SpectralEarth: Training Hyperspectral Foundation Models at Scale

Nassim Ait Ali Braham, Conrad M. Albrecht, Julien Mairal, Jocelyn Chanussot, *Fellow, IEEE*, Yi Wang, and Xiao Xiang Zhu, *Fellow, IEEE*

***Abstract*—Foundation models have triggered a paradigm shift in computer vision and are increasingly being adopted in remote sensing, particularly for multispectral imagery. Yet, their potential in hyperspectral imaging (HSI) remains untapped due to the absence of comprehensive and globally representative hyperspectral datasets. To close this gap, we introduce *SpectralEarth*, a large-scale multitemporal dataset designed to pretrain hyperspectral foundation models leveraging data from the environmental mapping and analysis program (EnMAP). SpectralEarth comprises 538 974 image patches covering 415 153 unique locations from 11 636 globally distributed EnMAP scenes spanning two years of archive. In addition, 17.5% of these locations include multiple timestamps, enabling multitemporal HSI analysis. Utilizing state-of-the-art self-supervised learning algorithms, we pretrain a series of foundation models on SpectralEarth, integrating a spectral adapter into classical vision backbones to accommodate the unique characteristics of HSI. In tandem, we construct nine downstream datasets for land-cover, crop-type mapping, and tree-species classification, providing benchmarks for model evaluation. Experimental results support the versatility of our models and their generalizability across different tasks and sensors. We also highlight computational efficiency during model fine-tuning.**

***Index Terms*—Hyperspectral imaging, Self-supervised learning, Foundation models.**

I. Introduction

HYPERSPECTRAL imaging (HSI) from space provides valuable information about the material composition of the Earth's surface and atmosphere. With hundreds of narrow bands, each a few nanometers in width, hyperspectral images record detailed electromagnetic information across a wide range of wavelengths, from long-wave ultraviolet to short-wave infrared [1]. This rich spectral information provides opportunities for numerous environmental applications such as soil and mineral mapping, pollution tracking, agricultural assessment, and forest monitoring [2], [3], [4], [5], [6].

In recent years, the availability of hyperspectral data has been significantly expanded by the launch of new satellite missions. Notable examples include Germany's Environmental Mapping and Analysis Program (EnMAP) [7] with precursor DLR Earth Sensing Imaging Spectrometer mission (DESIS) [8] mounted to the International Space Station (ISS), Italy's Hyperspectral Precursor and Application Mission (PRISMA) [9], and the European Space Agency's upcoming Copernicus Hyperspectral Imaging Mission for the Environment (CHIME) [10]. These developments open new avenues to employ deep learning and self-supervised learning (SSL) for large-scale HSI analysis.

Foundation models pretrained with SSL demonstrated remarkable generalization capabilities with minimal fine-tuning in computer vision [11]. Subsequently, there has been a surge in developing geospatial foundation models, particularly for multispectral and high-resolution RGB imagery [12]. This line of research benefits from the abundance of publicly available satellite data from missions like Copernicus Sentinel-2, which provides petabytes of archive data for model training. In contrast, progress in foundation models for the hyperspectral domain has been hindered by the lack of large-scale HSI datasets for pretraining. Historically, access to hyperspectral imagery has been costly, with most benchmarks restricted to small-scale aerial imagery datasets [13]. Although recent efforts [14], [15], [16] have begun to address this limitation, they still fall short in terms of data volume and geographic diversity to effectively pretrain general-purpose hyperspectral models.

To bridge this gap, we introduce *SpectralEarth*, a large-scale dataset derived from the EnMAP satellite mission, featuring over 538 974 hyperspectral image patches from 415 153 unique locations with global spatial distribution and cloud coverage below 10%. SpectralEarth, as illustrated in Fig. 1, represents an important leap in scale, being significantly larger than existing HSI datasets, and spans a wide variety of landscapes to reflect the full diversity of spectral signatures. Compiled from 11 636 EnMAP scenes, the dataset volume exceeds 3 TB in hyperspectral images. Notably, about 17.5% of its geospatial locations

Received 29 October 2024; revised 15 April 2025; accepted 7 June 2025. Date of publication 19 June 2025; date of current version 14 July 2025. The work of Nassim Ait Ali Braham was supported by the European Commission through the project "EvoLand" under the Horizon 2020 Research and Innovation Program under Grant 101082130). The work of Conrad M. Albrecht and Yi Wang was supported by the Helmholtz Association through the Framework of Helmholtz AI, under Grant: ZT-I-PF-5-01 — Local Unit Munich Unit Aeronautics, Space and Transport (MASTr). The work of Julien Mairal was supported by ERC under Grant 101087696 (APHELEIA Project). This work was supported in part by ANR 3IA MIAI@Grenoble Alpes (ANR-19-P3IA-0003), and in part by the Helmholtz Association's Initiative and Networking Fund on the HAICORE@FZJ partition. *(Corresponding author: Nassim Ait Ali Braham.)*

Nassim Ait Ali Braham is with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany, and also with Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany (e-mail: nassim.aitalibraham@dlr.de).

Conrad M. Albrecht is with Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany.

Julien Mairal and Jocelyn Chanussot are with University Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Yi Wang and Xiao Xiang Zhu are with the Chair of Data Science in Earth Observation, Technical University of Munich, 80333 Munich, Germany.

Data are available online at: https://geoservice.dlr.de/web/datasets/enmap_spectralearth, Code is available online at: https://github.com/AABNassim/spectral_earth.

Digital Object Identifier 10.1109/JSTARS.2025.3581451

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

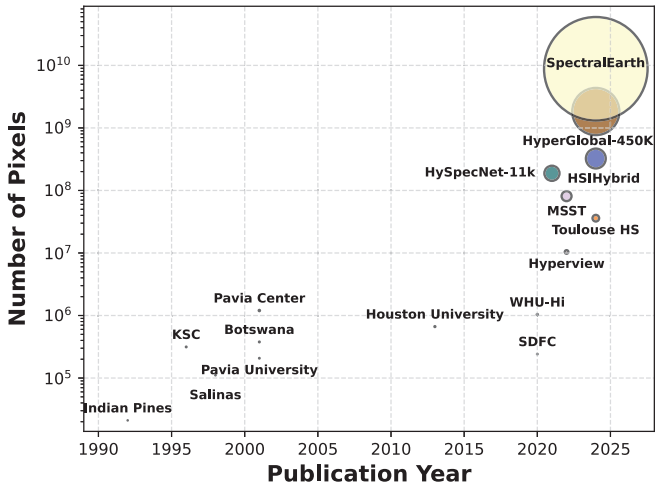


Fig. 1. Visualization of scale for various hyperspectral datasets published within the past three decades illustrating the volume of the SpectralEarth dataset (area of circles).

include time series data to enable multitemporal HSI analysis. To harness the rich information in hyperspectral data, we modify conventional vision backbones such as ResNet [17] and ViT [18] with a spectral adapter module. This enables such backbones to capture the unique spectral characteristics of HSI. Using three popular SSL algorithms, MoCo-V2 [19], [20], DINO [21], and masked autoencoder (MAE) [22], we train these backbones on SpectralEarth, to provide a plurality of pretrained models for hyperspectral image analysis. To benchmark the efficacy of our models, we introduce nine downstream datasets constructed by pairing EnMAP, DESIS, and Hyperion EO-1 imagery with land cover, crop type, and tree species labels. Our experiments demonstrate the potential of large-scale pretraining to efficiently generalize across various hyperspectral imaging contexts, including adapting to data from alternative sensors with different spectral characteristics. Our contributions are summarized as follows:

1) Assembly of *SpectralEarth*, a large-scale dataset comprising around 3.3 TB of nearly cloud-free EnMAP hyperspectral images. SpectralEarth features a global geospatial distribution encompassing over 538 974 image patches from 415 153 unique geo-locations, 73 307 including multiple timestamps.
2) Construction of nine downstream datasets designed to benchmark classification and semantic segmentation tasks for HSI.
3) An empirical evaluation of SSL in hyperspectral imaging, by pretraining various models on SpectralEarth and evaluating their generalizability across downstream tasks and sensors. In addition, we demonstrate the models' computational benefit through faster convergence in fine-tuning.

II. Related Work

A. Hyperspectral Datasets

Traditionally, hyperspectral benchmark datasets have been limited in geographical coverage, often confined to a single or a limited number of scenes [13], [29], [32], [33]. This limitation is primarily rooted in the expenses associated with airborne hyperspectral sensor surveys on the one hand and the collection of ground truth annotations and in situ measurements on the other hand. While collecting quality labels remains a challenge today, the increased accessibility to more cost-effective sensors and the launch of hyperspectral satellite missions enabled coverage of larger regions with HSI. Developments in SSL and foundation models have also driven interest in creating large-scale unlabeled HSI datasets. For instance, the HySpecNet-11k [14] collected about 11 500 unlabeled hyperspectral patches from the EnMAP satellite. HySpecNet-11 k is designed to benchmark (unsupervised) image compression techniques. Similarly, MSST [15] compiled a dataset of 19 792 EnMAP patches for SSL pretraining. While HySpecNet-11 k and MSST have contributed to increasing the scale of existing HSI datasets, they remain limited in size to train foundation models. The recently introduced HyperGlobal-450 K dataset [16] encompasses more than 450 000 HSI patches derived from 486 Hyperion EO-1 and 215 Gaofen-5B scenes. While this dataset significantly increases the scale of publicly available HSI data, it presents certain limitations. First, its geographical distribution is less diverse, with a substantial part of the data concentrated in a few provinces of China. Moreover, it relies on imagery from the decommissioned EO-1 satellite. In contrast, SpectralEarth offers more extensive global coverage using data from the currently operational EnMAP satellite, which is continuously collecting data. In addition, SpectralEarth is approximately five times larger in pixel count and includes a temporal dimension, increasing its value for pretraining of hyperspectral foundation models. A comparison of existing labeled and unlabeled HSI datasets is provided in Table I.

B. SSL for Remote Sensing

SSL emerged as a powerful paradigm for learning representations from unlabeled data [34]. Early SSL methods were based on engineered pretext tasks to encourage the model to learn distinct features by solving auxiliary problems, such as image colorization [35], jigsaw puzzles [36], and rotation angle prediction [37]. Recently, progress in SSL has been primarily focused on two families of methods.

1) *Joint-embedding architectures:* These methods train models to generate consistent representations for augmented views of the same input, enforcing invariance to these augmentations. To avoid convergence to trivial solutions, i.e., *model collapse* to a constant representation, several strategies have been developed: Contrastive approaches like MoCo [19] and SimCLR [38] use negative pairs to encourage learning discriminative representations. Clustering-based methods, such as SwAV [39], enforce balanced cluster assignment. Distillation-based frameworks like BYOL [40] and DINO [21] use an asymmetric teacher–student architecture. Methods like Barlow-Twins [41] and VICReg [42] explicitly use a loss regularization term to control the variance and feature correlation in the embedding space.

TABLE I
SUMMARY OF HYPERSPECTRAL REMOTE SENSING DATASETS: SPECTRALEARTH CONTAINS NEARLY NINE GIGAPIXELS OF 202 ENMAP BANDS EACH AT 30 M SPATIAL RESOLUTION

Dataset	# Img.	Size	Sensor	Bands	GSD	Multi-temporal
Indian Pines [23]	1	145×145	AVIRIS	224	20m	×
KSC [24]	1	512×614	AVIRIS	224	18m	×
Salinas [24]	1	512×217	AVIRIS	224	3.7m	×
Pavia University [24]	1	610×340	ROSIS	103	1.3m	×
Pavia Center [24]	1	1096×1096	ROSIS	102	1.3m	×
Botswana [24]	1	1476×256	EO-1	242	30m	×
Houston University [25]	1	349×1905	ITRES-CASI	144	2.5m	×
SDFC [26]	1	1201×201	HAHS	63	0.5m	×
Chikusei [27]	1	2517×2335	Hyperspec-VNIR-C	128	2.5m	×
WHU-Hi [28]	3	1217×303, 940×475, 550×400	Nano-Hyperspec	270-274	0.043-0.463m	×
Hyperview [29]	∼3k	avg. ∼60×60	HySpex	150	2m	×
Toulouse HS [30]	9	∼4000×1000	HySpex	310	1m	✓
HySpecNet-11k [14]	∼11k	128×128	EnMAP	202	30m	×
MSST [15]	∼20k	64×64	EnMAP	200	30m	×
HSIHybrid [31]	∼4m	9×9	Multiple	Variable	Variable	×
HyperGlobal-450K [16]	∼447k	64×64	EO-1 and Gaofen-5B	242 and 330	30m	×
SpectralEarth	∼538k	128×128	EnMAP	202	30m	✓

This equates to an area of a little over six million square kilometers. The rows above the dashed line concern manually annotated datasets, while the part below lists unlabeled datasets for SSL.

2) *Masked image modeling (MIM):* Inspired by masked language modeling, MIM methods predict missing parts of an image from a masked input. Existing methods differ in their reconstruction targets, loss functions, or network architectures. For example, MAE [22] and SimMIM [43] reconstruct raw pixels, BEiT [44] predicts discrete visual tokens, while Data2Vec [45] and MSN [46] employ a teacher–student architecture to regress targets generated from the teacher branch.

In remote sensing, SSL techniques have been applied and adapted to multispectral and high-resolution RGB imagery. SeCo [47] compiled a 200K-location dataset comprising Sentinel 2 images with multiple temporal views (one per season) and introduced seasonal contrast augmentation. The results supported the benefit of incorporating seasonal invariance in the training objective of contrastive methods. Building on this approach, SSL4EO-S12 [48] provided colocated Sentinel 2 and Sentinel 1 images and pretrained models for both sensors. Subsequently, SSL4EO-L [49] extended this approach to Landsat imagery, further expanding the available pretraining data, downstream datasets, and pretrained models. Several studies have adapted MAE for remote sensing imagery. SatMAE [50] tokenized inputs across spectral and temporal dimensions, while Scale-MAE [51] introduced ground-sampled distance positional encoding for multiscale imagery. GFM [52] assembled GeoPile, a pretraining dataset comprising high-resolution RGB images, by merging various existing datasets. GeoPile was used to train an MAE model with an additional distillation loss guided by an ImageNet-22 K pretrained teacher. SpectralGPT [53] and S2MAE [54] proposed a hierarchical MAE pretraining approach based on a spatial–spectral vision transformer for Sentinel 2 imagery. Prithvi [55] and Prithvi-EO-2.0 [56] built large-scale pretraining datasets from Harmonized Landsat-Sentinel-2 time-series and spatial/spectral/temporal ViT backbones using MAE. OmniSat [57] proposed a multisensor framework that combines masked autoencoding with contrastive learning, using high-resolution aerial imagery and multimodal time series from Sentinel-1 and Sentinel-2. CROMA [58] combined contrastive learning and MIM for SAR-Optical SSL. msGFM [59] proposed a multisensor MAE model with a cross-sensor reconstruction loss for paired images on RGB, Sentinel 1, Sentinel 2, and DSM data. SenPa-MAE [60] introduced a sensor-aware MAE pretraining strategy by incorporating metadata such as ground sampling distance and spectral response functions to allow sensor generalization. The model was trained with MAE on Sentinel-2, Planet-SuperDove, and Landsat-8/9. AnySat [61] leveraged I-JEPA [62] to learn joint representations across several sensors. The model was trained on a combination of multiple datasets, including high-resolution imagery, Sentinel 1 and 2, MODIS, Landsat-6/7/8, and ALOS-2 time series.

In the hyperspectral domain, most SSL works focus on pixel-level classification (with or without spatial context) where the pretraining scene is the same as the test image [63], [64], [65], [66], [67], [68], [69]. While this approach is effective for classical HSI datasets, the learned representations cannot generalize across scenes. To address this issue, HSIMAE [31] proposed HSIHybrid, a compilation of existing HSI datasets, and pretrained an MAE-like model. This approach effectively increases data diversity but poses challenges when combining sensors with different characteristics, requiring preprocessing steps such as PCA to unify the number of spectral bands. In addition, the scale of HSIHybrid remains limited when training HSI foundation models. Recent studies have explored collecting unlabeled HSI data for pretraining. For example, MSST [15] used MAE on EnMAP imagery, introducing decoupled spectral and spatial attention blocks to reduce the computational cost associated with spatial–spectral tokenization. DOFA [70] proposed a general-purpose architecture that dynamically generates patch embedding weights from sensor wavelength metadata. The model was trained using MAE combined with a distillation loss on RGB, multispectral, SAR, and EnMAP hyperspectral imagery. HyperSIGMA [16] proposed an MAE-inspired approach

with independent spectral and spatial transformers and evaluated their model on several hyperspectral downstream tasks. However, the data used for pretraining these models is less diverse than SpectralEarth. In addition, these studies focus exclusively on ViTs and MAE, excluding ConvNets and other pretraining methodologies. In contrast, our work explores a more diverse set of architectures and pretraining strategies and provides a library of pretrained models for HSI applications.

III. SpectralEarth and Benchmarks

This section introduces the SpectralEarth dataset and related downstream datasets for benchmarking.

A. SpectralEarth

1) Data Acquisition: The EnMAP data were sourced from the DLR GeoPortal. A total of 11 636 tiles were retrieved via the portal's graphical interface through a significant manual effort. All tiles were carefully selected based on human visual inspection. Cloud coverage was kept below ∼10%. The selection process encompassed the entire EnMAP archive from 27 April 2022 to 24 April 2024. All acquired tiles underwent radiometric, geometric, and atmospheric corrections using the L2A processor from the EnMAP mission [71].

2) Data Preprocessing: The EnMAP tiles have been split into geospatial patches—each of size 128 × 128 pixels where an individual pixel includes 224 bands. Bands dominated by water absorption, specifically [127–141] and [161–167], were excluded due to the frequent presence of missing values (`no-data`). Removing those bands is a common practice in hyperspectral dataset creation [14], [32]. As a result, the total number of bands per image was reduced to 202.

3) Temporal Views Extraction: To exploit all available tiles, we utilized overlaps between EnMAP data acquisitions to generate time-series of EnMAP patches.[1] At the same time, SpectralEarth patches for fixed geo-locations never overlap. This process required (a) the identification of tiles that overlap spatially and (b) extracting patches from the intersection of those tiles. Time series include valuable information on landscape evolution and provide characteristic signals for seasonal variation—an asset for contrastive learning [47], [48]. Algorithm 1 outlines the steps taken to generate temporal positives by high-level pseudocode. An *overlap graph* is first constructed to identify spatial intersections between tiles. For each tile, the algorithm forms candidate subsets of overlapping tiles, computes their joint intersection, and extracts patches of fixed size from these regions. We use an R-tree to efficiently track and store patch locations and avoid overlaps.

As a result of Algorithm 1, we identified a few long time series in the EnMAP archive. For those involving more than 25 intersecting tiles, calculating intersections of all subsets was computationally intractable due to combinatorial explosion. To address this, we implemented several heuristics to optimize the computation.

[1]EnMAP records hyperspectral imagery based on a schedule generated from user requests [7], i.e., the identification of spatially aligned patches for multiple timestamps is a valuable asset our data curation provides.

Algorithm 1: Temporal Views Extraction.

```
procedure Main Procedure
  tiles ← EnMAPData
  overlap_graph ← GetOverlaps(tiles)
  R_tree ← K_0 ▷ empty tree, for SpectralEarth patches
  for tile in tiles do
    combs ← BuildCombinations(tile, overlap_graph)
    for tile_subset in combs do
      intersection ← Intersection(tile_subset)
      patches ← Patchify(intersection)
      INSERT(R_tree, patches)
    end for
  end for
end procedure

function BuildCombinations(tile, overlap_graph)
  combinations ← GetEdges(tile, overlap_graph)[2]
  for subset size n in [3, 4, ...] do
    get n-tuples from (n-1)-tuples in combinations
    compute intersections of all n-tuples
    keep largest n-tuples by area
    if no valid n-tuple found then
      break
    end if
    add n-tuples to combinations
  end for
  return combinations
end Function
```

1) Consider only a subset of possible intersections using a breadth-first search approach.
2) Avoid redundant computations across tiles.
3) Parallelization of the code across all connected components of the overlap graph.

Following this approach, we extracted 73 307 nonoverlapping locations with multiple timestamps. For areas with a single timestamp, we patchified avoiding any spatial overlap. Correspondingly, in SpectralEarth, we extracted 415 153 locations with a spatial extent of 128 × 128 pixels, totaling 538 974 patches. The spatial coverage of SpectralEarth showcases the dataset's global diversity, in land-cover types Fig. 2 and geographical distribution Fig. 3. Notably, the data distribution varies across geo-locations, reflecting the flight request-based archive composition of EnMAP.

We analyze the temporal coverage of our dataset in Fig. 5. Fig. 5(a) illustrates the distribution of timestamp counts in SpectralEarth. Over the two-year period, 17.5% of locations were covered by more than one timestamp in the EnMAP archive—with the majority having only two timestamps. This highlights the challenge of collecting extensive hyperspectral time series data with broad geographical coverage today. Missions such as ESA's CHIME satellite constellation [10] will eventually resolve the issue. Fig. 5(b) shows the histogram of time gaps

Fig. 2. Mosaic of pseudo-RGB representatives from the SpectralEarth dataset to showcase various landscapes, including urban areas, agricultural land, deserts, forests, and water bodies. Each image is presented with a false-color composite using the first three principal components.

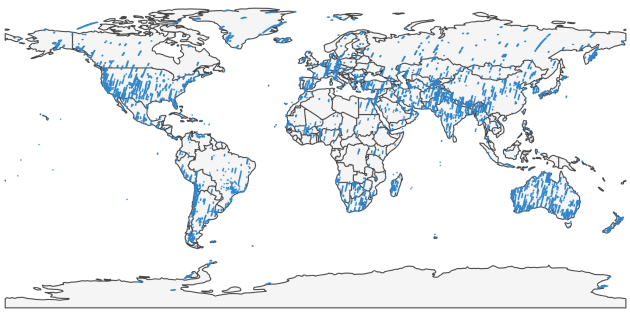


Fig. 3. Geographical distribution of SpectralEarth. The map depicts the global coverage of hyperspectral images within our dataset, demonstrating its extensive geographical scope.

between consecutive acquisitions, restricted to samples with at least two timestamps. We observe that the distribution is skewed toward smaller values. Notably, the highest peak is 4 days, matching EnMAP's revisit period. The median gap between two consecutive acquisitions across all samples is 27 days.

We also analyze the seasonal distribution of EnMAP acquisitions, as shown in Fig. 5(c). SpectralEarth includes data from all months, with a relatively balanced distribution. However, certain disparities can be observed, e.g., January is underrepresented while July, August, and March are over represented. This can be attributed to several factors, as follows:

1) Increased cloud cover during winter months, which inherently limits the availability of clear imagery.
2) A satellite outage in December-January 2023 and December-January 2024, which resulted in a temporary interruption of data acquisition.
3) The task scheduling nature of the EnMAP mission. As EnMAP is operationally tasked based on specific use cases, the data distribution inherently reflects real-world demands of satellite imagery, leading to a natural variation in seasonal coverage.

B. Downstream Tasks

To evaluate the models pretrained on SpectralEarth, we assembled nine downstream datasets for benchmarking. Each benchmark involves a subset of EnMAP/DESIS/EO-1 images aligned with specific geospatial products focusing on different aspects of land cover, agricultural, and forest analysis. While the labels in these datasets carry some inherent uncertainty, they are sufficiently accurate for model evaluation. Similar strategies have previously been employed to create remote sensing datasets and geospatial foundation model benchmarking [72], [72], [73]. An overview of the SpectralEarth downstream tasks is illustrated in Fig. 4

1) EnMAP-CORINE—Land Cover Classification: This dataset pairs EnMAP imagery with the *CORINE* land cover database for Europe. The CORINE land cover map has a 100 m spatial resolution, which is coarse-grained relative to EnMAP's 30 m resolution. The CORINE product reports an overall thematic accuracy exceeding 85%. Our SpectralEarth image patches correspond to approximately 38 $\times$ 38 CORINE land cover pixels. Consequently, we create a multilabel classification benchmark where each SpectralEarth patch is annotated with the set of CORINE classes covered. The resulting subset, called EnMAP-CORINE, includes 11 000 patches, each multilabeled with 19 distinct classes. These classes have been aggregated from the 44 categories defined in CORINE following the taxonomy proposed by BigEarthNet-MM [74]. The EnMAP-CORINE dataset exhibits a diverse range of land cover types. As Fig. 6(a) demonstrates, some classes such as "Arable Land" and "Complex cultivation patterns" are more frequent, while others like "Coastal/Inland wetlands" and "Beaches, dunes, sand" are comparatively rare. This class imbalance reflects the distribution of natural land cover in Europe.

2) EnMAP-CDL—Crop Type Segmentation: This dataset aligns EnMAP images with the cropland data layer (CDL) [75] product for the United States. The CDL map is published annually by the USDA's National Agricultural Statistics Service since 2008. It targets agricultural crop classification. In addition to cultivated crops such as Corn and Soybeans, CDL encompasses noncultivated areas like Grassland and Shrubland. According to the USDA metadata, classification accuracy is generally between 85% and 95% for major crop-specific classes.

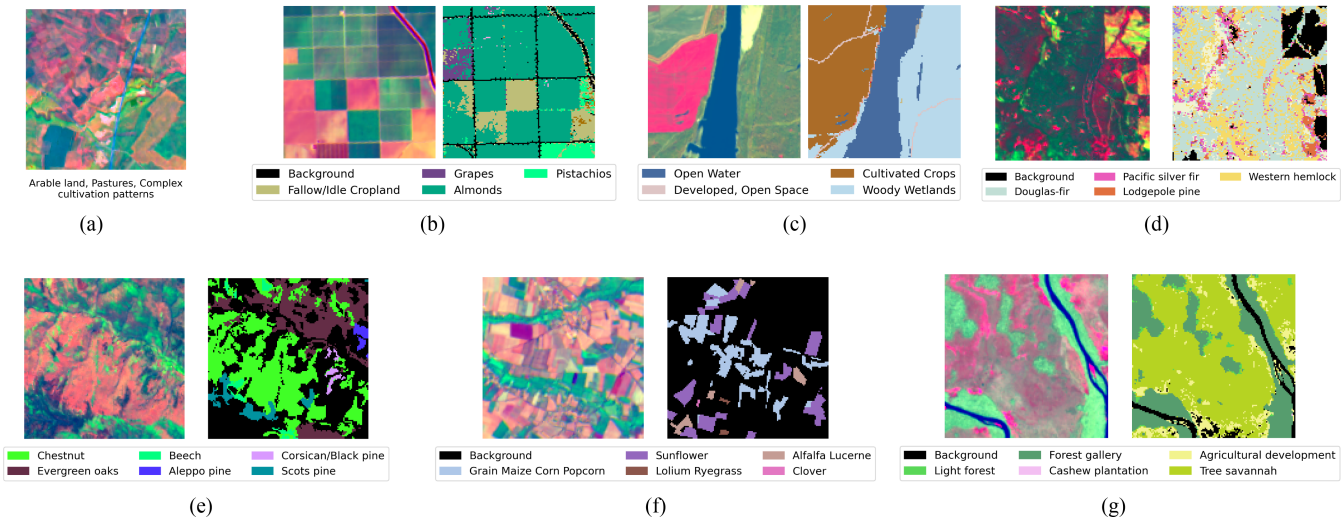


Fig. 4. Sample pseudo-RGB images from the proposed EnMAP benchmarks. (a) CORINE. (b) CDL. (c) NLCD. (d) TreeMap. (e) BD-Foret. (f) EuroCrops. (g) BNETD.

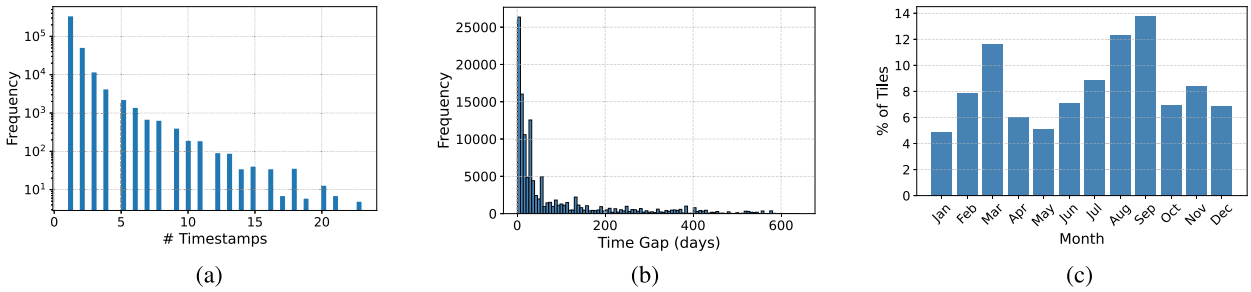


Fig. 5. Temporal distribution of SpectralEarth. (a) Number of timestamps per location. (b) Time gaps between acquisitions. (c) Monthly acquisition frequency. (a) Number of timestamps per geo-location (log scale). (b) Gap between consecutive timestamps. (c) Monthly distribution aggregated over 2 years.

To minimize the impact of seasonal variations on crop classification, the dataset utilizes SpectralEarth patches from the summers of 2022 and 2023. We mask out all nonagricultural classes, as well as winter crops (e.g., winter wheat), and crops that are likely to be harvested before the end of summer (e.g., Oats). Given the strong imbalance among CDL classes, we resampled the class distribution for a more balanced representation. The resulting dataset comprises 1600 patches with a total of 14 classes. The class distribution of the EnMAP-CDL dataset is shown in Fig. 6(d). Overall, the dataset exhibits a reasonable level of class balance among the different crop types, albeit some over represented classes such as Corn.

3) EnMAP-NLCD—Land Cover Segmentation: This downstream dataset leverages the National Land Cover Database (NLCD) [76] from the U.S. Geological Survey (USGS) to provide pixel-level land cover annotations. The NLCD map, a comprehensive land cover product, has been published every 2–3 years since 2001, with data available back to 2019. The NLCD 2019 product reports an overall accuracy of approximately 91%. We pair EnMAP patches over the US to the NLCD map, and manually filter out cloudy and snowy images. The resulting dataset contains 13 500 patches, each annotated with pixel-wise land cover class labels from one out of the 15 land cover classes. As depicted in Fig. 6(b), the class distribution is well-balanced.

4) EnMAP-TreeMap—Tree Species Segmentation: The EnMAP-TreeMap dataset maps EnMAP imagery from 2022 and 2023 to TreeMap [77] product, a 30 m resolution map of forest attributes across the continental United States. We use the forest-type layer, which encodes the dominant species group based on live tree stocking. To mitigate the temporal mismatch between TreeMap (2016) and EnMAP acquisitions, we mask regions affected by deforestation using the Hansen Global Forest Change dataset [78]. The resulting dataset includes 10 000 patches covering 24 tree species, enabling fine-grained classification of forest types. The class distribution is depicted in Fig. 6(c).

5) EnMAP-BDForet—Tree Species Segmentation: The EnMAP-BDForet dataset combines EnMAP hyperspectral imagery acquired between 2022 and 2024 with the BDForet V2 dataset from the Institut Géographique National (IGN) France.[2] We project the dominant tree species attribute onto the EnMAP grid to generate the labels. To mitigate land cover changes, we filter out areas affected by deforestation using the latest IGN

[2][Online]. Available: https://inventaire-forestier.ign.fr/spip.php?article646

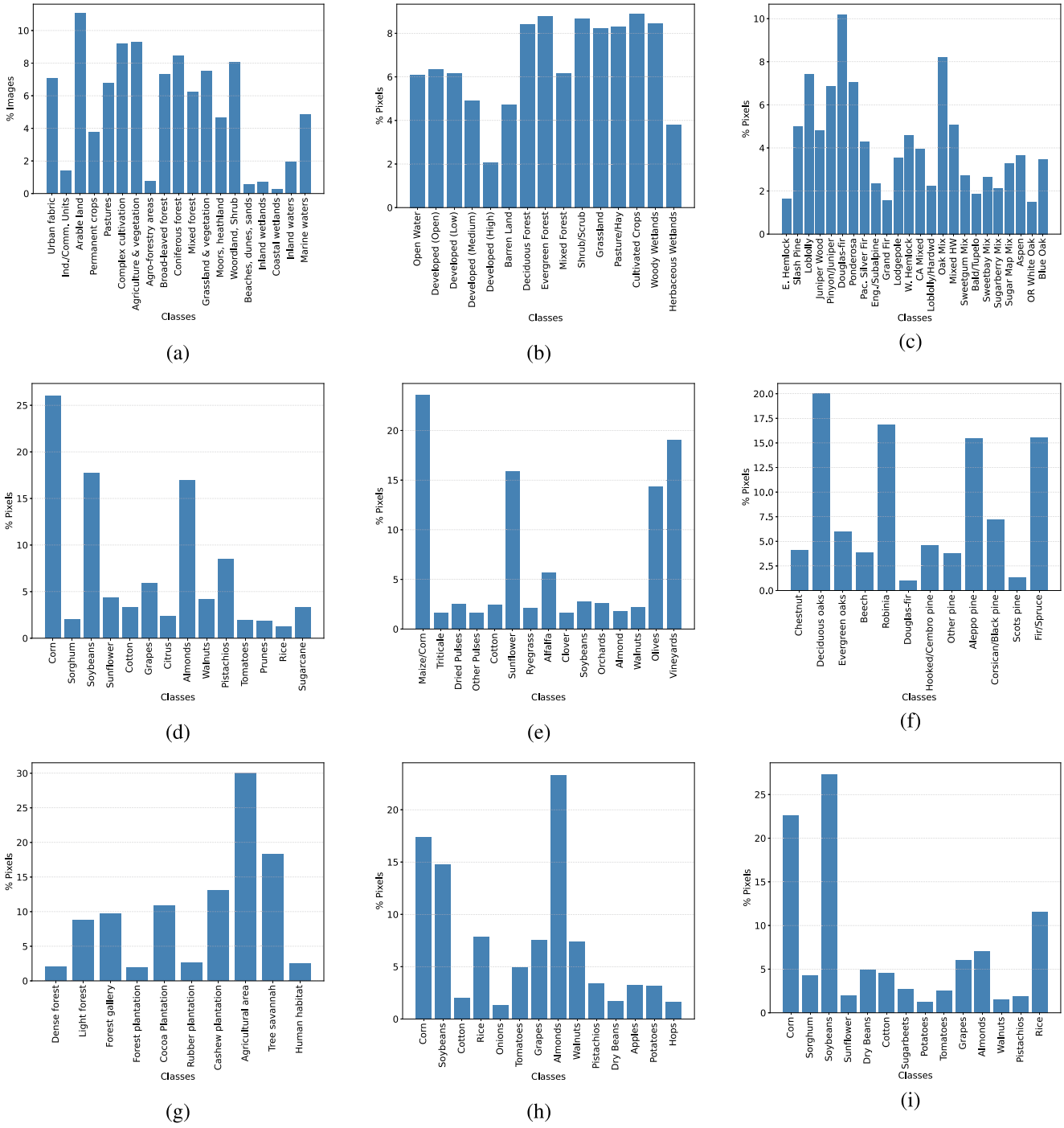


Fig. 6. Class distributions for the downstream datasets introduced in SpectralEarth. (a) EnMAP-CORINE. (b) EnMAP-NLCD. (c) EnMAP-Treemap. (d) EnMAP-CDL. (e) EnMAP-EuroCrops. (f) EnMAP-BDForet. (g) EnMAP-BNETD. (h) DESIS-CDL. (i) EO1-CDL.

forest mask from 2023. The resulting dataset comprises 2550 patches with 12 tree species. The class distribution is shown in Fig. 6(f).

6) EnMAP-EuroCrops—Crop Type Segmentation: The EnMAP-EuroCrops dataset aligns EnMAP imagery with crop type labels derived from the EuroCrops dataset [79], which is sourced from national agricultural inventories across European countries. This dataset focuses on crop-type classification for four countries—France, Germany (Brandenburg), Czechia, and Spain—for the year 2023, with additional coverage for France in 2022. Similar to CDL, we exclude crops which are likely to be harvested before our EnMAP acquisition. The resulting dataset consists of 1800 patches with 15 crop-type classes. Approximately 68% of the patches are located in France, 27% in Spain, and the remaining 5% in Germany and Czechia. The class distribution is illustrated in Fig. 6(e).

7) EnMAP-BNETD—Land Cover Segmentation: To explore the geographical generalizability of our models beyond pre-training regions, we introduce the EnMAP-BNETD dataset, covering land cover classification in Ivory Coast. This dataset

uses the BNETD 2020 Land Cover Map,[3] produced by the BNETD-CIGN with support from the European Union, and validated through extensive field campaigns in 2022 and 2023. The reported overall accuracy is 91%. We pair EnMAP hyperspectral imagery with land cover labels from this product, resulting in a dataset of 2100 patches with 10 classes. The class distribution is shown in Fig. 6(g).

8) DESIS-CDL—Crop Type Segmentation: The DESIS-CDL dataset contains images captured for the summers of 2018–2023 from the DLR Earth Sensing Imaging Spectrometer Mission (DESIS) instrument, matched with the corresponding CDL mask for crop segmentation. DESIS images are generated at 30 m spatial resolution with 235 bands covering wavelengths from 400 nanometers through a micron. The resulting dataset comprises 1000 images with 14 classes. Due to the limited availability of DESIS tiles, it was not possible to pair this dataset with EnMAP-CDL. Therefore, the class distribution [see Fig. 6(h)] and geographical coverage differ.

9) EO1-CDL—Crop Type Segmentation: The EO1-CDL dataset includes EO-1 Hyperion hyperspectral data paired with crop-type labels from the CDL. EO-1 Hyperion provides 220 spectral bands covering 0.357 to 2.576 μm, out of which we use 198. The dataset includes Hyperion scenes from 2002 to 2016, focusing on agricultural regions over the U.S. Due to the limited availability of Hyperion imagery in Google Earth Engine, the dataset consists of 550 patches distributed across 14 crop classes.

IV. Models

This section presents the design choices behind our models, including the choice of SSL algorithms and network architecture.

A. Pretraining Algorithms

Hyperspectral data has very different characteristics compared to natural images. Therefore, traditional SSL leaderboards in computer vision may not directly translate to the hyperspectral domain. Therefore, we explore a broad range of SSL methods to establish a set of baseline pretrained models for hyperspectral imagery. We use MoCo-V2 [19] and DINO [21] for their effectiveness in frozen encoder performance [11] and their compatibility with both CNNs and ViTs. Furthermore, we include MAE [22] for its strong performance in fine-tuning scenarios. Together, these methods form a representative set of algorithms to investigate SSL in the hyperspectral domain.

B. Network Architectures

Since we aim for large-scale coverage and pretraining with spaceborne imagery, our models must operate on large patches instead of pixels or small patches, as traditionally done in the hyperspectral literature. Most models developed at the pixel level in the HSI literature do not scale well with large patches [80], [81], [82]. On the other hand, models like ResNet [17] and ViT [18] are designed to handle larger spatial contexts. Therefore, we seek to adapt these architectures for hyperspectral imaging. A key challenge is that classical ResNet and ViTs treat spectral bands independently, failing to capture the correlations between adjacent bands and the overall characteristics of the spectrum. Our models should address the following requirements for hyperspectral imaging.

1) *Spectral feature extraction:* We aim to extract features from both the spatial and the spectral domain.
2) *Adaptability to diverse sensors:* The architecture should accommodate variability in the number of spectral bands. This property is important when transferring pretrained models to different sensors.
3) *Preserving fine-grained details:* Natural images often deal with large objects compared to medium-resolution remote sensing data. Therefore, the spatial downscaling in classical vision models can lead to a significant loss of details. Our architecture needs to preserve this fine-grained information.
4) *Computational efficiency:* 3-D convolution or spectral–spatial tokenization becomes computationally expensive for large patches, with many spectral bands, and deep architectures. In particular, self-attention scales quadratically with the number of tokens [83], and 3-D convolutions scale linearly with the number of spectral bands [84]. We, therefore, seek lightweight adaptation of classical architectures with minimal overhead.

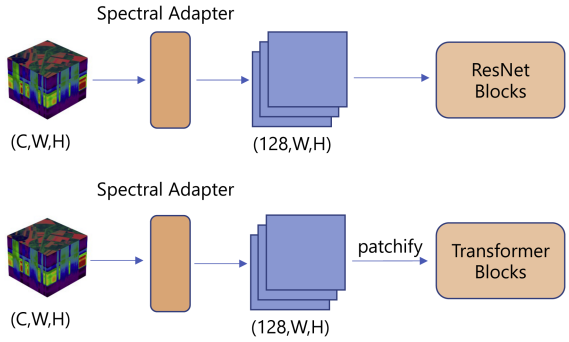


Fig. 7. Schematic representation of the proposed backbones tailored for HSI. We augment classical ResNet and ViT with a spectral adapter to process spectral information effectively.

To meet these requirements, we follow a pragmatic approach and implement a simple modular design, integrating 1-D convolutional layers as a spectral adapter at the onset of the standard vision backbones, as depicted in Fig. 7. These layers perform convolutions across the spectral dimension, generating feature maps for the core backbone. The spectral adapter consists of three (1-D Conv + BN + ReLu) layers. The kernel sizes for the 1-D convolutions are 7, 7, and 5, with strides of 5, 5, and 3, respectively. The resulting feature maps have 128 channels. To accommodate different sensors and feed the input into standard 2-D backbones, we use a global pooling layer to aggregate the remaining spectral dimensions (if any). Specifically, the modifications are as follows:

1) *Spectral ResNet:* We replace the stem layers with the spectral adapter. By removing the original stem layers

[3][Online]. Available: https://africageoportal.maps.arcgis.com/apps/webappviewer/index.html?id=88c2493e722546c09c2a0a8b394c4454

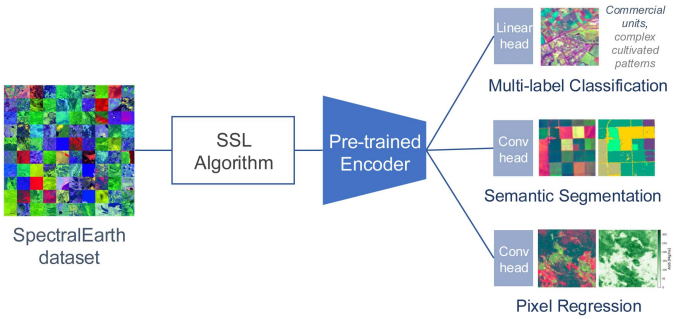


Fig. 8. Schematic overview of the pretraining/fine-tuning pipeline. An encoder is pretrained on SpectralEarth, and adapted to the several downstream tasks with task-specific heads.

in the ResNet architecture, we avoid the 4×downscaling factor of the image to better capture fine-grained details. In addition, the first bottleneck block is adjusted to take an input feature map of 128 channels instead of the ResNet's 64-channel input at this stage.

2) *Spectral ViT:* The spectral adapter is placed before the patch embedding layer, ensuring the transformer receives fixed-size spectral features as input. We use 4×4 patches instead of the typical 16×16 patches, which helps maintain fine spatial details and better preserves spectral information at the patch projection layer. Like the spectral ResNet model, we set the number of input channels in the projection layer to 128.

Although these simplifications may not capture all spectral and spatial interactions as explicitly as 3-D convolutions or spectral–spatial tokenization, they offer a practical tradeoff that scales better to large datasets and diverse sensors. Similar designs have been used effectively in HSI classification tasks [13], [85]. Our approach adapts this idea to large-scale SSL, providing a simple scalable solution for hyperspectral representation learning across sensors.

V. Experimental Setup

This section introduces the details of pretraining and evaluation protocols. A schematic overview of the pipeline is provided in Fig. 8.

A. Self-Supervised Pretraining

We pretrain a spectral ResNet-50 (Spec. RN50) and a spectral ViT-S (Spec. ViT-S) with MoCo-V2, DINO, and MAE. For larger models, Spec. ViT-Base(B)/Large(L)/Huge(H)/Giant(g), we restrict the experiments to MAE, given the high computational cost. For ViT-g, we follow the implementation from [11]. We use the augmentations from SimCLR [38] without color-jittering and gray-scale for DINO and MoCo-V2 to avoid distorting spectral information. When available, we use the temporal views as positive pairs for MoCo-V2 and DINO. For MAE, we mask 90% of the tokens. We pretrain over 100 epochs for MoCo-V2 and DINO, and 200 epochs for MAE with a batch size of 256 images. We use a `StepLR` scheduler for MoCo-V2 and `CosineAnnealingLR` for DINO and MAE.

B. Downstream Tasks

1) Multilabel Classification: We append a linear layer to the pretrained encoders and train the models with binary cross-entropy loss for 100 epochs with a batch size of 128. AdamW optimizer was employed with a cosine annealing scheduler. Data augmentation included random resized crop, horizontal, and vertical flipping.

2) Semantic Segmentation: We use a lightweight transfer protocol for the segmentation tasks, following previous works [19], [40]. The pretrained encoders are converted into segmentation models as follows.

1) *ResNet:* We remove all strides from convolutions, replace them with atrous convolutions, and append two final convolutional layers.
2) *ViT:* We reshape the tokens from the last layer and add two additional convolutional layers with upscaling to recover the full spatial resolution.

We train for 100 epochs with a cross-entropy loss. AdamW optimizer was used with a batch size of 64 for EnMAP-NLCD/TreeMap and 32 for EnMAP-CDL/BDForet/EuroCrops and DESIS-CDL. Data augmentation included random resized crop, horizontal, and vertical flipping.

3) Parameter Regression—Hyperview: We use the dataset from the hyperview challenge [29] to evaluate cross-sensor transferability. Hyperview is a soil parameter estimation dataset comprising 1732 hyperspectral patches for training and 1154 for testing, each with 150 spectral bands. Each patch is annotated with four soil parameters, including potassium (K), phosphorus pentoxide (P_2O_5), magnesium (Mg), and acidity pH. We resized all images in the Hyperview dataset to 128× 128pixels. For this downstream task, we append two fully connected layers to the backbones. The models are optimized based on the normalized MSE criterion defined in Section V-D with the AdamW optimizer for 400 epochs and a batch size of 32. As for the other tasks, we used random resized crop, horizontal, and vertical flipping during training.

4) Pixel-Wise Regression—HyBiomass: We evaluate our models on the above ground forest biomass dataset HyBiomass [86], which is a pixel regression task. The dataset consists of more than 34 000 patches distributed globally over forest areas. The above-ground biomass estimates are derived from the Global Ecosystem Dynamics Investigation lidar mission. We set the batch size to 128 and use the AdamW optimizer for 50 epochs. We use the same data augmentations as for the segmentation downstream tasks.

C. Evaluation Protocols

We consider multiple evaluation protocols with various levels of fine-tuning of the pretrained models.

1) *Frozen Encoder:* Evaluates the quality of learned representations without fine-tuning the backbone. It is the most common evaluation protocol in the SSL literature [11], [87].

2) *Fine-tuning:* Fine-tuning of all model parameters on downstream datasets to evaluate the transferability of the pretrained weights.
3) *Fine-tuning Adapter:* While frozen encoder evaluation is a common measure of the quality of learned representations, it does not yield optimal results. In remote sensing, it is often beneficial to adapt the model to the specific distribution of the downstream dataset, which can be influenced by several factors: geo-location, atmospheric conditions, seasonal changes, etc. In addition, the relevance of individual spectral bands may vary depending on the downstream task. Therefore, we consider an intermediate setting where only the initial spectral adapter block is fine-tuned.

D. Evaluation Metrics

Depending on the nature of the downstream task, we report the following metrics.

1) *Classification:* The multilabel F1-score is reported for the EnMAP-CORINE dataset.
2) *Semantic segmentation:* The mean intersection over union (mIoU) is reported for all semantic segmentation datasets.
3) *Regression:* For the Hyperview dataset, we follow the metric defined in the challenge [29]: let $\text{MSE}_i, i \in [1, 4]$ be the mean squared error of the predictor for soil parameter i. Let $\text{MSE}_i^{\text{base}}$ be the corresponding MSE of a base predictor returning the mean value of the soil parameter, calculated over the training set. The reported normalized MSE is defined as: $\text{MSE}_{\text{norm}} = \sum_{i=1}^{4} \text{MSE}_i / \text{MSE}_i^{\text{base}}$.
4) *Pixel-wise Regression:* we report the R^2 value for the HyBiomass dataset.

E. Comparison Methods

We compare our pretrained models against several baselines and recent foundation models for hyperspectral imagery.

1) *Random Frozen Encoder:* We use this baseline as a sanity check to assess whether the learned representations provide useful features for downstream tasks.
2) *Training from Scratch:* A strong baseline that has shown competitive performance when trained with sufficient data and epochs [88].
3) *DOFA-(B/L) [70]:* A foundation model capable of adapting to various sensors given the input wavelengths. The model includes EnMAP data in its pretraining corpus. We evaluate both the base (DOFA-B) and large (DOFA-L) variants.
4) *SpatSigma-(B/L) [16]:* HyperSigma is a recent foundation model for hyperspectral imagery pretrained on Gaofen and Hyperion EO-1 data. We use only the spatial encoder for simplicity, as this article reports limited benefits when combining the spatial and spectral branches. We report results for the base and large variants and refer to them as SpatSigma-B and SpatSigma-L respectively. To mitigate the mismatch in the number of spectral bands, we resample the patch embedding layers with linear interpolation for our experiments.

VI. Benchmark Results

Our main results are summarized in Tables II and III. We include random initialization baselines for each downstream task, along with our pretrained models. Due to hundreds of spectral channels of the SpectralEarth patches, a comparison with RGB-ImageNet weights or multispectral pretrained models is of little use.

A. Comparing SSL Algorithms

Table II summarizes our results for three SSL algorithms with the Spec. RN50 and Spec. ViT-S backbones.

1) Multilabel Classification: Pretrained models outperform random weights in linear evaluation for MoCo-V2 and DINO on EnMAP-CORINE, with DINO yielding the best results, demonstrating the ability of joint-embedding methods to learn robust representations, even in the absence of fine-tuning.

In addition, we observe that MAE underperforms in linear probing, which is consistent with the existing literature [89]. This can be attributed to the reconstruction pretraining objective, which focuses on low-level features, whereas joint-embedding methods learn more global and semantic representations. For full fine-tuning, CNN pretrained models with MoCo and DINO are on par with/slightly worse than training from scratch. In contrast, the Spec. ViT-S model benefits from pretraining in all settings compared to random initialization, with MAE yielding the best results.

Fine-tuning the spectral adapter yields a notable performance boost compared to linear evaluation. This observation aligns with reports in the literature [90]: fine-tuning early network layers is an efficient compromise between frozen weights versus complete fine-tuning. In terms of F1 score, only fine-tuning the adapter (0.3% of network parameters) is, in most settings, less than one point below training from scratch, highlighting the efficient adaptability of our pretrained models.

2) Semantic Segmentation: Pretrained models with a frozen encoder consistently outperform random weights across all segmentation tasks. This shows that, without any fine-tuning, the models have learned useful features for segmentation. MoCo-V2 and DINO pretrained models yield comparable results for both the CNN and the ViT backbones. However, MAE, as a feature extractor, consistently outperforms MoCo-V2 and DINO for Spec. ViT-S. This can be attributed to the pixel-level pretraining objective, which is more suitable for segmentation tasks (as opposed to image-level problems).

For complete fine-tuning, MoCo-V2 pretrained CNNs consistently outperform training from scratch on all segmentation tasks, while DINO is on par with the random initialization baseline. Notably, the gaps are more pronounced for BDForet and EuroCrops, where MoCo-V2 yields an improvement of 3 and 2 points of mIoU, respectively. For the ViT model, MAE + fine-tuning performs best in all settings, yielding consistent improvements of 2 to 3 points of mIoU. Depending on the dataset, the second-best models are either MoCo-V2 or DINO, with smaller improvements compared to training from scratch.

TABLE II
BENCHMARK RESULTS OF THE SPECTRALEARTH DOWNSTREAM TASKS FOR SPEC

Model Configuration			Classif. (F1)	Segmentation (mIoU)						Pix Reg. (R^2)
Arch.	Weights	Protocol	CORINE	CDL	NLCD	EUROCROPS	TREEMAP	BDFORET	BNETD	HYBIOMASS
Spec. RN50	Random	Frozen	69.92 ± 0.51	62.65 ± 0.80	35.94 ± 0.38	54.83 ± 0.83	35.87 ± 0.23	57.11 ± 0.80	41.35 ± 0.30	0.30 ± 0.01
		Full FT	**77.77 ± 0.38**	77.28 ± 0.13	47.95 ± 0.04	68.93 ± 1.41	46.47 ± 0.04	75.50 ± 0.42	48.97 ± 0.10	0.48 ± 0.00
	MoCo-V2	Frozen	74.07 ± 0.14	70.97 ± 0.19	41.87 ± 0.06	61.70 ± 0.08	41.67 ± 0.05	67.60 ± 0.34	44.84 ± 0.05	0.39 ± 0.00
		Full FT	77.43 ± 0.46	**77.53 ± 0.25**	**48.11 ± 0.03**	**70.95 ± 0.63**	46.94 ± 0.03	**78.56 ± 0.32**	49.54 ± 0.04	**0.50 ± 0.00**
		Adapter	76.82 ± 0.39	75.84 ± 0.11	44.62 ± 0.03	66.71 ± 0.23	44.84 ± 0.04	73.86 ± 0.19	47.46 ± 0.04	0.45 ± 0.00
	DINO	Frozen	76.61 ± 0.31	71.02 ± 0.13	41.83 ± 0.03	60.29 ± 0.29	42.03 ± 0.03	67.07 ± 0.11	44.32 ± 0.02	0.40 ± 0.00
		Full FT	77.71 ± 0.55	77.49 ± 0.23	47.62 ± 0.07	69.52 ± 1.00	**47.08 ± 0.13**	74.68 ± 0.90	**49.74 ± 0.31**	0.50 ± 0.00
		Adapter	76.96 ± 0.37	75.51 ± 0.15	44.38 ± 0.03	66.62 ± 0.35	44.70 ± 0.07	73.90 ± 0.11	47.21 ± 0.07	0.45 ± 0.00
Spec. ViT-S	Random	Frozen	70.59 ± 0.62	65.96 ± 0.72	36.94 ± 0.30	55.78 ± 0.52	38.72 ± 0.25	61.89 ± 0.85	41.55 ± 0.22	0.34 ± 0.00
		Full FT	77.63 ± 0.37	74.77 ± 0.22	45.31 ± 0.62	66.87 ± 0.62	44.43 ± 0.13	74.00 ± 0.49	46.57 ± 0.11	0.40 ± 0.02
	MoCo-V2	Frozen	73.99 ± 0.04	71.40 ± 0.32	39.99 ± 0.08	60.21 ± 0.39	41.76 ± 0.06	68.50 ± 0.12	43.50 ± 0.05	0.40 ± 0.00
		Full FT	78.37 ± 0.34	75.59 ± 0.14	45.92 ± 0.05	67.27 ± 0.29	45.03 ± 0.10	75.24 ± 0.28	47.02 ± 0.07	0.43 ± 0.01
		Adapter	76.72 ± 0.35	75.28 ± 0.19	43.18 ± 0.08	66.46 ± 0.21	44.09 ± 0.07	73.85 ± 0.23	45.73 ± 0.10	0.44 ± 0.00
	DINO	Frozen	75.11 ± 0.13	71.56 ± 0.35	40.47 ± 0.06	58.76 ± 0.27	41.89 ± 0.13	68.16 ± 0.43	43.12 ± 0.10	0.40 ± 0.00
		Full FT	78.67 ± 0.34	76.10 ± 0.31	45.64 ± 0.15	65.56 ± 0.76	45.44 ± 0.15	73.56 ± 0.41	47.11 ± 0.14	0.48 ± 0.00
		Adapter	77.00 ± 0.50	74.83 ± 0.07	42.86 ± 0.08	64.78 ± 0.65	43.54 ± 0.11	72.38 ± 0.71	44.99 ± 0.17	0.44 ± 0.00
	MAE	Frozen	72.85 ± 0.20	72.01 ± 0.28	41.08 ± 0.05	62.14 ± 0.18	41.19 ± 0.15	68.92 ± 0.73	43.34 ± 0.26	0.42 ± 0.00
		Full FT	**79.17 ± 0.55**	**77.16 ± 0.31**	**47.61 ± 0.10**	**69.87 ± 0.16**	**45.80 ± 0.15**	**76.54 ± 0.23**	**49.36 ± 0.07**	**0.51 ± 0.01**
		Adapter	76.75 ± 0.35	75.21 ± 0.21	43.64 ± 0.07	67.18 ± 0.58	43.45 ± 0.14	74.89 ± 0.47	45.89 ± 0.12	0.45 ± 0.01

RN50 and Spec ViT-S under different evaluation protocols: frozen backbone, full fine-tuning, and fine-tuning the spectral adapter only. The best score for *each dataset-backbone pair* is in bold, and the second best is underlined.

TABLE III
BENCHMARK RESULTS FOR LARGE VISION TRANSFORMERS PRETRAINED ON SPECTRALEARTH WITH MAE

Model Configuration		Classif. (F1)	Segmentation (mIoU)						Pix Reg. (R^2)
Arch.	Protocol	CORINE	CDL	NLCD	EUROCROPS	TREEMAP	BDFORET	BNETD	HYBIOMASS
Spec. ViT Models									
Spec. ViT-B	Supervised	77.17 ± 0.72	74.65 ± 0.60	45.57 ± 0.14	66.14 ± 0.43	44.51 ± 0.13	73.55 ± 0.39	46.76 ± 0.10	0.37 ± 0.02
	Frozen	75.04 ± 0.30	72.20 ± 0.16	40.82 ± 0.14	61.29 ± 0.22	39.74 ± 0.27	68.39 ± 0.20	43.37 ± 0.10	0.38 ± 0.01
	Full FT	79.49 ± 0.26	**77.44 ± 0.26**	47.59 ± 0.23	**69.34 ± 0.43**	46.10 ± 0.04	76.30 ± 0.52	**49.46 ± 0.06**	**0.51 ± 0.00**
	Adapter	77.63 ± 0.25	75.30 ± 0.32	43.81 ± 0.04	67.31 ± 0.43	43.65 ± 0.18	74.51 ± 0.15	45.52 ± 0.11	0.44 ± 0.01
Spec. ViT-L	Supervised	76.85 ± 0.67	74.28 ± 0.34	45.49 ± 0.13	65.66 ± 0.46	44.52 ± 0.08	72.93 ± 0.59	46.66 ± 0.04	0.38 ± 0.02
	Frozen	73.98 ± 0.21	72.72 ± 0.23	42.67 ± 0.03	62.13 ± 0.18	42.62 ± 0.19	71.07 ± 0.31	43.96 ± 0.18	0.42 ± 0.00
	Full FT	79.25 ± 0.65	77.41 ± 0.26	47.84 ± 0.03	68.73 ± 0.36	**46.10 ± 0.17**	76.26 ± 0.73	48.96 ± 0.39	0.49 ± 0.00
	Adapter	77.24 ± 0.07	75.94 ± 0.13	44.15 ± 0.07	67.69 ± 0.32	44.09 ± 0.09	74.89 ± 0.26	46.11 ± 0.12	0.45 ± 0.00
Spec. ViT-H	Supervised	77.18 ± 0.44	73.72 ± 0.47	45.33 ± 0.12	64.77 ± 0.48	44.45 ± 0.17	72.98 ± 0.45	46.62 ± 0.14	0.34 ± 0.05
	Frozen	73.96 ± 0.30	71.68 ± 0.21	41.87 ± 0.13	61.94 ± 0.11	41.95 ± 0.17	68.90 ± 0.12	43.48 ± 0.14	0.38 ± 0.00
	Full FT	**79.51 ± 0.30**	77.26 ± 0.18	**47.85 ± 0.14**	69.13 ± 0.49	46.05 ± 0.13	**76.57 ± 0.35**	49.38 ± 0.23	0.46 ± 0.01
	Adapter	77.23 ± 0.32	75.34 ± 0.22	43.90 ± 0.08	66.51 ± 0.14	43.75 ± 0.06	74.13 ± 0.23	45.56 ± 0.09	0.45 ± 0.01
Spec. ViT-g	Supervised	76.94 ± 0.61	73.44 ± 0.38	45.03 ± 0.04	64.40 ± 0.40	44.38 ± 0.11	72.83 ± 0.50	46.34 ± 0.07	0.39 ± 0.01
	Frozen	72.50 ± 0.26	68.39 ± 0.20	39.68 ± 0.03	57.11 ± 0.13	40.21 ± 0.07	62.99 ± 0.09	41.94 ± 0.08	0.36 ± 0.00
	Full FT	78.20 ± 0.66	76.14 ± 0.23	46.61 ± 0.17	66.75 ± 0.49	45.18 ± 0.12	73.47 ± 0.58	48.26 ± 0.06	0.35 ± 0.07
	Adapter	77.26 ± 0.50	75.25 ± 0.24	43.17 ± 0.09	66.84 ± 0.29	43.69 ± 0.05	74.06 ± 0.47	45.10 ± 0.06	0.41 ± 0.01
SOTA Models									
DOFA-B	Frozen	67.82 ± 0.15	58.99 ± 0.13	33.47 ± 0.07	45.44 ± 0.51	34.60 ± 0.20	53.22 ± 0.43	37.04 ± 0.10	0.28 ± 0.00
	Full FT	74.47 ± 0.66	69.11 ± 0.30	42.60 ± 0.06	56.99 ± 0.25	41.98 ± 0.05	64.90 ± 0.17	42.31 ± 0.06	0.40 ± 0.00
DOFA-L	Frozen	67.48 ± 0.13	60.64 ± 0.27	33.42 ± 0.08	46.11 ± 0.17	34.67 ± 0.14	51.26 ± 0.37	37.37 ± 0.08	0.28 ± 0.00
	Full FT	74.72 ± 0.32	69.40 ± 0.44	43.06 ± 0.07	56.54 ± 0.55	42.18 ± 0.17	64.80 ± 0.37	43.06 ± 0.05	0.41 ± 0.00
SpatSigma-B	Frozen	68.16 ± 0.33	63.49 ± 0.09	36.02 ± 0.06	50.67 ± 0.36	36.99 ± 0.09	58.20 ± 0.26	38.62 ± 0.12	0.31 ± 0.00
	Full FT	76.71 ± 0.32	70.63 ± 0.55	43.18 ± 0.07	58.32 ± 0.19	42.93 ± 0.09	69.27 ± 0.35	42.54 ± 0.06	0.41 ± 0.00
SpatSigma-L	Frozen	68.37 ± 0.33	61.72 ± 0.22	34.42 ± 0.30	47.86 ± 0.28	35.70 ± 0.09	54.94 ± 0.31	37.85 ± 0.06	0.27 ± 0.03
	Full FT	76.66 ± 0.29	70.29 ± 0.21	43.31 ± 0.07	56.83 ± 1.01	42.60 ± 0.10	67.98 ± 0.45	42.96 ± 0.08	0.42 ± 0.00

Models are evaluated under three protocols: frozen encoder, full fine-tuning, and fine-tuned adapter. Training from scratch is reported as a baseline. State-of-the-art models DOFA and HyperSigma are reported for comparison. The best score for each dataset is highlighted in bold, and the second best in underlined.

Fine-tuning the spectral adapter significantly outperforms frozen encoder performance in all settings, showing the importance of tailoring the spectral feature extraction to the characteristics of the downstream task. Notably, fine-tuning the spectral adapter outperforms training from scratch on CDL, EuroCrops, and BD-Foret.

Compared to CNNs, ViTs perform slightly worse on segmentation problems. This can be attributed to multiple factors.

1) Training ViTs from scratch results in substantially lower performance than CNNs due to the lack of strong inductive biases. While pretraining significantly mitigates this gap, CNNs still retain a slight edge after fine-tuning.
2) The spatial patterns of the masks, which present high-frequency features partly due to label noise, are harder to learn with a ViT due to the fixed patch size compared to CNNs, which can preserve full-resolution feature maps. This effect can be better observed in the patch size experiment (see Fig. 13).

3) Pixel-Wise Regression: The results on HyBiomass are in line with the segmentation downstream tasks. Our pretrained frozen encoders significantly outperform the random encoder baseline for all SSL methods and backbones. Fine-tuning the spectral adapter yields a significant performance boost compared to the frozen encoder setting, and full fine-tuning works best. MoCo and DINO provide consistent improvements with fine-tuning over training from scratch for both the CNN and ViT backbones. The largest gain from pretraining is observed for Spec. ViT-S with MAE, which achieves an R^2 improvement of 0.11 compared to training from scratch. This configuration also yields the overall best result. Moreover, both the frozen encoder and the adapter-only fine-tuning outperform training from scratch for Spec. ViT-S with MAE. This is consistent with our segmentation results: without pretrained weights, CNNs significantly outperform ViTs, which can be attributed to the lack of strong inductive biases in ViTs, requiring larger datasets to perform well.

B. Scaling-Up Pretrained Models

Recent studies have shown significant performance gains through model scaling in SSL [11], [22]. Consequently, recent geospatial foundation models are exploring ViT architectures with hundreds of millions of parameters [16], [53], [73]. We investigated model scaling on SpectralEarth by pretraining Spec. ViT architectures of varying sizes: Base (B), Large (L), Huge (H), and Giant (g). Our findings are presented in Table III.

We observe that MAE consistently outperforms training from scratch across all model sizes and datasets. Notably, scaling from Spec. ViT-B to ViT-L improves frozen encoder performance on multiple downstream tasks, including NLCD, EuroCrops, TreeMap, BDForet, and HyBiomass. Fine-tuning the adapter—only 56 K additional parameters—surpasses training from scratch on CORINE, CDL, EuroCrops, BDForet, and HyBiomass. While scaling provides limited performance gains for some downstream tasks up to Spec. ViT-H, we find that Spec. ViT-B offers the best tradeoff between performance and computational cost. Further increases in model size, as seen for Spec. ViT-g, degrades performance, possibly due to the size of the pretraining dataset. Although SpectralEarth is significantly larger than previous HSI datasets, it remains smaller than the datasets typically used to train billion-parameter models in computer vision [11]. This trend can also be observed for models such as DOFA and HyperSigma, where the *large* variants provide limited benefits over the *base* backbones. Our models outperform both DOFA and HyperSigma across all tasks and protocols, which we attribute to the architectural design of our backbones that better preserves fine-grained spatial information and the large EnMAP corpus in SpectralEarth. In particular, HyperSigma was trained on Hyperion EO-1 and Gaofen-5B, making it less competitive in EnMAP downstream tasks.

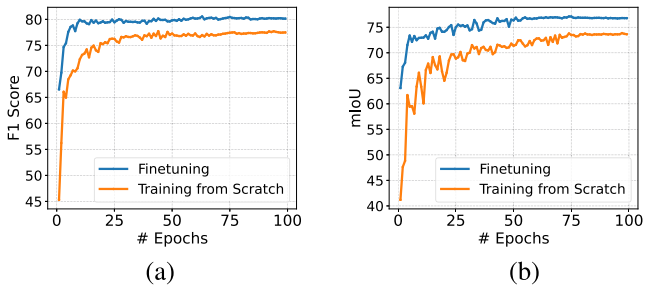


Fig. 9. Comparison of convergence speed for fine-tuning and training from scratch on EnMAP-CORINE and EnMAP-CDL for a Spec. RN50 backbone. (a) EnMAP-CORINE. (b) EnMAP-CDL.

C. Efficient Fine-Tuning

1) Model Convergence: Using a pretrained model can speed up convergence during fine-tuning, hence reducing the computational cost of training deep neural networks. To assess this for our models, we compare the training efficiency of Spec. RN50 models pretrained with DINO relative to training from scratch on the CORINE and CDL benchmarks. Fig. 9 displays the evolution of the validation metrics across training epochs. We observe that pretrained models converge more rapidly, particularly in the early stages of training. Notably, fine-tuning matches the 100 epochs performance of training from scratch in ~ 10 epochs for EnMAP-CORINE and EnMAP-CDL. These results highlight the computational advantage of using our pretrained models compared to training from scratch.

2) Parameter-Efficient Fine-Tuning: Foundation models are typically evaluated in linear probing to assess the quality of the learnt representations [19]. While a frozen encoder can achieve competitive performance, fine-tuning often surpasses it—at a higher computational cost. We experiment with the progressive unfreezing of network layers as a compromise between both scenarios. As depicted in Fig. 10, fine-tuning the first two blocks of a pretrained Spec. RN50 produces results that closely resemble full fine-tuning. This demonstrates the cost- and energy efficiency of our pretrained models for downstream tasks.

3) Training With Limited Labels: Pretrained models are very useful when labeled data is scarce [91]. To assess this, we evaluate the performance of the Spec. ViT-L model, pretrained with MAE, on varying subset sizes of EnMAP-CORINE and

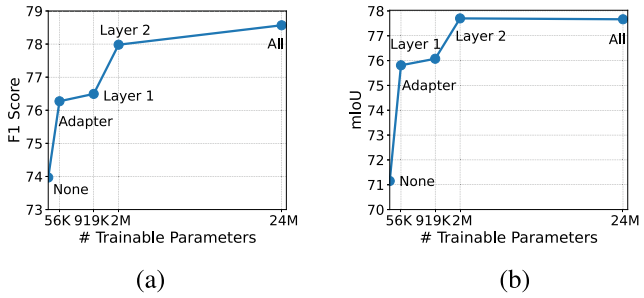


Fig. 10. Impact of incremental parameter unfreezing on EnMAP-CORINE and EnMAP-CDL with a pretrained Spec. RN50 model. (a) EnMAP-CORINE. (b) EnMAP-CDL.

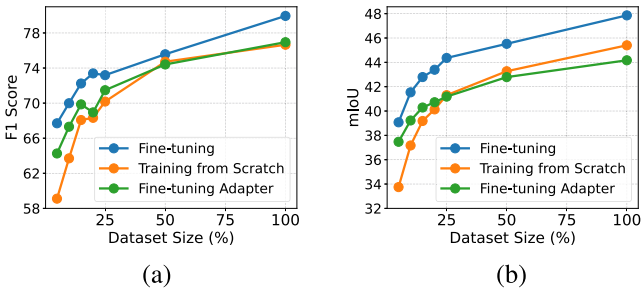


Fig. 11. Impact of the size of the downstream dataset evaluated on EnMAP-CORINE and EnMAP-NLCD for the Spec. ViT-L model. (a) EnMAP-CORINE. (b) EnMAP-NLCD.

EnMAP-NLCD datasets. The results, reported in Fig. 11, demonstrate the benefits of pretraining across all data regimes. In particular, we observe large gaps of 9 points in F1 score for EnMAP-CORINE, and 5 points in mIoU for EnMAP-NLCD, when using 5% of the labels compared to training from scratch. Interestingly, in lower data regime scenarios, fine-tuning the adapter only suffices to outperform training from scratch. This demonstrates the generalizability of the features learned from pretraining.

D. Cross-Sensor Transferability

To evaluate the generalizability of SpectralEarth-pretrained models across sensors, we assess their performance on three datasets acquired from sensors not seen during pretraining: EO-1 Hyperion (EO1-CDL), DESIS (DESIS-CDL), and Intuition-1 (Hyperview).[4] These datasets present different spectral characteristics from EnMAP.

We evaluate Spec. ViT-B and ViT-L models pretrained with MAE and compare them against training from scratch, as well as DOFA and HyperSigma. Our results are summarized in Table IV.

On the CDL-based tasks (EO-1 and DESIS), our models consistently outperform both DOFA and HyperSigma across all protocols. Notably, our frozen encoders outperform HyperSigma EO-1, despite not being pretrained on EO-1 imagery.

[4]The data for the hyperview challenge were collected using an airborne sensor that mimics the spectral characteristics of the Intuition-1 mission.

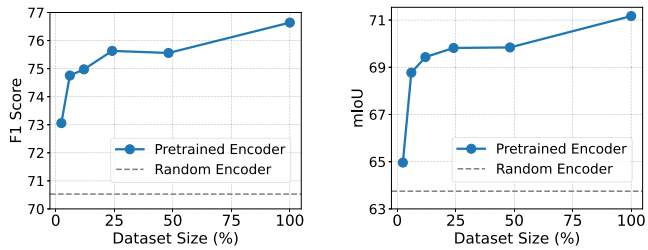


Fig. 12. Frozen encoder evaluation of Spec. RN50 on EnMAP-CORINE and EnMAP-CDL tasks with varying pretraining dataset sizes. (a) EnMAP-CORINE. (b) EnMAP-CDL.

This highlights the transferability of features learned from SpectralEarth. Given the relatively small size of the CDL datasets, fine-tuning the adapter performs on par with full fine-tuning and significantly outperforms training from scratch.

On the hyperview dataset, the best performance is obtained using the frozen Spec. ViT-L encoder. We attribute this to severe overfitting when fine-tuning due to the limited training data and the difficulty of the task. This result demonstrates the practicality of using our pretrained models as frozen feature extractors.

E. Ablation Studies

1) Impact of Temporal Positives: Temporal positives provide a natural data augmentation for joint-embedding methods and have demonstrated their benefits for multispectral imagery [47], [48], [49]. Given that only a subset of SpectralEarth includes temporal positives, it is unclear whether using temporal pairs significantly contributes to the overall performance of the pretrained models. To analyze their effect, we pretrain a Spec. RN50 using DINO with and without temporal pairs. Results in Table V confirm that excluding temporal positives degrades performance, with the effect most pronounced in frozen encoder evaluation.

2) How Much Data Are Needed for Pretraining?: To answer this question, we explore the impact of dataset size by pretraining a Spec. RN50 with DINO on randomly sampled subsets of SpectralEarth with sizes 10 K, 25 K, 50 K, 100 K, and 200 K locations. We conduct frozen-encoder evaluation on EnMAP-CORINE and EnMAP-CDL. Fig. 12 summarizes our findings. We observe a consistent improvement in performance metrics as the scale of the pretraining dataset increases from 10 k samples to the full size of SpectralEarth. This highlights the importance of large-scale datasets for hyperspectral SSL—even with small models such as Spec. RN50.

We also compare SpectralEarth with HySpecNet-11k [14] using the Spec. ViT-B backbone pretrained with MAE. Table VI reports performance under both frozen encoder and full fine-tuning settings. Pretraining on SpectralEarth consistently outperforms pretraining on HySpecNet-11 k, with the largest gains observed in the frozen encoder evaluation. Under complete fine-tuning, the performance gap narrows, as the downstream training offset differences in pretraining.

TABLE IV
CROSS-SENSOR TRANSFERABILITY EXPERIMENTS

Model Configuration		Hyperview (MSE_{Norm} ↓)	CDL (mIoU ↑)	
Backbone	Protocol	Intuition-1	EO-1	DESIS
SpecViT-B	From Scratch	0.88 ± 0.01	65.85 ± 0.78	66.18 ± 0.50
	Frozen	0.85 ± 0.01	59.02 ± 0.61	62.22 ± 0.61
	Full FT	0.86 ± 0.00	**66.69** ± **0.91**	68.50 ± 0.78
	Adapter	0.85 ± 0.00	66.17 ± 0.58	67.74 ± 0.29
SpecViT-L	From Scratch	0.89 ± 0.01	65.09 ± 0.76	66.06 ± 0.57
	Frozen	**0.81** ± **0.02**	59.46 ± 1.02	62.78 ± 0.31
	Full FT	0.87 ± 0.00	66.34 ± 0.77	**69.05** ± **0.50**
	Adapter	0.86 ± 0.00	66.58 ± 0.41	68.45 ± 0.16
DOFA-B	Frozen	0.89 ± 0.00	53.28 ± 0.82	57.92 ± 0.37
	Full FT	0.87 ± 0.01	61.09 ± 0.14	63.43 ± 0.72
DOFA-L	Frozen	0.88 ± 0.00	50.20 ± 0.47	58.26 ± 0.36
	Full FT	0.89 ± 0.01	61.15 ± 0.68	62.88 ± 0.34
SpatSigma-B	Frozen	0.88 ± 0.00	57.27 ± 0.24	59.15 ± 0.76
	Full FT	0.88 ± 0.00	62.91 ± 0.16	62.48 ± 0.29
SpatSigma-L	Frozen	0.87 ± 0.01	51.74 ± 0.54	56.72 ± 0.35
	Full FT	0.86 ± 0.01	62.90 ± 0.67	62.41 ± 0.35

We pretrain Spec. ViT-B and Spec. ViT-L with MAE on SpectralEarth and evaluate on three datasets from different sensors: Hyperview, EO1-CDL, and DESIS-CDL. We compare against training from scratch, DOFA and HyperSigma. Best results are in bold, second best are underlined.

TABLE V
IMPACT OF TEMPORAL POSITIVES

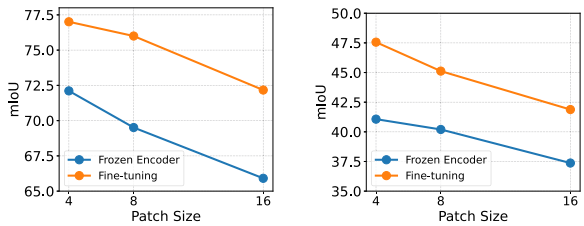
Model	EnMAP-CORINE (F1)		EnMAP-CDL (mIoU)	
	Frozen	Fine-tune	Frozen	Fine-tune
w/ TP	**76.64**	**77.98**	**71.15**	**77.66**
w/o TP	74.85	75.36	68.42	76.22

A Spec. RN50 model is pretrained on SpectralEarth with DINO and evaluated on EnMAP-CORINE and EnMAP-CDL.

TABLE VI
COMPARISON OF SPECTRALEARTH AND HYSPECNET-11 K

Model	EnMAP-CORINE (F1)		EnMAP-CDL (mIoU)	
	Frozen	Fine-tune	Frozen	Fine-tune
HyspecNet-11k	72.39	78.20	65.73	75.20
SpectralEarth	**74.98**	**79.65**	**72.42**	**77.46**

A Spec. ViT-B is pretrained using MAE and evaluated on EnMAP-CORINE and EnMAP-CDL.



(a) (b)

Fig. 13. Impact of the patch size for a pretrained Spec. ViT-S evaluated on the EnMAP-CDL and EnMAP-NLCD datasets. (a) EnMAP-CDL. (b) EnMAP-NLCD.

3) Importance of the Patch Size: Medium-resolution remote sensing imagery requires more fine-grained detail preservation than natural imagery, especially for segmentation tasks. This aspect is sometimes overlooked in the literature, where vision transformers with a patch size of 16×16 or 8×8 are commonly used as backbones to pretrain foundation models to reduce the computational cost [16], [50], [70], [73]. We argue that a smaller image size with a smaller patch size can lead better tradeoffs, particularly for HSI imagery where pixel information is rich and should not be heavily compressed in the embedding space. To analyze the impact of the patch size, we pretrained Spec. ViT-S using MAE with patch sizes of 4×4, 8×8, and 16×16. Results in Fig. 13 show that smaller patches consistently improve performance for segmentation tasks. This is further supported by qualitative results in Fig. 14, where models with smaller patches produce more refined segmentation maps. These results are consistent with the performance gap between ViT-based foundation models and ConvNet baselines observed in previous studies [92]. In practice, using additional trainable parameters to extract high-resolution features in combination with a pretrained ViT can mitigate this issue.

4) Masking Ratio in MAE: Given our small 4×4 patch size for Spec. ViT models, we expect a higher masking ratio is needed for MAE compared to the usual 75% used in computer vision [22]. To confirm this hypothesis, we pretrain Spec. ViT-S using MAE with different masking ratios, from 75% to 95%. Given MAE's poor linear probing results for the classification task, we focus on the segmentation downstream tasks. Results in Fig. 15 show that we obtain good representations even with an aggressive 95% masking, which enables faster pretraining as fewer tokens need to go through the encoder.

5) Impact of the Spectral Adapter: To assess the impact of the spectral adapter, we compare the classical ResNet-50 with our modified architecture, Spec. RN50. We adjust the first convolutional layer of the ResNet-50 model to accommodate

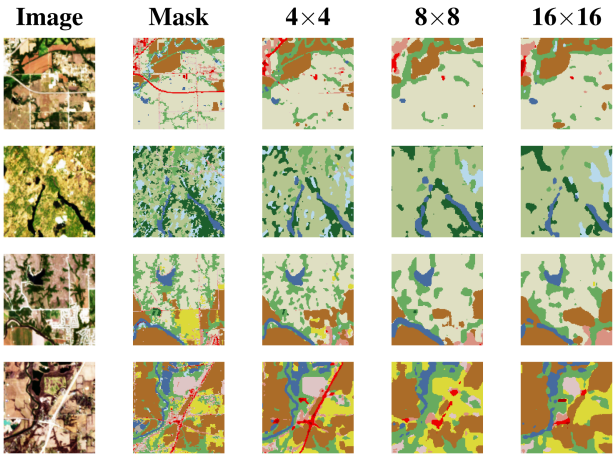


Fig. 14. Impact of the patch size in the Spec. ViT-S model on EnMAP-NLCD. From left to right, respectively: RGB visualization of the input image, ground truth (GT) mask, and predictions with 4×4, 8×8, and 16×16 patch sizes.

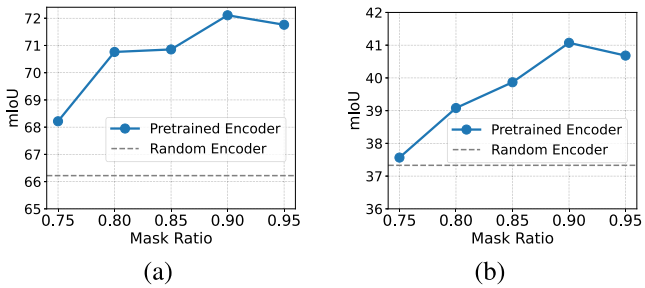


Fig. 15. Impact of the masking ratio in MAE evaluated on EnMAP-CDL and EnMAP-NLCD for the Spec. ViT-S model. (a) EnMAP-CDL. (b) EnMAP-NLCD.

TABLE VII
IMPACT OF THE SPECTRAL ADAPTER IN THE SPEC

Model	EnMAP-CORINE (F1)		EnMAP-CDL (mIoU)	
	Frozen	Fine-tune	Frozen	Fine-tune
w/o Adapter	72.82	77.10	67.50	74.31
w/ Adapter	**73.97**	**78.57**	**71.15**	**77.66**

RN50 architecture pretrained on spectralearth With MoCo-V2.

the 202 input bands. Both models are pretrained with MoCo-V2 and evaluated on the EnMAP-CORINE and EnMAP-CDL datasets. The results are summarized in Table VII. Spec. RN50 consistently outperforms the standard ResNet-50 in both frozen encoder evaluation and full fine-tuning. This result confirms the importance of a dedicated spectral feature extractor when dealing with hyperspectral images, particularly for tasks requiring fine-grained spectral discrimination.

VII. Conclusion

In this article, we introduce SpectralEarth, a large-scale dataset derived from EnMAP imagery as a basis for SSL methodologies in the hyperspectral domain. We leveraged SpectralEarth to pretrain foundation models based on popular SSL algorithms. Extensive evaluation on several downstream tasks benchmarked the pretrained models. Our results demonstrate the effectiveness of models pretrained on SpectralEarth to improve performance and reduce computational costs for downstream applications. We believe that the insights derived from this study serve as a basis for future development in SSL for hyperspectral imagery.

As it stands, we encourage further effort to construct additional hyperspectral benchmark datasets. Covering tasks such as unmixing enables a more comprehensive evaluation of hyperspectral foundation models. In particular, transferability across different hyperspectral sensors remains a challenging problem. Moreover, exploration of a wider range of architectures and SSL algorithms, beyond the Spectral ResNet-50 and ViT models employed in this work, may provide a more comprehensive toolbox for practical deep learning on HSI. Last but not least, exploration and benchmarking of multisensor SSL methods is a promising avenue of research toward more general geospatial foundation models.

Acknowledgment

The authors would like to thank the computational and data resources provided through the joint high-performance data analytics (HPDA) project "terrabyte" of the German Aerospace Center (DLR) and the Leibniz Supercomputing Center (LRZ).

References

[1] M. J. Khan, H. S. Khan, A. Yousaf, K. Khurshid, and A. Abbas, "Modern trends in hyperspectral image analysis: A review," *IEEE Access*, vol. 6, pp. 14118–14129, 2018.

[2] S. Chabrillat, E. Ben-Dor, J. Cierniewski, C. Gomez, T. Schmid, and B. van Wesemael, "Imaging spectroscopy for soil mapping and monitoring," *Surv. Geophys.*, vol. 40, pp. 361–399, 2019.

[3] F. D. Van der Meer et al., "Multi-and hyperspectral geologic remote sensing: A review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 14, no. 1, pp. 112–128, 2012.

[4] J. Jia, Y. Wang, J. Chen, R. Guo, R. Shu, and J. Wang, "Status and application of advanced airborne hyperspectral imaging technology: A review," *Infrared Phys. Technol.*, vol. 104, 2020, Art. no. 103115.

[5] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sens.*, vol. 12, no. 16, 2020, Art. no. 2659.

[6] V. Upadhyay and A. Kumar, "Hyperspectral remote sensing of forests: Technological advancements, opportunities and challenges," *Earth Sci. Informat.*, vol. 11, pp. 487–524, 2018.

[7] H. Kaufmann et al., "Science plan of the environmental mapping and analysis program (ENMAP)," 2012.

[8] G. Kerr et al., "The hyperspectral sensor desis on muses: Processing and applications," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 268–271.

[9] S. Pignatti et al., "The prisma hyperspectral mission: Science activities and opportunities for agriculture and land monitoring," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2013, pp. 4558–4561.

[10] M. Rast, J. Nieke, J. Adams, C. Isola, and F. Gascon, "Copernicus hyperspectral imaging mission for the environment (CHIME)," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 108–111.

[11] M. Oquab et al., "DinoV2: Learning robust visual features without supervision," 2023, *arXiv:2304.07193*. [Online]. Available: https://openreview.net/forum?id=a68SUt6zFt

[12] Y. Wang, C. Albrecht, N. A. A. Braham, L. Mou, and X. Zhu, "Self-supervised learning in remote sensing: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 4, pp. 213–247, Dec. 2022.

[13] N. Audebert, B. Le Saux, and S. Lefèvre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, Jun. 2019.

[14] M. H. P. Fuchs and B. Demir, "Hyspecnet-11 k: A large-scale hyperspectral dataset for benchmarking learning-based hyperspectral image compression methods," 2023, *arXiv:2306.00385*.

[15] L. Scheibenreif, M. Mommert, and D. Borth, "Masked vision transformers for hyperspectral image classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2165–2175.

[16] D. Wang et al., "HyperSigma: Hyperspectral intelligence comprehension foundation model," 2024. [Online]. Available: https://arxiv.org/abs/2406.11519

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[18] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.

[20] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," 2020, *arXiv:2003.04297*.

[21] M. Caron et al., "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9650–9660.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[23] M. F. Baumgardner, L. L. Biehl, and D. A. Landgrebe, "220 band aviris hyperspectral image data set: Jun.12, 1992 indian pine test site 3," *Purdue Univ. Res. Repository*, vol. 10, no. 7, 2015, Art. no. 991.

[24] M. V. B. A. M. Grana, "Hyperspectral datasets," 2022. [Online]. Available: http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[25] C. Debes et al., "Hyperspectral and lidar data fusion: Outcome of the 2013 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2405–2418, Jun. 2014.

[26] L. Sun, J. Zhang, J. Li, Y. Wang, and D. Zeng, "SDFC dataset: A large-scale benchmark dataset for hyperspectral image classification," *Opt. Quantum Electron.*, vol. 55, no. 2, 2023, Art. no. 173.

[27] N. Yokoya and A. Iwasaki, "Airborne hyperspectral data over Chikusei," Space Appl. Lab., Univ. Tokyo, Tokyo, Japan, Tech. Rep. SAL-2016-05-27, vol. 5, no. 5, 2016, Art. no. 5.

[28] Y. Zhong, X. Hu, C. Luo, X. Wang, J. Zhao, and L. Zhang, "Whu-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF," *Remote Sens. Environ.*, vol. 250, 2020, Art. no. 112012.

[29] J. Nalepa et al., "The hyperview challenge: Estimating soil parameters from hyperspectral images," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 4268–4272.

[30] R. Thoreau, L. Risser, V. Achard, B. Berthelot, and X. Briottet, "Toulouse hyperspectral data set: A benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques," *ISPRS J. Photogrammetry Remote Sens.*, vol. 212, pp. 323–337, 2024.

[31] Y. Wang et al., "HSIMAE: A unified masked autoencoder with large-scale pre-training for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 14064–14079, 2024.

[32] University of the Basque Country, "Hyperspectral remote sensing scenes." Accessed: 2023. [Online]. Available: https://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes

[33] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856.

[34] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 11, pp. 4037–4058, Nov. 2020.

[35] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 649–666.

[36] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6707–6717.

[37] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. Int. Conf. Learn. Representations*, 2018.

[38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[39] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 9912–9924.

[40] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 21271–21284.

[41] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12310–12320.

[42] A. Bardes, J. Ponce, and Y. Lecun, "VICREG: Variance-invariance-covariance regularization for self-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2022.

[43] Z. Xie et al., "Simmim: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 9653–9663.

[44] H. Bao, L. Dong, S. Piao, and F. Wei, "BEIT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Representations*, 2021.

[45] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2Vec: A general framework for self-supervised learning in speech, vision and language," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 1298–1312.

[46] M. Assran et al., "Masked siamese networks for label-efficient learning," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 456–473.

[47] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.

[48] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "SSL4EO-S12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation," 2022, *arXiv:2211.07044*.

[49] A. Stewart et al., "SSL4EO-L: Datasets and foundation models for landsat imagery," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 59787–59807.

[50] Y. Cong et al., "SatMAE: Pre-training transformers for temporal and multi-spectral satellite imagery," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 197–211.

[51] C. J. Reed et al., "Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4088–4099.

[52] M. Mendieta, B. Han, X. Shi, Y. Zhu, C. Chen, and M. Li, "GFM: Building geospatial foundation models via continual pretraining," 2023, *arXiv:2302.04476*. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/Mendieta_Towards_Geospatial_Foundation_Models_via_Continual_Pretraining_ICCV_2023_paper.html

[53] D. Hong et al., "SpectralGPT: Spectral remote sensing foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5227–5244, Aug. 2024.

[54] X. Li, D. Hong, and J. Chanussot, "S2MAE: A spatial-spectral pretraining foundation model for spectral remote sensing data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 24088–24097.

[55] J. Jakubik et al., "Foundation models for generalist geospatial artificial intelligence," 2023, *arXiv:2310.18660*.

[56] D. Szwarcman et al., "Prithvi-eo-2.0: A versatile multi-temporal foundation model for Earth observation applications," 2024, *arXiv:2412.02732*.

[57] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, "Omnisat: Self-supervised modality fusion for Earth observation," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 409–427.

[58] A. Fuller, K. Millard, and J. Green, "Croma: Remote sensing representations with contrastive radar-optical masked autoencoders," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 5506–5538.

[59] B. Han, S. Zhang, X. Shi, and M. Reichstein, "Bridging remote sensors with multisensor geospatial foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2404.01260

[60] J. Prexl and M. Schmitt, "Senpa-mae: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining," 2024, *arXiv:2408.11000*.

[61] G. Astruc, N. Gonthier, C. Mallet, and L. Landrieu, "AnySat: One earth observation model for any resolutions, scales, and modalities," 2024, *arXiv:2412.14123*.

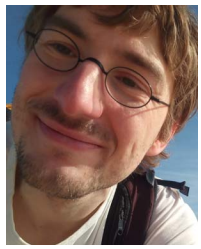
[62] M. Assran et al., "Self-supervised learning from images with a joint-embedding predictive architecture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15619–15629.

[63] Z. Li et al., "Few-shot hyperspectral image classification with self-supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5517917.

[64] S. Hou, H. Shi, X. Cao, X. Zhang, and L. Jiao, "Hyperspectral imagery classification based on contrastive learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5521213.
[65] W. Liu et al., "Self-supervised feature learning based on spectral masking for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4407715.
[66] N. A. A. Braham, J. Mairal, J. Chanussot, L. Mou, and X. X. Zhu, "Enhancing contrastive learning with positive pair mining for few-shot hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 8509–8526, 2024.
[67] X. Cao, H. Lin, S. Guo, T. Xiong, and L. Jiao, "Transformer-based masked autoencoder with contrastive loss for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5524312.
[68] D. Ibanez, R. Fernandez-Beltran, F. Pla, and N. Yokoya, "Masked auto-encoding spectral–spatial transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542614.
[69] R. Tian, D. Liu, Y. Bai, Y. Jin, G. Wan, and Y. Guo, "Swin-MSP: A shifted windows masked spectral pretraining model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2024, Art. no. 4509114.
[70] Z. Xiong et al., "Neural plasticity-inspired foundation model for observing the Earth crossing modalities," 2024, *arXiv:2403.15356*.
[71] R. de Los Reyes, M. Langheinrich, and M. Bachmann, "Enmap ground segment-level 2A processor (atmospheric correction over land) atbd," *DLR*, Cologne, Germany, EN-PCV-TN-6007, vol. 2, no. 4, 2023.
[72] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "BigEarthNet: A large-scale benchmark archive for remote sensing image understanding," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5901–5904.
[73] J. Jakubik et al., "Foundation models for generalist geospatial artificial intelligence," 2023. [Online]. Available: https://arxiv.org/abs/2310.18660
[74] G. Sumbul et al., "BigEarthNet-MM: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval [software and data sets]," *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 3, pp. 174–180, Sep. 2021.
[75] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring us agriculture: The us department of agriculture, national agricultural statistics service, cropland data layer program," *Geocarto Int.*, vol. 26, no. 5, pp. 341–358, 2011.
[76] J. Dewitz and U. G. Survey, "National Land Cover Database (NLCD) 2019 Products (ver. 2.0)," *U.S. Geological Surv. Data Release, Reston, VA, USA*, Jun. 2021, doi: 10.5066/P9KZCM54.
[77] K. L. Riley, I. C. Grenfell, J. D. Shaw, and M. A. Finney, "Treemap 2016 dataset generates conus-wide maps of forest characteristics including live basal area, aboveground carbon, and number of trees per acre," *J. Forestry*, vol. 120, no. 6, pp. 607–632, 2022.
[78] M. C. Hansen et al., "High-resolution global maps of 21st-century forest cover change," *Science*, vol. 342, no. 6160, pp. 850–853, 2013.
[79] M. Schneider, T. Schelte, F. Schmitz, and M. Körner, "EuroCrops: The largest harmonized open crop dataset across the European union," *Sci. Data*, vol. 10, no. 1, Sep. 2023, Art. no. 612, doi: 10.1038/s41597-023-02517-0.
[80] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-D deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 847–858, Feb. 2018.
[81] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
[82] D. Hong et al., "SpectralFormer: Rethinking hyperspectral image classification with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5518615.
[83] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.
[84] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.
[85] M. Ahmad, S. Distifano, M. Mazzara, and A. M. Khan, "Traditional to transformers: A survey on current trends and future prospects for hyperspectral image classification," 2024, *arXiv:2404.14955*.
[86] A. Banze, T. Stassin, N. A. A. Braham, R. S. Kuzu, S. Besnard, and M. Schmitt, "Hybiomass: Global hyperspectral imagery benchmark dataset for evaluating geospatial foundation models in forest aboveground biomass estimation," 2025. [Online]. Available: https://arxiv.org/abs/2506.11314
[87] A. Bardes, J. Ponce, and Y. LeCun, "Vicregl: Self-supervised learning of local visual features," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 8799–8810.
[88] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4918–4927.
[89] J. Lehner, B. Alkin, A. Fürst, E. Rumetshofer, L. Miklautz, and S. Hochreiter, "Contrastive tuning: A little help to make masked autoencoders forget," 2023, *arXiv:2304.10520*.
[90] Y. Lee et al., "Surgical fine-tuning improves adaptation to distribution shifts," 2022, *arXiv:2210.11466*.
[91] A. Newell and J. Deng, "How useful is self-supervised pretraining for visual tasks?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7345–7354.
[92] Y. Xie et al., "When are foundation models effective? Understanding the suitability for pixel-level classification using multispectral imagery," 2024, *arXiv:2404.11797*.

Nassim Ait Ali Braham received the M.Sc. degree in computer science from Ecole nationale Supérieure d'Informatique (ESI), Algiers, Algeria, in 2019 and the M.Sc. degree in artificial intelligence and data science from Université Paris Dauphine-PSL, Paris, France, in 2020. He is currently working toward the Ph.D. degree AI for earth observation with the German Aerospace Center, Wessling, Germany, and with the Technical University of Munich, Munich, Germany.

In 2019, he spent six months as a research intern with the LIRIS-CNRS laboratory, Lyon, France. In 2020, he spent six months with the LAMSADE-CNRS Laboratory, PSL Research University, Paris, France. In 2023, he spent six months as a visiting Researcher with INRIA THOTH, Grenoble, France. His research interests include deep learning, computer vision, self-supervised learning, and remote sensing.



Conrad M. Albrecht received the Graduate degree in physics (International Max-Planck Research School for Quantum Dynamics in Physics, Chemistry, and Biology) from the Institute for Theoretical Physics, Heidelberg, Germany, with an extra certification in computer science (Cluster- and Detector Management team, CERN, Geneva, Switzerland) in 2010, and the Ph.D. degree, focus on distributed computing to study physics at low temperatures, in physics from Heidelberg University, Heidelberg, Germany, in 2014.

Since April 2021, he has been a Researcher with the Earth Observation Center of the German Aerospace Center (DLR), Oberpfaffenhofen, Germany. As PI of the HelmholtzAI Young Investigator Group (YIG) "Large-Scale Data Mining in Earth Observation" (DM4EO), he advanced weakly-supervised deep learning methodologies for remote sensing. In July 2023, he was a visiting Associate Professor with the Institute of Nasca, Yamagata University, Yamagata, Japan, contributing to research in machine learning for the UNESCO World Heritage of the Nasca culture in Peru. For over 6 years, he was a Research Scientist with the Physical Sciences Department, IBM T. J. Watson Research Center, Yorktown, NY, USA. Beginning in 2024, and initiated by Helmholtz Imaging and Data Science, he is with the German Department, Princeton University, Princeton, NJ, USA, for cross-atlantic undergraduate internships. In fall 2024, he was an Adjunct Professor of Earth and Environmental Engineering with Columbia University, Broadway, NY, USA, and in spring 2025, he supervises Columbia University students for environmental monitoring. His research interests include physical models and numerical analysis, employing big data technologies and machine learning through open-science research.

Dr. Albrecht's work was the recipient of the 2024 Cozzarelli Prize by the National Academy of Sciences, jointly with his collaborators. He co-organized workshops at the IEEE BigData conference, IGARSS conferences, the EGU General Assembly, the CVPR conference, and the AAAS annual meeting. Some of his transatlantic initiatives include: In 2023, he was invited by the Alexander-von-Humboldt Foundation to present at the German-American Frontiers of Engineering Symposium.

Julien Mairal received the Graduate degree from Ecole Polytechnique, Palaiseau, France, in 2005 and the Ph.D. in electrical engineering degree from Ecole Normale Superieure, Cachan, France, in 2010.

After that, he joined the Statistics Department, UC Berkeley, Berkeley, CA, USA, as a Postdoctoral Researcher. In 2012, he joined Inria, Grenoble, France, where he is currently a Research Director and head of the Thoth team. His research interests include machine learning, computer vision, mathematical optimization, and statistical image and signal processing.

Dr. Mairal was the recipient of a starting grant and a consolidator grant from the European Research Council in 2016 and 2022, respectively. He was also the recipient of the Cor Baayen prize in 2013, the IEEE PAMI young research award in 2017, and the test-of-time award at ICML 2019.

Jocelyn Chanussot (Fellow, IEEE) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995 and the Ph.D. degree from the Université de Savoie, Annecy, France, in 1998.

From 1999 to 2023, he has been with Grenoble INP, Grenoble, France, where he was a Professor of Signal and Image Processing. He is currently a Research Director with INRIA, Grenoble, France. He has been a Visiting Scholar with Stanford University, Stanford, CA, USA, KTH Royal Institute of Technology, Stockholm, Sweden, and National University of Singapore, Singapore. Since 2013, he has been an Adjunct Professor of the University of Iceland, Reykjavik, Iceland. In 2015–2017, he was a Visiting Professor with the University of California, Los Angeles, Los Angeles, CA, USA. He holds the AXA Chair in Remote Sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information research Institute, Beijing, China. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007–2010), which got the 2010 IEEE GRS-S Chapter Excellence Award. He was the recipient of multiple outstanding paper awards. He was the Vice-President of the IEEE Geoscience and Remote Sensing Society, in charge of meetings and symposia (2017–2019). He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote sensing (WHISPERS). He was the Chair (2009–2011) and Cochair of the GRS Data Fusion Technical Committee (2005–2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is currently an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE Transactions on Image Processing and the Proceedings of the IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015). In 2014, he was a Guest Editor for the *IEEE Signal Processing Magazine*. He is an ELLIS Fellow, a Fellow of the Asia-Pacific Artificial Intelligence Association, a member of the Institut Universitaire de France (2012–2017), and a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters, since 2018).

Yi Wang received the B.E. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2018 and the M.Sc. degree in geomatics engineering from the University of Stuttgart, Stuttgart, Germany, in 2021. He is currently working toward the Ph.D. degree n AI for earth observation with the Technical University of Munich (TUM), Munich, Germany.

From 2021 to 2024, he was a Research Associate with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Cologne, Germany. In 2020, he spent three months at the perception system group, Sony Corporation, Stuttgart, Germany. His research interests include self-supervised learning, weakly-supervised learning, and multimodal representations.

Xiao Xiang Zhu (Fellow, IEEE) received the master's degree (M.Sc.) in earth oriented space science and technology, and the Ph.D. degree, in remote sensing, the Habilitation degree in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is currently the Chair Professor for Data Science in Earth Observation with TUM and was the founding Head of the Department "EO Data Science" with the Remote Sensing Technology Institute, German Aerospace Center (DLR), Cologne, Germany. Since May 2020, she has been the PI and Director of the international future AI lab "AI4EO—Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has also been a Director of the Munich Data Science Institute (MDSI), TUM. From 2019 to 2022, she has been a co-coordinator of the Munich Data Science Research School (www.mu-ds.de) and the head of the Helmholtz Artificial Intelligence—Research Field "Aeronautics, Space and Transport." She was a Guest Scientist or visiting Professor with the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, the University of Tokyo, Tokyo, Japan, and University of California, Los Angeles, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is currently a visiting AI Professor with ESA's Phi-lab, Frascati, Italy. Her main research interests are remote sensing and Earth observation, signal processing, machine learning and data science, with their applications in tackling societal grand challenges, e.g. Global Urbanization, UN's SDGs and Climate Change.

Dr. Zhu has been a member of young academy (Junge Akademie/Junges Kolleg) with the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She is a Fellow of the Academia Europaea (the Academy of Europe). She serves in the scientific advisory board in several research organizations, among others the German Research Center for Geosciences (GFZ, 2020–2023) and Potsdam Institute for Climate Impact Research (PIK). She is currently an Associate Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition* and was the Area Editor responsible for special issues of *IEEE Signal Processing Magazine* (2021–2023). She is currently a Fellow of AAIA and ELLIS.