SPECTRALEARTH-FM: BRINGING HYPERSPECTRAL IMAGERY INTO MULTIMODAL EARTH OBSERVATION PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Earth observation (EO) foundation models (FMs) are increasingly trained on multisensor data, spanning multispectral imagery (MSI), synthetic aperture radar (SAR), and derived geospatial layers, but hyperspectral imagery (HSI) remains underrepresented. Conversely, most existing hyperspectral FMs are trained on HSI alone, leaving joint pretraining and fusion of HSI with co-located EO sensors underexplored. We introduce SpectralEarth-FM, a hierarchical transformer for multisensor EO input with heterogeneous spectral dimensionality. The architecture combines spectral tokenization for HSI, sensor-specific encoders, a cross-sensor fusion module, and a shared encoder. To pretrain SpectralEarth-FM, we curate SpectralEarth-MM, a multimodal dataset that co-locates HSI from three spaceborne sensors (EnMAP, EMIT, DESIS) with Sentinel-2, Landsat-8/9 optical imagery, Landsat land surface temperature, and Sentinel-1 SAR. It comprises approximately 2 M globally distributed locations, 25 M georeferenced patches, and over 40 TB of data. Pretraining uses a Joint-Embedding Predictive Architecture (JEPA)-style objective that matches representations between global views and single-sensor local views from the same location. We evaluate SpectralEarth-FM on hyperspectral downstream tasks and standard EO benchmarks following the PANGAEA protocol, achieving results that are competitive with state-of-the-art models.

1 INTRODUCTION

Earth observation (EO) foundation models (FMs) increasingly combine multispectral imagery (MSI), synthetic aperture radar (SAR), and geospatial layers such as elevation, vegetation indices, and land-cover maps (Fuller et al., 2023; Jakubik et al., 2023; Astruc et al., 2025; Jakubik et al., 2025; Wang et al., 2025b). In parallel, large-scale pretraining on hyperspectral imagery (HSI) learns transferable spectral–spatial representations for fine-grained spectral discrimination tasks (Wang et al., 2025a; Braham et al., 2025; Baumann et al., 2026). However, joint pretraining of HSI with spatially aligned MSI, SAR, and thermal imagery remains underexplored.

HSI contains high-dimensional, spectrally contiguous measurements with strong inter-band correlations and sensor-dependent noise. HSI models therefore typically use spectral grouping or band-aware tokenization (Hong et al., 2022; Wang et al., 2025a; Braham et al., 2025). Combining HSI with MSI, SAR, and land surface temperature (LST) requires accommodating these spectral characteristics alongside inputs with fewer channels.

We present SpectralEarth-FM, a hierarchical transformer combining HSI spectral tokenizers, sensor-specific branches, multisensor fusion, and a shared encoder. Its Joint-Embedding Predictive Architecture (JEPA)-style objective matches representations of global multisensor views and single-sensor local crops from the same location. Learning in latent space avoids direct reconstruction across sensors with different spectral dimensionalities. For pretraining, we introduce SpectralEarth-MM, pairing EMIT (Green et al., 2020), EnMAP (Guanter et al., 2015), and DESIS (Krutz et al., 2019) HSI with spatially aligned Sentinel-2, Landsat-8/9 optical imagery, Landsat-8/9 LST, and Sentinel-1 SAR. The dataset provides 25 million co-located patches over two million globally distributed locations and over 40 TB of data, with multitemporal coverage where available.

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.

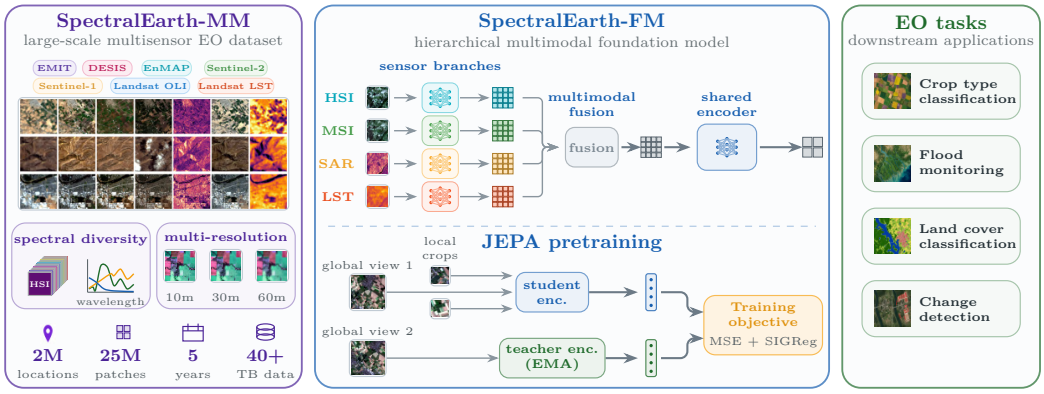

Figure 1: Overview of SpectralEarth-FM architecture and pretraining objective (center), trained on the multimodal corpus SpectralEarth-MM (left) and evaluated on EO benchmarks (right).

We evaluate SpectralEarth-FM on hyperspectral tasks and the PANGAEA benchmarks (Marsocci et al., 2026), obtaining competitive results against recent EO FMs. We also examine transfer to unseen HSI sensors through spectral interpolation to pretrained sensor branches.

This paper makes the following contributions, illustrated in Figure 1:

- **SpectralEarth-MM**, a globally distributed, multitemporal corpus spanning HSI, MSI, SAR, and LST, centered on spaceborne hyperspectral missions.
- **SpectralEarth-FM**, an EO FM based on a hierarchical vision transformer with dedicated spatial–spectral processing and fusion of optical, radar, and thermal observations.
- A **JEPA-style pretraining strategy** that learns from global multisensor views and single-sensor local crops, with evaluation on ten hyperspectral downstream tasks and the PANGAEA benchmarks.

2 Related Work

EO foundation models Early EO self-supervised learning used single optical sensors with seasonal, geographic, or temporal contrastive objectives (Ayush et al., 2021; Manas et al., 2021; Wang et al., 2022). Masked reconstruction is used in SatMAE for temporal multispectral imagery (Cong et al., 2022), Prithvi and Prithvi-EO-2.0 for harmonized Landsat-Sentinel data (Jakubik et al., 2023; Szwarcman et al., 2026), and SpectralGPT with Sentinel-2 spectral–spatial tokenization (Hong et al., 2024). RingMo partially masks image patches to retain information about small objects during pretraining (Sun et al., 2023), while RingMo-Sense predicts future observations from image sequences (Yao et al., 2023). CROMA combines cross-modal contrastive learning with masked reconstruction on paired Sentinel-1/2 observations (Fuller et al., 2023). OmniSat uses sensor-specific encoders and resolution-aware tokenization (Astruc et al., 2024), and AnySat extends this architecture with JEPA-style pretraining (Astruc et al., 2025). MMEarth incorporates derived products and auxiliary tasks (Nedungadi et al., 2024), while TerraMind learns cross-modal prediction across optical, radar, and geospatial layers (Jakubik et al., 2025). TerraFM combines Sentinel-1/2 with modality augmentation and cross-attention fusion (Danish et al., 2026). We extend multisensor pretraining to multiple spaceborne HSI sensors paired with spatially aligned MSI, SAR, and LST.

Hyperspectral foundation models Early work on HSI pretraining combines masked image reconstruction with blockwise patch embeddings, spectral positional embeddings, and spatial–spectral attention (Scheibenreif et al., 2023). HyperSIGMA (Wang et al., 2025a) fuses features from spatial and spectral transformers pretrained on EO-1 Hyperion and Gaofen-5B imagery. SpectralEarth (Braham et al., 2025) combines convolutional spectral adapters, standard vision backbones, and several self-supervised objectives, providing EnMAP pretraining data and HSI benchmarks. HyperSL learns directly from pixel spectra without spatial context (Kong et al., 2025). HyperFM combines

spectral grouping with lightweight within-group and cross-group attention (Tushar & Purushotham, 2026).

Any-sensor foundation models A separate line of work generalizes across sensors using sensor metadata. DOFA generates embeddings from wavelength metadata through a dynamic hypernetwork (Xiong et al., 2024). Copernicus-FM extends this approach to non-spectral Sentinel modalities (Wang et al., 2025b). RAMEN adds scale-conditioned convolutional experts that resample each modality to a common target resolution (Houdré et al., 2026). Panopticon uses channel sub-sampling and early band aggregation (Waldmann et al., 2025). SpecAware conditions HSI token encoding on sensor attributes and image content (Ji et al., 2026). CARL learns camera-agnostic representations across RGB, MSI, and HSI through wavelength-aware encoding (Baumann et al., 2026), and SMARTIES maps heterogeneous MSI bands into a shared spectrum-aware space (Sumbul et al., 2025). SpectralEarth-FM instead learns dedicated branches with explicit spectral tokenization and transfers to unseen HSI sensors by spectral interpolation to pretrained branches.

3 SpectralEarth-MM dataset

SpectralEarth-MM is a large-scale multisensor EO dataset used to pretrain SpectralEarth-FM.

3.1 Dataset overview

SpectralEarth-MM uses spaceborne HSI as the reference observation. It includes HSI from three sensors: EnMAP, EMIT, and DESIS. These sensors provide complementary hyperspectral coverage: DESIS observes the visible and near-infrared (VNIR) range at high spectral resolution, while EnMAP and EMIT cover VNIR and shortwave infrared (SWIR) wavelengths, with different spatial resolutions and acquisition geometries. For each selected HSI footprint, we retrieve spatially aligned Sentinel-2 and Landsat-8/9 optical imagery, Landsat LST, and Sentinel-1 SAR observations.

All observations are organized over a common geographic footprint of $(3.84\,\mathrm{km})^2$. This footprint corresponds to 128^2 pixels for EnMAP, DESIS and Landsat-8/9 at 30 m resolution, 64^2 pixels for EMIT at 60 m resolution, and 384^2 pixels for Sentinel-1 and Sentinel-2 at 10 m resolution. The resulting dataset spans approximately 2 M globally distributed locations, 25 M georeferenced patches, and over 40 TB of data. Multiple HSI acquisition dates are retained for the same geolocation when available, although the temporal density varies across regions and sensors. The selected HSI anchor footprints cover approximately 20 % of the global land surface (Figure 4 in Section A.1).

Table 1: EO pretraining datasets (details in Table 7).

Dataset	HSI	MSI	SAR	LST	Loc. [M]
SatlasPretrain	✗	✓	✗	✗	0.86
SSL4EO-S12	✗	✓	✓	✗	0.25
TerraMesh	✗	✓	✓	✗	9.00
MMEarth	✗	✓	✓	✗	1.20
MajorTOM-Core	✗	✓	✓	✗	1.47
Copernicus-Pretrain	✗	✓	✓	✗	0.31
HySpecNet-11k	✓	✗	✗	✗	0.01
HyperFM250k	✓	✗	✗	✗	0.25
Hyper-400K	✓	✗	✗	✗	0.40
HyperGlobal-450K	✓	✗	✗	✗	0.45
SpectralEarth	✓	✗	✗	✗	0.42
SpectralEarth-MM	✓	✓	✓	✓	2.00

As summarized in Table 1, existing multisensor EO pretraining datasets generally do not include spaceborne HSI, while HSI datasets are typically not paired with other EO sensors over the same locations. SpectralEarth-MM addresses both by pairing spaceborne HSI with multispectral, radar, and LST observations at scale. A detailed comparison with existing EO pretraining datasets is provided in Table 7 (Section A.2).

3.2 Dataset construction

HSI preprocessing and footprint extraction We start from approximately 60 K EnMAP, 33 K EMIT, and 11 K DESIS tiles collected from mission archives. We apply sensor-specific quality control, including georeferencing checks, cloud screening, and removal of noisy spectral bands. The remaining scenes are patchified into non-overlapping $(3.84\,\mathrm{km})^2$ footprints. Patches with missing data are discarded, and the remaining samples are grouped by geographic footprint and timestamp.

Rebalancing HSI locations The raw HSI archive is imbalanced across regions and sensor combinations. We retrieve annual AlphaEarth embeddings (Brown et al., 2025), cluster them hierarchically,

and subsample overrepresented configurations following Vo et al. (2024), preserving scarce configurations and all timestamps at selected locations. The retained archive contains approximately 1.4 M EnMAP, 1.4 M EMIT, and 275 K DESIS locations, with overlap across sensors. Appendix A details the sensor coverage constraints and rebalancing procedure.

Multisensor pairing and temporal filtering We pair the selected HSI footprints with additional observations, namely Sentinel-2, Landsat-8/9 MSI, Landsat LST, and Sentinel-1 SAR. Optical observations are restricted to at most 5% cloud cover. For each HSI timestamp, we retrieve the closest observation that satisfies the cloud cover constraint within predefined time intervals. After pairing, we apply temporal filtering to reduce redundant acquisitions while preserving multiple dates when available. The search windows and temporal filtering criteria are detailed in Appendix A.

4 SpectralEarth-FM

The sensors in SpectralEarth-MM differ in spectral dimensionality and spatial resolution. We model the HSI spectrum before spatial encoding to avoid mixing all bands in a single patch projection. Sensor-specific patch sizes, chosen according to each sensor's spatial resolution, give each token the same ground extent. A hierarchical encoder captures local structure for dense prediction. SpectralEarth-FM combines these components with cross-sensor fusion and supports single- and multisensor inference over pretrained sensors (Figure 2).

4.1 Network architecture

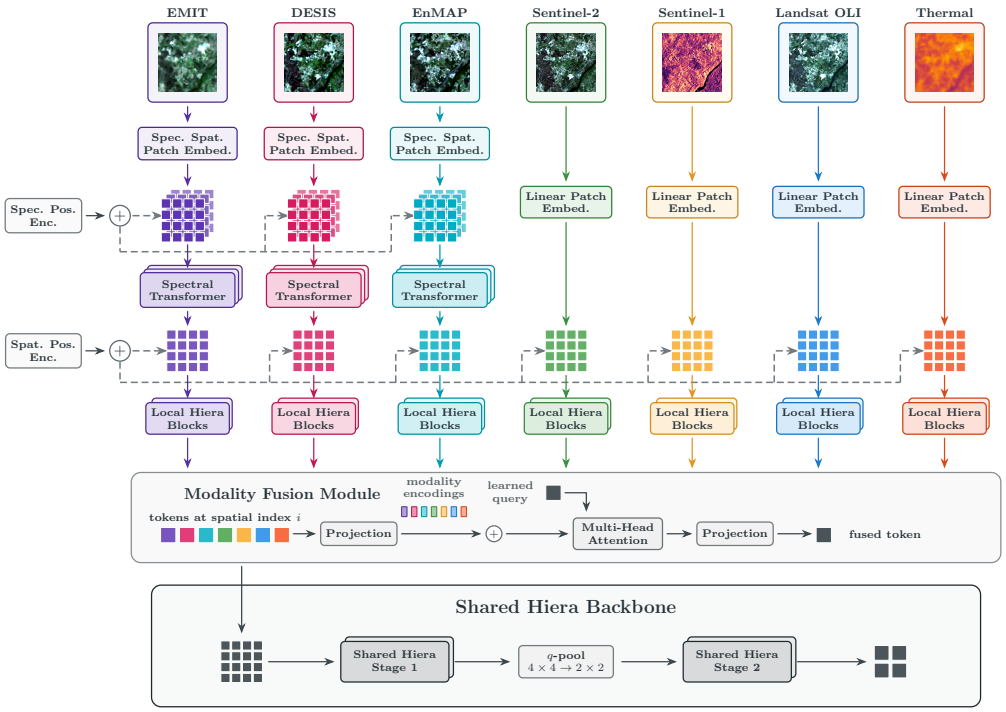

Figure 2: *SpectralEarth-FM architecture.* Sensor-specific branches map inputs to a common spatial token grid using spectral tokenization for HSI and projections for other sensors. Local hierarchical features are fused before entering a shared hierarchical backbone.

Input data Let $\mathcal{S}$ denote the set of pretraining sensors and $\mathcal{A} \subseteq \mathcal{S}$ those available at a location. Input x_s has C_s spectral channels and sensor-specific spatial dimensions. Patch sizes give all branches the same spatial token layout.

Spectral tokenization Each HSI sensor has a spectral tokenizer that partitions the C_s bands per spatial patch into G_s contiguous groups. Dedicated patch projections E_s^{spec} produce G_s tokens per spatial position. Learnable spectral position encodings P_λ preserve band order, and a spectral transformer SpecTr processes the sequence. A learned query attends over the spectral transformer outputs to aggregate them into one spatial token (Agg):

$$T_s = \mathrm{Agg}(\mathrm{SpecTr}(E_s^{\text{spec}}(x_s) + P_\lambda)) \in \mathbb{R}^{L \times D}, \tag{1}$$

where L is the number of spatial patches and D the embedding dimension. Other sensors (Sentinel-2, Landsat-8/9 optical, Sentinel-1, Landsat LST) use a patch projection, $T_s = E_s^{\text{patch}}(x_s) \in \mathbb{R}^{L \times D}$.

Sensor-specific local branches After tokenization, shared spatial position embeddings P_p are added and sensor-specific local hierarchical blocks LHB_s process the tokens to produce branch embeddings:

$$H_s = \mathrm{LHB}_s(T_s + P_p) \in \mathbb{R}^{L_\ell \times D_\ell}. \tag{2}$$

Here $L_\ell \leq L$ is the pooled token count and D_ℓ the branch output and shared-encoder input width. Hiera-style blocks (Ryali et al., 2023) use early window-based attention followed by spatial pooling.

Cross-sensor fusion After local encoding, the branch embeddings $\{H_s : s \in \mathcal{A}\}$ are fused at each spatial position. Each branch embedding is first projected to a fusion dimension D_f and augmented with a learned sensor-specific embedding. A learned query then attends over the resulting sensor tokens, and the output is layer-normalized and mapped back to dimension D_ℓ, producing a single fused token grid $F \in \mathbb{R}^{L_\ell \times D_\ell}$. The fusion module reuses the projected-attention design of the spectral tokenizer, replacing spectral groups with sensor branches.

Shared hierarchical encoder The fused token grid $F \in \mathbb{R}^{L_\ell \times D_\ell}$ is passed to a shared Hiera encoder (Ryali et al., 2023), producing a representation Z. For classification, the final tokens are mean-pooled. For dense prediction, intermediate feature maps from each hierarchical stage can be passed to task-specific decoders.

4.2 Multisensor pretraining

Objective design Pixel-level reconstruction (He et al., 2022) is a challenging pretraining target: HSI measurements are affected by atmospheric conditions, calibration, and sensor noise, and these challenges multiply across modalities that differ in spectral dimensionality, radiometric scale, and noise statistics. Instead, we adopt a latent prediction objective (Assran et al., 2023; Balestriero & LeCun, 2025), where the model predicts latent representations rather than pixel measurements, focusing on semantic structure over low-level sensor characteristics.

View construction Each location in SpectralEarth-MM provides between five and seven modalities depending on HSI sensor coverage. We sample V global views and N local views per location. Global views are spatial crops containing four randomly sampled modalities, with consistent spatial transforms across modalities. Since this subset is resampled at every step, the model is exposed to many sensor combinations rather than a fixed configuration. Local views are smaller spatial crops that retain one optical modality each, excluding SAR and LST. Multi-crop training (Caron et al., 2021) matches local crops to the representation of a larger spatial context. We extend this from spatial to sensor context: each branch learns to match the fused multisensor target from its own input alone. This keeps single-sensor representations useful when only one sensor is available at inference.

Teacher–student objective Learning joint and single-sensor representations requires targets that remain informative despite differences in the views. If the same encoder produced both targets and predictions, the loss could be reduced by discarding information from the fused representation that single-sensor views cannot recover. We address this with an asymmetric teacher–student design (Grill et al., 2020; Chen & He, 2021), illustrated in Figure 3. A teacher network maintained as an exponential moving average (EMA) of the student processes only the global views and defines a stable target. The student processes all views and is trained to match this target.

Let f_θ and p_ϕ denote the student encoder and projector, and let $f_{\bar{\theta}}$ and $p_{\bar{\phi}}$ be the EMA encoder and projector. The teacher target for sample b is the stop-gradient mean of teacher projections over the V

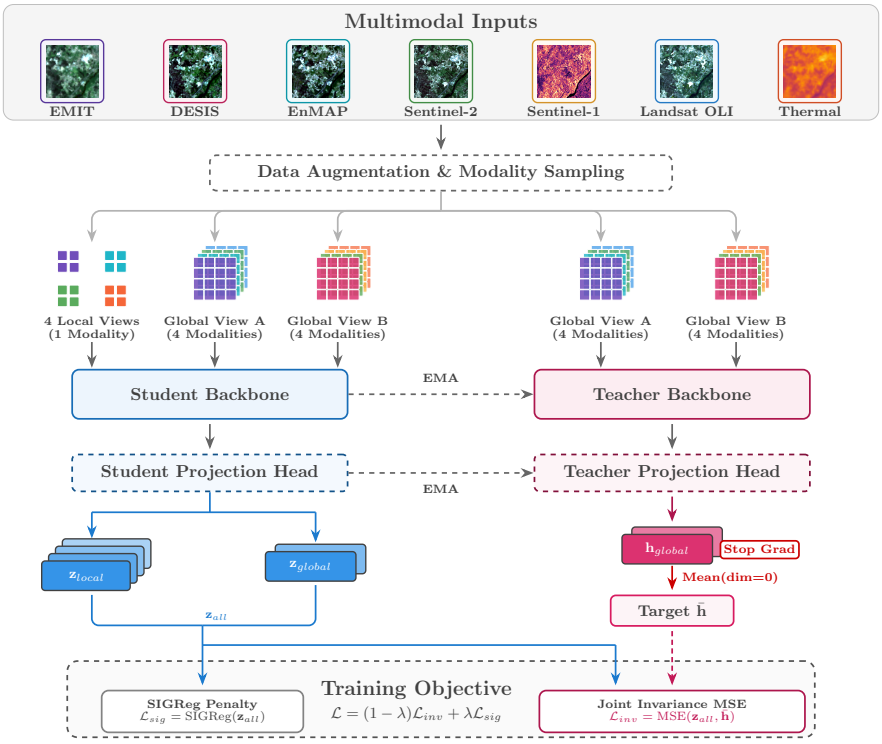

Figure 3: *SpectralEarth-FM pretraining.* Global teacher views define a stop-gradient target. The student matches it from global and single-sensor local views. SIGReg regularizes student projections, with statistics computed separately for each view.

global views:

$$\bar{h}_b = \mathrm{sg}\left(\frac{1}{V}\sum_{v=1}^{V} p_{\bar{\phi}}(f_{\bar{\theta}}(v_b^g))\right). \tag{3}$$

The student produces projections $z_{v,b} = p_\phi(f_\theta(v_b))$ for each view. With projection dimension d, the invariance loss is:

$$\mathcal{L}_{\text{inv}} = \frac{1}{B\,|\mathcal{V}|\,d}\sum_{b=1}^{B}\sum_{v\in\mathcal{V}}\left\|z_{v,b} - \bar{h}_b\right\|_2^2, \tag{4}$$

where $\mathcal{V}$ denotes all views, global and local, and B is the batch size. Matching the fused global representation from a single modality aligns single-sensor representations with the joint representation space. SIGReg (Balestriero & LeCun, 2025) encourages non-degenerate embeddings by matching random one-dimensional projections of their distribution to an isotropic Gaussian. It acts on the student projections $z_{v,b}$, computing statistics across the distributed minibatch separately for each view and averaging over views. The final objective is:

$$\mathcal{L} = (1-\lambda)\,\mathcal{L}_{\text{inv}} + \lambda\,\mathcal{L}_{\text{SIGReg}}(\{z_{v,b}\}_{v,b}), \tag{5}$$

where $\lambda \in [0,1]$ controls the regularization weight.

5 Experiments

5.1 Experimental setup

Pretraining We pretrain SpectralEarth-FM on SpectralEarth-MM with two global views and four single-modality local views per location for 50 epochs, using AdamW, a learning rate of 2×10^{-5}, a global batch size of 224 on 32 GPUs, projection dimension 128, and SIGReg weight $\lambda = 0.005$. Appendices B and C provide the architecture, view construction, and full-data and ablation training recipes.

Downstream evaluation We evaluate frozen pretrained encoders on hyperspectral benchmarks (Section 5.2) and general EO benchmarks under PANGAEA (Section 5.3). Hyperspectral segmentation and co-located fusion use micro intersection over union (IoU). OxHyperMinerals uses macro multilabel IoU. PANGAEA uses class-macro mean IoU (mIoU). All scores are percentages, and reported gains are percentage-point differences.

Sensor branch routing A downstream input can be routed to one or more pretrained sensor branches. In the *single-branch* setting (SpectralEarth-FM-s), inputs are routed to the branch with the closest spectral coverage and spatial resolution (e.g., EO-1 Hyperion $\rightarrow$ EnMAP, GF-5 $\rightarrow$ EnMAP). In the *multi-branch* setting (SpectralEarth-FM-a), the same acquisition is passed to all pretrained optical branches, and the resulting features are concatenated before the downstream head. HSI inputs are remapped by wavelength interpolation for HSI branches or band averaging for MSI branches (Section D.1). For PANGAEA benchmarks, SpectralEarth-FM-s routes to the Sentinel-2 branch and SpectralEarth-FM-a routes to the Sentinel-2 and OLI branches. Both models use Sentinel-1 data when available.

5.2 Hyperspectral benchmarks

We evaluate on ten benchmarks: eight from the SpectralEarth suite (Braham et al., 2025), covering EnMAP (CDL, BD-Forêt, BNTD, EuroCrops, NLCD, TreeMap), DESIS (DESIS-CDL), and EO-1 Hyperion (EO-1 CDL), plus GF-5 Wuhan (Hong et al., 2023) and OxHyperMinerals (Růžička & Markham, 2025). For the latter, we use EMIT Level-2A reflectance instead of the original Level-1 product and sample a random $10\,\%$ subset of the full dataset. We compare against CARL (Baumann et al., 2026), DOFA-B/L (Xiong et al., 2024), HyperSIGMA-B/L (Wang et al., 2025a), Panopticon (Waldmann et al., 2025), SpecAware (Ji et al., 2026), and Spectral ViT-B/L (Braham et al., 2025). For segmentation, all models use a lightweight SegFormer-style head (Xie et al., 2021) and a learning rate of 10^{-4}.

Results In Table 2, SpectralEarth-FM-s achieves an average rank of 2.30 and SpectralEarth-FM-a achieves 1.30, ranking first overall. On the six EnMAP benchmarks, SpectralEarth-FM-a outperforms all baselines, including Spec. ViT-L. It also leads on DESIS-CDL and EO-1 CDL, the latter reaching 69.1 after interpolation to pretrained optical branches. On GF-5 Wuhan, both variants outperform HyperSIGMA-B/L by 1.6–2.7 points despite HyperSIGMA being pretrained on Gaofen-5 imagery, although Panopticon performs best on this task. On OxHyperMinerals, both variants substantially outscore all baselines. Multi-branch routing improves over single-branch routing on all ten benchmarks, with the largest gains on OxHyperMinerals ($+2.1$), CDL ($+1.7$), and EuroCrops ($+1.6$), thanks to the complementary features from the different pretrained branches.

Sensor branch routing ablation Table 3 isolates branch selection for the same acquisition. EnMAP gives the highest single-branch score on seven of ten tasks, DESIS leads on DESIS-CDL (68.0), and EMIT leads on OxHyperMinerals (76.0). OLI gives the highest single-branch score on GF-5 Wuhan (42.6), but generally scores lower than the HSI branches on the other tasks. Combining all HSI branches improves the best single-branch result on eight tasks and matches it on DESIS-CDL. Adding S2 and OLI improves nine tasks and ties on TreeMap.

Co-located multisensor fusion Table 4 evaluates separate co-located acquisitions on multisensor versions of three downstream tasks (Section D.1), whose splits differ from the main HSI evaluation. EnMAP leads the single-sensor comparison on BD-Forêt, while S2 leads on EuroCrops. On CDL, their scores differ by 0.2 points. OLI scores lower on all three tasks. Fusing EnMAP with S2 improves over either input alone, by 1.8, 1.9, and 2.5 points over the better single sensor on CDL, BD-Forêt, and EuroCrops, respectively. Adding OLI yields further gains of 0.1, 0.6, and 0.7 points.

5.3 General EO benchmarks

We evaluate on seven PANGAEA tasks (Marsocci et al., 2026): HLS Burns (Jakubik et al., 2023), MADOS (Kikaki et al., 2024), PASTIS-R (Garnot et al., 2022; Garnot & Landrieu, 2021), Sen1Floods11 (Bonafilia et al., 2020), DynEarthNet (Toker et al., 2022), SN7 (Van Etten et al., 2021), and AI4Farms (Persello et al., 2023), using locally rerun evaluations with the PANGAEA framework and its baselines,

Table 2: Frozen-encoder hyperspectral evaluation. Scores are micro IoU (%) for segmentation and macro multilabel IoU (%) for OxHyp. Best and second-best distinct scores are bold and underlined. Lower average rank is better.

Model	CDL	BD-F	BNTD	EuCrops	NLCD	TreeMap	D-CDL	EO1-CDL	GF5-WH	OxHyp	Avg. rank
CARL	57.4	52.3	36.3	47.0	34.3	35.6	58.5	52.0	40.4	61.1	7.30
DOFA-B	55.6	47.1	35.6	44.9	32.7	33.8	55.2	55.6	41.9	61.2	8.30
DOFA-L	54.6	45.1	35.4	44.4	32.1	32.2	55.6	53.7	41.5	56.2	9.20
HyperSIGMA-B	58.2	52.4	35.9	47.6	34.6	35.2	49.4	57.7	40.4	65.6	6.80
HyperSIGMA-L	58.4	50.0	35.3	46.0	33.9	34.6	48.2	57.3	40.3	64.4	8.10
Panopticon	61.5	60.4	38.5	49.6	38.6	38.4	57.7	63.8	**44.1**	62.6	4.90
SpecAware	51.9	43.5	33.3	44.2	31.7	29.6	46.8	50.8	36.1	37.9	11.00
Spec. ViT-B	70.9	67.2	41.3	60.7	41.1	39.0	62.4	64.0	43.8	66.1	3.70
Spec. ViT-L	70.5	69.7	42.0	60.9	42.9	41.3	64.0	64.5	<u>43.9</u>	65.8	3.10
SpectralEarth-FM-s	<u>73.4</u>	<u>72.6</u>	<u>45.1</u>	<u>64.7</u>	<u>44.3</u>	<u>42.7</u>	<u>68.0</u>	<u>68.5</u>	42.0	<u>76.0</u>	<u>2.30</u>
SpectralEarth-FM-a	**75.1**	**73.4**	**45.6**	**66.3**	**44.9**	**43.2**	**68.4**	**69.1**	43.0	**78.1**	**1.30**

Table 3: Branch routing of a single acquisition through the full-data 50-epoch student. Scores are micro IoU (%) for segmentation and macro multilabel IoU (%) for OxHyp. Best and second-best distinct scores are bold and underlined.

Target branches	CDL	BD-F	BNTD	EuCrops	NLCD	TreeMap	D-CDL	EO1-CDL	GF5-WH	OxHyp
EnMAP	73.4	72.6	45.1	64.7	44.3	<u>42.7</u>	64.3	68.5	42.0	75.8
DESIS	71.8	66.2	43.1	60.5	41.6	41.0	<u>68.0</u>	64.3	40.9	65.1
EMIT	71.2	69.6	43.0	61.1	41.4	41.2	60.8	66.9	42.3	76.0
S2	71.8	68.2	43.9	61.7	42.4	41.4	64.2	68.4	42.2	70.0
OLI	66.8	64.5	43.0	56.2	41.1	40.3	56.2	61.5	<u>42.6</u>	67.4
All HSI	<u>74.6</u>	<u>73.0</u>	<u>45.5</u>	<u>65.8</u>	<u>44.7</u>	**43.2**	<u>68.0</u>	<u>68.6</u>	42.5	<u>78.0</u>
All HSI + S2 + OLI	**75.1**	**73.4**	**45.6**	**66.3**	**44.9**	**43.2**	**68.4**	**69.1**	**43.0**	**78.1**

Table 4: Co-located fusion on joint splits, evaluated with micro IoU (%). The best and second-best scores for each dataset are bold and underlined.

Dataset	EnMAP	S2	OLI	EnMAP+S2	EnMAP+OLI	S2+OLI	EnMAP+S2+OLI
CDL	72.7	72.5	67.8	<u>74.5</u>	73.9	73.2	**74.6**
BD-F	74.6	73.6	68.8	<u>76.5</u>	75.7	74.7	**77.1**
EuCrops	70.3	70.5	63.7	<u>73.0</u>	71.2	71.4	**73.7**

plus THOR (Forgaard et al., 2026). We were unable to reproduce results for three tasks (Crop Type Mapping-SS (Rustowicz et al., 2019), BioMassters (Nascetti et al., 2023) and FiveBillionPixels (Tong et al., 2023)) and therefore exclude them. The runs use 100 % training data.

Results Table 5 reports class-macro test mIoU. SpectralEarth-FM-a achieves the best average rank (3.00), followed by SpectralEarth-FM-s (4.43). The multi-branch variant improves on five tasks and yields similar results on SN7 and AI4Farms. It achieves the best scores on HLS Burns (85.51), MADOS (69.70), and PASTIS (44.49). Together with Section 5.2, these results show strong transfer across both hyperspectral and standard EO benchmarks.

5.4 Pretraining ablations

We pretrain each variant on a random 200 K-location subset sampled from SpectralEarth-MM for 50 epochs, changing only the named component (Appendix C). Table 6 reports frozen-encoder, single-branch evaluation on eight tasks. Removing local views, SIGReg, or the EMA teacher reduces every score, with average drops of 3.77, 3.40, and 3.15 points, respectively. Replacing spectral tokenizers with linear HSI encoders also hurts every task (average 1.86 points), supporting the benefit

Table 5: PANGAEA class-macro test mIoU (%) using 100% training data. Models use S1+S2 on PASTIS and Sen1Floods11 where supported († marks optical-only input). U-Net and ViT are trained from scratch. Best and second-best distinct scores are bold and underlined.

Model	HLS Burns	MADOS	PASTIS	Sen1Floods11	DynEarthNet	SN7	AI4Farms	Avg. rank
CROMA	82.57	60.72	32.86	88.12	37.90	59.40	25.92	9.00
DOFA	80.62	62.67	30.07	88.01	<u>39.58</u>	61.47	27.08	8.00
GFM-Swin†	76.89	66.61	21.95	72.67	33.50	61.07	26.36	12.00
Prithvi†	83.61	53.16	34.36	90.31	36.02	58.53	27.09	8.57
RemoteCLIP†	76.65	62.16	17.64	74.03	33.76	58.20	25.12	14.43
SatlasNet†	79.95	58.15	29.72	87.36	33.24	<u>61.71</u>	24.40	12.57
Scale-MAE†	76.71	58.10	23.76	75.23	35.62	**62.84**	21.86	13.14
SpectralGPT†	80.80	53.96	36.05	88.45	37.95	58.87	26.27	9.57
S12-MoCo†	81.61	47.83	34.28	89.28	36.40	57.62	25.36	11.71
S12-DINO†	81.72	56.63	36.16	88.24	35.66	56.69	25.42	11.00
S12-MAE†	81.88	55.23	31.78	89.00	36.78	56.66	24.96	11.29
S12-Data2Vec†	82.19	54.70	34.55	88.24	36.05	58.02	24.44	11.43
TerraMind-B	82.45	<u>67.10</u>	40.69	89.50	39.45	60.02	27.07	5.29
THOR-B	80.72	53.05	42.22	<u>90.49</u>	38.09	59.47	<u>27.79</u>	7.00
U-Net Baseline	84.93	51.20	27.12	**91.18**	**40.21**	61.37	**45.81**	5.86
ViT Baseline†	83.21	45.67	39.78	87.09	36.35	55.02	24.58	12.43
SpectralEarth-FM-s	<u>85.03</u>	63.58	<u>42.76</u>	89.84	36.61	60.93	27.64	<u>4.43</u>
SpectralEarth-FM-a	**85.51**	**69.70**	**44.49**	89.90	39.18	60.93	27.48	**3.00**

Table 6: Pretraining ablations on 200 K locations for 50 epochs, using frozen encoders and single-branch routing. Reference IoU (%) and changes (pp) are shown.

Variant	CDL	BD-F	BNTD	EuCrops	D-CDL	EO1-CDL	GF5-WH	OxHyp	Mean Δ
Reference	72.2	69.9	44.4	61.7	65.0	65.7	43.2	75.9	—
No EMA teacher	−2.3	−4.6	−1.9	−3.1	−0.9	−7.5	−2.3	−2.6	−3.15
No SIGReg	−4.0	−6.1	−2.0	−4.4	−1.0	−4.4	−1.3	−3.9	−3.40
2 local views	−1.1	−1.2	−0.2	−0.3	−0.1	−0.5	+0.1	−0.8	−0.51
No local views	−4.0	−4.9	−1.5	−4.1	−3.9	−4.5	−1.3	−5.8	−3.77
SAR/LST as local views	−1.5	−0.8	−0.4	−0.3	−0.2	−1.3	−0.2	−0.7	−0.67
Linear HSI encoder	−1.7	−3.1	−1.8	−2.1	−1.1	−2.2	−0.8	−2.2	−1.86

of explicit spectral processing. Using two instead of four local views reduces the average score by 0.51 points. Allowing SAR/LST in local views reduces it by 0.67 points and lowers every score, suggesting that matching a fused target is not effective when a local view carries too little optical information.

6 Conclusions & Future Work

We introduced SpectralEarth-FM, a multisensor transformer-based EO FM, and SpectralEarth-MM, a large-scale dataset co-locating HSI from three spaceborne sensors with multispectral, SAR, and LST observations. SpectralEarth-FM combines spectral tokenization, sensor-specific branches, hierarchical spatial processing, and a shared encoder pretrained with a JEPA-style objective. Evaluated across ten hyperspectral benchmarks and seven PANGAEA tasks, SpectralEarth-FM achieves the best average rank in both evaluation settings, and transfers to HSI sensors unseen during pretraining (EO-1 Hyperion, GF-5) through spectral interpolation. Both SpectralEarth-FM and SpectralEarth-MM will be released to support future research in EO FMs.

Limitations Unlike any-sensor models, SpectralEarth-FM dedicates explicit processing to each pretraining sensor, trading out-of-the-box sensor flexibility for stronger in-distribution spectral representations. As shown in Table 4 and Section 5.2, fusing sensors with varying spatio-spectral characteristics involves a complex interplay of spatial and spectral features that is not yet fully understood. Future work should investigate explainability methods to understand how cross-sensor interactions are encoded in the shared representation, and extend the framework beyond land surface imagery to oceanic and atmospheric applications.

AI USE STATEMENT

Generative AI tools were used to assist with editing and formatting the manuscript and with code documentation. The authors take full responsibility for the text, claims, and artifacts.

ETHICS STATEMENT

SpectralEarth-MM combines Earth observation products whose access and reuse are governed by the source missions' terms. The release will respect these conditions (Appendix A.5). Geographic coverage is uneven and temporal correspondence varies across sensors. These limitations should be considered when applying the representations, particularly in regions and settings not represented by the downstream evaluation.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe the data construction and source products. Section 4 and Appendices B and C specify the architecture and pretraining procedure, and Section 5 and Section D describe downstream evaluation. Code, pretrained checkpoints, and data will be released subject to the source missions' access conditions.

REFERENCES

Kevin Alonso, Martin Bachmann, Kara Burch, Emiliano Carmona, Daniele Cerra, Raquel De los Reyes, Daniele Dietrich, Uta Heiden, Andreas Hölderlin, Jack Ickes, et al. Data products, quality and validation of the DLR Earth sensing imaging spectrometer (DESIS). *Sensors*, 19(20):4471, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a Joint-Embedding predictive architecture. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. OmniSat: Self-supervised modality fusion for Earth observation. In *European Conference on Computer Vision*, pp. 409–427. Springer, 2024.

Guillaume Astruc, Nicolas Gonthier, Clement Mallet, and Loic Landrieu. AnySat: One Earth observation model for many resolutions, scales, and modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19530–19540, 2025.

Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10161–10170, 2021. doi: 10.1109/ICCV48922.2021.01002.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. SatlasPretrain: A large-scale dataset for remote sensing image understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16726–16736, 2023. doi: 10.1109/ICCV51070.2023.01538.

Alexander Baumann, Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Alexander Studier-Fischer, Berkin Özdemir, Lena Maier-Hein, and Slobodan Ilic. CARL: Camera-agnostic representation learning for spectral image analysis. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=TpbhS1yfz0.

Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabe-Moreno, et al. TerraMesh: A planetary mosaic of multimodal Earth observation data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2025.

Derrick Bonafilia, Beth Tellman, Tyler Anderson, and Erica Issenberg. Sen1Floods11: A georeferenced dataset to train and test deep learning flood algorithms for Sentinel-1. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:16780–16797, 2025. doi: 10.1109/JSTARS.2025.3581451.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. AlphaEarth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745–15753, 2021. doi: 10.1109/CVPR46437.2021.01549.

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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=WBhqzpF6KYH.

Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Muhammad Haris Khan, Rao Muhammad Anwer, Jorma Laaksonen, Fahad Shahbaz Khan, and Salman Khan. TerraFM: A scalable foundation model for unified multisensor Earth observation. In *The Fourteenth International Conference on Learning Representations*, 2026.

European Union. How to access data (Copernicus), 2026. URL https://www.copernicus.eu/en/terms-use/how-access-data. Accessed September 2026.

Theodor Forgaard, Jarle H Reksten, Anders U Waldeland, Valerio Marsocci, Nicolas Longépé, Michael Kampffmeyer, and Arnt-Børre Salberg. THOR: A versatile foundation model for Earth observation climate and society applications. *arXiv preprint arXiv:2601.16011*, 2026.

Alistair Francis and Mikolaj Czerkawski. Major TOM: Expandable datasets for Earth observation. In *2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2935–2940, 2024. doi: 10.1109/IGARSS53475.2024.10640760.

Martin Hermann Paul Fuchs and Begüm Demir. HySpecNet-11k: a large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods. In *2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1779–1782, 2023. doi: 10.1109/IGARSS52108.2023.10283385.

Anthony Fuller, Koreen Millard, and James Green. CROMA: Remote sensing representations with contrastive radar-optical masked autoencoders. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 5506–5538. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/11822e84689e631615199db3b75cd0e4-Paper-Conference.pdf.

Vivien Sainte Fare Garnot and Loic Landrieu. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4872–4881, 2021.

Vivien Sainte Fare Garnot, Loic Landrieu, and Nesrine Chehata. Multi-modal temporal attention models for crop mapping from satellite time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 187:294–305, 2022.

German Aerospace Center (DLR). License agreement on the use of EnMAP data, version 1.1, 2022. URL https://www.enmap.org/data/resources/EnMAP_Data_License.pdf.

German Aerospace Center (DLR). DESIS license terms for scientific use, 2024a. URL https://geoservice.dlr.de/resources/licenses/desis/DESIS_License_Agreement_for_Scientific_Use.pdf.

German Aerospace Center (DLR). EOWEB GeoPortal. https://eoweb.dlr.de/egp/, 2024b. Accessed: 2025.

German Aerospace Center (DLR). *EnMAP Frequently Asked Questions (FAQ)*. German Remote Sensing Data Center, 2.7 edition, 2026. URL https://www.enmap.org/data/doc/EnMAP_FAQ.pdf.

Robert O Green, Natalie Mahowald, Charlene Ung, David R Thompson, Lori Bator, Matthew Bennet, Michael Bernas, Natalie Blackway, Christine Bradley, Jeff Cha, et al. The Earth surface mineral dust source investigation: An Earth science imaging spectroscopy mission. In *2020 IEEE aerospace conference*, pp. 1–15. IEEE, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent – a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Luis Guanter, Hermann Kaufmann, Karl Segl, Saskia Foerster, Christian Rogass, Sabine Chabrillat, Theres Kuester, André Hollstein, Godela Rossner, Christian Chlebek, Christoph Straif, Sebastian Fischer, Stefanie Schrader, Tobias Storch, Uta Heiden, Andreas Mueller, Martin Bachmann, Helmut Mühle, Rupert Müller, Martin Habermeyer, Andreas Ohndorf, Joachim Hill, Henning Buddenbaum, Patrick Hostert, Sebastian Van der Linden, Pedro J. Leitão, Andreas Rabe, Roland Doerffer, Hajo Krasemann, Hongyan Xi, Wolfram Mauser, Tobias Hank, Matthias Locherer, Michael Rast, Karl Staenz, and Bernhard Sang. The enmap spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, 7(7):8830–8857, 2015. ISSN 2072-4292. doi: 10.3390/rs70708830. URL https://www.mdpi.com/2072-4292/7/7/8830.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2022. doi: 10.1109/TGRS.2021.3130716.

Danfeng Hong, Bing Zhang, Hao Li, Yuxuan Li, Jing Yao, Chenyu Li, Martin Werner, Jocelyn Chanussot, Alexander Zipf, and Xiao Xiang Zhu. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks. *Remote Sensing of Environment*, 299:113856, 2023.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. SpectralGPT: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27838–27848, 2026.

Johannes Jakubik, Sujit Roy, CE Phillips, Paolo Fraccaro, Denys Godwin, Bianca Zadrozny, Daniela Szwarcman, Carlos Gomes, Gabby Nyirjesy, Blair Edwards, et al. Foundation models for generalist geospatial artificial intelligence. *arXiv preprint arXiv:2310.18660*, 2023.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. TerraMind: Large-scale generative multimodality for Earth observation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394, 2025. doi: 10.1109/ICCV51701.2025.00693.

Renjie Ji, Xue Wang, Chao Niu, Wen Zhang, Yong Mei, and Kun Tan. SpecAware: A spectral-content aware foundation model for unifying multi-sensor learning in hyperspectral remote sensing mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 234:242–260, 2026. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2026.02.024. URL https://www.sciencedirect.com/science/article/pii/S0924271626000754.

Katerina Kikaki, Ioannis Kakogeorgiou, Ibrahim Hoteit, and Konstantinos Karantzalos. Detecting marine pollutants and sea surface features with deep learning in Sentinel-2 imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 210:39–54, 2024.

Weili Kong, Baisen Liu, Xiaojun Bi, Changdong Yu, Xinyao Li, and Yushi Chen. HyperSL: A spectral foundation model for hyperspectral image interpretation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–19, 2025. doi: 10.1109/TGRS.2025.3566205.

David Krutz, Rupert Müller, Uwe Knodt, Burghardt Günther, Ingo Walter, Ilse Sebastian, Thomas Säuberlich, Ralf Reulke, Emiliano Carmona, Andreas Eckardt, et al. The instrument design of the DLR Earth sensing imaging spectrometer (DESIS). *Sensors*, 19(7):1622, 2019.

Oscar Manas, Alexandre Lacoste, Xavier Giró-i Nieto, David Vazquez, and Pau Rodriguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9414–9423, 2021.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, Heng Fang, Yifang Ban, Maarten Vergauwen, Nicolas Audebert, and Andrea Nascetti. PANGAEA: Assessing geospatial foundation models capabilities through a global and inclusive benchmark. *IEEE Geoscience and Remote Sensing Magazine*, 14(1):245–285, 2026. doi: 10.1109/MGRS.2025.3628194.

NASA Earth Science Data and Information System. ESDIS project data use and citation guidance (ESDS-RFC-055), 2025. URL https://doi.org/10.5067/DOC/ESCO/ESDS-RFC-055.

NASA LP DAAC. EMIT L2A estimated surface reflectance and uncertainty and masks 60 m V001. NASA Earthdata Search.

Andrea Nascetti, Ritu Yadav, Kirill Brodt, Qixun Qu, Hongwei Fan, Yuri Shendryk, Isha Shah, and Christine Chung. BioMassters: A benchmark dataset for forest biomass estimation using multi-modal satellite time-series. *Advances in Neural Information Processing Systems*, 36:20409–20420, 2023.

Vishal Nedungadi, Ankit Kariryaa, Stefan Oehmcke, Serge Belongie, Christian Igel, and Nico Lang. MMEarth: Exploring multi-modal pretext tasks for geospatial representation learning. In *European Conference on Computer Vision*, pp. 164–182. Springer, 2024.

Claudio Persello, Jeroen Grift, Xinyan Fan, Claudia Paris, Ronny Hänsch, Mila Koeva, and Andrew Nelson. AI4SmallFarms: A dataset for crop field delineation in Southeast Asian smallholder farms. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. doi: 10.1109/LGRS.2023.3323095.

Rose M. Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, and David Lobell. Semantic segmentation of crop type in Africa: A novel dataset and analysis of deep learning methods. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition workshops*, pp. 75–82, 2019.

Vít Růžička and Andrew Markham. HyperspectralViTs: General hyperspectral models for on-board remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:10241–10253, 2025. doi: 10.1109/JSTARS.2025.3557527.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2166–2176, 2023. doi: 10.1109/CVPRW59228.2023.00210.

Gencer Sumbul, Chang Xu, Emanuele Dalsasso, and Devis Tuia. SMARTIES: Spectrum-aware multi-sensor auto-encoder for remote sensing images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5569–5578, 2025.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732.

Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Orsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique De Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-EO-2.0: A versatile multitemporal foundation model for Earth observation applications. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–20, 2026. doi: 10.1109/TGRS.2025.3642610.

Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. DynamicEarthNet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21158–21167, 2022.

Xin-Yi Tong, Gui-Song Xia, and Xiao Xiang Zhu. Enabling country-scale land cover mapping with meter-resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196: 178–196, 2023.

Zahid Hassan Tushar and Sanjay Purushotham. HyperFM: An efficient hyperspectral foundation model with spectral grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 6955–6965, 2026.

U.S. Geological Survey. Are there any restrictions on the use or redistribution of Landsat data? USGS FAQ, 2026. URL `https://www.usgs.gov/faqs/are-there-any-restrictions-use-or-redistribution-landsat-data`. Accessed September 2026.

Adam Van Etten, Daniel Hogan, Jesus Martinez Manso, Jacob Shermeyer, Nicholas Weir, and Ryan Lewis. The multi-temporal urban development SpaceNet dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6398–6407, 2021.

Huy V. Vo, Vasil Khalidov, Timothée Darcet, Théo Moutakanni, Nikita Smetanin, Marc Szafraniec, Hugo Touvron, camille couprie, Maxime Oquab, Armand Joulin, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. Automatic data curation for self-supervised learning: A clustering-based approach. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=G7p8djzWOl`.

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 Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025a. doi: 10.1109/TPAMI.2025.3557581.

Yi Wang, Conrad M. Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *IEEE Geoscience and Remote Sensing Magazine*, 10(4):213–247, 2022. doi: 10.1109/MGRS.2022.3198244.

Yi Wang, Nassim Ait Ali Braham, Zhitong Xiong, Chenying Liu, Conrad M. Albrecht, and Xiao Xiang Zhu. SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [software and data sets]. *IEEE Geoscience and Remote Sensing Magazine*, 11 (3):98–106, 2023. doi: 10.1109/MGRS.2023.3281651.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9888–9899, 2025b. doi: 10.1109/ICCV51701.2025.00922.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12077–12090, 2021.

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

Fanglong Yao, Wanxuan Lu, Heming Yang, Liangyu Xu, Chenglong Liu, Leiyi Hu, Hongfeng Yu, Nayu Liu, Chubo Deng, Deke Tang, Changshuo Chen, Jiaqi Yu, Xian Sun, and Kun Fu. RingMo-Sense: Remote sensing foundation model for spatiotemporal prediction via spatiotemporal evolution disentangling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–21, 2023. doi: 10.1109/TGRS.2023.3316166.

A Dataset details

This appendix provides additional details on the construction of SpectralEarth-MM. It describes the geographic and spectral coverage of SpectralEarth-MM, compares it with existing EO pretraining datasets, details the construction pipeline, and summarizes the characteristics and licenses of each data source.

A.1 Geographic coverage

Figure 4 shows the global distribution of the HSI anchor patches in SpectralEarth-MM, colored by HSI sensor. The selected footprints cover approximately 20% of the global land surface and span all continents except Antarctica. Coverage differs across sensors because of their orbits and acquisition strategies. EMIT and DESIS operate from the International Space Station, whose orbit restricts their coverage to low and mid latitudes. EMIT samples concentrate on arid and semi-arid regions (Green et al., 2020), including the southwestern United States, North Africa, the Arabian Peninsula, Central Asia, and Australia. EnMAP operates from a sun-synchronous orbit and extends coverage to high northern latitudes. Persistently cloudy regions, such as the Amazon, are sparsely covered.

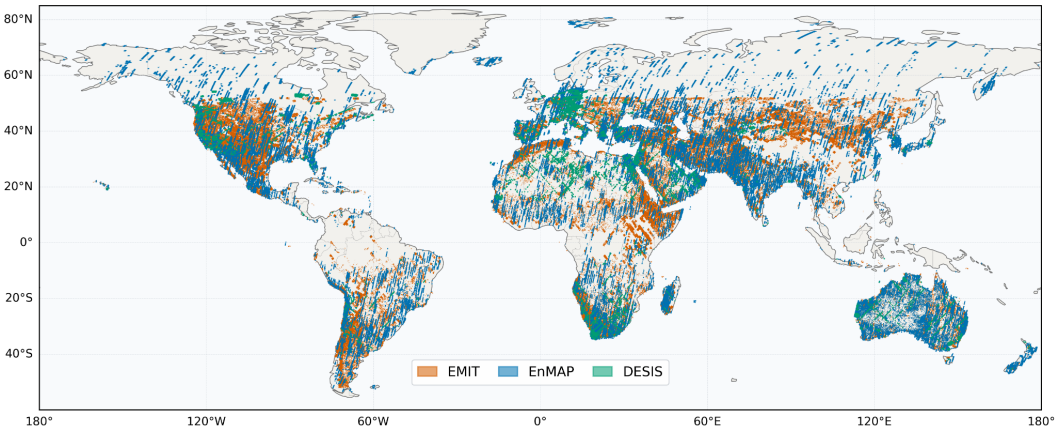

Figure 4: *Spatial coverage of SpectralEarth-MM.* Global distribution of HSI anchor patches in SpectralEarth-MM. Colors indicate the HSI sensor associated with each acquisition.

A.2 Comparison with existing datasets

Table 7 compares the modalities, spatial and temporal coverage, and scale of representative EO pretraining datasets. Most multisensor datasets pair multispectral imagery with SAR, but none includes HSI or LST. HSI datasets are limited to one or two hyperspectral sensors without co-located observations from other EO sensors. SpectralEarth-MM combines all four observation types from seven sensors over common footprints, with multitemporal coverage. It also includes significantly more locations than the largest HSI dataset.

A.3 Spectral coverage

Figure 5 shows the spectral bands of the optical sensors in SpectralEarth-MM. EnMAP and EMIT sample the VNIR and SWIR ranges with hundreds of narrow, contiguous bands, while DESIS is limited to the VNIR range. Sentinel-2 and Landsat-8/9 cover a similar range with a few broad bands. The Landsat thermal bands, from which the LST product is derived, extend the coverage to the thermal infrared. Sentinel-1 is omitted because it is non-optical.

Table 7: *Representative EO pretraining datasets.* We group datasets into general multisensor remote sensing and HSI-specific pretraining datasets. "Spat. align" indicates whether multiple observation types are provided over common geographic footprints. "Multi-temp." indicates the availability of multiple acquisition dates for the same geolocation.

Dataset	HSI	MSI	SAR	LST	Spat. align	Multi-temp.	Size [TB]	Loc. [M]	Patches [M]	Sensors†
SatlasPretrain (Bastani et al., 2023)	✗	✓	✗	✗	○	✓	30.0	0.86	∼9.0	2
SSL4EO-S12 (Wang et al., 2023)	✗	✓	✓	✗	✓	✓	3.7	0.25	1.5	2
TerraMesh (Blumenstiel et al., 2025)	✗	✓	✓	✗	✓	✗	17.0	9.00	64.0	3
MMEarth (Nedungadi et al., 2024)	✗	✓	✓	✗	✓	✗	0.6	1.20	7.2	4
MajorTOM-Core (Francis & Czerkawski, 2024)	✗	✓	✓	✗	✓	✗	62.0	1.47	6.0	4
Copernicus-Pretrain (Wang et al., 2025b)	✗	✓	✓	✗	✓	✓	20.7	0.31	18.7	5
HySpecNet-11k (Fuchs & Demir, 2023)	✓	✗	✗	✗	✗	✗	0.1	0.01	0.01	1
HyperFM250k (Tushar & Purushotham, 2026)	✓	✗	✗	✗	✗	✗	3.0	0.25	0.25	1
Hyper-400K (Ji et al., 2026)	✓	✗	✗	✗	✗	✗	17.0	0.40	0.40	1
HyperGlobal-450K (Wang et al., 2025a)	✓	✗	✗	✗	✗	✗	0.4	0.45	0.45	2
SpectralEarth (Braham et al., 2025)	✓	✗	✗	✗	✗	✓	3.3	0.42	0.54	1
SpectralEarth-MM (ours)	✓	✓	✓	✓	✓	✓	40.0	2.00	25.00	7

Legend: ✓ available, ○ partially available, ✗ unavailable.

†: derived products from the same instrument count as one. Geospatial layers and non-EO data such as climate and weather data are excluded.

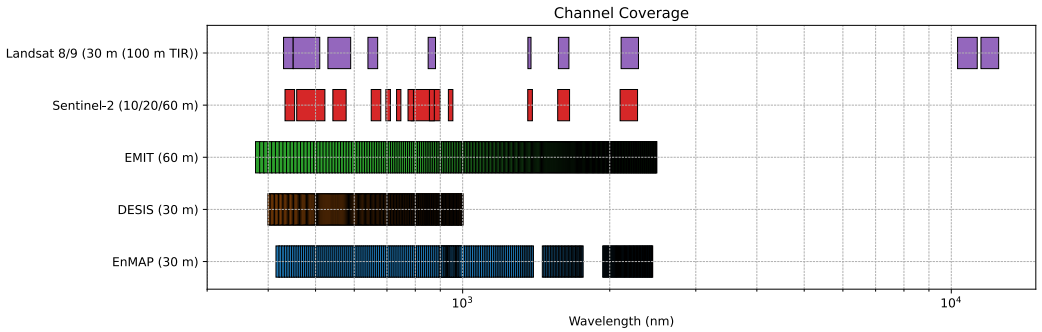

Figure 5: Spectral coverage of the optical sensors and Landsat thermal bands (long wavelength to the far right) associated with the LST product in SpectralEarth-MM. EnMAP, EMIT, and DESIS provide dense VNIR/SWIR measurements, while Sentinel-2 and Landsat-8/9 provide sparse multispectral bands. Sentinel-1 is omitted because it is non-optical.

A.4 Construction pipeline

Figure 6 provides an overview of the construction pipeline of SpectralEarth-MM, which proceeds in three stages: HSI preprocessing, rebalancing, and multisensor pairing.

HSI acquisition and quality control We assemble the HSI archive from EnMAP, EMIT, and DESIS mission data. For EnMAP, we use Level-2A products acquired between May 2022 and September 2025 and apply an initial 10% cloud filter. For DESIS, we select cloud-free tiles from the archive available through the EOWEB GeoPortal (German Aerospace Center (DLR), 2024b). For EMIT, we collect Level-2A products acquired up to September 2025 with cloud coverage below 10%. The resulting raw HSI archive contains approximately 60 K EnMAP tiles, 33 K EMIT tiles, and 11 K DESIS tiles.

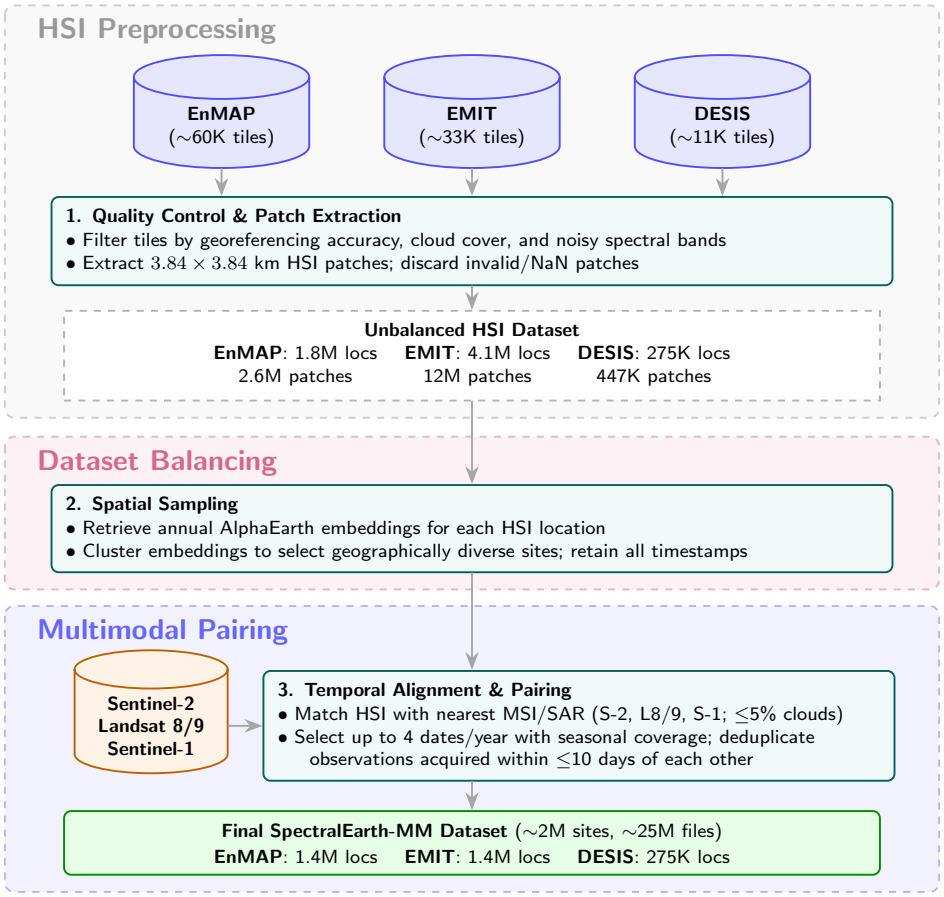

Figure 6: *Construction pipeline of SpectralEarth-MM.* HSI acquisitions serve as anchors: they are filtered, patchified, grouped by location, rebalanced, and paired with co-located MSI, SAR, and LST observations.

Before patch extraction, we apply sensor-specific quality control. For EnMAP and DESIS, we discard scenes whose orthorectification RMSE, as reported in the metadata, exceeds 60 m. For EnMAP and EMIT, we additionally remove, after visual inspection, scenes with heavy clouds that the metadata does not capture. For all three HSI sensors, we remove noisy spectral bands, including bands affected by water absorption. EMIT scenes are reprojected to local UTM zones to match the spatial reference of EnMAP and DESIS.

Patch extraction and grouping After quality control, HSI scenes are patchified into non-overlapping $3.84\,\mathrm{km} \times 3.84\,\mathrm{km}$ footprints. Patches with more than 1% missing or invalid pixels are discarded. The remaining patches are grouped by geographic footprint and timestamp. Observations from multiple HSI sensors that intersect the same footprint are grouped as co-located HSI views. This yields locations with a single HSI sensor, two HSI sensors, or all three, depending on acquisition availability.

The initial HSI archive is strongly imbalanced across sensor combinations. Before rebalancing, it contains approximately 4.1 M EMIT locations, 1.8 M EnMAP locations, and 275 K DESIS locations, corresponding to roughly 12 M, 2.6 M, and 447 K HSI patches, respectively. EMIT-only locations and EMIT–EnMAP pairs dominate the archive, while DESIS coverage is comparatively limited.

Balancing HSI locations We rebalance the HSI archive before pairing it with the other EO modalities. Rebalancing is performed separately for each HSI co-location configuration rather than with a single global subsampling rule. We retain all DESIS-containing configurations because of their

limited availability, and all EnMAP-only locations because they cover regions outside the ISS orbit. We downsample the two largest configurations: EMIT-only locations are reduced from approximately 2.9 M to 500 K, and EMIT–EnMAP locations from approximately 1.1 M to 750 K.

Subsampling relies on AlphaEarth embeddings (Brown et al., 2025). For each HSI location, we select a reference year, using the majority year when multiple sensors or timestamps are present. We then retrieve the corresponding annual AlphaEarth embeddings and average them to obtain one 64-dimensional embedding per location. These embeddings are clustered with a three-level hierarchical clustering procedure, and samples are selected within the resulting hierarchy following Vo et al. (2024). This reduces redundancy in overrepresented regions while preserving broad geographic and semantic coverage. All timestamps of the selected locations are retained.

After rebalancing, the selected HSI anchor patches comprise approximately 1.4 M EnMAP locations, 1.4 M EMIT locations, and 275 K DESIS locations. These counts are not additive, because a location can contain observations from multiple HSI sensors and dates. Figure 4 shows the geographic distribution of these anchor patches.

Multisensor pairing After rebalancing, each selected HSI footprint is paired with Sentinel-2, Landsat-8/9 optical imagery, Landsat LST, and Sentinel-1 observations, using the HSI timestamps as temporal anchors. For Sentinel-2 and Landsat-8/9 optical images, we select observations within a window of 6 months around the HSI anchor and apply a 5% cloud threshold. For Sentinel-1, we use a wider window of up to 4 years to account for gaps in the acquisitions. We retain only locations with at least one HSI anchor and paired observations for all additional sensors. Depending on HSI sensor availability, a location therefore contains observations from five to seven sensors. Figure 7 shows four examples (rows) with all SpectralEarth-MM observation types (columns).

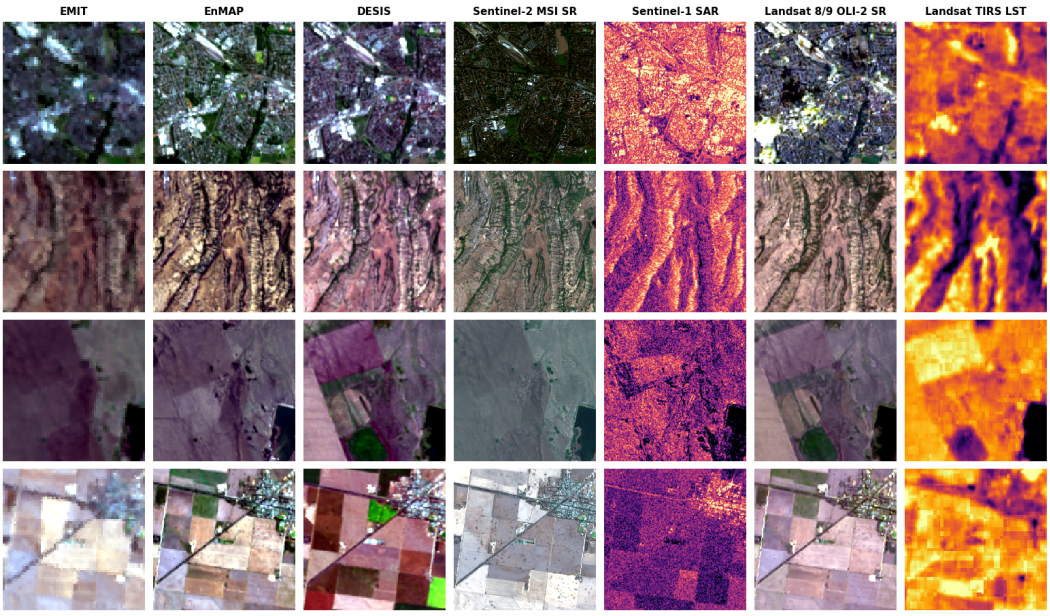

Figure 7: Examples of co-located observations in SpectralEarth-MM. Each row corresponds to a geographic footprint and each column to an observation type. Images are visualized for inspection and do not imply identical acquisition time or pixel-level spatial alignment.

Temporal filtering and de-duplication Temporal coverage is uneven across locations and sensors: some locations have dense time series, while others have only one or a few acquisitions. To reduce redundancy and bound storage and training cost, we retain at most four timestamps per sensor and location. Selection favors temporal proximity between paired cross-sensor observations and, when available, temporal spread across the year. We also remove near-duplicate observations of the same sensor acquired within 10 days of each other. The final dataset contains approximately 2 M globally distributed locations, 25 M georeferenced patches, and over 40 TB of data.

A.5 SPATIO-SPECTRAL CHARACTERISTICS AND LICENSING

Table 8 summarizes the spatial resolution, spectral range, revisit time, and data license of each source in SpectralEarth-MM. Ground sampling distance ranges from 10 m to 60 m. EnMAP and DESIS acquisitions are tasked, and the ISS-based EMIT and DESIS follow the station's non-sun-synchronous, precessing orbit, so their revisit and local overpass time vary with latitude. All sources are freely available for research. DESIS data are restricted to non-commercial scientific use, while the other sources also permit commercial use.

Table 8: Summary of SpectralEarth-MM's data sources and licensing conditions. Resolution denotes spatial ground sampling distance. Spectral ranges are given as wavelength intervals. SAR denotes synthetic aperture radar, and TIR denotes thermal infrared. ISS-based instruments (EMIT and DESIS) have no fixed orbital revisit. Reported values correspond to effective revisit determined by orbital geometry and acquisition tasking constraints. Revisit times for satellite constellations refer to average global repeat cycles and may vary with latitude and observation mode.

Mission	Operator	Spat. Res. [m]	Spectrum [μm]	Revisit [d] (note)	License
EMIT	NASA	60	0.38–2.5	$\sim$4 (effective, ISS orbit and tasking constrained)	NASA Earth Science Data and Information Policy: fully open access without use restrictions (NASA Earth Science Data and Information System, 2025)
DESIS	DLR	30	0.4–1.0	$\sim$4 (effective, ISS orbit and tasking constrained)	DLR/Teledyne scientific-use license: free for non-commercial scientific research only (German Aerospace Center (DLR), 2024a)
EnMAP	DLR	30	0.42–2.45	27 in nadir mode 4 in off-nadir mode (tasked acquisitions)	Free and open access via DLR geoservice upon registration. Scientific and commercial use permitted (German Aerospace Center (DLR), 2022)
Sentinel-2	ESA	10/20/60	0.44–2.19	5 (constellation, 10-day/unit)	Copernicus free, full and open access data policy (European Union, 2026)
Landsat-8/9	NASA/ USGS	30 (100 TIR)	0.43–2.29 (OLI) 10.6–12.5 (TIR)	8 (constellation, 16-day/unit)	USGS Landsat Collection open data policy: unrestricted public access and redistribution (U.S. Geological Survey, 2026)
Sentinel-1	ESA	10–20	5.55×10^4 (C-band SAR, 5.55 cm)	6 (constellation, 12-day/unit)	Copernicus free, full and open access data policy (European Union, 2026)

B MODEL DETAILS

This appendix details the architecture of SpectralEarth-FM described in Section 4.1.

Spectral grouping and tokenization For each HSI sensor, the retained bands are ordered by center wavelength and partitioned into contiguous groups of approximately equal wavelength span, skipping the gaps left by removed absorption bands. The number of groups is a sensor-specific hyperparameter: EnMAP (German Aerospace Center (DLR), 2026) and EMIT (NASA LP DAAC) use 15 groups, and DESIS (Alonso et al., 2019) uses 10 groups. The average group sizes are 13.5, 16.3, and 21.5 bands, respectively, with group sizes ranging from 8 to 23 bands across sensors.

Each spectral group has its own patch projection layer into a spectral token of dimension $d_\lambda = 96$. Learned positional embeddings encode the order of the groups along the spectrum. A four-layer spectral transformer with two attention heads processes the sequence of group tokens independently at each spatial position, so spectral interactions are modeled before any spatial attention. The group tokens are then aggregated with projected-attention pooling: they are projected to a spectral fusion dimension of 192, attended by a learned query with two heads, and mapped to the common spatial embedding dimension $D = 192$.

Patch embedding and token layout Sensor-specific patch projections map the inputs of all sensors to a common 64×64 token grid before the local Hiera branches. The patch strides, in native sensor pixels, are 2×2 for EnMAP, DESIS, Landsat optical, and Landsat LST, 1×1 for EMIT, and 6×6 for Sentinel-2 and Sentinel-1. The corresponding input crops are 128×128, 64×64, and 384×384 pixels, respectively. Each token therefore covers 60 m on the ground for every sensor. The patch projections use overlapping kernels of 3×3 for the 2×2 strides and 7×7 for the 6×6 strides, and a 1×1 kernel for EMIT.

Hiera configuration The backbone uses a three-stage Hiera (Ryali et al., 2023) layout with depths $[2, 6, 10]$. The initial token embedding size is $D = 192$, with two attention heads, and width and head count are doubled at each pooling transition. The first two stages form the sensor-specific local branches. Stage 1 keeps $D = 192$ and operates on the 64×64 token grid. Stage 2 begins with a 2×2 query-pooling transition, which reduces the grid to 32×32 and increases the width to $D_\ell = 384$ with four heads. Therefore, each local branch outputs a $32^2 \times 384$ token grid.

The shared Hiera trunk corresponds to the final stage. Its first block applies the second 2×2 query-pooling transition, which reduces the grid to 16×16 and increases the width to 768 with eight heads. The remaining shared blocks operate at this resolution and width. No class token is used: image-level representations are obtained by mean-pooling the final spatial tokens, and dense prediction heads use the intermediate feature maps of each stage.

Fusion configuration Cross-sensor fusion is applied between the local branches and the shared Hiera trunk. At each spatial position, the available sensor tokens $H_s \in \mathbb{R}^{D_\ell}$ are projected from $D_\ell = 384$ to a fusion dimension $D_f = 768$, and learned sensor embeddings are added. A single learned query, shared across positions, aggregates them with two-head attention. The output is layer-normalized and projected back to $D_\ell = 384$, which gives the $32^2 \times 384$ input of the shared trunk. Single-sensor views pass through the same fusion module.

Model size The encoder has approximately 156 M trainable parameters. Most of them are in the sensor-specific local branches and the shared trunk, while the spectral tokenizers and the fusion module are comparatively small.

C TRAINING DETAILS

Optimization We pretrain for 50 epochs with AdamW, weight decay 0.05, and mixed precision. Full-data pretraining runs for approximately 7 days on 32 NVIDIA A100 GPUs ($5{,}400$ GPU-hours) with a peak learning rate of 2×10^{-5}, a global batch size of 224, and five warmup epochs. Ablations use a random 200 K-location subset, a peak learning rate of 10^{-5}, a batch size of 84 on 12 GPUs, and ten warmup epochs. The learning rate is warmed up linearly and then follows a cosine decay.

Objective The projector is a three-layer MLP that maps mean-pooled encoder outputs from 768 to 128 dimensions through two hidden layers of width 2048, each followed by BatchNorm and ReLU. We use SIGReg weight $\lambda = 0.005$. The EMA momentum follows a cosine schedule from 0.996 to 1.

View construction Each location yields two global views with four sampled modalities each, and four local views. Each local view contains one optical modality drawn from those in the global views, with an independently sampled crop. Crops cover $[0.25, 1.0]$ of the location area for global views and $[0.05, 0.25]$ for local views. Global inputs use the crop sizes in Section B, and local inputs one quarter of that side length. Spatial transformations are shared across modalities within a view. Augmentations are horizontal and vertical flips, Gaussian blur, and radiometric jitter.

Normalization Optical inputs are converted to reflectance. SAR backscatter in $[-50, 0]$ dB and land surface temperature in $[250, 350]$ K are mapped linearly to $[0, 1]$, and all inputs are clipped to $[0, 1]$. Sentinel-1, Sentinel-2, and Landsat optical inputs are then standardized per band. HSI inputs are not standardized, to preserve the shape of the spectrum.

D EVALUATION DETAILS

D.1 HYPERSPECTRAL BENCHMARKS

Benchmarks The hyperspectral evaluation in Section 5.2 covers ten tasks from five HSI sensors. Eight tasks come from the SpectralEarth benchmark suite (Braham et al., 2025): six EnMAP tasks (CDL, BD-Forêt, BNTD, EuroCrops, NLCD, and TreeMap), one DESIS task (DESIS-CDL), and one EO-1 Hyperion task (EO-1 CDL). GF-5 Wuhan (Hong et al., 2023) adds a second sensor that is not part of SpectralEarth-MM. OxHyperMinerals (Růžička & Markham, 2025) is a multilabel mineral identification task on EMIT imagery, for which we use Level-2A reflectance instead of the original Level-1 product, consistent with the EMIT data in SpectralEarth-MM. Segmentation tasks are evaluated with micro IoU, and OxHyperMinerals with macro multilabel IoU.

Multisensor downstream tasks To evaluate SpectralEarth-FM with all optical branches, we extend three SpectralEarth tasks, CDL (U.S. agriculture), BD-Forêt (French forests), and EuroCrops (EU agriculture), to multisensor tasks by adding co-located Sentinel-2 and Landsat-8/9 imagery. Consistent with the HSI filtering in SpectralEarth, we keep MSI observations with cloud cover below 10% and a maximum temporal lag to the HSI acquisition of 10 days for CDL and EuroCrops, and 120 days for BD-Forêt. These datasets are used for the fusion experiments in Table 4.

Spectral mapping for branch routing When an HSI input is routed to a different HSI branch, it is linearly interpolated to the center wavelengths of that branch. When it is routed to an MSI branch, each target band is computed by overlap-weighted averaging of the input bands, which approximates integration over the target band. The mapping is fixed, requires no fitting, and preserves spectral order.

D.2 DOWNSTREAM TRAINING

For hyperspectral segmentation, encoder feature maps are projected to dimension 128, upsampled to a common resolution, concatenated, and processed by a 3×3 convolution and a 1×1 classifier. The head is trained for 100 epochs using AdamW with learning rate 10^{-4}, cosine decay, and weight decay 10^{-3}.