HYPERSMARTIES: PHYSICS-AWARE SPECTRAL REPRESENTATION LEARNING FOR REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Unifying image representation learning is essential for enabling self-supervised pretraining in remote sensing to scale across diverse sensors. However, existing models are typically constrained to fixed, predefined band sets, or do not scale seamlessly to new generation hyperspectral sensors, where hundreds of spectrally narrow channels densely sample the electromagnetic spectrum. To address this, we introduce hyperSMARTIES, a unified foundation model for wavelength-continuous masked data modeling. hyperSMARTIES learns a unified spectral projection that embeds arbitrary bands using a single shared projection function, and employs a transformer block that adaptively aggregates per-band embeddings into a single spectral token conditioned by the central wavelengths of bands. This lifts the need for sensor/range-specific projection functions and enables unified projection across diverse spectral regimes. Furthermore, spectral consistency is enforced via flow matching over wavelengths: a wavelength-indexed velocity field is learned on band embeddings to enforce smooth, locally consistent evolution across neighboring wavelengths, hence reproducing the physical characteristics of hyperspectral signals. We pretrain hyperSMARTIES on hyperspectral imagery acquired by the EnMAP sensor, which provides fine-grained spectral sampling, and apply downstream transfer on multiple tasks with diverse sensors. hyperSMARTIES not only outperforms previous models with improved unseen-sensor generalization, but also shows wavelength-dependent consistency in the learned band embeddings.

1 INTRODUCTION

Remote sensing (RS) image archives are large and diverse, spanning imagery sensors with different spectral properties, resolutions, and acquisition conditions. Accordingly, geospatial foundation models (GFMs) hold the promise of encoding general-purpose image representations to strengthen Earth observation (EO) capabilities: a single pretrained backbone that can transfer across sensors, tasks, and geographic regions (Zhu et al., 2026). However, scaling self-supervised pretraining across such heterogeneous data, and thus unifying RS image representation learning, remains challenging. In particular, spectral heterogeneity—differences in band definitions, band counts, and spectral sampling density—leads to a fundamental mismatch between how RS imagery is acquired and how most GFMs are parameterized.

Most of the existing GFMs assume a predefined set of input bands and a corresponding set of projection parameters that map these bands into a token space. Although this is convenient for widely used multispectral imagery (*e.g.*, Sentinel-2), it becomes unfeasible to leverage such models for downstream transfer on unseen sensors with different spectral properties. To address this, several multisensor GFMs have been proposed (*e.g.*, SMARTIES (Sumbul et al., 2025), DOFA (Xiong et al., 2025), Panopticon (Waldmann et al., 2025) etc.). Yet, how to scale multisensor pretraining to many more sensors remains an open challenge. This is particularly problematic for new generation hyperspectral sensors, where hundreds of narrow bands densely sample the electromagnetic spectrum.

In this work, we aim to learn unified RS image representations by encoding arbitrary combinations of image bands in a wavelength-continuous space that preserves local properties of spectral relations across bands of potentially heterogeneous sensors. To this end, we introduce hyperSMARTIES, a unified foundation model for wavelength-continuous masked data modeling. It is built on two principles: 1) unified projection of arbitrary bands via a wavelength-continuous spectral projection;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and 2) spectral consistency enforced by flow matching (Lipman et al., 2023) over the pretraining spectral range. The former principle removes the dependency on predefined band sets by replacing sensor/range-specific projection functions with a single shared projection function parametrized by the central wavelengths of bands. This function first maps each band to a band-wise patch embedding in a common latent space. For the tokenization of a given patch with multiple band embeddings, we employ a transformer block that adaptively aggregates per-band patch embeddings into a single spectral token through attention conditioned on the wavelengths of bands. This unifies tokenization across diverse spectral regimes with variable number of bands, and enables to perform masked data modeling with vision transformers. The latter principle preserves physical consistency of the band embeddings: we apply flow matching over wavelengths by learning a wavelength-indexed velocity field on band embeddings. This encourages smooth, locally consistent evolution of band embeddings across neighboring wavelengths. Unlike conventional flow matching, our formulation does not rely on synthetic noise paths. Instead, it leverages band embeddings directly and treats central wavelength as the path parameter in place of time. Our flow matching objective acts as a regularizer preventing discontinuities across adjacent bands in the electromagnetic spectrum.

We pretrain hyperSMARTIES on EnMAP data (Chabrillat et al., 2022). EnMAP is a recently launched hyperspectral satellite that provides a fine-grained spectral sampling (*i.e.*, the continuous signal sensed is quantized into distinct spectral bands), which is necessary for fully leveraging hyperSMARTIES' spectral projection to learn the continuity of pretraining spectrum. We evaluate downstream transfer on multiple downstream tasks with diverse sensors well beyond EnMAP to assess unseen-sensor generalization, while evaluating on both accuracy and physical fidelity. hyperSMARTIES contributes significantly to the efforts on unified RS image representation learning since it is: 1) scalable (as it is based on a single, wavelength adaptive, shared projection function); 2) physically consistent (since regularized by wavelength-indexed flow matching); and 3) successful in transferring to sensors unseen during pretraining, therefore keeping the promise of multi-sensor foundation models for EO.

2 Related Work

2.1 Geospatial Foundation Models

Geospatial foundation models (GFMs) through self-supervised learning (Noman et al., 2024; Hong et al., 2024; Sumbul et al., 2025; Li et al., 2024; Fuller et al., 2023; Xiong et al., 2025; Jakubik et al., 2025; Waldmann et al., 2025) have become one of the central paradigms in EO (Zhu et al., 2026; Xiao et al., 2025; Lu et al., 2025; Huang et al., 2025; Zhang et al., 2025). This is motivated by their ability to leverage the scale and diversity of RS imagery for learning general image representations, which can be exploited for various downstream tasks (*e.g.*, scene classification, land-cover mapping, semantic segmentation etc.) with improved performance and reduced annotation costs. A large body of research on GFMs is concentrated on adapting masked image modeling with masked autoencoders (MAEs, He et al. (2022)) for RS imagery since, in combination with vision transformers (ViTs, Dosovitskiy et al. (2021)), MAEs can be effectively scaled into a vast amount of RS data. Many of such models, however, assume a fixed input bands from a predefined image modality, therefore focusing on multi-scale robustness for very high-resolution RGB imagery (Reed et al., 2023; Tang et al., 2023), ViT adaptation for multispectral imagery (*e.g.*, Sentinel-2) as in Cong et al. (2022); Noman et al. (2024), and more recently scaling to hyperspectral imagery (*e.g.*, EnMAP) like Braham et al. (2025); Wang et al. (2025a). Although such models are meant to be task-agnostic, the dependency on the input image modality limits their reuse across sensors with different spectral properties.

2.2 Unifying RS Representation Learning Across Sensors

A key challenge for GFMs is unifying RS image representation learning across various sensors with spectral heterogeneity: sensors differ in the number of bands, band definitions (mostly associated with corresponding central wavelength and bandwidth), and spectral sampling density across bands, making fixed-sensor models hard to reuse seamlessly. Initial attempts addressing this limitation have focused on blending multiple image modalities as paired tuples for multi-sensor GFM pretraining in combination with using sensor-specific encoders and projectors (Astruc et al., 2025; Houdré et al., 2026; Jakubik et al., 2025; Nedungadi et al., 2024; Fuller et al., 2023; Guo et al., 2024). While projecting RS data into tokens, these models utilize a fixed number of projection functions for a fixed & predefined set of bands from multiple sensors. This can be regarded as wavelength

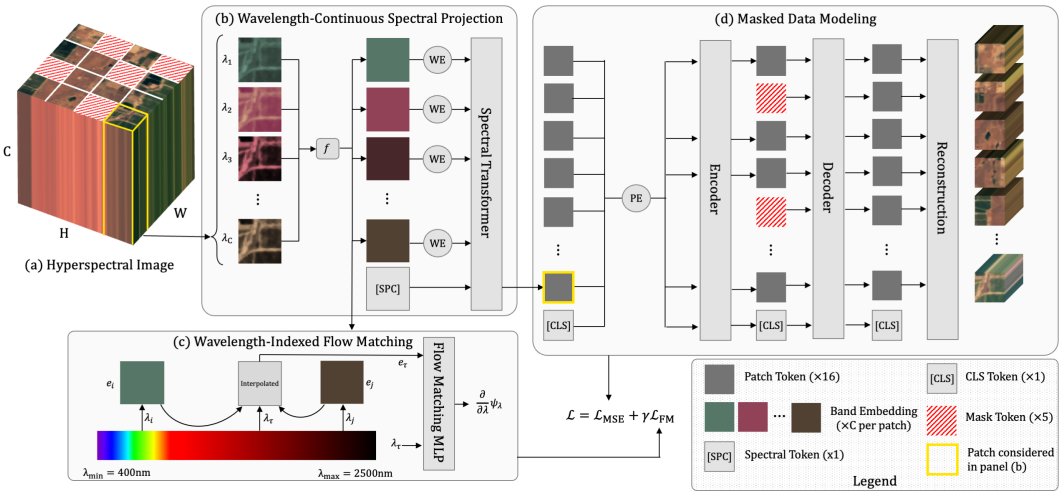


Figure 1: The hyperSMARTIES model: **(a)** given a hyperspectral datacube, we partition it into patches (4×4 in the example) with C channels associated with the set $\{\lambda_i\}_{i=1}^{C}$ of wavelengths. **(b)** Each channel of a patch (highlighted in yellow in panel (a)) is encoded with a unified projection function and then passed in a spectral transformer. The resulting [SPC] token becomes the representation of the patch. **(c)** Flow matching regularization is applied between pairs of bands' representations. **(d)** Finally, the patch representations, each one resulting from the process in panel (b), are fed to a transformer using masked data modeling and leading to the reconstruction of the original signal for each patch. WE and PE denote wavelength encoding and positional encoding, respectively.

discrete projection with fixed bands, as they utilize only a small subset of discrete partitions in the electromagnetic spectrum (ES). Although such models significantly strengthen the multi-modal RS data processing, their generalization to unseen ranges of ES, and thus unseen sensors, is still limited.

Toward unified RS image representation learning with masked data modeling, SMARTIES (Sumbul et al., 2025) learns ES range-specific projection functions rather than defining them based on the corresponding sensors beforehand. Its design is effective for processing arbitrary number of bands, while allowing to utilize learned projection functions for unseen image bands with closer frequencies to pretraining spectra. However, due to its discrete partitioning of the ES and the fact that SMARTIES uses one projection function for each partition, scaling this model to many more sensors with high spectral granularity (*e.g.*, bands of hyperspectral imagery densely sampled from the ES) is challenging. As an alternative to this discretization, DOFA (Xiong et al., 2025), and more recently Copernicus-FM (Wang et al., 2025b) propose wavelength-conditioned dynamic weight generation for wavelength-continuous projection of multi-sensor RS imagery, still supporting arbitrary band combinations via a hypernetwork that produces projection weights as a function of wavelength. FoMo-Net (Bountos et al., 2023), and more recently Panopticon (Waldmann et al., 2025) employ shared projection function across bands of multiple sensors conditioned by band frequencies. The unified projection of arbitrary spectral ranges in a continuous space, as in these models, has the risk of not only underperforming wavelength discrete projection, but also losing the physical consistency in embedding space for the bands associated with nearby wavelengths. As RS spectra typically vary smoothly for bands with close wavelengths (Wang et al., 2020; Wang & Hu, 2025), preserving this in the resulting embedding space is important for various applications (*e.g.*, biophysical parameter estimation (Verrelst et al., 2019), water quality assessment (Liu et al., 2021), wildfire monitoring (van Gerrevink & Veraverbeke, 2021), etc.).

To recapitulate, previous (multi-sensor) GFMs either assume fixed, predefined band sets, discretize ES into a small number of ranges, or enable processing of arbitrary bands in a wavelength-continuous space without explicitly enforcing spectral consistency. Our work builds on these directions by combining: (i) unified wavelength-conditioned spectral projection; and (ii) wavelength-indexed flow matching for spectral consistency, using hyperspectral pretraining as a natural regime to learn wavelength-continuous embeddings.

3 HYPERSMARTIES

To unify RS image representation learning, we introduce a foundation model named hyperSMARTIES (Fig. 1) that employs: i) a unified wavelength-conditioned spectral projection with attention-based spectral tokenization (Section 3.1); and ii) wavelength-indexed flow matching as a self-supervised spectral consistency regularizer (Section 3.2), for masked data modeling (Section 3.3). Thanks to these, hyperSMARTIES: i) enables unified projection of arbitrary bands across diverse spectral regimes, and thus eliminates the need for sensor-dependent data processing of GFMs; ii) becomes scalable to a high number bands; and iii) encodes physical characteristics of RS image signals by enforcing local consistency of bands with respect to their wavelengths in the embedding space.

3.1 WAVELENGTH-CONTINUOUS SPECTRAL PROJECTION

We consider RS images acquired by heterogeneous spectral sensors. For each image $\mathbf{I} \in \mathbb{R}^{W \times H \times C}$, the set of available spectral bands is denoted by $\mathcal{B}_{\mathbf{I}} = \{b_1, \ldots, b_C\}$, where each band is associated with a central wavelength λ_c expressed in nanometers and locating the band in the electromagnetic spectrum (ES). We note that C may highly vary across sensors and samples (between 3 and 20 for multispectral imaging and in the hundreds for hyperspectral imagery).

To use $\mathbf{I}$ in a vision transformer (ViT), the image needs to be tokenized. For this, hyperSMARTIES employs a single wavelength-continuous spectral projection, which is illustrated in Fig. 1b. To this end, we first divide the image into N non-overlapping image patches $\mathcal{P}_{\mathbf{I}} = \{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$, where each patch $\mathbf{p}_n \in \mathbb{R}^{S \times S \times C}$ is the size of $S \times S$ with C spectral bands in $\mathcal{B}_{\mathbf{p}_n}$. Then, we define a single shared projection function $f : \mathbb{R}^{S \times S} \to \mathbb{R}^D$ that maps any band into a common D-dimensional embedding space through a linear layer, leading to the set of band embeddings $\mathcal{E}_{\mathbf{p}_n} = \{e_1, \ldots, e_C\}$.

Wavelength Conditioning. To encode the physical characteristics of bands, we condition each band embedding e_c of a given patch with its central wavelength λ_c (denoted as wavelength hereafter). To this end, we first separately encode wavelengths with a 1D sinusoidal encoding $\text{WE}(\lambda_c) \in \mathbb{R}^D$ as follows:

$$\text{WE}(\lambda_c, 2k) = \sin\left(\frac{1000 \times \tilde{\lambda}_c}{10000^{2k/D}}\right), \text{WE}(\lambda_c, 2k+1) = \cos\left(\frac{1000 \times \tilde{\lambda}_c}{10000^{2k/D}}\right) \tag{1}$$

where λ_c is min-max normalized over the pretraining spectral range $[\lambda_{\text{min}}, \lambda_{\text{max}}]$, leading to $\tilde{\lambda}_c$. We note that we increase the resolution of sinusoidal encoding by scaling it with 1000, enabling it to represent fine-scale changes in ES. To condition bands with their wavelengths, $\text{WE}(\lambda_c)$ is added to each band embedding e_c, leading to $z_c = e_c + \text{WE}(\lambda_c)$.

Attention-Based Spectral Tokenization. To adaptively aggregate band embeddings into a single token for a given patch, we employ a learnable spectral token [SPC] together with a single spectral transformer block ST that processes a sequence of wavelength-conditioned band embeddings with [SPC] and updates them through attention ($z_c \mapsto z_c^a$):

$$\text{ST} : (z_1, \ldots, z_C, [\text{SPC}]) \to (z_1^a, \ldots, z_C^a, t_{\mathbf{p}_n}) \tag{2}$$

where the updated spectral token $t_{\mathbf{p}_n} \in \mathbb{R}^D$ is used as the final token of the patch $\mathbf{p}_n$ to be processed by the ViT encoder. Projecting all bands of all patches $\mathcal{P}_{\mathbf{I}}$ results in the set of image tokens $\mathcal{T}_{\mathbf{I}} = \{t_{\mathbf{p}_1}, \ldots, t_{\mathbf{p}_N}\}$ to be exploited for masked data modeling (Section 3.3).

3.2 WAVELENGTH-INDEXED FLOW MATCHING FOR SPECTRAL CONSISTENCY

The wavelength-continuous spectral projection of hyperSMARTIES alone does not guarantee physical consistency with respect to wavelengths in the band embedding space despite the fact that RS spectra typically shows smooth, locally consistent evolution across the ES. To preserve such smoothness, we introduce wavelength-indexed flow matching over ES as a spectral consistency regularizer (Fig. 1c). Unlike conventional flow matching used in generative modeling Lipman et al. (2023), our formulation operates directly on band embeddings $\mathcal{E}$ instead of synthetic noise paths. Moreover, the vector field of our flow matching is wavelength-dependent (*i.e.*, wavelength-continuous) rather than time-dependent.

Our formulation learns a wavelength-continuous velocity field $v : [0, 1] \times \mathbb{R}^D \to \mathbb{R}^D$ that defines a wavelength-dependent flow $\psi : [0, 1] \times \mathbb{R}^D \to \mathbb{R}^D$, satisfying:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\psi_\lambda(e) = v_\lambda\left(\psi_\lambda(e)\right) \tag{3}$$

where $\psi_\lambda := \psi(\lambda, e)$. The velocity field $v_\lambda(e_\tau, \lambda_\tau)$ estimates the velocity of the band embedding e_τ given the wavelength λ_τ encoded by sinusoidal encoding $\mathrm{WE}(\lambda_\tau)$ (see Eq. (1)), while v_λ can be modeled by a MLP. To achieve this, for a given C band embeddings of N patches in a mini-batch composed of L samples, we first randomly draw $L \times N$ wavelength samples $\mathcal{W} = \{\lambda_\tau \mid \lambda_\tau \sim U[\lambda_{\text{min}}, \lambda_{\text{max}}]\}_{\tau=1}^{L\times N}$, where min-max wavelengths are defined over the mini-batch. For each λ_τ, we randomly sample two bands (b_i, b_j) whose wavelengths span λ_τ ($\lambda_i < \lambda_\tau < \lambda_j$). Then, the interpolated band embedding e_τ and the corresponding target velocity $v_\lambda(e_\tau, \lambda_\tau)$ (*i.e.*, the derivative over the flow) can be written as follows:

$$e_\tau = \frac{\lambda_j - \lambda_\tau}{\lambda_j - \lambda_i}e_i + \frac{\lambda_\tau - \lambda_i}{\lambda_j - \lambda_i}e_j, \quad v_\lambda(e_\tau, \lambda_\tau) = \frac{e_j - e_i}{\lambda_j - \lambda_i} \tag{4}$$

Accordingly, the flow matching loss $\mathcal{L}_{\text{FM}}$ can be written as the mean squared error (MSE) between all the target and estimated derivatives over $\mathcal{W}$:

$$\mathcal{L}_{\text{FM}} = \frac{1}{N \times L}\sum_{\mathcal{W}} \text{MSE}(v_\lambda, \hat{v}_\lambda) \tag{5}$$

where $\hat{v}_\lambda$ is the prediction of the flow matching network. This objective acts as a regularizer for preventing discontinuities across adjacent bands in the ES. Thus, it enforces smooth, locally consistent evolution of band embeddings across neighboring wavelengths that allows to preserve spectral consistency of image bands in the band embedding space.

It is worth noting that in this section we explain wavelength-indexed flow matching over $\mathcal{W}$ sampled on the bands of all the patches for the sake of simplicity. However, $\mathcal{L}_{\text{FM}}$ is applied on the $\mathcal{W}^{\text{vis}}$ sampled only on the visible patches passed to the ViT backbone (see Section 3.3). We refer readers to Eq. (6) for $\mathcal{L}_{\text{FM}}$ actually used in pretraining with masked data modeling.

3.3 Masked Data Modeling

Once the tokenization is achieved through wavelength-continuous spectral projection, we use a standard encoder-decoder transformer architecture for masked data modeling (Fig. 1d). To this end, by following the vanilla masked autoencoder, we employ *masking*, *encoding*, *decoding*, and *reconstruction* steps, which are described below.

For the set $\mathcal{T}_{\mathbf{I}} = \{t_{\mathbf{p}_1}, \ldots, t_{\mathbf{p}_N}\}$ of image tokens, we apply random masking with a ratio R, leading to $(1-R)N$ visible tokens $\mathcal{T}_{\mathbf{I}}^{\text{vis}}$. We use a ViT backbone (*e.g.*, ViT-B, ViT-L etc.) to encode these visible tokens with the learnable [CLS] token, while sinusoidal positional encoding is used for patch positions. Once the tokens are processed by the backbone, the transformer decoder is applied on both the visible tokens and the special [MASK] tokens, which are used in place of masked patches. Then, the decoder exploits [MASK] tokens with the positional encodings for decoding the corresponding masked patches. Finally, the decoded masked patch representations are reconstructed back to the original spectral channels ($\hat{\mathcal{P}}_{\mathbf{I}}^{\text{mask}}$) through linear mapping. We would like to highlight that, as explained in Section 3.1, we aggregate band-level embeddings of a patch into a single patch token for masked data modeling which significantly reduces the computational cost of transformer data processing. However, this makes the problem of converting a decoded patch representation back to C spectral channels through a single shared function ill-posed. Accordingly, by following SMARTIES (Sumbul et al., 2025), we use one linear reconstruction function per unique spectral range of pretraining spectra, leading to the set of C linear layers $\mathcal{R} = \{r_1, \ldots, r_C\}$, where $r_c : \mathbb{R}^D \to \mathbb{R}^{S\times S}$ takes the decoded token $t_{\mathbf{p}_n}$, and reconstructs the cth spectral channel of a given patch $\mathbf{p}_n$. To learn model parameters of hyperSMARTIES in a self-supervised way, we utilize the Mean Squared Error (MSE) loss $\mathcal{L}_{\text{MSE}}$ between the reconstructed patches of a given mini-batch and the corresponding original masked patches. When the flow matching loss from Eq. (5) over the visible patches is included, the overall pretraining loss $\mathcal{L}$ of a given mini-batch can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \gamma\mathcal{L}_{\text{FM}} = \frac{\sum(\mathcal{P}_{\mathbf{I}}^{\text{mask}} - \hat{\mathcal{P}}_{\mathbf{I}}^{\text{mask}})^2}{RNL} + \gamma\frac{\sum_{\mathcal{W}^{\text{vis}}} \text{MSE}(v_\lambda, \hat{v}_\lambda)}{(1-R)NL} \tag{6}$$

where γ is used for weighting the flow matching loss (see Appendix D for the sensitivity analysis).

4 EXPERIMENTS

4.1 PRETRAINING DATA & EXPERIMENTAL SETUP

Since we aim to learn image representations respecting the smooth and continuous properties of RS spectra, we pretrain hyperSMARTIES on data with maximal spectral resolution (*i.e.*, the highest number of bands C): we therefore focus on hyperspectral imagery, which provides the highest fine-grained spectral sampling to date. Specifically, we use the SpectralEarth (Braham et al., 2025) dataset, which includes 539K images acquired by the EnMAP sensor, covering 415K unique locations, globally distributed. For each location including multiple images with different timestamps, we randomly sample one image per epoch. Each SpectralEarth image contains 202 spectral channels with spatial resolution of 30m and pixel size of 128×128. Wavelengths span frequencies λ between 418nm and 2445nm. For pretraining (PT), we first apply min-max normalization with the 1% and 99% percentile values, and then zero mean, unit variance standardization.

Following the SpectralEarth setup, we pretrain three models of hyperSMARTIES with different depths of backbones (ViT-S, ViT-B and ViT-L) for 200 epochs with patch size of 4 and the fixed seed of 42. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with batch size of 1024 (distributed over 8 H100 GPUs) and base learning rate of 1.5e-4. For each model, one additional transformer block is used as the spectral transformer of hyperSMARTIES. While ViT-B/L models are pretrained by using all the unique locations of SpectralEarth, ViT-S model is pretrained on 5% randomly selected unique locations and used only for ablation and sensitivity analyses. To reduce computational complexity and the risk of overfitting on PT spectra, we randomly select 50, 25 and 50 bands for the backbones of ViT-S, ViT-B and ViT-L, respectively, on each mini-batch during PT. For the flow matching (FM) architecture of all the models, we use 4-layers MLP with the hidden dimension of 128 and the ELU activation function. For PT, the masking ratio R and the $\mathcal{L}_{\text{FM}}$ weighting hyperparameter γ are set to 75% and 0.01, respectively. We refer readers to: 1) Appendix B for the details of PT; and 2) Appendix D for our sensitivity analysis on PT band and time series sampling, γ, FM dimension and depth.

To systematically evaluate hyperSMARTIES across diverse sensors, downstream tasks and scenarios, we perform the following experiments: 1) *Seen sensor generalization* to assess the task-agnostic generalization of hyperSMARTIES, where performance is evaluated across eight different datasets on the same type of imagery (EnMAP) seen during PT; 2) *Unseen sensor generalization* to analyze whether hyperSMARTIES generalizes on new, unseen sensors not present during PT; and 3) *Spectral consistency analysis* to assess the ability of hyperSMARTIES to learn wavelength-dependent spectral consistency over band embeddings. For downstream transfer results, we report mean $\pm$ standard deviation evaluated over three independent runs with seeds 42, 1 and 999. For the details of these experiments and the corresponding datasets, we refer readers to Appendix C.

4.2 SEEN SENSOR GENERALIZATION

In Table 1, we evaluate the downstream generalization capability of hyperSMARTIES (ViT-B), where the target imagery is acquired by the same sensor type as during PT (EnMAP). We consider the tasks of: 1) multi-label scene classification on the CORINE dataset (Braham et al., 2025); 2) pixel-wise regression on the HyBiomass dataset (Banze et al., 2025); and 3) semantic segmentation on the CDL, NLCD, EuroCrops, BDForet, TreeMap, and BNETD datasets (Braham et al., 2025). We follow widely used transfer learning strategies: frozen backbone probing (FP) and finetuning (FT). We compare our model with recent hyperspectral GFMs (SpatSIGMA (Wang et al., 2025a) and SpectralEarth (Braham et al., 2025)) in addition to DOFA (Xiong et al., 2025) and Panopticon (Waldmann et al., 2025) multi-sensor GFMs (whose PT data includes EnMAP images) and consider ViT-B backbone variants only. In addition, we also consider the fully supervised ViT-B model (Braham et al., 2025) for a reference. Results show that hyperSMARTIES consistently outperforms these models on various downstream tasks under both FP and FT. In particular, hyperSMARTIES outperforms SpatSIGMA, DOFA and Panopticon on all datasets and tasks with a significant margin. Compared to SpectralEarth, hyperSMARTIES provides the best results in all experiments except CDL in FP and TreeMap in FT, where our model shows highly competitive performance (*e.g.*, only 0.3% lower mIoU on TreeMap). These results show that the wavelength-continuous spectral projection of hyperSMARTIES, coupled with its attention-based spectral tokenization, significantly improve hyperspectral image representations, leading to effective transfer to various and diverse downstream tasks. One can also observe

Table 1: Seen sensor (EnMAP) generalization results of hyperSMARTIES (ViT-B) through frozen backbone probing and finetuning on CORINE dataset for multi-label scene classification in F_1 score (%↑), HyBiomass dataset for pixel-wise regression in R^2 (↑), and CDL, NLCD, EuroCrops, BDForet, TreeMap, BNETD datasets for semantic segmentation in mIoU (%↑). The highest results are written in bold, while the second best results are underlined.

Model	CORINE	HyBiomass	CDL	NLCD	EuroCrops	BDForet	TreeMap	BNETD
Supervised ViT	77.2 ± 0.7	0.37 ± 0.02	74.7 ± 0.6	45.6 ± 0.1	66.1 ± 0.4	73.6 ± 0.4	44.5 ± 0.1	46.8 ± 0.1
	Frozen Backbone Probing							
SpatSIGMA	68.2 ± 0.3	0.31 ± 0.00	63.5 ± 0.1	36.0 ± 0.1	50.7 ± 0.4	58.2 ± 0.3	37.0 ± 0.1	38.6 ± 0.1
SpectralEarth	75.0 ± 0.3	0.38 ± 0.01	**72.2** ± 0.2	40.8 ± 0.1	61.3 ± 0.2	68.4 ± 0.2	39.7 ± 0.3	43.4 ± 0.1
DOFA	67.8 ± 0.2	0.28 ± 0.00	59.0 ± 0.1	33.5 ± 0.1	45.4 ± 0.5	53.2 ± 0.4	34.6 ± 0.2	37.0 ± 0.1
Panopticon	75.9 ± 0.3	0.34 ± 0.01	58.9 ± 0.1	38.3 ± 0.1	47.9 ± 0.1	59.2 ± 0.2	39.2 ± 0.2	39.5 ± 0.1
hyperSMARTIES	**76.5** ± 0.3	**0.42** ± 0.01	71.7 ± 0.1	**42.4** ± 0.0	**62.5** ± 0.0	**70.1** ± 0.1	**40.9** ± 0.1	**45.4** ± 0.0
	Finetuning							
SpatSIGMA	76.7 ± 0.3	0.41 ± 0.00	70.6 ± 0.6	43.2 ± 0.1	58.3 ± 0.2	69.3 ± 0.4	42.9 ± 0.1	42.5 ± 0.1
SpectralEarth	79.5 ± 0.3	**0.51** ± 0.00	77.4 ± 0.3	47.6 ± 0.2	69.3 ± 0.4	76.3 ± 0.5	**46.1** ± 0.0	49.5 ± 0.1
DOFA	74.5 ± 0.7	0.40 ± 0.00	69.1 ± 0.3	42.6 ± 0.1	57.0 ± 0.3	64.9 ± 0.2	42.0 ± 0.1	42.3 ± 0.1
Panopticon	80.2 ± 0.3	0.36 ± 0.01	68.4 ± 0.2	40.8 ± 0.1	58.0 ± 0.4	65.9 ± 0.3	42.5 ± 0.1	43.9 ± 0.0
hyperSMARTIES	**80.8** ± 0.3	**0.51** ± 0.00	**77.5** ± 0.2	**48.4** ± 0.1	**69.4** ± 0.3	**77.1** ± 0.3	45.8 ± 0.0	**49.9** ± 0.0

Table 2: Results of the models with ViT-L backbone and their relative performance changes (±%) compared to those with ViT-B backbone (ViT-B → ViT-L) under seen sensor generalization with frozen backbone probing on the CORINE, HyBiomass, CDL and NLCD datasets.

Model	CORINE	HyBiomass	CDL	NLCD
SpatSIGMA	68.4 ± 0.3 (+0.3%)	0.27 ± 0.03 (−12.9%)	61.7 ± 0.2 (−2.8%)	34.4 ± 0.3 (−4.4%)
SpectralEarth	74.0 ± 0.2 (−1.3%)	0.42 ± 0.00 (+10.5%)	**72.7** ± 0.2 (+0.7%)	42.7 ± 0.0 (+4.7%)
DOFA	67.5 ± 0.1 (−0.4%)	0.28 ± 0.00 (0.0%)	60.6 ± 0.3 (+2.7%)	33.4 ± 0.1 (−0.3%)
hyperSMARTIES	**77.2** ± 0.3 (+0.9%)	**0.44** ± 0.01 (+4.8%)	72.6 ± 0.3 (+1.3%)	**43.4** ± 0.0 (+2.4%)

that FT allows hyperSMARTIES to characterize band embeddings and aggregate them based on the specific needs of a task, since the improvements observed under FP largely persist also under FT.

Backbone Scaling. We analyze scaling from ViT-B to ViT-L under FP on four datasets. Table 2 shows the performance of ViT-L models and the relative change with respect to ViT-B results. hyperSMARTIES is the only model always benefiting from backbone scaling: it improves the results of all datasets, while achieving the strongest absolute results across the compared models on three datasets. These results indicate that hyperSMARTIES shows favorable scaling behavior when increasing backbone capacity, consistent with its goal of unified image representation learning that remains effective as the model size grows (see the effect of PT band sampling on this in Appendix D).

4.3 Unseen Sensor Generalization

To evaluate whether hyperSMARTIES can transfer beyond the hyperspectral EnMAP PT sensor, in Table 3 we assess unseen-sensor generalization on four downstream datasets spanning distinct sensors and spectral characteristics: 1) Hyperview (Nalepa et al., 2022) including HySpex hyperspectral images for scene-level regression; 2) BigEarthNet-S2 (BEN-S2) (Sumbul et al., 2021) including Sentinel-2 multispectral images for multi-label scene classification; 3) m-chesapeake (m-CP) (Lacoste et al., 2023) including NAIP very high resolution (VHR) RGB images for semantic segmentation; and 4) EO-1 CDL (Braham et al., 2025) including EO-1 Hyperion hyperspectral images for semantic segmentation. We follow the FP transfer learning strategy to study only the contribution of the representations learned, without potential effects due to finetuning on the downstream task. In addition to the GFMs from Table 1, we also consider SMARTIES (Sumbul et al., 2025) and Copernicus-FM (Wang et al., 2025b) multi-sensor GFMs for comparison in Table 3, since they have also shown unseen sensor transfer capabilities. One can observe from the results that hyperSMARTIES achieves the best scores on Hyperview, EO-1 CDL and m-CP and the second best score on BEN-S2, showing a high unseen sensor generalization across hyperspectral, multispectral and VHR RGB imagery without a high increase in number of parameters. In particular, hyperSMARTIES provides a 0.4% higher mAP on BEN-S2 compared to SMARTIES, which included BEN-S2 in its PT data. Similarly, compared to DOFA, hyperSMARTIES improves results by more than 4% in mIoU on m-CP with NAIP imagery, which is included in DOFA PT. We note that hyperSMARTIES without Sentinel-2 (S2) PT achieves

Table 3: Unseen sensor generalization results of hyperSMARTIES (ViT-B) through frozen backbone probing on: Hyperview dataset with HySpex hyperspectral images for scene-level regression in MSE_{norm} (↓); BEN-S2 dataset with Sentinel-2 multispectral images for multi-label scene classification in mAP (%↑); m-CP dataset with NAIP RGB images; and EO-1 CDL dataset with EO-1 Hyperion hyperspectral images for semantic segmentation in mIoU (%↑). PT denotes pretraining.

Model	Params (M)	Hyperview	EO-1 CDL	BEN-S2	m-CP
		Same Sensor PT / Same Modality PT			
SpatSIGMA	119.0	✗ / ✓	✓ / ✓	✗ / ✗	✗ / ✗
SpectralEarth	87.5	✗ / ✓	✗ / ✓	✗ / ✗	✗ / ✗
DOFA	113.1	✗ / ✓	✗ / ✓	✓ / ✓	✓ / ✓
Panopticon	99.0	✗ / ✓	✗ / ✓	✓ / ✓	✗ / ✓
Copernicus-FM	139.5	✗ / ✗	✗ / ✗	✓ / ✓	✗ / ✗
SMARTIES	88.6	✗ / ✗	✗ / ✗	✓ / ✓	✗ / ✓
hyperSMARTIES	93.0	✗ / ✓	✗ / ✓	✗ / ✗	✗ / ✗
SpatSIGMA		0.88 ± 0.00	57.3 ± 0.2	69.4 ± 0.0	65.1 ± 0.1
SpectralEarth		0.85 ± 0.01	59.0 ± 0.6	76.4 ± 0.1	73.0 ± 0.2
DOFA		0.89 ± 0.00	53.3 ± 0.8	79.4 ± 0.1	71.8 ± 0.1
Panopticon		0.83 ± 0.01	54.8 ± 0.6	**84.5** ± 0.0	75.6 ± 0.2
Copernicus-FM		0.92 ± 0.00	51.4 ± 0.1	82.6 ± 0.0	73.1 ± 0.1
SMARTIES		0.83 ± 0.00	57.1 ± 0.4	82.3 ± 0.0	73.9 ± 0.1
hyperSMARTIES		**0.82** ± 0.00	**63.9** ± 0.2	82.7 ± 0.0	**76.3** ± 0.1

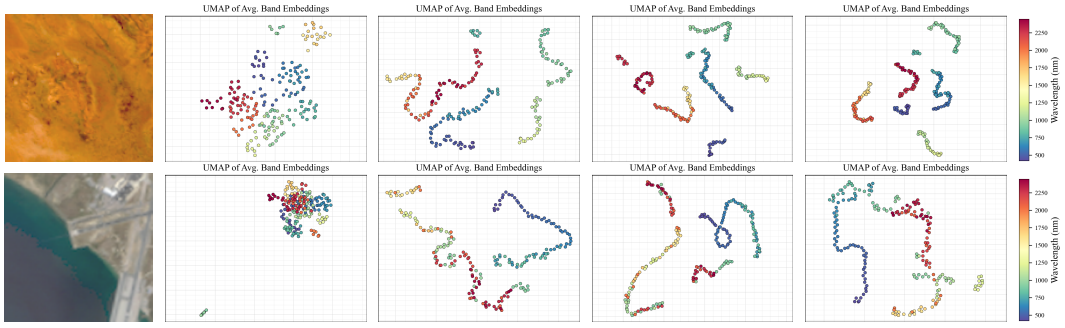
only less than 2% lower mAP on BEN-S2 compared to Panopticon, which is pretrained with 2+ million S2 images. These results show that once hyperSMARTIES is pretrained on images with high spectral granularity (hyperspectral), it can use its spectral projection to leverage the continuity of the spectra, which is further supported by its wavelength-indexed FM. This allows the encoded dense spectral structure to be re-used when transferring to other sensors with lower spectral granularity.

4.4 Ablation & Spectral Consistency Analyses

Table 4: Ablation for hyperSMARTIES (ViT-S) through frozen backbone probing. FM, FDR, PCC and PT denote flow matching, finite difference regularization, Pearson correlation coefficient and pretraining, respectively.

Projection	Spectral Consistency	Wavelength Conditioning	PT Params (M) / GFLOPs	PT PCC	CDL	BEN-S2
Discrete	✗	✗	28.6 / 22.2	0.12	64.2	77.1
Continuous	✗	✗	28.1 / 69.9	0.44	60.0	76.9
Continuous	✗	✓	28.1 / 69.9	0.41	64.9	75.7
Continuous	FDR	✓	28.1 / 69.9	0.39	64.6	76.6
Continuous	FM	✓	28.2 / 69.9	**0.47**	**65.8**	**77.3**

In Table 4, we compare discrete vs. wavelength-continuous projection and, within the continuous setting, we also isolate wavelength conditioning and spectral consistency. The latter is enforced either by our wavelength-indexed flow matching (FM) or by finite difference regularization (FDR), a simpler alternative that directly penalizes the embedding distance between bands with neighboring wavelengths. All variants are evaluated with frozen backbone probing on CDL (seen sensor) and BEN-S2 (unseen sensor). To measure how the band embedding space is organized with respect to the electromagnetic spectrum (ES), we also report the Pearson correlation coefficient (PCC) between band embedding and wavelength distances among the pairs of different EnMAP bands in the PT data. Discrete projection, corresponding to SMARTIES trained under identical settings, is competitive in terms of accuracy, while leading to an almost unstructured band embedding space (PCC of 0.12): since each ES partition has its own projection function, the embeddings of bands falling into different partitions are not related. Replacing it with a single shared, wavelength-continuous projection considerably increases the wavelength alignment of the embedding space (PCC of 0.44), yet it reduces seen sensor performance by 4.2% on CDL. This confirms the risk discussed in Section 2.2 that a continuous projection can underperform a discrete one without any indication of band placement in the ES. Wavelength conditioning removes this gap and even surpasses the discrete variant on CDL (64.9%). However, conditioning alone does not constrain the geometry of the resulting space (PCC from 0.44 to 0.41), and the unseen sensor performance drops (75.7% on BEN-S2). Explicitly enforcing spectral consistency addresses this: FDR recovers part of the unseen sensor performance (76.6%) but improves neither CDL nor the PCC. In contrast, our wavelength-indexed FM achieves the best results on all three metrics. Modeling how band embeddings change along the ES, rather than only how far apart they are, improves the physical fidelity of the embedding space and the downstream accuracy on both seen and unseen sensors. This implies that wavelength-continuous projection, wavelength conditioning, and spectral consistency are complementary to each other.



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

Figure 2: Band embedding geometry analysis through UMAP projections of band embeddings (averaged over patches) of (a) EnMAP images obtained by: (b) discrete projection; (c) continuous projection; (d) continuous projection with wavelength conditioning; and (e) hyperSMARTIES, continuous projection with wavelength conditioning and spectral consistency through flow matching.

To further assess the spectral consistency of hyperSMARTIES through the band embedding geometry, Fig. 2 shows UMAP (McInnes et al., 2018) projections of the band embeddings (averaged over all the patches) of two EnMAP images with different characteristics, where each point corresponds to a band colored by its central wavelength. With discrete projection (Fig. 2b), the embeddings form a diffuse cloud in which wavelengths are strongly intermixed, especially for the bottom image, consistent with the low PCC. Continuous projection (Fig. 2c) already produces partially ordered structures, but the trajectories are fragmented and the wavelength ordering is frequently broken. Adding wavelength conditioning (Fig. 2d) leads to cleaner wavelength gradients, although several bands remain scattered away from the main structures. Finally, hyperSMARTIES (Fig. 2e) organizes the bands into smooth, filament-like trajectories along which the wavelength varies gradually, so that bands with neighboring wavelengths are also neighbors in the embedding space, which is consistent with the aim of the FM. These examples, as well as the additional ones given in Appendix F, supports that, due to its wavelength-indexed FM objective, hyperSMARTIES learns a band embedding geometry aligned with the central wavelengths of bands. This is essential since many RS applications rely on subtle spectral information: enforcing local consistency in the embedding space encodes physically consistent spectral relationships during PT, and thus facilitates transfer to new tasks with seen/unseen sensors.

Regarding computational cost, wavelength-continuous projection variants slightly reduce the number of PT parameters, while FM comes with negligible increase. The higher PT GFLOPs of the continuous variants stems from the self-attention used in the spectral transformer of hyperSMARTIES. This is an intentional choice rather than, for instance, cross-attention (see Table 6 for comparison), allowing interactions among bands. To assess the effectiveness of spectral transformer through [SPC] token attention analysis, we refer readers to Appendix E: [SPC] attention profiles are consistent with established remote sensing physics, while they differ between tasks in a physically consistent way.

5 Conclusion

Learning general purpose RS image representations that can transfer across not only tasks and geographic regions, but also multiple sensors with variable spectral characteristics is of great importance to empower EO capabilities. In this paper, we introduce hyperSMARTIES, a unified foundation model for wavelength-continuous masked data modeling in RS. It allows to: 1) lift the dependency on embedding RS data based on predefined spectra through its wavelength-continuous spectral projection; and 2) preserve physical characteristics of RS spectra by enforcing locally consistent changes of band embeddings with respect to the central wavelengths of bands. Pretraining hyperSMARTIES on high spectral resolution hyperspectral imagery enables it to leverage the dense spectral sampling, leading to highly transferable representations, effective in addressing diverse and challenging downstream tasks under multiple seen/unseen sensor setups. hyperSMARTIES is also capable of encoding physically consistent fine-grained spectral relationships, essential for many EO applications relying on subtle spectral information.

AI USE STATEMENT

We used Large Language Models (LLMs) as assistive tools during the writing of this paper. In detail, LLMs were only employed to polish grammar and to improve clarity. LLMs were not involved in research idea conceptualization, methodology and experimental design.

REPRODUCIBILITY STATEMENT

We have taken the following steps to ensure our results are reproducible. The main downstream transfer results are reported as mean $\pm$ standard deviation over three independent runs with seeds 42, 1, and 999, while the pretraining stage uses the single fixed seed of 42 as described in Section 4.1. The architectural details are given in Sections 3 and 4.1, while further details and the complete hyperparameter settings for both pretraining and downstream transfer are given in Appendices B and C. Upon acceptance, we will make a public code release under an open-source license, with fixed random seeds and pinned dependency versions so that our experimental setup can be reproduced exactly. We will additionally release not only the pretrained model weights but also the finetuned weights for all downstream datasets.

REFERENCES

Guillaume Astruc, Nicolas Gonthier, Clément Mallet, and Loic Landrieu. AnySat: One earth observation model for many resolutions, scales, and modalities. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 19530–19540, 2025.

Grayson Badgley, Christopher B. Field, and Joseph A. Berry. Canopy near-infrared reflectance and terrestrial photosynthesis. *Science Advances*, 3(3):e1602244, 2017. doi: 10.1126/sciadv.1602244.

Aaron Banze, Timothée Stassin, Nassim Ait Ali Braham, Rıdvan Salih Kuzu, Simon Besnard, and Michael Schmitt. Hybiomass: Global hyperspectral imagery benchmark dataset for evaluating geospatial foundation models in forest aboveground biomass estimation. *IEEE Geosci. Remote Sens. Lett. (GRSL)*, 22:1–5, 2025.

Claire Boryan, Zhengwei Yang, Rick Mueller, and Mike Craig. Monitoring US agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program. *Geocarto International*, 26(5):341–358, 2011.

Nikolaos Ioannis Bountos, Arthur Ouaknine, and David Rolnick. Fomo-bench: a multi-modal, multi-scale and multi-task forest monitoring benchmark for remote sensing foundation models. *arXiv preprint arXiv:2312.10114*, 2023.

Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, Yi Wang, and Xiao Xiang Zhu. SpectralEarth: Training hyperspectral foundation models at scale. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. (JSTARS)*, 18:16780–16797, 2025.

S. Chabrillat, Luis Guanter, Hermann Kaufmann, S. Förster, Alison Beamish, Arlena Brosinsky, Hendrik Wulf, Saeid Asadzadeh, M. Bochow, Niklas Bohn, Nina Bösche, A. Bracher, Maximilian Brell, Henning Buddenbaum, D. Cerra, S. Fischer, T. Hank, U. Heiden, B. Heim, W. Heldens, J. Hill, Andre Hollstein, P. Hostert, H. Krasemann, L. LaPorta, P. J. Leitão, S. van der Linden, W. Mauser, Robert Milewski, M. Mottus, A. Okujeni, N. Oppelt, N. Pinnel, S. Roessner, R. Röttgers, A. Schickling, T. Schneiderhan, M. Soppa, K. Staenz, and K. Segl. EnMAP science plan. Technical report, GFZ Data Services, Potsdam, 2022.

Jan G. P. W. Clevers and Anatoly A. Gitelson. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation*, 23:344–351, 2013.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *Int. Conf. Learn. Represent. (ICLR)*, 2016.

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)*, pp. 197–211, 2022.

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.

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)*, pp. 5506–5538, 2023.

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 Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 27662–27673, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 16000–16009, 2022.

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.

Nicolas Houdré, Diego Marcos, Hugo Riffaud de Turckheim, Dino Ienco, Laurent Wendling, Camille Kurtz, and Sylvain Lobry. RAMEN: Resolution-adjustable multimodal encoder for earth observation. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2026.

Ziyue Huang, Hongxi Yan, Qiqi Zhan, Shuai Yang, Mingming Zhang, Chenkai Zhang, YiMing Lei, Zeming Liu, Qingjie Liu, and Yunhong Wang. A survey on remote sensing foundation models: From vision to multimodality. *arXiv preprint arXiv:2503.22081*, 2025.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *Int. Conf. Comput. Vis. (ICCV)*, 2025.

Shruti Khanna, Alicia Palacios-Orueta, Michael L. Whiting, Susan L. Ustin, David Riaño, and Javier Litago. Development of angle indexes for soil moisture estimation, dry matter detection and land-cover discrimination. *Remote Sensing of Environment*, 109(2):154–165, 2007. doi: https://doi.org/10.1016/j.rse.2006.12.018.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan David Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Andrew Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, Mehmet Gunturkun, Gabriel Huang, David Vazquez, Dava Newman, Yoshua Bengio, Stefano Ermon, and Xiao Xiang Zhu. GEO-Bench: Toward foundation models for earth monitoring. *arXiv preprint arXiv:2306.03831*, 2023.

Xuyang Li, Danfeng Hong, and Jocelyn Chanussot. S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data. In *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 27696–27705, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Int. Conf. Learn. Represent. (ICLR)*, 2023.

Xiaohan Liu, Christopher Steele, Stefan Simis, Mark Warren, Andrew Tyler, Evangelos Spyrakos, Nick Selmes, and Peter Hunter. Retrieval of chlorophyll-a concentration and associated product uncertainty in optically diverse lakes and reservoirs. *Remote Sensing of Environment*, 267:112710, 2021.

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

Siqi Lu, Junlin Guo, James R. Zimmer-Dauphinee, Jordan M. Nieusma, Xiao Wang, Parker VanValkenburgh, Steven A. Wernke, and Yuankai Huo. Vision foundation models in remote sensing: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 13(3):190–215, 2025.

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)*, pp. 9414–9423, October 2021.

L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Jakub Nalepa, Bertrand Le Saux, Nicolas Longépé, Lukasz Tulczyjew, Michal Myller, Michal Kawulok, Krzysztof Smykala, and Michal Gumiela. The hyperview challenge: Estimating soil parameters from hyperspectral images. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 4268–4272, 2022.

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)*, pp. 164–182, 2024.

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 Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 27811–27819, 2024.

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)*, pp. 4088–4099, 2023.

Karin L Riley, Isaac C Grenfell, John D Shaw, and Mark A Finney. TreeMap 2016 dataset generates conus-wide maps of forest characteristics including live basal area, aboveground carbon, and number of trees per acre. *Journal of Forestry*, 120(6):607–632, 2022.

Maja Schneider, Tobias Schelte, Felix Schmitz, and Marco Körner. EuroCrops: The largest harmonized open crop dataset across the european union. *Scientific Data*, 10:612, 2023.

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.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *Int. Conf. Comput. Vis. (ICCV)*, pp. 5569–5578, 2025.

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

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *Int. Conf. Mach. Learn. (ICML)*, 2021.

Max J. van Gerrevink and Sander Veraverbeke. Evaluating the hyperspectral sensitivity of the differenced normalized burn ratio for assessing fire severity. *Remote Sensing*, 13(22), 2021.

Jochem Verrelst, Zbyněk Malenovský, Christiaan Van der Tol, Gustau Camps-Valls, Jean-Philippe Gastellu-Etchegorry, Philip Lewis, Peter North, and Jose Moreno. Quantifying vegetation biophysical variables from imaging spectroscopy data: A review on retrieval methods. *Surveys in Geophysics*, 40(3):589–629, 2019.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pp. 2204–2214, 2025.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, Chuan Fu, Hongruixuan Chen, Chengxi Han, Naoto Yokoya, Jing Zhang, Minqiang Xu, Lin Liu, Lefei Zhang, Chen Wu, Bo Du, Dacheng Tao, and Liangpei Zhang. HyperSIGMA: Hyperspectral intelligence comprehension foundation model. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, pp. 1–18, 2025a.

Qi Wang, Fahong Zhang, and Xuelong Li. Hyperspectral band selection via optimal neighborhood reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8465–8476, 2020.

Ruike Wang and Jing Hu. Fewer interpretable bands via self-supervision for hyperspectral anomaly detection. *Neurocomputing*, 616:128964, 2025.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision. In *Int. Conf. Comput. Vis. (ICCV)*, pp. 9888–9899, 2025b.

Aoran Xiao, Weihao Xuan, Junjue Wang, Jiaxing Huang, Dacheng Tao, Shijian Lu, and Naoto Yokoya. Foundation models for remote sensing and earth observation: A survey. *IEEE Geosci. Remote Sens. Magazine (GRSM)*, 13(4):297–324, 2025.

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

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 foundation model for observing the earth crossing modalities. *arXiv preprint arXiv:2403.15356*, 2025.

Hao Zhang, Jin-Jian Xu, Hong-Wei Cui, Lin Li, Yaowen Yang, Chao-Sheng Tang, and Niklas Boers. When geoscience meets foundation models: Toward a general geoscience artificial intelligence system. *IEEE Geosci. Remote Sens. Magazine (GRSM)*, 13(4):79–118, 2025.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J. Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of Earth foundation models. *Nature Communications Earth & Environment*, 7:103, 2026. doi: 10.1038/s43247-025-03127-x.

A APPENDIX

The following appendices provide detailed information for pretraining (PT) in Appendix B and downstream transfer across different datasets in Appendix C. We also provide our sensitivity analysis in Appendix D. In addition, we assess the effectiveness of the spectral transformer of hyperSMARTIES through [SPC] token attention analysis in Appendix E, while the additional results for the band embedding geometry analysis through UMAP projections are given in Appendix F. Our code and the pretrained models of hyperSMARTIES will be publicly available upon acceptance.

B PRETRAINING

We pretrain three models of hyperSMARTIES by using ViT-S, ViT-B and ViT-L backbones (Touvron et al., 2021; Dosovitskiy et al., 2021): hyperSMARTIES (ViT-L), hyperSMARTIES (ViT-B) and hyperSMARTIES (ViT-S) (which is used only for ablation and sensitivity analyses).

While hyperSMARTIES (ViT-B) and hyperSMARTIES (ViT-L) are pretrained on all the unique locations of SpectralEarth by using the same decoder architecture with the vanilla MAE (He et al., 2022), hyperSMARTIES (ViT-S) is pretrained on randomly selected 5% of all unique locations by using a lightweight decoder with the depth of 1. For each model, as the spectral transformer block of hyperSMARTIES, one additional transformer block is used by following the same embedding dimension and number of heads of the corresponding backbone. We note that self-attention is preferred instead of cross-attention in the spectral transformer block, allowing hyperSMARTIES to explicitly model interactions among wavelength-conditioned band embeddings with the cost of increased computing time (see Table 6 for the comparison of self-attention vs. cross-attention). Accordingly, to not only reduce computational complexity but also prevent the risk of overfitting on PT spectra (see Table 5a), we randomly select 25 bands of each mini-batch for hyperSMARTIES (ViT-B) PT. For hyperSMARTIES (ViT-L), we increase the number of randomly selected bands to 50 in parallel with the increased backbone and spectral transformer capacity. However, we intentionally use 50 bands for hyperSMARTIES (ViT-S) PT to prevent potential underfitting, as only 5% of SpectralEarth locations are used for PT. For the flow matching (FM) architecture of all the models, we use 4-layers MLP with the hidden dimension of 128 and the ELU activation function (Clevert et al., 2016), while its input is the summation of band embeddings and the sinusoidal encoding of band wavelengths (see Eq. (1)).

We pretrain all the models for 200 epochs using AdamW optimizer (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 0.05), mixed precision (FP16) and the single fixed seed of 42. For PT, we use the batch size of 1024 distributed over 8 H100/H200 GPUs. The base learning rate is set to 1.5e-4, which is linearly scaled with respect to the base batch size of 256. In terms of learning rate scheduling, we consider warmup of 20 epochs and cooldown by half-cosine decay schedule for 180 epochs. We set the patch size to 4 by following SpectralEarth model PT. The masking ratio R and the $\mathcal{L}_{\text{FM}}$ weighting hyperparameter γ are set to 75% and 0.01, respectively. For data augmentation during PT, we apply random vertical and horizontal flipping with 50% probability in addition to random rotation by $k\times 90^\circ$ degrees with $k \in \{0, 1, 2, 3\}$. Then, by randomly sampling scaling parameter between 0.25 and 1, we randomly crop and resize images while keeping the original width-height ratio.

C DOWNSTREAM TRANSFER

Once the pretraining of hyperSMARTIES models is done, the resulting models except decoders and the FM MLPs are used for downstream transfer. To this end, we consider frozen backbone probing (FP) and full finetuning (FT) as the transfer learning strategies for 100 epochs using AdamW optimizer (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay of 1e-4) on 1 H100/H200 GPU with the the batch size of 8. The base learning rate is set to 1e-3, which is linearly scaled only for FT with respect to the base batch size of 256. For FT, we warmup for 5 epochs and cooldown by 95 epochs with half-cosine decay schedule. For FP, however, the learning rate is decayed $10\times$ at epochs 60 and 80. For FT, we utilize drop path with the rate of 0.2. For the datasets whose image pixel size does not match with the input image size of hyperSMARTIES (128×128), we resize the corresponding images to the input image size with bi-cubic interpolation. While we

apply random cropping and resizing (as in PT) only for FT, we apply random vertical and horizontal flipping in addition to random rotation for both FP and FT.

Throughout the experiments, we consider 4 different tasks on 12 different datasets: 1) multi-label scene classification on the BigEarthNet-S2 (BEN-S2) (Sumbul et al., 2021) and CORINE (Braham et al., 2025) datasets; 2) pixel-wise regression on the HyBiomass (Banze et al., 2025) dataset; 3) scene-level regression on the Hyperview (Nalepa et al., 2022) dataset; and 4) semantic segmentation on m-chesapeake (m-CP) (Lacoste et al., 2023), CDL, NLCD, EuroCrops, BDForet, TreeMap, BNETD and EO-1 CDL (Braham et al., 2025) datasets.

For a fair comparison with other models, we follow the previous works to use same evaluation protocols, the same task heads and the same splits of datasets by following the official splits. In detail, for the task head of BEN-S2 and CORINE datasets we use a linear classifier. As the task head for HyBiomass, CDL, NLCD, EuroCrops, BDForet, TreeMap, BNETD and EO-1 CDL datasets, we use two layers of upscaling and convolution as in Braham et al. (2025). We utilize UPerNet segmentation head (Xiao et al., 2018) for m-CP by following Xiong et al. (2025). For the Hyperview dataset, we use 3-layers MLP (with the backbone embedding dimension as the first hidden dimension, which is divided by two for the second hidden dimension) by following Braham et al. (2025). The same task head is used across all the compared models for each dataset. For the compared hyperspectral models accepting fixed number of bands (SpatSIGMA and SpectralEarth), we duplicate the bands of BEN-S2 and m-CP as many as needed to match the input number of bands. For downstream transfer results, we report mean $\pm$ standard deviation evaluated over three independent runs with seeds 42, 1 and 999. Below, we explain each dataset and, when applicable we provide dataset specific experimental setup:

- **BEN-S2** includes over 590K Sentinel-2 multispectral images with the pixel size of 120×120. Each BEN-S2 image is associated with a subset of 19 land-use land-cover classes, while containing 12 spectral bands with 10m, 20m and 60m spatial resolutions. We note that by following initial benchmarks on this dataset (*e.g.*, SatMAE (Cong et al., 2022), SeCO (Mañas et al., 2021) *etc.*), we apply FT and FP by using the 10% of the complete training set and evaluating on the entire validation set. For BEN-S2, we set the batch size to 128.
- **CORINE** contains 11K EnMAP hyperspectral images with the pixel size of 128×128. Each image in this dataset is associated with a subset of 19 land-use land-cover classes obtained from the CORINE Land Cover product over Europe. We note that we did not apply random cropping and resizing for FT on this dataset to prevent under-fitting.
- **HyBiomass** is an above-ground forest biomass estimation dataset, including 34K globally distributed EnMAP hyperspectral images with the pixel size of 128×128. Each pixel of a HyBiomass image is associated with its above-ground biomass estimate from the Global Ecosystem Dynamics Investigation (GEDI) mission.
- **Hyperview** is a soil parameter estimation dataset, including 2886 HySpex hyperspectral images with variable pixel sizes. Each image in Hyperview dataset is associated with four soil parameters, while containing 150 spectral channels with wavelengths between 462nm and 938nm.
- **m-CP** consists of 5K NAIP RGBN images with the pixel size of 256×256 and the spatial resolution of 1m. Each pixel in a m-CP image is associated with one of 7 land-cover classes.
- **CDL** is a crop type segmentation dataset, including 1600 EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a CDL image is associated with one of 14 crop classes derived from the USDA Cropland Data Layer (Boryan et al., 2011) over the United States.
- **NLCD** is a land-cover segmentation dataset, including 13500 EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a NLCD image is associated with one of 15 land-cover classes derived from the National Land Cover Database (NLCD) 2019 of the U.S. Geological Survey (USGS).
- **EuroCrops** is a crop type segmentation dataset, including 1800 EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a EuroCrops image is associated with one of 15 crop classes derived from the national agricultural inventories of European countries in EuroCrops product (Schneider et al., 2023).
- **BDForet** is a tree species segmentation dataset, including 2550 EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a BDForet image is associated with one of 12 tree species classes derived from the BDForet V2 product of the Institut Geographique National (IGN), France.

Table 5: Sensitivity analysis for hyperSMARTIES (ViT-S) through frozen backbone probing on CDL and BEN-S2 datasets. FM and PT denote flow matching and pretraining, respectively.

(a) PT band sampling

# of Bands	GFLOPs	CDL	BEN-S2
12	33.4	63.1	76.8
25	45.8	65.0	**77.5**
50	69.9	65.8	77.3
100	119.8	66.9	76.1
202	227.8	**68.6**	71.9

(b) Time series sampling

Sampling	# of Images	CDL
✓	20.5K	65.8
✗	26.6K	**66.5**

(c) $\mathcal{L}_{\text{FM}}$ weight

γ	CDL
0.01	**65.8**
0.1	65.1
1.0	64.6

(d) FM dimension

Dimension	CDL
64	65.3
128	**65.8**
256	65.2

(e) FM depth

Depth	CDL
2	65.0
3	65.7
4	**65.8**
5	65.4

- **TreeMap** is a tree species segmentation dataset, including 10K EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a TreeMap image is associated with one of 24 tree species classes derived from the TreeMap product (Riley et al., 2022) of the USDA Forest Service Rocky Mountain Research Station. We note that we increase the value of the weight decay to 0.05 for FT on this dataset to prevent over-fitting.
- **BNETD** is a land-cover segmentation dataset, including 2100 EnMAP hyperspectral images with the pixel size of 128×128. Each pixel in a BNETD image is associated with one of 10 land-cover classes derived from the BNETD 2020 Land Cover Map of the Center for Geographic Information and Digital from the National Study Office Techniques and Development (BNETD-CIGN) in Ivory Coast. We note that we did not apply random cropping and resizing for FT on this dataset to prevent under-fitting.
- **EO-1 CDL** is a crop type segmentation dataset, including 550 EO-1 Hyperion hyperspectral images with the pixel size of 128×128, the spatial resolution of 30m and 198 spectral channels with wavelengths between 426nm and 2395nm.. Each pixel in a EO-1 CDL image is associated with one of 14 crop classes derived from the USDA Cropland Data Layer (Boryan et al., 2011) over the United States.

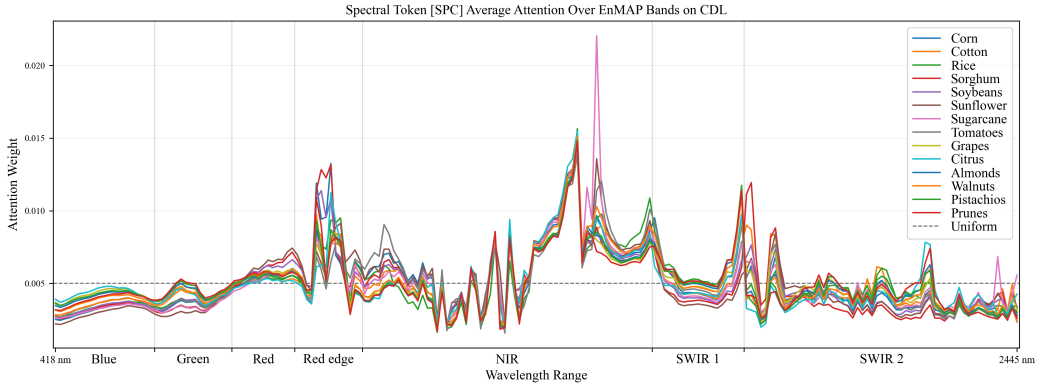
D Sensitivity Analysis

We conduct sensitivity analysis using hyperSMARTIES (ViT-S) pretrained on 5% of the SpectralEarth locations to reduce computational cost. We report FP performance in Table 5.

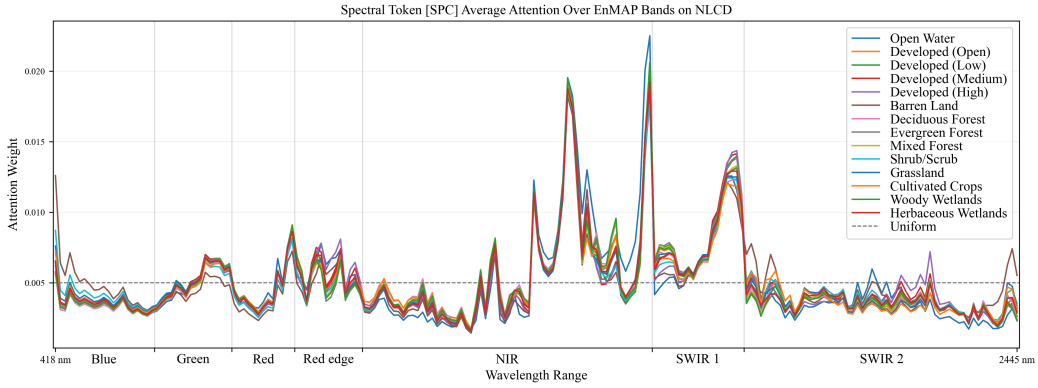
Pretraining band sampling. In Table 5a we analyze the effect of band sampling during PT: seen sensor performance consistently improves together with the increased complexity in terms of GFLOPs as the number of sampled bands increases, while unseen sensor transfer performance on BEN-S2 starts to decrease after a certain number sampled bands. This indicates that moderate sampling (*e.g.*, 25-50) can already provide a strong learning signal and can significantly reduce computational cost of PT, while excessive sampling may encourage overfitting to the PT spectrum. We note that band sampling is applied only during PT, while hyperSMARTIES can still ingest the full band set of any sensor during downstream transfer. Being able to randomly sample and process hyperspectral bands during PT rather than being restricted to the full band set is one of the key features of hyperSMARTIES (thanks to its continuous projection, wavelength conditioning, and FM) that allows scalable PT and unseen sensor generalization at the same time. These are complementary to each other as long as moderate sampling of the PT spectrum is used in hyperSMARTIES. In detail, this also allows hyperSMARTIES to benefit from using more bands during PT in parallel with the increased model capacity (e.g., ViT-B → ViT-L), and thus to show favorable scaling behavior when increasing backbone capacity as shown in Table 2.

Time series sampling. Table 5b compares PT with time-series sampling enabled vs. disabled, showing that time sampling strategy does not significantly decrease the performance, but reduces computational overhead by using more than 20% less images during PT.

Flow matching sensitivity. We assess sensitivity to hyperparameters loss weight γ, hidden dimension, and depth. Table 5c shows that a small loss weight (γ=0.01) performs best among the tested values, while larger weights degrade performance—consistent with the role of FM as a regularizer. Table 5d and Table 5e show that FM hidden dimensionality and depth exhibit mild sensitivity, with the 4-layer MLP and hidden size 128 leading to the highest scores. Overall, these results indicates



(a)



(b)

Figure 3: Average attention weights of the spectral token [SPC] over EnMAP bands in the spectral transformer of hyperSMARTIES, averaged over all test samples of a given class, for ViT-B backbone finetuned independently on: **(a)** CDL crop-type labels; and **(b)** NLCD land-cover labels. The dashed line denotes uniform attention. Wavelength ranges are indicated on the x-axis.

that hyperSMARTIES is relatively stable with respect to moderate changes in FM architecture, but sensitive to overly strong FM loss weighting.

E Spectral Transformer Analysis: [SPC] Token Attention

To assess the effectiveness of the spectral transformer, we conduct [SPC] token attention analysis for hyperSMARTIES (ViT-B) finetuned on: i) CDL in Fig. 3a; and ii) NLCD in Fig. 3b. To this end, we report the attention weights of the [SPC] token in the spectral transformer over the EnMAP bands, averaged over all the test patches of a given class of the corresponding dataset, together with the uniform attention level as a reference.

From both figures, two main observations can be made. First, the attention profiles are strongly non-uniform: attention varies by roughly an order of magnitude across the spectrum, from below 0.002 to above 0.020. Accordingly, the spectral tokenization does not treat the aggregation of bands as an averaging operation, but learns which portions of the electromagnetic spectrum (ES) should dominate the [SPC] token. Second, the profiles are class-dependent: all the classes of each dataset share a common overall shape, which reflects sensor and acquisition characteristics shared across the dataset, but differ in the relative magnitude of the peaks, and these differences are largest exactly in the ranges where attention departs the most from uniform. This shows that a single token per patch remains informative because band aggregation is adaptive to the observed spectral content, while considerably reducing the sequence length processed by the ViT backbone.

On CDL (see Fig. 3a), the red edge and NIR ranges are significantly elevated and carry the largest inter-class spread of any region, while the large portions of the visible and SWIR-2 ranges stay below uniform. On NLCD (see Fig. 3b), relative mass shifts away from the red edge towards the NIR and SWIR-1 ranges with an average increase in the visible range. Although hyperSMARTIES does not receive any spectral prior beyond the central wavelengths of bands, the attention profiles are consistent with established remote sensing physics, while they differ between tasks in a physically consistent way. For the crop-type classification on CDL, the most prominent peaks on attention weights are concentrated on: 1) the red edge, which is the ES region long established as the most diagnostic for chlorophyll content (Clevers & Gitelson, 2013); and 2) the NIR, which is decisive for canopy structure (Badgley et al., 2017). For the land-cover classification on NLCD, by contrast, NIR-SWIR transition region in the ES is associated with the highest attention weights due to its importance for vegetation and soil moisture (Khanna et al., 2007), allowing to discriminate vegetated from non-vegetated, and water from land. Considering together, these figures indicate that the spectral transformer attention does not converge to a single fixed band prior inherited from PT, but re-weights the importance of the ES regions per task in a physically coherent way.

We would like to note that self-attention is preferred instead of cross-attention in the spectral transformer block, which is an intentional choice for allowing the interaction within band embeddings. In Table 6, we compare self-attention vs. cross-attention in the spectral transformer block of hyperSMARTIES (ViT-S). We report the frozen backbone probing performance on CDL (seen sensor) and BEN-S2 (unseen sensor), in addition to PT params and GFLOPs. As it can be seen from the table, both seen and unseen sensor performances significantly drop with cross-attention even though it provides more efficient PT in terms of GFLOPs. This is due to the fact that self-attention of spectral transformer block allows hyperSMARTIES to explicitly model interactions among band embeddings before forming the final patch token, which is necessary for not only downstream performance but also ensuring physical consistency.

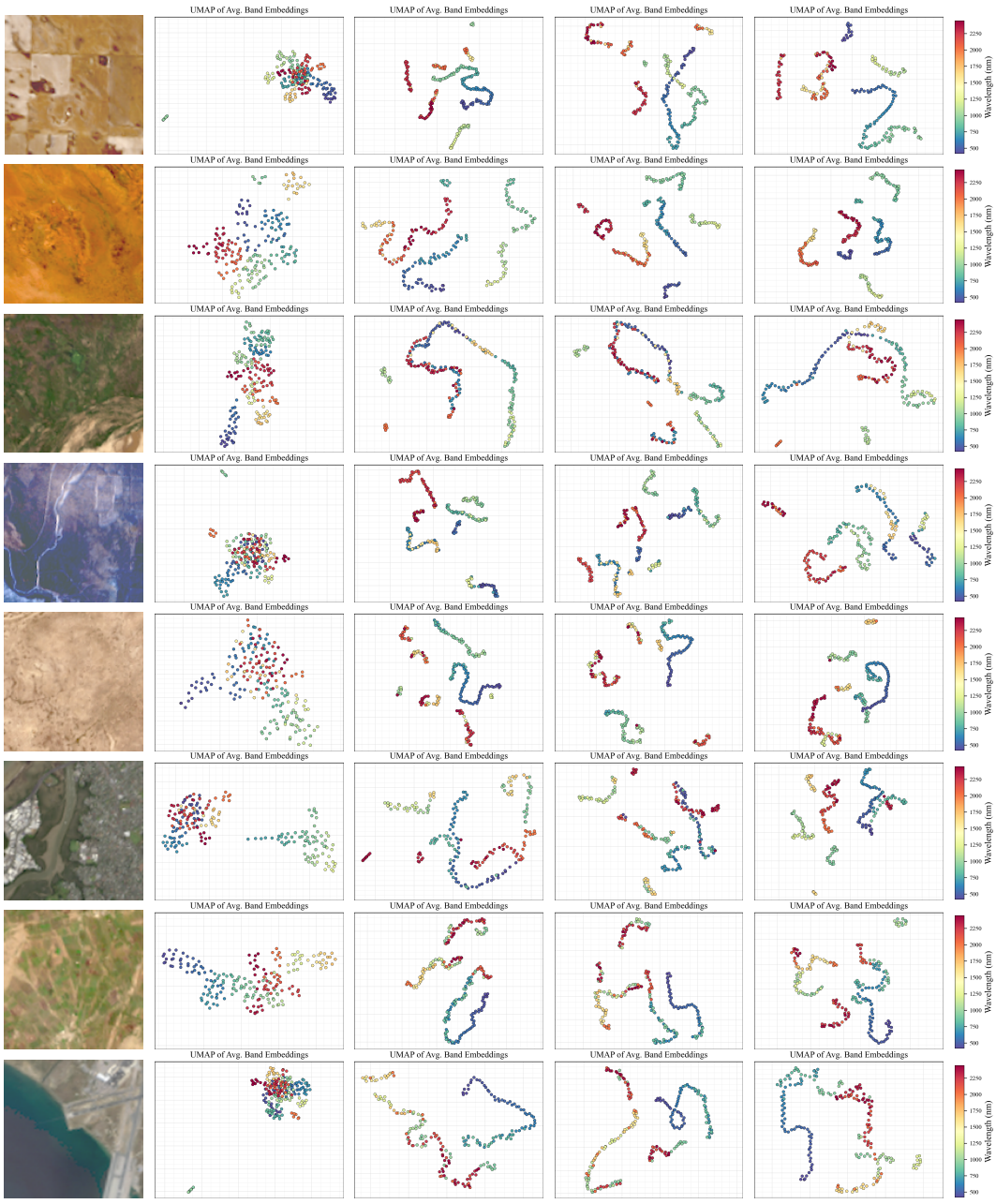
Table 6: Ablation for the spectral transformer block in hyperSMARTIES (ViT-S) through frozen backbone probing. PT denotes pretraining.

Attention Type	PT Params (M) / GFLOPs	CDL	BEN-S2
Cross-attention	28.2 / 57.0	46.4	65.4
Self-attention	28.2 / 69.9	**65.8**	**77.3**

F Additional Results for Band Embedding Geometry Analysis

In the main body of the paper, we analyze whether hyperSMARTIES learns wavelength-dependent local consistency in its band embedding space with the results given in Fig. 2, showing UMAP projections of the band embeddings (averaged over all the patches) of two SpectralEarth images.

In Fig. 4, we provide additional images with higher variation in semantic and spectral characteristics. One can see from the figures that even though band embedding geometry vary depending on image characteristics, these images also support our conclusion from the main body: hyperSMARTIES learns a band embedding geometry aligned with the central wavelengths of bands, rather than treating bands independently, and thus preserves spectral consistency.



(a) Image (b) Discrete Proj. (c) Continuous Proj. (WC ✗; SC: ✗) (d) Continuous Proj. (WC ✓; SC: ✗) (e) Continuous Proj. (WC ✓; SC: ✓)

Figure 4: Additional examples for our band embedding geometry analysis through UMAP projections of band embeddings (averaged over patches) of (a) EnMAP images obtained by: (b) discrete projection; (c) continuous projection; (d) continuous projection with wavelength conditioning (denoted as WC); and (e) hyperSMARTIES, continuous projection with WC and spectral consistency (denoted as SC) through flow matching. All models are based on ViT-S backbone.