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

Supplementary Material

In the supplementary material, we provide detailed information for pretraining and evaluation across different datasets. Besides, we provide additional analyses, including the pretraining efficiency of SMARTIES, more ablation results on the use of pretraining data and projection extrapolation to unseen spectral ranges. Our code and pretrained models are available at https://gsumbul.github.io/SMARTIES.

S1. Implementation

S1.1. Spectrum-Aware Projection Layers

We provide the details for our projection layers f_i in Tab. S1. The spectral range of each layer is defined according to the bands of different sensors. Specifically, f_1 to f_{12} follows the bands in Sentinel-2, f_{13} to f_{15} are based on RGB images from Maxar, and f_{16}, f_{17} corresponds to Sentinel-1. Each reprojection layer r_i takes charge of the same wavelength range as its corresponding projection layer f_i.

Sensor	Dataset	Layer	Band	Wavelength (nm)
S2	BEN-S2 fMoW-S2	f_1	B01	422 - 463
		f_2	B02	427 - 558
		f_3	B03	524 - 595
		f_4	B04	634 - 696
		f_5	B05	689 - 719
		f_6	B06	726 - 755
		f_7	B07	761 - 802
		f_8	B08	728 - 938
		f_9	B8A	843 - 886
		f_{10}	B09	923 - 964
		f_{11}	B11	1516 - 1704
		f_{12}	B12	2002 - 2376
Maxar	fMoW-RGB	f_{13}	Blue	430 - 545
		f_{14}	Green	466 - 620
		f_{15}	Red	590 - 710
S1	BEN-S1	f_{16}	VV	$5.5{\times}10^7$ - $5.6{\times}10^7$
		f_{17}	VH	$5.5{\times}10^7$ - $5.6{\times}10^7$

Table S1. Spectral ranges for the projection layers used in SMARTIES pretraining. S2 denotes Sentinel-2, S1 denotes Sentinel-1, BEN is the abbreviation for BigEarthNet.

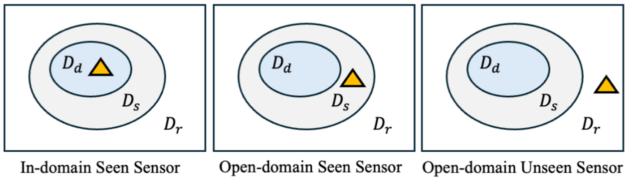

Figure S1. Different inference modes for downstream transfer to diverse sensors: (1) *in-domain, seen sensor*: transfer in the domain D_d of the same datasets seen during pretraining, (2) *open domain, seen sensor*: transfer in the domain D_s of new task, observed by the same sensors used during pretraining and (3) *open domain, unseen sensor*: transfer in the domain D_r of new tasks observed by any sensor. The yellow triangle denotes the position of each inference mode. From D_d to D_r, an increasing degree of generalization is required.

S1.2. Pretraining

We pretrain two versions of SMARTIES by using ViT-B and ViT-L [2] backbones: SMARTIES (ViT-B) and SMARTIES (ViT-L), while we use the same decoders with the vanilla MAE [5]. For both versions, we pretrain for 300 epochs, using AdamW optimizer [8] ($\beta_1 = 0.9$, $\beta_1 = 0.95$ and weight decay of 0.05) and mixed precision (FP16) with the batch size of 2048 (distributed over 8 A100 GPUs), the base learning rate of 1.5e-4, warmup of 20 epochs and cooldown by half-cosine decay schedule. For data augmentation, we randomly apply vertical flipping, horizontal flipping and rotation in order. After this, by randomly sampling scale parameter between 0.25 and 1 and keeping the same width-height ratio, we crop images, which are then resized to the input image size with bi-cubic interpolation. For a given pair of images, we apply identical transformations to both images.

S1.3. Evaluation

Once SMARTIES is pretrained with ViT-B and ViT-L backbones, the resulting encoders and spectrum-aware projection layers are used for single/multi-modal downstream transfer of single/multi-label classification and semantic segmentation with RS images from diverse sensors. For downstream transfer, we consider all the possible inference modes: (1) *in-domain, seen sensor inference*, (2) *open-domain, seen sensor inference*, and (3) *open-domain, unseen sensor inference* that are illustrated in Fig. S1.

For a fair comparison with other foundation models, we follow the same evaluation protocols and the splits of datasets with previous works by using non-parametric *k*NN

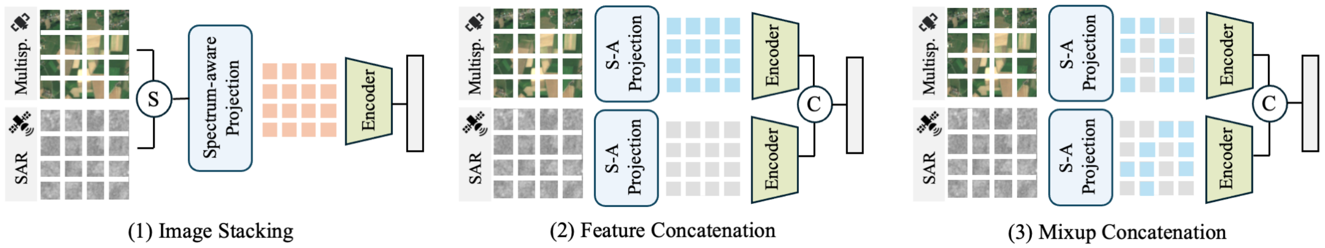

Figure S2. Different multi-modal fusion strategies for downstream transfer on multi-modal input images. S-A Projection, S and C denote Spectrum-aware Projection, stacking and concatenation, respectively.

classification, linear probing, non-linear frozen backbone finetuning and full finetuning. *k*NN classification allows to directly assess the learned representations without additional training, while linear probing or frozen backbone finetuning require to train a linear classifier or nonlinear task head, respectively, on top of the frozen backbone. Finetuning requires to train the entire backbone with the task head on the downstream dataset. Below, we provide the evaluation details for each dataset.

BigEarthNet-S1. By following CROMA [3], we apply linear probing by using the 10% of the complete training set and evaluating on the entire validation set without data augmentation. Linear probing is applied for 100 epochs by using AdamW optimizer with the batch size of 1024 and the base learning rate of 1e-3, which is decayed 10× at epochs 60 and 80. We resize images to the input image size with bi-cubic interpolation.

BigEarthNet-S2. By following SatMAE (S2) [1] and SeCO [9], we apply full finetuning by using the 10% of the complete training set and evaluating on the entire validation set. We finetune for 100 epochs, using AdamW optimizer with the batch size of 256, the base learning rate of 5e-5, the weight decay of 0.05, the drop path rate of 0.2, warmup of 5 epochs and cooldown by half-cosine decay schedule. For data augmentation, we first randomly apply vertical flipping, horizontal flipping and rotation in order. Then, we resize images to the input image size with bi-cubic interpolation.

BigEarthNet-MM. By following CROMA [3], we apply linear probing by using 10% of the complete training set and evaluating on the entire validation set without data augmentation. Linear probing is applied for 100 epochs by using AdamW optimizer with the batch size of 1024 and the base learning rate of 1e-3, which is decayed 10× at epochs 60 and 80. We resize image pairs to the input image size with bi-cubic interpolation. To operate SMARTIES on multi-modal input, as shown in Fig. S2, we consider three multi-modal fusion strategies: 1) image stacking; 2) feature concatenation; and 3) mixup concatenation. Compared to first two strategies, mixup concatenation, where we concatenate features extracted from the backbone from both modalities after applying mixup with spectrum-aware projections, is introduced for the first time in our paper.

EuroSAT. We apply linear probing, *k*NN classification and fine-tuning on EuroSAT by using the same splits as SatMAE (S2) [1]. For linear probing, we use AdamW optimizer for 100 epochs with the batch size of 1024 and the base learning rate of 1e-3, which is decayed 10× at epochs 60 and 80. For finetuning, we use AdamW optimizer for 150 epochs with the batch size of 256, the weight decay of 0.05, the base learning rate of 2e-4, the drop path rate of 0.1, warmup of 5 epochs and cooldown by half-cosine decay schedule. We also apply CutMix ($\alpha = 1$) and MixUp ($\alpha = 0.8$) with between images and labels. Only for finetuning, we use data augmentation with random vertical flipping, horizontal flipping and rotation in order. We resize images to the input image size with bi-cubic interpolation.

RESISC-45. We apply fine-tuning on RESISC-45 by using the same splits as Scale-MAE [12]. To this end, we use AdamW optimizer for 200 epochs with the batch size of 64, the weight decay of 0.05, the base learning rate of 6.25e-5, the drop path rate of 0.2, warmup of 5 epochs and cooldown by half-cosine decay schedule. For data augmentation, we first apply random vertical flipping, horizontal flipping and rotation in order. Then, by randomly sampling scale parameter between 0.25 and 1 and keeping the same width-height ratio, we crop images, which are then resized to the input image size with bi-cubic interpolation during training. During evaluation, we first resize images to 256×256, and then apply center cropping with the input image size.

WHU-RS19. For *k*NN classification, we only resize images to the input image size with bi-linear interpolation.

UCMerced. We first resize images to 256×256, and then apply center cropping with the input image size for *k*NN classification.

BurnScars, DynamicEarthNet, SpaceNet7 experiments are conducted by following the default evaluation protocol of the PANGAEA [10] benchmark for a fair comparison with other methods. In detail, for all these datasets, frozen backbone UPerNet probing is applied: model weights of

Model	Backbone	PT Epochs	PT Data Size		mAP (%)	Acc. (%)
			S2	RGB	BEN-S2 10%	RESISC-45
SatMAE (S2) [1]	ViT-L	50	713K	-	82.1	N/A
SatMAE (S2) [1]	ViT-L	200	713K	-	86.2	N/A
SatMAE (RGB) [1]	ViT-L	800	-	364K	N/A	94.8
Scale-MAE [12]	ViT-L	800	-	364K	N/A	95.7
CROMA [3]	ViT-B (×2)	300	1M	-	87.6	N/A
SpectralGPT [6]	ViT-L	200	713K	-	86.9	N/A
SpectralGPT$^+$ [6]	ViT-L	300	1M	-	89.0	N/A
S2MAE [7]	ViT-L	200	713K	-	86.5	N/A
S2MAE* [7]	ViT-L	300	1M	-	88.5	N/A
SatMAE++ (RGB) [11]	ViT-L	800	-	364K	N/A	97.5
SatMAE++ (S2) [11]	ViT-L	50	713K	-	85.1	N/A
SMARTIES (Ours)	ViT-L	300	248K	60K	87.7	95.8
CROMA [3]	ViT-L (×2)	600	1M	-	88.3	N/A
SpectralGPT [6]	ViT-H	200	713K	-	89.2	N/A
SpectralGPT$^+$ [6]	ViT-H	300	1M	-	91.4	N/A
S2MAE [7]	ViT-H	200	713K	-	88.8	N/A
S2MAE* [7]	ViT-H	300	1M	-	90.7	N/A
SkySense [4]	ViT-L (×2) + Swin-H	780	21.5M	21.5M	88.7	96.3*

Table S2. Pretraining efficiency comparison of the existing foundation models, including 1) the considered backbones, 2) pretraining (PT) epochs, 3) PT data size in terms of numbers of Sentinel-2 (S2) and RGB images, 4) BEN multi-label scene classification results (mAP) when finetuning (FT) is applied with 10% of the training set, 5) RESISC-45 scene classification results (top-1 accuracy) under FT. *20% of the training set is used. N/A indicates *not applicable* due to either the lack of publicly available models or sensor mismatch between models and datasets (which could lead to unfair comparisons).

our pretrained encoder are frozen, while UPerNet segmentation head is learned on top of it. As DynamicEarthNet includes multi-temporal images, Lightweight Temporal Attention Encoder (L-TAE) [13] is utilized between the encoder and the segmentation head to map each image time-series into an aggregated feature map. To learn the segmentation head parameters for all the datasets, AdamW optimizer is used for 80 epochs with the batch size of 8, the weight decay of 0.05 and the base learning rate of 1e-4, which is decayed 10× at 60% and 90% of the total steps. We refer readers to [10] for the details of the PANGAEA evaluation protocol.

SICKLE. We apply non-linear frozen backbone finetuning by freezing the parameters of our pretrained model, while learning a segmentation head on top it. We use the same segmentation head with [14] by using a single convolutional layer followed by bi-linear upsampling. For zero-shot sensor transfer to Landsat-8 images, we apply interpolation to unseen spectrum ranges of: 1) blue band (B2) via the weighted average of the projection layers dedicated to Sentinel-2 blue (B02) and aerosol (B01) bands; and 2) thermal infrared band (B10) via the weighted average of the projection layers dedicated to Sentinel-2 SWIR band (B12) and Sentinel-1 VV band. For the rest of the bands, we select the relevant projection layers dedicated to Sentinel-2 bands, where the same spectral ranges are shared with Landsat-8 bands. To learn the segmentation head, we use AdamW optimizer for 200 epochs with the batch size of 32 and the base learning rate of 8e-3, which is decayed 10× at epochs 120 and 160. Without any data augmentation, we resize images to the input image size with nearest-neighbor interpolation.

S2. Pretraining Efficiency

In the main body of our paper, we test our models against the existing foundation models, which use as similar pretraining (PT) data size and epochs as possible, for a fair comparison. To further compare the PT efficiency of SMARTIES, in Tab. S2 we provide an extended comparison of the existing models in terms of PT data size and epochs together with BEN-S2 and RESISC-45 results under finetuning. Results demonstrate that SMARTIES shows a significantly higher PT efficiency compared to previous methods in terms of both PT data size and epochs (which is associated with PT time). In detail, by comparing the results in the first block of Tab. S2, one can see that SMARTIES uses the fewest Sentinel-2 (S2) images for PT (248K) to achieve highly competitive performance (87.7%) on BEN-S2 compared to the state-of-the-art SpectralGPT$^+$ model (89.0%), which uses four times more of S2 images during PT. Meanwhile, SMARTIES also shows high efficiency in terms of the use of RGB data. By using only 60K RGB PT data, SMARTIES surpasses most of the RGB-specific mod-

Backbone	PT Epochs	PT Data Split		Acc.
		BEN	fMoW	
ViT-B	50	✓	✗	91.1
		✗	✓	92.1
		✓	✓	93.2
	100	✓	✓	94.3
ViT-L	100	✓	✓	94.6

Table S3. *k*NN classification accuracy (%) on EuroSAT when different subsets of the pretraining (PT) data are used for SMARTIES.

Method	Backbone	PT Epochs	SAR PT	mAP
SMARTIES (w/o PE)	ViT-B	50	✗	62.1
SMARTIES (w PE)	ViT-B	50	✗	64.0
SMARTIES (Ours)	ViT-B	50	✓	73.6
SpectralGPT [6]	ViT-B	200	✗	57.1
SatMAE (S2) [1]	ViT-L	200	✗	67.4
SMARTIES (Ours)	ViT-B	300	✓	78.9

Table S4. BEN-S1 multi-label classification results (mAP) when linear probing is applied with 10% of the training set. PE: projection extrapolation; PT: pretraining.

els pretrained with 6 times more data and over 2 times more PT epochs. The high data efficiency of SMARTIES can be attributed to: 1) the sensor-agnostic design, which explicitly represents data into transferable spectrum-aware spaces instead of learning shared representations from heterogeneous sensors implicitly; and 2) the implicit data augmentation brought by cross-sensor token mixup. We would like to note that masked data modeling in combination with ViTs can be effectively scaled into larger models with higher amount of PT data [5]. This can be seen from Tab. S2: 50 vs. 200 epochs PT of SatMAE (S2) and S2MAE (ViT-L) vs. S2MAE (ViT-H). Thus, by feeding more PT data with more epochs, the performance of SMARTIES can be further scaled up. To further analyze this, we assess the effect of different subsets of our PT data together with different PT epochs and backbones in Tab. S3 under *k*NN classification of EuroSAT. One can observe from the table that the higher the number of images and epochs used by SMARTIES for PT the better it performs. In addition, by using a larger ViT model, SMARTIES is capable of achiving higher *k*NN accuracy.

S3. Extrapolation to Unseen Spectral Ranges

For downstream transfer to a sensor unseen during pretraining (i.e., open-domain, unseen sensor inference), SMARTIES can be adapted to unseen spectral ranges by applying interpolation to the learned projection layers as it is explained in Sec. 3.4 and shown with the SICKLE results (cf. Sec. 4.4). Here, we further evaluate the generalization ability of SMARTIES to unseen regions out of the (min, max) of the pretraining spectra through extrapolation. This setting is significantly more challenging than the SICKLE experiments, where the thermal infrared bands falls within the pretraining range. We simulate an unseen spectral range out of the pretraining spectra by excluding SAR data during pretraining. Then, we perform linear probing on BEN-S1 by extrapolating the learned projection layers of SMARTIES. In addition, we also apply linear probing with SpectralGPT and SatMAE (S2), for which SAR is already excluded from pretraining. To do this, we duplicate VV and VH bands of BEN-S1 six times, which are given as model inputs in place of the original input (Sentinel-2 image). Table S4 shows the corresponding results. One can see from the table that SMARTIES pretrained without SAR data yields lower BEN-S1 results than the full pretraining even though projection extrapolation yields modest +2% mAP. This shows that the downstream transfer capability of SMARTIES to an unseen sensor is valid: 1) for the the unseen ranges falling inside the pretraining spectra through projection interpolation; 2) not for the ranges out of the limits of pretraining spectra through projection extrapolation. Once SAR data is included in pretraining, however, SMARTIES with even 50 pretraining epochs provides 16.5% higher mAP than SpectralGPT, which rely on sensor-specific pretraining. These results indicate that the generalization capability of SMARTIES highly depends on the spectral range seen during pretraining.

References

[1] 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. 2, 3, 4

[2] 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. 1

[3] 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. 2, 3

[4] 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. 3

[5] 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. 1, 4

[6] 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. 3, 4

[7] 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. 3

[8] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent. (ICLR)*, 2019. 1

[9] 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. 2

[10] 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. 2, 3

[11] Mubashir Noman, Muzammal Naseer, Hisham Cholakkal, Rao Muhammad Anwar, Salman Khan, and Fahad Shahbaz Khan. Rethinking transformers pre-training for multispectral satellite imagery. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pages 27811–27819, 2024. 3

[12] 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, 3

[13] Vivien Sainte Fare Garnot and Loic Landrieu. Lightweight temporal self-attention for classifying satellite images time series. *arXiv preprint arXiv:2007.00586*, 2020. 3

[14] Depanshu Sani, Sandeep Mahato, Sourabh Saini, Harsh Kumar Agarwal, Charu Chandra Devshali, Saket Anand, Gaurav Arora, and Thiagarajan Jayaraman. SICKLE: A multisensor satellite imagery dataset annotated with multiple key cropping parameters. In *IEEE/CVF Winter Conf. on App. of Comput. Vis. (WACV)*, pages 5995–6004, 2024. 3