets seen during pretraining; *open-domain, seen sensor* transfer generalization on datasets unseen during pretraining from known sensors.

- **Unseen Sensor Transfer Experiments.** In this *open-domain, unseen sensor* transfer, we evaluate the generalization capability of SMARTIES on new sensors not present in the pretraining data.

We try to compare our models with all the existing models involving similar number of parameters and backbones on the modality they were designed for (e.g. Scale-MAE on RGB) to ensure fairness[4]. Experimental comparison is conducted on the mostly used datasets and tasks of the previous studies, while following the same evaluation protocol: kNN ($k = 20$) classification, linear probing (LP), full finetuning (FT) and frozen backbone finetuning. We refer reader to the Sec. S1.3 in the supplementary material for the details of downstream transfer on each dataset and task.

4.3. Multispectral, Radar, and RGB Experiments

BigEarthNet. In Tab. 1, we test the performance of SMARTIES on BigEarthNet-S1 (BEN-S1), BigEarthNet-S2 (BEN-S2), and multi-modal BigEarthNet (BEN-MM) [38] by following the finetuning strategies mostly used in previous papers. By doing so, we assess the *in-domain* representation ability of SMARTIES on datasets seen during pretraining, while focusing on multi-label RS scene classification task. The BEN setup, involving both S1 and S2, enables to study the ability of SMARTIES on handling known sensors variability. Results in Tab. 1 show that SMARTIES excels as a unified FM, demonstrating superior performance with diverse sensor inputs (single-modal/multi-modal) and distinguishing itself from previous methods that require sensor-specific pretraining. With various single-sensor input (columns S1 and S2 of Tab. 1), SMARTIES consistently outperforms sensor-specific competitors on both BEN-S1 LP and BEN-S2 FT for ViT-B and ViT-L backbones. Since previous models are mainly customized for optical data (BEN-S2), most of them perform poorly under the BEN-S1 setting, composed of SAR data. When considering multi-sensor fused inputs (column MM of Tab. 1), SMARTIES outperforms previous state-of-the-art model CROMA (which was specifically pretrained with SAR-Optical image pairs), by significant 1.5% mAP (ViT-L) through LP. In addition, the results on BEN-MM LP reflect the complementary benefits of leveraging multi-modal data using a single model: compared to LP with only SAR input (BEN-S1 LP), multi-modal input (BEN-MM LP) boosts the performance by 6.2% mAP with our model.

Downstream Transfer. We test the *open-domain, seen sensor* representation ability of SMARTIES on the downstream tasks of scene classification with datasets RESISC-45 [6], EuroSAT [18], WHU-RS19 [10] and UCMerced [47] which are not seen during pretraining. Remote Sensing Image Scene Classification (RESISC-45) is a very high-resolution RGB imagery dataset, which contains 31,500 images and 45 scene classes in total. EuroSAT is a multispectral dataset for scene classification, including 27K Sentinel-2 images with 10 classes. Results on the EuroSAT dataset are provided in Tab. 2, where SMARTIES surpasses previous FMs both in LP (98.9% *vs.* 97.7%) and FT (99.6% *vs.* 99.2%). Tab. 3 shows the results on RESISC-45, where SMARTIES demonstrates highly competitive performance, even against

[4] For a broader comparison, see Tab. S2 in the supplementary material.

Method	Backbone	Top-1 Acc.
MAE [17]	ViT-L	93.3
SatMAE (RGB) [9]	ViT-L	94.8
MCMAE [13]	Conv ViT-L	95.0
Scale-MAE [34]	ViT-L	95.7
SatMAE++ (RGB) [32]	ViT-L	**97.5**
SMARTIES (Ours)	ViT-L	<u>95.8</u>

Table 3. Top-1 accuracy (%) of finetuning on RESISC-45.

Method	EuroSAT	WHU-RS19	UCMerced
SatMAE (RGB) [9]	84.4	69.9	69.7
Scale-MAE [34]	86.7	79.5	<u>75.0</u>
Cross-Scale MAE [40]	<u>87.8</u>	<u>79.8</u>	74.5
SMARTIES (Ours)	**93.7**	**80.4**	**77.0**

Table 4. *k*NN classification accuracies averaged over different scale ratios (12.5%, 25%, 50%, 100%).

models specifically trained on RGB data. Compared to previous methods, SMARTIES also showcases high data efficiency, achieving these results with only 496K images for pretraining, of which 60K RGB images represent only a small fraction (see Sec. S2 in the supplementary material for a detailed analysis). These results in the *open-domain, seen sensor* setup further demonstrate the remarkable generalization and scalability of SMARTIES.

Multi-scale Transfer. RS sensors exhibit pronounced differences in terms of spatial resolution. We mimic this variability in sensors characteristics by resizing images for different scale ratios (12.5%, 25%, 50%, 100%) for model evaluation. Experiments are performed on three datasets which are all unseen when pretraining: EuroSAT, WHU-RS19 [10], UCMerced [47] to assess model's ability on handling scale variances. WHU-RS19 and UCMerced are both very high-resolution RGB image datasets. More precisely, WHU-RS19 contains 19 typical RS scene classes and images with up to 0.5m spatial resolution, while UCMerced contains 21 land use classes of urban locations around the United States. Tab. 4 presents the *k*NN classification accuracies on this multi-scale setup for different FMs. Without scale-specific pretraining, SMARTIES significantly outperforms previous state-of-the-art FMs (SatMAE, Scale-MAE, and Cross-scale MAE), which are specifically designed to tackle multi-scale input, with an improvement 5.9% on EuroSAT, 0.6% on WHU-RS19, and 2.0% on UCMerced. Results from this setup verifies the robustness of SMARTIES against scale variability. We argue that SMARTIES owes this feature to the multi-scale nature of the pretraining data.

4.4. Unseen Sensor Transfer Experiments

The sensor-agnostic design of SMARTIES allows adapting to new sensors that were not present during pretraining (Sec. 3.4). We verify this *sensor transfer* capability in the *open domain, unseen sensor* mode (see Fig. S1 in the supplementary material) through frozen backbone semantic segmentation experiments on SICKLE [36] for crop-type mapping, BurnScars [20] for burn scars detection, DynamicEarthNet (DEN) for land-use and land-cover mapping [41] and SpaceNet7 [43] for building detection. When encoder and projection layers are kept frozen, the unseen sensor transfer of SMARTIES can be achieved via: 1) the classical naive approach of feeding the new sensor's bands to the closest projection layers; and 2) the proposed projection interpolation (cf. Sec. 3.4 and Fig. 4). We verify the former on the Planet images of DEN and SpaceNet7, and the Harmonized Landsat Sentinel-2 (HLS) images of BurnScars. Even though Planet and HLS sensors have been never seen during SMARTIES pretraining, wavelength ranges of their bands highly overlap with the pretraining spectra that allows to directly leverage the already learned projection layers. Tab. 5 shows the results under a comparison with state-of-the-art multi-sensor FMs by using the evaluation protocol of the PANGAEA [28] benchmark with frozen backbone UPerNet [45] probing. SMARTIES surpasses previous multi-sensor FMs on BurnScars and SpaceNet7 with a highly competitive performance on DEN. These results shows the success of SMARTIES for unseen sensor transfer in the case of overlap with pretraining spectra. SICKLE contains Landsat-8 satellite images, which are acquired by an optical sensor (OLI) and a thermal sensor (TIRS). TIRS sensor with its thermal infrared bands is characterized by wavelength ranges unseen during pretraining (i.e., non-overlapping with pretraining spectra). To test SMARTIES performance on the SICKLE's Landsat-

Method	Backbone	BurnScars	DEN	SpaceNet7
GFM [29]	Swin-B	76.9	34.1	60.9
CROMA [12]	ViT-B (×2)	81.8	38.3	59.9
SenPa-MAE [33]	ViT-B	80.8	30.2	58.5
DOFA [46]	ViT-B	80.6	**39.3**	<u>61.8</u>
TerraMindv1 [21]	ViT-B	<u>82.4</u>	37.9	60.6
SMARTIES (Ours)	ViT-B	**82.8**	<u>38.5</u>	**62.2**

Table 5. Semantic segmentation results (mIoU) with frozen backbone UPerNet probing on BurnScars, DynamicEarthNet (DEN) and SpaceNet7, using the PANGAEA benchmark [28].

Method	Training	mIoU	Acc.	F1
U-Net 2D [35]	Scratch	47.7	69.7	62.7
DeepLapV3+ [5]	Scratch	<u>48.5</u>	<u>71.2</u>	<u>63.2</u>
SMARTIES (w/o PI)	Frozen	35.4	55.8	50.6
SMARTIES (Ours)	Frozen	**50.2**	**75.5**	**63.7**

Table 6. Unseen sensor transfer of SMARTIES (ViT-B) for crop-type segmentation on SICKLE. PI: projection interpolation; Frozen: a segmentation head is finetuned with frozen backbone.

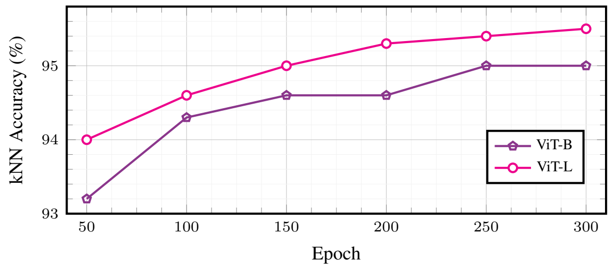

Figure 5. kNN classification accuracy (%) on EuroSAT versus pretraining epoch for ViT-B and ViT-L.

8 segmentation task by using both OLI and TIRS bands, we apply interpolation for the thermal infrared bands, and learn a single layer segmentation head to assess its unseen sensor transfer capability. Tab. 6 shows a comparison of our model's performance with fully supervised models trained specifically on this dataset. SMARTIES with projection interpolation surpasses fully supervised models in all the metrics, and by a margin of 1.7% in mIoU, even with frozen backbone parameters. By combining spectrum-aware projection layers (Fig. 3 and Fig. 4), SMARTIES achieves strong generalization and adaptability, showing its potential as a versatile FM for new sensor types without requiring additional sensor-specific finetuning.

4.5. Ablations

Model Size and Pretraining Epochs. Our model is built based on the standard ViT backbones with the lightweight projectors, so that SMARTIES does not significantly increase parameter count compared to the standard MAE (+5.9M for ViT-L backbone). Fig. 5 presents the effects of ViT backbone and pretraining epochs for kNN classification on EuroSAT. Overall, using ViT-L brings continuous improvements over ViT-B of about 0.5% for the same epochs, showing the scalability of SMARTIES.

Cross-sensor Token Mixup. We study the robustness of the learned representations by ablating the use of cross-sensor token mixup with models pretrained for 50 epochs (Tab. 7). With mixup, the model's performance in processing multi-modal data is significantly enhanced: the LP performance increases by 2.2% on BEN-MM. We argue that mixup adds variability in the input data that helps to prevent overfitting, even with BEN only pretraining.

Fusion Strategies for Multi-modal Input. We also examine the effects of feature fusion strategies for the multi-modal downstream transfer. As shown in Tab. 8, we design three fusion strategies to adapt to multi-modal inputs: *Image Stacking* denotes a strategy directly stacking images from different modalities as input to the model; *Feature Concatenation* means concatenating features obtained by the encoder for different modalities; finally, *Mixup Concatenation* is our proposed strategy, which concatenates features extracted from the encoder from different modalities after applying mixup with spectrum-aware projections. An illustration of these strategies is given in Fig. S2 in the supplementary material. Results in Tab. 8 show that Mixup Concatenation achieves the highest mAP in both 1% and 10% training data of BEN-MM for LP.

Method	Backbone	Acc.
SMARTIES (w/o mixup)	ViT-B	91.0
SMARTIES (w mixup, BEN only)	ViT-B	91.1
SMARTIES (w mixup)	ViT-B	**93.2**

Table 7. kNN classification accuracy (%) on EuroSAT under different pretraining settings (pretraining epochs are set to 50).

Strategy	Backbone	1%	10%
Image Stacking	ViT-L	75.9	83.1
Feature Concatenation	ViT-L	77.0	84.7
Mixup Concatenation	ViT-L	**79.2**	**86.7**

Table 8. mAP on BEN-MM for linear probing under different feature fusion strategies with 1% and 10% of the training sets.

5. Conclusion

The variety of sensors that acquire a continuous stream of information characterizing the Earth's surface makes RS data multi-modal by nature. In this paper, we introduce SMARTIES, a unified and versatile foundation model that achieves sensor-agnostic representations by projecting diverse sensory data into shared spectrum-aware space and training with the masked reconstruction objective with cross-sensor token mixup. This strategy has the advantage of seamlessly handling diverse sensory inputs at pretraining, exhibiting scalability to RS sensors characterized by different spectral properties. SMARTIES can learn transferable representations not only for pretraining sensors but also for unseen ones, demonstrating unprecedented generalization capabilities. SMARTIES outperforms existing foundation models for RS in 10 datasets on both single-modal and multi-modal tasks, including experiments testing the downstream transfer to the sensor never seen during pretraining. Unifying multi-sensor RS image interpretation with a single foundation model has the potential to leverage the synergistic advantages of different sensors for Earth observation, while eliminating the need for isolated efforts in training sensor-specific models. SMARTIES makes a further step in that direction, and extensions to the temporal domain and to more diversified downstream tasks will be our future efforts toward unified, physics-inspired foundation models for RS.

Acknowledgment

This work is supported by the European Space Agency (ESA) through the Discovery and Preparation Program, and is part of the project Toward a Foundation Model for Multi-Sensor Earth Observation Data with Language Semantics.

References

[1] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 15619–15629, 2023. 6

[2] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Int. Conf. Comput. Vis. (ICCV)*, pages 10181–10190, 2021. 6

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 348–367. Springer, 2022. 2

[4] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Int. Conf. Comput. Vis. (ICCV)*, pages 16772–16782, 2023. 6

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 801–818, 2018. 7

[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. 6

[7] Mingmin Chi, Antonio Plaza, Jón Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu. Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11):2207–2219, 2016. 1

[8] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6172–6180, 2018. 5

[9] Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David B. Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 197–211, 2022. 1, 2, 5, 6, 7

[10] Dengxin Dai and Wen Yang. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens. Lett. (GRSL)*, 8(1):173–176, 2010. 6, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent. (ICLR)*, 2021. 5

[12] Anthony Fuller, Koreen Millard, and James R. Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 5506–5538, 2023. 1, 2, 6, 7

[13] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. MCMAE: Masked convolution meets masked autoencoders. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 35632–35644, 2022. 7

[14] Luis Gómez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015. 1

[15] Xin Guo, Jiangwei Lao, Bo Dang, Yingying Zhang, Lei Yu, Lixiang Ru, Liheng Zhong, Ziyuan Huang, et al. SkySense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 27662–27673, 2024. 1, 2

[16] Boran Han, Shuai Zhang, Xingjian Shi, and Markus Reichstein. Bridging remote sensors with multisensor geospatial foundation models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 27852–27862, 2024. 6

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 16000–16009, 2022. 2, 5, 7

[18] 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 J. Sel. Top. Appl. Earth Obs. Remote Sens. (JSTARS)*, 12(7): 2217–2226, 2019. 6

[19] Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, et al. SpectralGPT: Spectral remote sensing foundation model. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 46(8):5227–5244, 2024. 1, 2, 6

[20] Johannes Jakubik, Sujit Roy, C. E. Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023. 7

[21] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, et al. TerraMind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025. 7

[22] Marrit Leenstra, Diego Marcos, Francesca Bovolo, and Devis Tuia. Self-supervised pre-training enhances change detection in Sentinel-2 images. In *Int. Conf. Pattern Recog. (ICPR), Workshop Pattern Recog. in Remote Sens.*, pages 578–590, 2021. 2

[23] Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2mae: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 27696–27705, 2024. 1, 2, 6

[24] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent. (ICLR)*, 2019. 5

[26] Oscar Mañas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Int. Conf. Comput. Vis. (ICCV)*, pages 9414–9423, 2021. 6

[27] Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Change-aware sampling and contrastive learning for satellite images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 5261–5270, 2023. 6

[28] Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, et al. PANGAEA: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint 2412.04204*, 2024. 7

[29] Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, and Chen Chen. Towards geospatial foundation models via continual pretraining. In *Int. Conf. Comput. Vis. (ICCV)*, pages 16806–16816, 2023. 6, 7

[30] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 58363–58408, 2023. 2

[31] Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 164–182, 2024. 2

[32] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multi-spectral satellite imagery. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 27811–27819, 2024. 1, 6, 7

[33] Jonathan Prexl and Michael Schmitt. SenPa-MAE: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *German Conf. Pattern Recog. (GCPR)*, 2024. 2, 7

[34] 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 *Int. Conf. Comput. Vis. (ICCV)*, pages 4088–4099, 2023. 2, 7

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Int. Conf. Medical Image Comput. Computer-assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015. 7

[36] Depanshu Sani, Sandeep Mahato, Sourabh Saini, Harsh Kumar Agarwal, Charu Chandra Devshali, Saket Anand, Gaurav Arora, and Thiagarajan Jayaraman. SICKLE: A multi-sensor satellite imagery dataset annotated with multiple key cropping parameters. In *IEEE/CVF Winter Conf. on App. of Comput. Vis. (WACV)*, pages 5995–6004, 2024. 7

[37] Linus Scheibenreif, Michael Mommert, and Damian Borth. Contrastive self-supervised data fusion for satellite imagery. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 3: 705–711, 2022. 2

[38] Gencer Sumbul, Arne de Wall, Tristan Kreuziger, Filipe Marcelino, Hugo Costa, Pedro Benevides, Mário Caetano, Begüm Demir, and Volker Markl. BigEarthNet-MM: A large scale multi-modal multi-label benchmark archive for remote sensing image classification and retrieval. *IEEE Geosci. Remote Sens. Magazine (GRSM)*, 9(3):174–180, 2021. 5, 6

[39] Gencer Sumbul, Markus Müller, and Begüm Demir. A novel self-supervised cross-modal image retrieval method in remote sensing. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2426–2430, 2022. 2

[40] Maofeng Tang, Andrei Liviu Cozma, Konstantinos Georgiou, and Hairong Qi. Cross-scale MAE: A tale of multi-scale exploitation in remote sensing. In *Adv. Neural Inform. Process. Syst. (NeurIPS)*, pages 20054–20066, 2023. 7

[41] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Camero Andres, Jingliang Hu, Ariadna Hoderlein, Caglar Senaras, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022. 7

[42] Devis Tuia, Ribana Roscher, Jan Dirk Wegner, Nathan Jacobs, Xiao Xiang Zhu, and Gustua Camps-Valls. Towards a collective agenda on AI for earth science data analysis. *IEEE Geosci. Remote Sens. Magazine (GRSM)*, 9(2):88–104, 2021. 1

[43] Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development spacenet dataset. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 6394–6403, 2021. 7

[44] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geosci. Remote Sens. Magazine (GRSM)*, 10(4):213–247, 2022. 2

[45] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, 2018. 7

[46] Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024. 1, 2, 7

[47] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *SIGSPATIAL Int. Conf. Advances in Geographic Inf. Systems*, page 270–279, 2010. 6, 7